Building a Large-Scale Repository of Textual Entailment Rules

Milen Kouylekov[*] [†], **Bernardo Magnini**[†]

[*]University of Trento
via Sommarive 14, 38050, Povo (Trento), Italy
milen@kouylekov.net
[†] ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
via Sommarive 18, 38050, Povo (Trento), Italy
magnini@itc.it

Abstract

Entailment rules are rules where the left hand side (LHS) specifies some knowledge which *entails* the knowledge expressed in the RHS of the rule, with some degree of confidence. Simple entailment rules can be combined in complex entailment chains, which in turn are at the basis of entailment-based reasoning, which has been recently proposed as a pervasive and application independent approach to Natural Language Understanding. We present the first release of a large-scale repository of entailment rules at the lexical level, which have been derived from a number of available resources, including WordNet and a word similarity database. Experiments on the PASCAL-RTE dataset show that this resource plays a crucial role in recognizing textual entailment.

1. Introduction

Recognizing Textual Entailment (RTE) is attracting increasing interest in Computational Linguistic. Among other events, the PASCAL-RTE evaluation campaign in 2005 and the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment, showed that Textual Entailment is a core task in several application scenarios (e.g. question answering, information extraction, automatic summarization) where textual inferences are necessary to address the language variability problem (i.e. the fact that the same meaning can be expressed with different lexical and syntactic structures).

The RTE task takes as input a T/H pair and consists in automatically determining whether an entailment relation holds between T and H or not. The task covers almost all the phenomena in language variability: entailment can be due to lexical variations, to syntactic variation, to semantic inferences or to complex combinations of all such levels. As a consequence of the complexity of the task, one of the crucial aspects for any RTE system is the amount of linguistic and world knowledge required for filling the gap between T and H.

A crucial role in textual entailment is played by entailment rules (Dagan and Glickman, 2004), which consist of an entailing template (the left hand side of the rule) and an entailed template the (right hand side of the rule), sharing the same variable scope. Prior or contextual (i.e. posterior) probabilities are assigned to the rule. As an example, a lexical entailment rule such as:

$$[a\ shootout\ that\ killed\ Y] \xrightarrow{entails} [Y\ died] \qquad (1)$$

will help to detect an entailment relation at the lexical-syntactic level between the following portions of text:

1. *T* - The two suspect belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.
   H - Cardinal Juan Jesus Posadas Ocampo died in 1993.

However, for concrete applications, a huge amount of such entailment rules is necessary. To this aim, we have been investigating a number of techniques to automatically derive entailment rules from already existing linguistic resources, including WordNet and a word similarity database. The contribution of the paper is twofold: on the one side, we we present the first release of a large-scale repository of entailment rules at the lexical level; on the other side, from the perspective of textual entailment algorithms, we provide a clear and homogeneous framework for the evaluation of the contribution of each resource to provide entailment rules.

The paper is structured as follows. Section 2 reports on the two resources we have considered, showing how we extract entailment rules from them. In Section 3 we address the issue of estimating a confidence score for entailment rules, showing that a probabilistic framework based on the impact of rules on entailment recognition is an effective solution. Section 4 and 5, respectively, show the experimental setting and the results we have obtained. Finally, Section 6 sum up our work and suggest future directions of work.

2. Resources for Entailment Rules

This section presents the resources we have considered in order to build a database of entailment rules. We have derived entailment rules from two available lexical resources, i.e. WordNet and a word-similarity database.

2.1. WordNet

WordNet (Fellbaum, 1998) is a lexical database which includes lexical and semantic relations among word senses. Originally developed for English, versions of WordNet are currently available also for other languages (e.g. Spanish, German and Italian).

We have defined entailment rules over WordNet (we used version 2.0) considering a subset of the relations among

synsets. More precisely, if A and B are synsets in WordNet, then we have derived an entailment rule in the following cases:

- if A is hypernym of B; as an example, the following rules are derived, following the hypernym chain:

  $[terorist] \xrightarrow{entails} [radical] \xrightarrow{entails} [person]$

- if A is synonym of B, as shown in the example below:

  $[kill] \xrightarrow{entails} [shoot_down]$

- if A entails B, as shown in the example below:

  $[kill] \xrightarrow{entails} [perish]$

- if A pertains to B, , as shown in the example below:

  $[terrorist_cell] \xrightarrow{entails} [terrorist]$

2.2. Word Similarity Database

As for the second source of entailment rules we have used the use of a thesaurus of dependency-based relations available at *http://www.cs.ualberta.ca/lindek/downloads.htm*. Dependency relations are represented as triples, each consisting of a head, a dependency type and a modifier. When estimating the similarity among two words, such triples can be viewed as features for the head and the modifiers in the triples.
For each word, the thesaurus lists up to 200 most similar words and their similarities, estimated on a parsed corpus using frequency counts of the dependency triples.
A complete review of the method, including a comparison with different approaches, is presented in (Lin, 1949).
As an example of the rules extracted from the similarity database, the following rule has been derived from ....

$[terrorist] \xrightarrow{entails} [LebaneseShiiteMoslem]$

3. Estimating Rule Prior Probabilities

In this section we propose an approach for calculating the prior probabilities of the entailment rules extracted from both the available resources. The approach is based on the intuition that the confidence of a rule is proportional to the benefit it brings in recognizing an entailment relation in a T/H pair: the higher the contribution of the rule, the higher the probability that the rule represents a good entailment relation. According to this intuition we have set up an experiment where the entailment rules of the two resources have been given to a RTE system (described in Section 3.2), whose performance are evaluated on the PASCAL-RTE1 dataset (described in Section 3.1).
We consider an entailment rule as correct if it facilitates the system to correctly identify whether an entailment relation exists between an *T-H* pair. In this way the entailment rules are splitted into two categories: correct, if they improve the system accuracy, and incorrect, if they do not. Table 1 lists the different ways in which an entailment rule can affect an entailment recognition system. The first column represents the entailment relation between the *T-H* pair. The second and third columns represent the result assigned by the system, respectively without and with entailment rules. The last column is the category we assign to a rule according to its behavior.

T-H pair	No-rules	With-rules	Category
true	true	true	correct
true	true	false	incorrect
true	false	true	correct
true	false	false	incorrect
false	true	true	incorrect
false	true	false	correct
false	false	true	incorrect
false	false	false	correct

Table 1: Rule classification according to system results.

As an example, a rule is a correct one if it allows the system to classify a good *T-H* pair as positive, supposed that without such rule the system would have judged the *T-H* pair as non-entailment (see line number 3 of Table 1).
We define the *accuracy* of a set of entailment rules derived from a resource as the proportion of the correct entailment rules against the total number of rules. Accordingly, we define the *prior probability* of an entailment rule extracted from a resource R as:

$$P_{prior}(rule_R) = P_{correct(D,S)}(rule_R) \quad (2)$$

where D is a dataset of T/H pairs and S is an entailment recognition system that uses entailment rules. This probability is proportional to the accuracy of the resource.

3.1. Dataset

To collect a set of T/H pairs we used an the dataset provided from the Pascal Recognizing Textual Entailment Challenge (PASCAL-RTE). The PASCAL-RTE challenge is a recent evaluation campaign which attracted considerably attention (16 different groups participated to the 2005 campaign). The view underlying the RTE challenge (Dagan et al., 2005) is that different natural language processing applications, including Question Answering (QA), Information Extraction (IE), (multi-document) summarization, and Machine Translation (MT), have to address the language variability problem and would benefit from textual entailment in order to recognize that a particular target meaning can be inferred from different text variants. The different applications address the problem with application-oriented manners and methods and the impact of RTE is evaluated on the final application performance.
The PASCAL-RTE campaign was based on a human annotated dataset of T H pairs, collected from different text processing applications. Each pair corresponds to a success or failure case of an actual application. The collected examples represent a range of different levels of entailment reasoning based on lexical syntactic logical and word knowledge, at different levels of difficulty. The pairs are taken from seven different application scenario:

- Information Retrieval - queries selected by examining prominent sentences in news stories.
- Comparable Documents - comparable news articles that cover a common story.
- Reading Comprehension - exercises in human language teaching.
- Question Answering - Question from CLEF-QA (Cross Language evaluation Forum) and TREC (Text Retrieval Conference).
- Information Extraction - dataset of annotated relations *kill* and *birth place*
- Machine Translation - automatic translations.
- Paraphrase Acquisition.

3.2. RTE system

The entailment recognition system we used is described in (Kouleykov and Magnini, 2005). In this system we adopted a tree edit distance algorithm applied to the syntactic representations (i.e. dependency trees) of both T and H.
According to our approach, T entails H if there exists a sequence of transformations applied to T such that we can obtain H with an overall cost below a certain threshold. The underlying assumption is that pairs that exhibits an entailment relation have a low cost of transformation. The kind of transformations we can apply (i.e. deletion, insertion and substitution) are determined by a set of predefined entailment rules, which also determine a cost for each editing operation.
We have implemented the tree edit distance algorithm described in the paper from (Zhang and Shasha, 1990) and apply it to the dependency trees derived from T and H. Edit operations are defined at the level of single nodes of the dependency tree (i.e. transformations on subtrees are not allowed in the current implementation). Since the (Zhang and Shasha, 1990) algorithm does not consider labels on edges, while dependency trees provide them, each dependency relation R from a node A to a node B has been re-written as a complex label B-R concatenating the name of the destination node and the name of the relation. All nodes except the root of the tree are relabeled in this way. The algorithm is directional: we aim to find the better (i.e. less costly) sequence of edit operation that transform T (the source) into H (the target). According to the constraints described above, the following transformations are allowed:

- **Insertion**: insert a node from the dependency tree of H into the dependency tree of T. When a node is inserted it is attached with the dependency relation of the source label.
- **Deletion**: delete a node N from the dependency tree of T. When N is deleted all its children are attached to the parent of N. It is not required to explicitly delete the children of N as they are going to be either deleted or substituted on a following step.
- **Substitution**: change the label of a node N1 in the source tree (the dependency tree of T) into a label of a node N2 of the target tree (the dependency tree of H). Substitution is allowed only if the two nodes share the same part-of-speech. In case of substitution the relation attached to the substituted node is changed with the relation of the new node.

4. Experiments and Results

In this section we describe the experimental setting used to estimate prior probabilities for the entailment rules derived both from WordNet and the word similarity database. We have experimented three system settings.

System 1: Tree Edit Distance Baseline. In this configuration, considered as a baseline for the Tree Edit Distance approach, no entailment rule is used, and the cost of the three edit operations is set as follows:
Deletion: always 0;
Insertion: the idf of the word to be inserted;
Substitution: 0 if $w_1 = w_2$, infinite in all the other cases.
In this configuration the system just needs a non-annotated corpus for estimating the idf of the word to be inserted. The corpus is composed of 4.5 million news documents from the CLEF-QA (Cross Language evaluation Forum) and TREC (Text Retrieval Conference) collections. Deletion is 0 because we expect much more deletions that insertions, due to the fact that *T* is longer than *H*.

System 2: Word Similarity Database. This is the same as System 1, but we estimate the cost of substitution using the entailment rules extracted from the word similarity database described in Section 2.1.
Deletion: always 0;
Insertion: the idf of the word to be inserted;
Substitution: 0 if $w_1 = w_2$ or if an entailment rule between w_1 and w_2 exists in the similarity database, infinite in all the other cases.

System 3: WordNet. This is the same as System 1, but we estimate the cost of substitutions using the entailment rules extracted from WordNet relations, as described in Section 2.2.
Deletion: always 0;
Insertion: the idf of the word to be inserted;
Substitution: 0 if $w_1 = w_2$ or if an entailment rule between w_1 and w_2 can be found in WordNet, infinite in all the other cases.

4.1. Results

Table 2 reports on the results obtained by the three systems on the PASCAL-RTE1 dataset:
The usability of the entailment rules is calculated for both resources. The results show that rules extracted from WordNet have an higher usability, which is reflected on the performance of the system using them.
The similarity database used in System 2 increased the performance of the baseline system with successful substitutions using 113 rules made by the algorithm from. The

System	Accuracy	#rules	#correct	rule usability
1	0.560	0	0	0
2	0.566	224	144	0.59
3	0.72	104	70	0.67

Table 2: Results on the PASCAL-RTE 1 dataset.

usability of the resource low as the number of the correct rules is close to the number of incorrect.
The system based on Wordnet entailment rules, i.e. System 3, also increases the performance against the baseline system. Although the number of entailment rules used is lower than in System 2 (i.e. only 104), the accuracy is the highest achived using tree edit distance and it is 0.012 more than the baseline. In comparison with System 2 it makes less substitutions. This shows that increasing the number of substations does not mean an automatic increase of the performance. But acquiring highier quility rules can increase the performance of the system.

5. Conclusions and Future Work

We have presented the first release of a large-scale repository of entailment rules, which have been automatically derived both from WordNet and a word similarity database. We have presented an approach for assigning to each rule a prior probability, representing the strength of the entailment relation.
As for future work, a statistical validation of the extracted entailment rules, either using a large scale text corpus or the Internet, can improve the quality of the extracted rules. In this line, (Szpektor et al., 2004) presents an approach based on search engines for template validation.
We also plan to extend the usage of WordNet as an entailment resource. The potential of Extended WordNet (Harabagiu et al., 1999) as an entailment resource is discussed in (Moldovan et al., 2003) and (Moldovan and Rus, 2001). Other resources (e.g.paraphrases in (Lin and Pantel, 2001), entailment patterns as acquired in (Szpektor et al., 2004)) could significantly widen the application of entailment rules and, consequently, improve performances. We estimated that for about 40% of the true positive pairs the system could have used entailment rules found in entailment and paraphrasing resources. As an example, the pair 565:

> T - Soprano's Square: Milan, Italy, home of the famed La Scala opera house, honored soprano Maria Callas on Wednesday when it renamed a new square after the diva.
>
> H - La Scala opera house is located in Milan, Italy.

could be successfully solved using a paraphrase pattern such as *Y home of X <=> X is located in Y*, which can be found in (Lin and Pantel, 2001). However, in order to use this kind of entailment rules and calculate their prior probabilities , it would be necessary to extend the "single node" implementation of tree edit distance to address editing operations among sub-trees.

6. References

Ido Dagan and Oren Glickman. 2004. Generic applied modeling of language variability. *In Proceedings of PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble

Ido Dagan, Oren Glickman and Bernardo Magnini. 2005. The PASCAL Recognizing Textual Entailment Challenge *In Proceedings of PASCAL Workshop on Recognizing Textual Entailment* Southampton, UK

Christiane Fellbaum. 1998. WordNet, an electronic lexical database *MIT Press, 1998*

Sandra Harabagiu, George Miller and Dan Moldovan. 1999. WordNet 2 - A morphologically and Semantically Enhanced Resource. *In proceeding of ACL-SIGLEX99*, Marylend

Milen Kouleykov and Bernardo Magnini. 2005. Combining Lexical Resources with Tree Edit Distance for Recognizing Textual Entailment. *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop*,LNAI, Springer

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin

Dekang Lin and Patrick Pantel. 2001 Discovery of inference rules for Question Answering *Natural Language Engineering, 7(4)*, pages 343-360

Dan Moldovan and Vasile Rus. 2001. Logic Form Transformation and it's Applicability in Question Answering. *In proceedings of ACL*

Dan Moldovan, Sandra Harabagio, Roxana Girju, Paul Morarescu, Lacatsu and Adrian Novischi. 2003. LCC Tools for Question Answering. *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002-2003)*.

Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling Web-based Acquisition of Entailment Relations. *In Proceedings of EMNLP-04 - Empirical Methods in Natural Language Processing*, Barcelona

Kaizhong Zhang ,Dennis Shasha. 1990. Fast algorithm for the unit cost editing distance between trees. *Journal of algorithms, vol. 11, p. 1245-1262*