Mining Wikipedia for Large-scale Repositories of Context-Sensitive Entailment Rules

Milen Kouylekov[1], **Yashar Mehdad**[1,2], **Matteo Negri**[1]

FBK-Irst[1], University of Trento[2]
Trento, Italy
kouylekov@fbk.eu, mehdad@fbk.eu, negri@fbk.eu

Abstract

This paper focuses on the central role played by lexical information in the task of Recognizing Textual Entailment. In particular, the usefulness of lexical knowledge extracted from several widely used static resources, represented in the form of *entailment rules*, is compared with a method to extract lexical information from Wikipedia as a dynamic knowledge resource. The proposed acquisition method aims at maximizing two key features of the resulting entailment rules: *coverage* (*i.e.* the proportion of rules successfully applied over a dataset of TE pairs), and *context sensitivity* (*i.e.* the proportion of rules applied in appropriate contexts). Evaluation results show that Wikipedia can be effectively used as a source of lexical entailment rules, featuring both higher coverage and context sensitivity with respect to other resources.

1. Introduction

Textual Entailment Recognition (RTE) is a semantic inference task that consists in recognizing whether the meaning of one text (called the *hypothesis* H) can be inferred from another text (the *text* T) (Dagan and Glickman, 2004). Current approaches to the task usually rely on lexical information mined from a relatively small set of lexical-semantic resources, as an additional source of evidence about the existence of entailment relations between T-H pairs.

In spite of the substantial agreement on the usefulness of these prominent resources for the extraction of lexical knowledge, only few works attempted to provide a quantitative measure of their actual utility. Among these, the results recently reported by Mirkin et al. (Mirkin et al., 2009) demonstrate that: *i)* the most widely used resources for lexical knowledge (*e.g.* WordNet (Fellbaum, 1998)) allow for limited recall figures, *ii)* resources built considering distributional evidence (*e.g.* Lin Dependency and Proximity thesauri (Lin, 1998)) are suitable to capture more entailment relationships, *iii)* the application of rules in inappropriate contexts severely impacts on performance.

Based on these findings, (Mirkin et al., 2009) also proposes a simple method to extract lexical knowledge from Wikipedia. This method is based on the the first step of the algorithm used by (Kazama and Torisawa, 2007) to improve Named Entity Recognition with Wikipedia as a source of external knowledge. The idea is to generate inference rules by extracting from the first sentence of each page a noun that appears in a *is-a* pattern referring to the title of the page. Though recall scores achieved with this method are the lowest compared to other resources, relatively high precision motivates further research on how to profitably use Wikipedia as a source of lexical knowledge for the RTE task.

Along such direction, this paper proposes an alternative methodology to extract from Wikipedia large amounts of lexical entailment rules (*i.e.* rules, such as $limousine \Rightarrow car$, that assign an entailment probability to a term t in T, and a term H in H). The main difference with respect to the approach proposed by (Mirkin et al., 2009) is that, instead of considering only the first sentence of Wikipedia articles for knowledge acquisition, we compute Latent Semantic Analysis scores over the entire articles, and use a relatedness threshold estimated over training data to reduce noise. This solution maximizes the number of the extracted rules, without affecting precision. The benefits of the proposed methodology come in terms of *coverage* (*i.e.* the amount of rules successfully applied over a dataset of RTE pairs), and *context sensitivity* (*i.e.* the amount of rules applied in appropriate contexts).

The paper is structured as follows. First, Section 2. focuses on the role of lexical knowledge in the RTE task, and overviews the resources typically used to support the matching between T-H pairs. Then, Section 3. describes our methodology to extract lexical entailment rules from Wikipedia. Section 4. summarizes the results of comparative evaluations of rule repositories created from the different resources. A first experiment was done by running an RTE system using the Tree Edit Distance algorithm (TED) over the RTE-5 dataset. Further analysis was carried out to compare the different rule repositories considering their potential coverage on the same dataset, independently from a specific RTE algorithm. Finally, Section 5. concludes the paper providing directions for future work.

2. The role of lexical knowledge in RTE

Lexical knowledge is widely used by current approaches to RTE, in order to support mappings between T-H pairs. As regards knowledge sources, there is a substantial agreement on the usefulness of some prominent resources, including: WordNet (Fellbaum, 1998), eXtended WordNet (Moldovan and Novischi, 2002), the dependency and proximity thesauri described in (Lin, 1998), and VerbOcean (Chklovski and Pantel, 2004).

As emerges from the past RTE Evaluation Campaigns, WordNet is undoubtedly the mostly used lexical resource. This is confirmed by the ablation tests carried out within the last RTE-5 challenge (Bentivogli et al., 2009), which

showed that all the analysed systems (19 out of 20 participants) use it as an external knowledge source to perform semantic alignments between content words in T and H (Iftene and Moruz, 2009), (Wang et al., 2009), (Mehdad et al., 2009). For instance, WordNet 3.0 *synonymy* and *hyponymy* relations can be effectively used to support lexical mappings between all the entailing examples shown in Table 2.

Similarly, the use of VerbOcean to facilitate mappings between verbs in T and H is documented in several recent works (6 RTE-5 participants). Among these, while (Wang et al., 2009) used all the VerbOcean relations between verbs, other participants used only some of them, such as [*stronger-than*] (Mehdad et al., 2009), or [*opposite-of*] (Iftene and Moruz, 2009).

The use of eXtended WordNet is reported in (Tatu et al., 2006), where it is effectively used to automatically construct lexical chains between words in T to Hs constituents, thus producing entailment axioms supporting the entailment checking.

Apart from the previously mentioned work by (Mirkin et al., 2009), the potential of Dekang Lin's thesauri in the RTE task has been partially explored by (Marsi et al., 2006), which uses word similarity measures for the alignment of dependency trees.

2.1. Lexical entailment rules

Following the approach proposed in (Kouylekov and Magnini, 2006), our use of lexical knowledge builds on the creation of repositories of lexical *entailment rules*, that can be extracted from any source of external knowledge (*e.g.* thesauri, the Web, a local corpus). Each rule has a left hand side (a word in T) and a right hand side (a word in H), associated to a probability that indicates if the left hand side entails or contradicts the right hand side. As an example, a rule [$phobia \Rightarrow disorder$] can be extracted from WordNet, and applied to the first example in Table 2. to increase the probability to discover the entailment relation between T and H.

Given the RTE-5 dataset, entailment rules for our experiments have been automatically acquired from some of the previously mentioned resources. Rule extraction is carried out as follows:

WordNet rules are collected for each pair of terms (w_1 in T and w_2 in H) that are connected by the *synonym* or *hypernym* relations. More specifically, given a word w_1 in T, a new rule [$w_1 \Rightarrow w_2$] is created for each word w_2 in H that is a synonym or an hypernym of w_1.

VerbOcean rules are collected for each pair of verbs (v_1 in T and v_2 in H) that are connected by the [*stronger-than*] relation. More specifically, given a verb v_1 in T, a new rule [$v_1 \Rightarrow v_2$] is created for each verb v_2 in H that is connected to v_1 by the [*stronger-than*] relation (*i.e.* when [v_1 *stronger-than* v_2]). Though potentially useful, transitive closures are not considered due to the high level of noise introduced by verb ambiguities.

Lin Dependency/Proximity Similarity rules are collected from the thesauri of dependency and proximity based similarities described in (Lin, 1998), and available at Dekang Lin's website[1]. More specifically, given a word w_1 in T, a new rule [$w_1 \Rightarrow w_2$] is created for each word w_2 in H that is related to w_1 in the thesauri.

3. Mining entailment rules from Wikipedia

Wikipedia, as a source of lexical entailment rules, offers at least two advantages over other resources. The first one relates to *coverage*: with more than 3.000.000 articles, Wikipedia covers the vast majority of concepts potentially appearing in any RTE dataset. This is particularly evident with named entities (*e.g.* instances of the categories PERSON or LOCATION), whose coverage in Wikipedia is much larger than in any other source of lexical knowledge commonly used by RTE systems. The second advantage relates to *context sensitivity*, as it allows to consider the context (*i.e.* the actual content of the articles) in which rule elements tend to appear.

To embed context sensitivity in our rules, our approach is based on computing over Wikipedia a Latent Semantic Analysis (LSA) score between all possible word pairs that appear in the T-H pairs of an RTE dataset. To this aim we use the jLSI (java Latent Semantic Indexing) tool[2] to measure the relatedness between all the terms in a T-H pair. We created a model from the 200,000 most visited Wikipedia articles, after cleaning unnecessary markup tags. Cleaned articles are used as documents for creating the term-by-document matrix. Then, we empirically estimate over the training data a relatedness threshold in order to filter out all the pairs of terms featuring low similarity, thus obtaining a set of pairs where the first term entails the second one with a high probability.

As a result, starting from the RTE5 dataset, we obtained around 199K rules, which reports the highest number of rules compared to other resources, leading to a higher coverage and better performance.

4. Comparing repositories over RTE-5 data

This section summarizes the experiments carried out to compare rule repositories obtained from different resources in the RTE task, focusing on: the RTE system we used, the overall experimental settings, and the achieved results.

4.1. EDITS

Our experiments have been carried out using EDITS (Edit Distance Textual Entailment Suite) (Negri et al., 2009), a freely available open source tool for recognizing textual entailment developed by FBK-irst. EDITS implements a distance-based approach, which assumes that the distance between T and H is a characteristic that separates the positive T-H pairs, for which the entailment relation holds, from the negative pairs, for which the entailment relation does not hold. The system allows for different configurations of its three basic components, namely: *i)* the **edit distance algorithm**, that computes the T-H distance as the overall cost of the edit operations (*i.e.* insertion, deletion and substitution) that are necessary to transform T into H; *ii)* the

[1] http://www.cs.ualberta.ca/\~lindek/downloads.htm

[2] Available at http://tcc.itc.it/research/textec/tools-resources/jLSI.html

1	T: Agoraphobia means fear of open spaces and is one of the most common phobias. H: Agoraphobia is a widespread disorder.	*HasHypernym(phobia,disorder)*
2	T: Everest summiter David Hiddleston has passed away in an avalanche of Mt. Tasman. H: A person died in an avalanche.	*Synonym(pass away, die)*
3	T: In the Italian Alps, four climbers died in an avalanche in the Argentera valley on Sunday afternoon. H: Humans died in an avalanche.	*HasHypernym(climber,human)*
4	T: El Nino usually begins in December and lasts a few months. H: El Nino usually starts in December.	*Synonym(begin, start)*
5.	T: There are currently eleven (11) official languages of the European Union in number. H: There are 11 official EU languages.	*Synonym(European Union, EU)*

Table 1: RTE samples, and useful lexical knowledge from WordNet

	VO		WN		PROX		DEP		WIKI	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Accuracy	61.8	58.8	61.8	58.6	61.8	58.8	62	57.3	62.6	60.3

Table 2: results

cost scheme, that defines the cost associated to each edit operation involving an element of T and an element of H; and *iii)* an (optional) **entailment rules repository** that provides specific knowledge (*e.g.* lexical, syntactic, semantic) about the allowed transformations between portions of T and H. Each rule has a left hand side (an element of T) and a right hand side (an element of H), associated to a probability which indicates if the left hand side entails or contradicts the right hand side. Rules can be manually defined, or they can be extracted from any external resource available (*e.g.* WordNet, Wikipedia, the Web, a local corpus).

Since our objective is to compare the utility of the lexical knowledge extracted from different sources, each experiment has been carried out with our best configuration of EDITS (the one used for our RTE-5 submission, which is thoroughly described in (Mehdad et al., 2009)), except for the rule repository used. This configuration is based on using Tree Edit Distance (TED) as the core algorithm to perform transformations between syntactic representations[3]) of the T-H pairs.

4.2. Rule repositories

Our experiments have been carried out comparing results achieved over the RTE-5 dataset by exploiting the following rule repositories:

WIKI: Out of the original 199217 rules extracted from Wikipedia, we estimated a threshold[3] over training data to filter out those with lower reliability. As a result, 58278 rules have been retained.

WN: 1106 rules have been extracted from WordNet as described in Section 2.1.

VO: similarly, 192 rules have been extracted from VerbOcean.

DEP: Out of the 5432 rules extracted from Lin's dependency thesaurus, as described in Section 2.1., we estimated a threshold[4] to filter out those with lower reliability. As a result, 2468 rules have been retained.

PROX: in the same way, out of 8029 original rules extracted from the Lin's proximity thesaurus, only 236 have been retained.

4.3. Results

Table 2 reports the accuracy results we achieved over RTE-5 data (both on the development and test sets), showing that Wikipedia rules outperform all the other rule repositories, with performance increase over the test set ranging from 2.5% to 5.2% Accuracy improvement.

These results demonstrate that applying entailment rules extracted from Wikipedia, we gain a higher coverage as well as a better performance in our entailment framework. As an example, the entailment relations between "*Apple*" and "*Macintosh*", or between "*Iranian*" and "*IRIB*" can be represented by rules which could not be extracted using WordNet or any other resource. This confirms our hypothesis that increasing the coverage using a context sensitive approach in rule extraction, may result in a better performance in the RTE task.

It's worth mentioning that these results can be easily replicated by downloading EDITS[5], thus representing a valuable starting point for further research and potential improvements.

Though encouraging and substantially confirming our working hypothesis, the observed performance increase is lower than expected. This might be due to the difficulty of exploiting lexical information when the TED algorithm is used. Often, valid and reliable rules that could be potentially applied to reduce the distance between T and H are often ignored because of the syntactic constraints imposed

[3]dependency trees obtained by using the Stanford Parser available at `http://nlp.stanford.edu/software/lex-parser.shtml`.

[4]The threshold was empirically estimated running a set of experiments to select the subset of rules that best performs on training data. This could result in a good trade-off between precision and coverage of the extracted rules. Though higher thresholds could increase precision, leading to more accurate rules, the reduced amount of extracted rules would directly affect coverage, causing an overall performance decrease.

[5]The current release of the EDITS package, together with all the rule repositories and the cost scheme we used for our experiments are available at `http://edits.fbk.eu`

	VO		WN		PROX		DEP		WIKI	
	Extracted	Retained	Extr.	Ret.	Extr.	Ret.	Extr.	Ret.	Extr.	Ret.
Coverage	0.08%	0.08%	0.4%	0.4%	3%	0.09%	2%	1%	83%	24%

Table 3: Coverage of rule repositories over the RTE-5 dataset.

by the algorithm. To verify this hypothesis we performed another experiment, comparing the different resources in terms of potential coverage of the pairs in the same dataset, independently from any RTE algorithm.

4.4. Estimating coverage over RTE-5 data

As previously mentioned, one of the main motivations to use Wikipedia as a source to extract entailment rules is the large coverage of the available RTE data. To prove our claim, we performed an analysis on the coverage in the rules extracted and retained from each resource. To this aim, we count the number of pairs in the RTE-5 data which contain rules present in the WordNet, VerbOcean, Lin Dependency/Proximity, and Wikipedia repositories.

Following our definition of entailment rules, we compute the total of RTE-5 T-H pairs for which rules such as $w_1 \Rightarrow w_2$ match a word w_1 in T and a word w_2 in H. Then, we estimated the number of rules that can be extracted from each resource and the rules that were retained in our experiments. Table 3 shows the coverage of the content words of the extracted rules for RTE-5 from the different resources. As can be seen, the coverage of Wikipedia is the highest amongst other resources.

5. Conclusion and future work

In this paper we focused on the role played by lexical knowledge in the RTE task. After a short overview of the main knowledge sources used in current RTE systems, we proposed a method for extracting lexical entailment rules from Wikipedia. The proposed acquisition method aims at maximizing two key features of the resulting entailment rules: *coverage* (*i.e.* the proportion of rules successfully applied over a dataset of TE pairs), and *context sensitivity* (*i.e.* the proportion of rules applied in appropriate contexts). As regards evaluation, we first reported the results of a comparison over the RTE-5 dataset between rule repositories acquired from Wikipedia, WordNet, VerbOcean, and Lin's dependency and proximity thesauri.

Though accuracy improvement is smaller than expected, the Wikipedia repository showed to systematically outperform the others. To check if the small accuracy improvement was due a to sub-optimal use of the rules by the Tree Edit Distance algorithm available in EDITS, we carried out a second experiment comparing the different resources in terms of potential coverage of the pairs in the same dataset, independently from any RTE algorithm. Also in this case Wikipedia rules achieved the highest result observed, demonstrating the effectiveness of our rule extraction method. Building on these findings, future work will concentrate on defining more flexible algorithms, capable of exploiting the full potential (both in terms of coverage and in terms of context sensitivity) offered by Wikipedia rules.

6. References

L. Bentivogli, I. Dagan, H. Trang Dang, D. , Giampiccolo, and B. Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC 2009 Notebook Papers*.

T. Chklovski and P. Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

I. Dagan and O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

A. Iftene and M.A. Moruz. 2009. Uaic participation at rte5. In *TAC 2009 Notebook Papers*.

J. Kazama and K. Torisawa. 2007. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of EMNLP-CoNLL*.

M. Kouylekov and B. Magnini. 2006. Building a large-scale repository of textual entailment rules. In *Proceedings of LREC*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

E. Marsi, E. Krahmer, W. Bosma, and M. Theune. 2006. Normalized alignment of dependency trees for detecting textual entailment. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop*.

Y. Mehdad, M. Negri, Kouylekov M., E. Cabrio, and B. Magnini. 2009. Using lexical resources in a distance based approach to rte. In *TAC 2009 Notebook Papers*, Gaithersburg, MD, US.

S. Mirkin, I. Dagan, and E. Shnarch. 2009. Evaluating the inferential utility of lexical-semantic resources. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

D. Moldovan and A. Novischi. 2002. Lexical chains for question answering. In *Proceedings of COLING*.

M. Negri, M. Kouylekov, B. Magnini, Mehdad Y., and E. Cabrio. 2009. Towards extensible textual entailment engines: the edits package. In *Proceedings of AI*IA*.

M. Tatu, B. Iles, J. Slavick, A. Novischi, and D. Moldovan. 2006. Cogex at the second recognizing textual entailment challenge. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop*.

R. Wang, Zhang Y., and G. Neumann. 2009. A joint syntactic-semantic representation for recognizing textual relatedness. In *TAC 2009 Notebook Papers*.