Extracting Context-Rich Entailment Rules from Wikipedia Revision History

Elena Cabrio
INRIA
2004, route de Lucioles BP93
06902 Sophia Antipolis, France.
elena.cabrio@inria.fr

Bernardo Magnini
FBK
Via Sommarive 18
38100 Povo-Trento, Italy.
magnini@fbk.eu

Angelina Ivanova
University of Oslo
Gaustadalléen 23B
Ole-Johan Dahls hus
N-0373 Oslo, Norway.
angelii@ifi.uio.no

Abstract

Recent work on Textual Entailment has shown a crucial role of knowledge to support entailment inferences. However, it has also been demonstrated that currently available entailment rules are still far from being optimal. We propose a methodology for the automatic acquisition of large scale context-rich entailment rules from Wikipedia revisions, taking advantage of the syntactic structure of entailment pairs to define the more appropriate linguistic constraints for the rule to be successfully applicable. We report on rule acquisition experiments on Wikipedia, showing that it enables the creation of an innovative (i.e. acquired rules are not present in other available resources) and good quality rule repository.

1 Introduction

Entailment rules have been introduced to provide pieces of knowledge that may support entailment judgments (Dagan *et al.*, 2009) with some degree of confidence. More specifically, an entailment rule is defined (Szpektor *et al.*, 2007) as a directional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees). The left-hand side (LHS) of the pattern entails the right-hand side (RHS) of the same pattern under the same variable instantiation. Given the Text-Hypothesis pair (T-H) in Example 1:

Example 1.

T: *Dr. Thomas Bond established a hospital in Philadelphia for the reception and cure of poor sick persons.*

H: *Dr. Bond created a medical institution for sick people.*

a (directional) lexical rule like:

1) **LHS:** *hospital* ⇒ **RHS:** *medical institution*
probability: 0.8

brings to a TE system (aimed at recognizing that a particular target meaning can be inferred from different text variants in several NLP application, e.g. Question Answering or Information Extraction) the knowledge that the word *hospital* in Text can be aligned, or transformed, into the word *medical institution* in the Hypothesis, with a probability 0.8 that this operation preserves the entailment relation among T and H. Similar considerations apply for more complex rules involving verbs, as:

2) **LHS:** *X establish Y* ⇒ **RHS:** *X create Y*
probability: 0.8

where the variables may be instantiated by any textual element with a specified syntactic relation with the verb. Both kinds of rules are typically acquired either from structured sources (e.g. WordNet (Fellbaum, 1998)), or from unstructured sources according for instance to distributional properties (e.g. DIRT (Lin and Pantel, 2001)). Entailment rules should typically be applied only in specific contexts, defined in (Szpektor *et al.*, 2007) as *relevant contexts*. Some existing paraphrase and entailment acquisition algorithms add constraints to the learned rules (e.g. (Sekine, 2005), (Callison-Burch, 2008)), but most do not. Because of a lack of an adequate representation of the linguistic context in which the

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 34–43,
Jeju, Republic of Korea, 8-14 July 2012.

rules can be successfully applied, their concrete use reflects this limitation. For instance, rule 2 (extracted from DIRT) fails if applied to "The mathematician established the validity of the conjecture", where the sense of *establish* is not a synonym of *create* (but of *prove, demonstrate*), decreasing system's precision. Moreover, these rules often suffer from lack of directionality, and from low accuracy (i.e. the strength of association of the two sides of the rule is often weak, and not well defined). Such observations are also in line with the discussion on ablation tests carried out at the last RTE evaluation campaigns (Bentivogli *et al.*, 2010).

Additional constraints specifying the variable types are therefore required to correctly instantiate them. In this work, we propose to take advantage of Collaboratively Constructed Semantic Resources (CSRs) (namely, Wikipedia) to mine information useful to context-rich entailment rule acquisition. More specifically, we take advantage of material obtained through Wikipedia revisions, which provides at the same time real textual variations from which we may extrapolate the relevant syntactic context, and several simplifications with respect to alternative resources. We consider T-H pairs where T is a revision of a Wikipedia sentence and H is the original sentence, as the revision is considered more informative then the revised sentence.

We demonstrate the feasibility of the proposed approach for the acquisition of context-rich rules from Wikipedia revision pairs, focusing on two case studies, i.e. the acquisition of entailment rules for *causality* and for *temporal expressions*. Both phenomena are highly frequent in TE pairs, and for both there are no available resources yet. The result of our experiments consists in a repository that can be used by TE systems, and that can be easily extended to entailment rules for other phenomena.

The paper is organized as follows. Section 2 reports on previous work, highlighting the specificity of our work. Section 3 motivates and describes the general principles underlying our acquisition methodology. Section 4 describes in details the steps for context-rich rules acquisition from Wikipedia pairs. Section 5 reports about the experiments on causality and temporal expressions and the obtained results. Finally, Section 6 concludes the paper and suggests directions for future improvements.

2 Related work

The use of Wikipedia revision history in NLP tasks has been previously investigated by a few works. In (Zanzotto and Pennacchiotti, 2010), two versions of Wikipedia and semi-supervised machine learning methods are used to extract large TE data sets similar to the ones provided for the RTE challenges. (Yatskar *et al.*, 2010) focus on using edit histories in Simple English Wikipedia to extract lexical simplifications. Nelken and Yamangil (2008) compare different versions of the same document to collect users' editorial choices, for automated text correction, sentence compression and text summarization systems. (Max and Wisniewski, 2010) use the revision history of French Wikipedia to create a corpus of natural rewritings, including spelling corrections, reformulations, and other local text transformations. In (Dutrey *et al.*, 2011), a subpart of this corpus is analyzed to define a typology of local modifications.

Because of its high coverage, Wikipedia is used by the TE community for lexical-semantic rules acquisition, named entity recognition, geographical information[1] (e.g. (Mehdad *et al.*, 2009), (Mirkin *et al.*, 2009), (Iftene and Moruz, 2010)), i.e. to provide TE systems with world and background knowledge. However, so far it has only been used as source of factual knowledge, while in our work the focus is on the acquisition of more complex rules, concerning for instance spatial or temporal expressions.

The interest of the research community in producing specific methods to collect inference and paraphrase pairs is proven by a number of works in the field, which are relevant to the proposed approach.

As for paraphrase, Sekine's Paraphrase Database (Sekine, 2005) is collected using an unsupervised method, and focuses on phrases connecting two Named Entities. In the Microsoft Research Paraphrase Corpus[2], pairs of sentences are extracted from news sources on the web, and manually annotated. As for rule repositories collected using distributional properties, DIRT (Discovery of Inference Rules from Text)[3] is a collection of inference rules

[1] http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources

[2] http://research.microsoft.com/en-us/downloads

[3] http://www.aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection

(Lin and Pantel, 2001), obtained extracting binary relations between a verb and an object-noun (or a small clause) from dependency trees. Barzilay and Lee (2003) present an approach for generating sentence level paraphrases, learning structurally similar patterns of expression from data and identifying paraphrasing pairs among them using a comparable corpus. Since the data sets cited so far are paraphrase collections, rules are bidirectional, while one of the peculiarities of the entailment relation is the directionality, addressed in our work.

Aharon *et al.* (2010) presented FRED, an algorithm for generating entailment rules between predicates from FrameNet. Moreover, the TEASE collection of entailment rules (Szpektor *et al.*, 2004) consists of 136 templates provided as input, plus all the learned templates. Their web-based extraction algorithm is applied to acquire verb-based expressions. No directionality of the pairs is specified, but additional guessing mechanisms it are proposed. In (Szpektor and Dagan, 2008), two approaches for unsupervised learning of *unary* rules (i.e. between templates with a single variable) are investigated.

In (Zhao *et al.*, 2009), a pivot approach for extracting paraphrase patterns from bilingual parallel corpora is presented, while in (Callison-Burch, 2008) the quality of paraphrase extraction from parallel corpora is improved by requiring that phrases and their paraphrases have the same syntactic type. Our approach is different from theirs in many respects: their goal is paraphrase extraction, while we are extracting directional entailment rules; as textual resources for pattern extraction they use parallel corpora (using patterns in another language as pivots), while we rely on monolingual Wikipedia revisions (taking benefit from its increasing size); the paraphrases they extract are more similar to DIRT, while our approach allows to focus on the acquisition of rules for specific phenomena frequent in entailment pairs, and not covered by other resources.

3 General methodology

The general approach we have implemented is based on the idea that, given a *seed word*, we extract all the entailment rules from Wikipedia revision pairs where the seed word appears as the head of the rule either in T or H. The head is the non-variable part of the rule on which the other parts depend (i.e. the word *establish* is the head of rule 2).

Entailment judgment. A Wikipedia revision may be consistent with the original sentence, bringing to an entailment relation, or it may introduce inconsistency, expressing a contradiction w.r.t. the original sentence. We manually checked a sample of revision pairs ($\sim$200), and we found out that in about 95% of the revisions entailment is preserved, in line with (Zanzotto and Pennacchiotti, 2010). We assume this one as the default case in our experiments.

Monothematic pairs. The capability of automatic extraction of entailment rules is affected by the complexity of the pairs from which we extract the rules. In our experiments we take advantage of revision pairs with minimal difference between T and H, and we assume that for such pairs we have only one rule to extract. Under this perspective, T-H pairs derived from Wikipedia revisions have strong similarity with *monothematic pairs* (i.e. pairs where the entailment judgment is due to only one linguistic phenomenon, as suggested in (Bentivogli *et al.*, 2010)). Section 4.2 describes the algorithm for filtering out revision pairs with more than one phenomenon.

Directionality. A Wikipedia revision, in principle, may be interpreted as either T entailing H, or as H entailing T. However, through a manual inspection of a revision sample ($\sim$200 pairs), it came out that in most of the cases the meaning of the revised sentence (T) entails the meaning of the original one (H). Given such observation, for our experiments (Sections 4 and 5) we assume that for all revision pairs, the revised sentence (T) entails the original one (H).

Context of a rule. We have defined the notion of context of a rule R as a set of morpho-syntactic constraints C over the application of R in a specific T-H pair. Ideally, the set of such constraints should be the minimal set of constraints over R such that the proportion of successful applications of R is maximized (e.g. the precision-recall mean is highest). Intuitively, given an entailment rule, in absence of constraints we have the highest recall (the rule is always applied when the LHS is activated in T and the RHS is activated in H), although we may find cases of wrong application of the rule (i.e. low precision). On the other side, as syntactic constraints are

required (e.g. the subject of a verb has to be a noun) the number of successful applications increases, although we may find cases where the constraints prevent the correct application (e.g. low recall).

In the absence of a data set where we can empirically estimate precision and recall of rule application, we have approximated the ideal context on the basis of linguistic intuitions. More specifically, for different syntactic heads of the rules, we define the most appropriate syntactic constraints through a search algorithm over the syntactic tree produced on T and H (see Section 4.4 for a detailed explanation).

4 Entailment rules acquisition

In the next sections, the steps for the acquisition of rules from Wikipedia pairs are described in detail.

4.1 Step 1: preprocessing Wikipedia dumps

We downloaded two dumps of the English Wikipedia (one dated 6.03.2009, *Wiki 09*, and one dated 12.03.2010, *Wiki 10*).[4] We used the script *WikiExtractor.py*[5] to extract plain text from Wikipedia pages, discarding any other information or annotation, but keeping the reference to the original document. For our goal, we consider only non-identical documents present in both *Wiki 09* and *Wiki 10* (i.e. 1,540,870 documents).

4.2 Step 2: extraction of entailment pairs

For both *Wiki 09* and *Wiki 10* each document has been sentence-splitted, and the sentences of the two versions have been aligned to create pairs. To measure the similarity between the sentences in each pair, we adopted the *Position Independent Word Error Rate (PER)* (Tillmann *et al.*, 1997), a metric based on the calculation of the number of words which differ between a pair of sentences (*diff* function in (1)). Such measure is based on Levenshtein distance, but works at word level, and allows for reordering of words and sequences of words between the two texts (e.g. a translated text *s* and a reference translation *r*). It is expressed by the formula:

$$PER(s,r) = \frac{diff(s,r)+diff(r,s)}{\|r\|} \quad (1)$$

[4] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

Pairs are clustered according to different thresholds:

- Pairs composed by identical sentences were discarded; if only one word was different in the two sentences, we checked if it was a typo correction using (Damerau, 1964) distance. If that was the case, we discarded such pairs as well.
- Pairs in which one of the sentences contains the other one, meaning that the users added some information to the new version, without modifying the old one (set *a*: 1,547,415 pairs).
- Pairs composed by very similar sentences, where users carried out minor editing ($PER < 0.2$) (set *b*: 1,053,114 pairs). We filtered out pairs where differences were correction of misspelling and typos, and two-word sentences.
- Pairs composed by sentences where major editing was carried out ($0.2 < PER < 0.6$), but still describe the same event (set *c*: 2,566,364).
- Pairs in which the similarity between sentences is low ($PER > 0.6$) were discarded.

To extract entailment rules, we consider only the pairs contained in *set b*. For each pair, we intuitively set the sentence extracted from *Wiki 10* as the Text, since we assume that it contains more (and more precise) information w.r.t. the sentence extracted from *Wiki 09*. We set the sentence extracted from *Wiki 09* as the Hypothesis (see Examples 2 and 3).

Example 2.

T: *The Oxford Companion to Philosophy says "there is no single defining position that all anarchists hold [...]"*
H: *According to the Oxford Companion to Philosophy "there is no single defining position that all anarchists hold [...] "*

Example 3.

T: *Bicycles are used by all socio-economic groups because of their convenience [...].*
H: *Bicycles are used by all socio-economic groups due to their convenience [...].*

4.3 Step 3: extraction of entailment rules

Pairs in *set b* are collected in a data set, and processed with the Stanford parser (Klein and Manning,

2003); chunks are extracted from each pair using the script *chunklink.pl*.[6] The assumption underlying our approach is that the difference between T and H (i.e. the editing made by the user on a specific structure) can be extracted from such pairs and identified as an entailment rule. The *rule extraction* algorithm was implemented to this purpose. In details, for each sentence pair the algorithm iteratively compares the chunks of T and H to extract the ones that differ. It can be the case that several chunks of H are identical to a given chunk of T, as in:

```
T:<NP>[The DT][Oxford NNP][Companion NNP]
 </NP><PP>[to TO]</PP> <NP>[Philosophy NNP]
 </NP><VP>[says VBZ]</VP>...

H:<PP>[According VBG]</PP><PP>[to TO]</PP>
 <NP>[the DT][Oxford NNP][Companion NNP]</NP>
 <PP>[to TO]</PP><NP>[Philosophy NNP]</NP>...
```

Therefore, to decide for instance which chunk <PP>[to TO]</PP> from H corresponds to the identical chunk in T, the algorithm checks if the previous chunks are equal as well. If this is the case, such chunks are matched. In the example above, the second chunk <PP>to</PP> from H is considered as a good match because previous chunks in T and H are equal as well (<NP>the Oxford Companion</NP>). If the previous chunks in T and H are not equal, the algorithm keeps on searching. If such match is not found, the algorithm goes back to the first matching chunk and couples the chunk from T with it. Rules are created setting the unmatched chunks from T as the left-hand side of the rule, and the unmatched chunks from H as the right-hand side of the rule. Two consecutive chunks (different in T and H) are considered part of the same rule. For instance, from Examples 2 and 3:

2) <LHS> says </LHS>
<RHS> according to </RHS>

3) <LHS> because of </LHS>
<RHS> due to </RHS>

On the contrary, two non consecutive chunks generate two different entailment rules.

[6] http://ilk.uvt.nl/team/sabine/chunklink/README.html

4.4 Step 4: rule expansion with minimal context

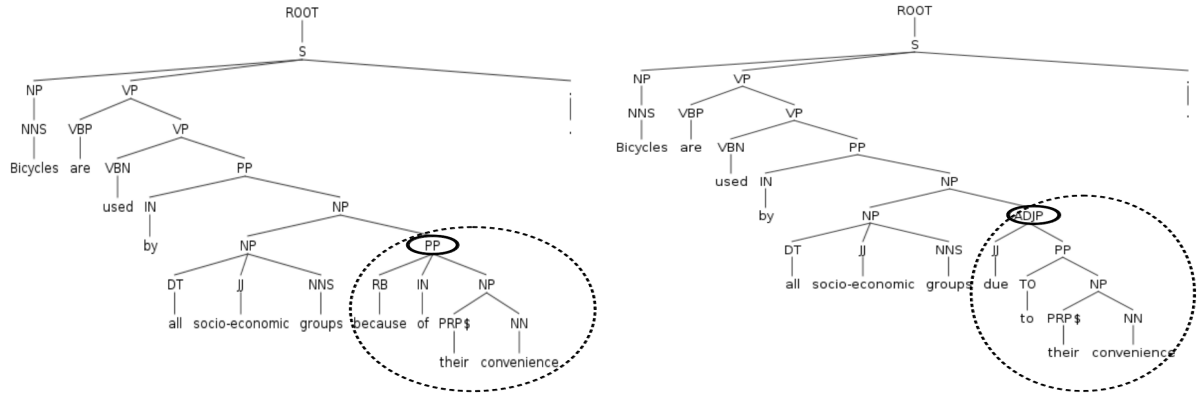
As introduced before, our work aims at providing precise and context-rich entailment rules, to maximize their correct application to RTE pairs. So far, rules extracted by the rule extraction algorithm (Section 4.3) are too general with respect to our goal.

To add the minimum context to each rule (as discussed in Section 3), we implemented a *rule expansion* algorithm: both the file with the syntactic representation of the pairs (obtained with the Stanford parser), and the file with the rules extracted at Step 3 are provided as input. For every pair, and separately for T and H, the words isolated in the corresponding rule are matched in the syntactic tree of that sentence, and the common subsumer node is detected. Different strategies are applied to expand the rule, according to linguistic criteria. In details, if the common subsumer node is *i)* a Noun Phrase (NP) node, the rule is left as it is; *ii)* a Prepositional Phrase node (PP), all the terminal nodes of the subtree below PP are extracted; *iii)* a clause introduced by a subordinating conjunction (SBAR), all the terminal nodes of the subtree below SBAR are extracted; *iv)* an adjectival node (ADJP), all the terminal nodes of the tree below the ADJP node are extracted; *v)* a Verbal Phrase node (VP), the dependency tree under the VP node is extracted.

For Example 3 (see Figure 1), the LHS of the rule *because of* is matched in the syntactic tree of T and the prepositional phrase (PP) is identified as common subsumer node. All the terminal nodes and the PoS of the tree below PP are then extracted. The same is done for the RHS of the rule, where the common subsumer node is an adjectival phrase (ADJP).

5 Experiments and results

In the previous section, we described the steps carried out to acquire context-rich entailment rules from Wikipedia revisions. To show the applicability of the adopted methodology, we have performed two experiments focusing, respectively, on entailment rules for *causality* and *temporal expressions*. In particular, as case studies we chose two seeds: the conjunction *because* to derive rules for causality, and the preposition *before* for temporal expressions.



(a) LHS rule

(b) RHS rule

Figure 1: Rule expansion with minimal context (Example 3)

causality (*because*)	**temporal exp. (*before*)**
(PP(RB because)(IN of)(NP(JJ)(NNS)) ⇒ (ADJP(JJ due)(PP(TO to)(NP(JJ)(NNS)))) e.g.: *because of contractual conflicts ⇒ due to contractual conflicts*	(SBAR(IN before)(S)) ⇒ (ADVP(RB prior)(PP(TO to)(S) e.g.: *before recording them ⇒ prior to recording them*
(SBAR(IN because)(S)) ⇒ (VP(PP(IN on)(NP(DT the) (NNS grounds)))(SBAR (IN that)(S) e.g.: *because it penalized people ⇒ on the grounds that it penalized people*	(ADVP(RB prior)(PP(TO to)(NP(DT)(NN)))) ⇒ (SBAR(IN before)(NP(DT)(NN))) e.g.: *prior to the crash ⇒ before the crash*
(PP(RB because)(IN of)(NP(DT)(NN))) ⇒ (PP(IN as)(NP (NP(DT a)(NN result))(PP(IN of)(NP(DT)(NN))))) e.g.: *because of an investigation ⇒ as a result of an investigation*	(SBAR(IN until)(NP(CD))) ⇒ (SBAR(IN before)(NP(CD))) e.g.: *until 1819 ⇒ before 1819*

Table 1: Sample of extracted entailment rules.

Accordingly, we extracted from *set b* only the pairs containing one of these two seeds (either in T or in H) and we built two separate data sets for our experiments. We run the rule extraction algorithm, and then we filtered again the rules acquired, to collect only those containing one of the two seeds (either in the LHS or in the RHS). This second filtering has been done because there could be pairs in which either *because* or *before* are present, but the differences in T and H do not concern those seeds. The algorithm for rule expansion has then been applied to the selected rules to add the minimal context. The resulting rule for Example 3 is:

```
<rule ruleid="23" docid="844" pairid="15">
<LHS> (PP
     (RB  8 because) (IN 9 of)(NP
         (PRP 10 their)
         (NN 11 convenience))) </LHS>
<RHS> (ADJP
     (JJ  8 due)(PP
         (TO 9 to) (NP
             (PRP 10 their)
             (NN 11 convenience)))) </RHS>
</rule>
```

To create entailment rules balancing high-precision with their recall (Section 3), when the words of the context added to the rule in Step 4 are identical we substitute them with their PoS. For Example 3, the rule is generalized as follows:

```
<rule ruleid="23" docid="844" pairid="15">
<LHS> (PP
     (RB because) (IN of)(NP
         (PRP)
         (NN))) </LHS>
<RHS> (ADJP
     (JJ due)(PP
         (TO to) (NP
             (PRP)
             (NN)))) </RHS>
</rule>
```

The intuition underlying the generalization phase is to allow a more frequent application of the rule, while keeping some constraints on the allowed context. The application of the rule from Example 3 is

allowed if the subtrees below the seed words are the same (the rule can be applied in another T-H pair as, e.g. *because of his status ⇒ due to his status*).

Contradictions (e.g. antonyms and semantic oppositions) are generally very infrequent, but in certain cases they can have high impact (one of the most frequent rule collected for temporal expression is *before S ⇒ after S*). For this reason, we used WordNet (Fellbaum, 1998) to identify and filter antonyms out during the generalization phase. We also checked for awkward inconsistencies due to mistakes of the algorithm on noisy Wikipedia data (e.g. rules with the same seed word in both the LHS and the RHS), and we automatically filtered them out. Table 1 reports a sample of rules extracted for each seed word. Statistics about the resulting data sets, i.e. the number of acquired rules both before and after the generalization phase are shown in Table 2. Identical rules are collapsed into a unique one, but the value of their frequency is kept in the header of that rule. Such index can then be used to estimate the correctness of the rule and, according to our intuition, the probability that the rule preserves the entailment relation.[7]

	causality	temporal exp.
# rules before gen.	1671	813
# rules after gen.	**977**	**457**
rules frequency ≥ 2	66	27

Table 2: Resulting sets of entailment rules

5.1 Evaluation

Due to the sparseness of the phenomena under consideration (i.e. causality and temporal expressions) in RTE data sets, evaluating the acquired rules on such data does not provide interesting results.

For this reason, (following (Zhao *et al.*, 2009), (Callison-Burch, 2008), (Szpektor *et al.*, 2004)), we opted for a manual analysis of a sample of 100 rules per set, including all the rules whose frequency is $\geq$2 (Table 2), plus a random set of rules with frequency equal to 1. Two annotators with skills in linguistics annotated such rules according to five possible values (rules have been presented with the sentence pairs from which they have been acquired): *entailment=yes* (YES), i.e. correctness of the rule; *entailment=more-phenomena* (+PHEN), i.e. the rule is correct, but more than one phenomenon is involved, see Section 5.2; *entailment=unknown* (UNK), i.e. there is no entailment between the LHS and the RHS of the rule, often because the editing changed the semantics of the proposition; *entailment=unknown:reverse_entailment* (REV), wrong directionality, i.e. the RHS of the rule entails the LHS; *entailment=error* (ERR), i.e. the rule is wrong, either because the editing in *Wiki10* was done to correct mistakes, or because the rule is not well-formed due to mistakes produced by our algorithm.

The inter-annotator agreement has been calculated, counting when judges agree on the assigned value. It amounts to 80% on the sample of rules for causality, and to 77% on the sample of rules for temporal expressions. The highest inter-annotator agreement is for correct *entailment* rules, whereas the lowest agreement rates are for *unknown* and *error* judgments. This is due to the fact that detecting correct rules is straightforward, while it is less clear whether to consider a wrong rule as well-formed but with an *unknown* judgment, or to consider it as not appropriate (i.e. *error*). Table 3 shows the outcomes of the analysis of the two sets of rules, as resulting after a reconciliation phase carried out by the annotators. Such results, provided both for the whole samples[8] and for the rules whose frequency is $\geq$2 only, are discussed in the next section.

		YES	+PHEN	UNK	REV	ERR
caus.	all	67	2	13	8	10
	fr$\geq$2	80.3	0	16.7	1.5	1.5
temp.	all	36	6	23	7	28
	fr$\geq$2	52	3.7	37	7.3	0

Table 3: Accuracy (%) of the extracted sets of rules.

5.2 Discussion and error analysis

Due to the amount of noisy data present in Wikipedia, on average 19% of the collected rules

[7] It is difficult to compare our results with related work, since such phenomena are not covered by other resources. The correct comparison would be with the subset of e.g. DIRT paraphrases dealing with causality and temporal relations, if any.

[8] We are aware of the fact that including all the most frequent rules in the sample biases the results upwards, but our choice is motivated by the fact that we aim at verifying that with redundancy the accuracy is actually improved.

include editing done by the users for spelling and typos corrections, or are just spam (Table 3). To discard such cases, spell-checkers or dictionary-based filters should be used to improve our filtering techniques. Moreover, to select only reliable rules we consider making use of their frequency in the data to estimate the confidence that a certain rule maintains the entailment. The accuracy of the rules occurring more than once is indeed much higher than the accuracy estimated on the whole sample. Also the percentage of incorrect rules is strongly reduced when considering redundant rules. Our assumption about the directionality of entailment rules extracted from Wikipedia versions is also verified (less than 10% of the rules per set are tagged as *reverse-entailment*).

However, since the acquisition procedure privileges precision, only a few rules appear very frequently (Table 2), and this can be due to the constraints defined for the context extraction. This fact motivates also the lower precision of the rules for temporal expressions, where 73% of the sample we analyzed involved rules with frequency equal to 1. Moreover, in most of the rules annotated as *unknown*, the editing of *Wiki10* changed the semantics of the pair, e.g. *before 1990 ⇒ 1893*, or *when x produced ⇒ because x produced*. Further strategies to empirically estimate precision and recall of rule application should be experimented as future work. Indeed, several rules appearing only once represent correct rules, and should not be discarded a priori.

Finally, the idea of using only very similar pairs to extract entailment rules is based on the assumption that such rules should concern one phenomenon at a time (Bentivogli *et al.*, 2010). Despite the strategies adopted to avoid multiple phenomena per rule, in about 10% of the cases two phenomena (e.g lexical and syntactic) are collapsed on consecutive tokens, making it complex to separate them automatically: e.g. in *because of his divorce settlement cost ⇒ due to the cost of his divorces settlement*, the causative (*because of x ⇒ due to x*) and the argument realization (*x cost ⇒ cost of x*) rules should be separated.

6 Conclusion and future work

We have presented a methodology for the automatic acquisition of entailment rules from Wikipedia revision pairs. The main benefits are the following: *i)* potential large-scale acquisition, given the increasing size of Wikipedia revisions; *ii)* new coverage, because Wikipedia revisions contain linguistic phenomena (e.g. causality, temporal expressions), which are not covered by existing resources: as a consequence, the coverage of current TE systems can be significantly extended; *iii)* quality: we introduce the notion of context of a rule as the minimal set of syntactic features maximizing its successful application, and we have implemented it as a search over the syntactic representation of revision pairs.

Results obtained on two experimental acquisitions on causality and temporal expressions (seeds *because* and *before*) show both good quality and coverage of the extracted rules. The obtained resources[9]: *i)* cover entailment and paraphrasing aspects not represented in other similar sets of rules, *ii)* can be easily extended by applying the algorithms to automatically collect rules for other phenomena relevant to inference; and *iii)* are periodically updated, as Wikipedia revisions change continuously. We consider such aspects as part of our future work.

These results encourage us to further improve the approach, considering a number of directions. First, we plan to improve our filtering techniques to exclude revision pairs containing more than one phenomenon considering the syntactic structure of the sentence. Moreover, we are planning to carry out more extended evaluations, according to two possible strategies: *i)* applying the instance-based approach (Szpektor *et al.*, 2007) on the Penn Treebank data (i.e. for each PTB sentence that contains the LHS of an entailment rule from our set, a pair sentence will be generated by replacing the LHS of the rule with its RHS. Human judges will then judge each pair); *ii)* integrating the extracted rules into existing TE systems. However, this evaluation has to be carefully designed, as the ablation tests carried on at the RTE challenges show. In particular, as RTE tasks are moving towards real applications (e.g. summarization) we think that knowledge reflecting real textual variations produced by humans (as opposed to knowledge derived from linguistic resources) may introduce interesting and novel hints.

[9] Available at `http://www.aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`. We encourage its integration into TE systems, to obtain feedback on its utility in TE tasks.

Acknowledgments

This work has been partially supported by the EC-funded project EXCITEMENT (FP7 ICT-287923).

References

Roni Ben Aharon, Idan Szpektor, Ido Dagan. 2010. *Generating Entailment Rules from FrameNet.* Proceedings of the ACL 2010 Conference Short Papers. July 11-16. Uppsala, Sweden.

Regina Barzilay, Lillian Lee. 2003. *Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.* Proceedings of the HLT-NAACL. May 27-June 1. Edmonton, Canada.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa T. Dang, Danilo Giampiccolo. 2010. *The Sixth PASCAL Recognizing Textual Entailment Challenge.* Proceedings of the TAC 2010 Workshop on TE. November 15-16. Gaithersburg, Maryland.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, Bernardo Magnini. 2010. *Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference.* Proceedings of the Seventh conference on International Language Resources and Evaluation. May 19-21. Malta.

Chris Callison-Burch. 2008. *Syntactic constraints on paraphrases extracted from parallel corpora.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2008) October 25-27. Honolulu, Hawaii.

Ido Dagan, Bill Dolan, Bernardo Magnini, Dan Roth. 2009. *Recognizing textual entailment: Rational, evaluation and approaches.* Natural Language Engineering (JNLE). Special Issue 04, volume 15, i-xvii. Cambridge University Press.

Fred J. Damerau. 1964. *A technique for computer detection and correction of spelling errors.* Commun. ACM, 7 (3), pages 171–176. ACM, New York, NY, USA.

Camille Dutrey, Houda Bouamor, Delphine Bernhard and Aurelien Max 2011. *Local modifications and paraphrases in Wikipedia's revision history.* SEPLN journal (Revista de Procesamiento del Lenguaje Natural), 46:51-58.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* Language, Speech and Communication. MIT Press.

Adrian Iftene, Mihai-Alex Moruz. 2010. *UAIC Participation at RTE-6.* Proceedings of the TAC 2010 Workshop on TE. November 15-16. Gaithersburg, Maryland.

Dan Klein, Christopher D. Manning. 2003. *Accurate Unlexicalized Parsing.* Proceedings of the 41st Meeting of the Association for Computational Linguistics. July 7-12. Sapporo, Japan.

Dekang Lin, Patrick Pantel. 2001. *Discovery of Inference Rules for Question Answering.* Natural Language Engineering 7(4):343-360.

Rowan Nairn, Cleo Condoravdi, Lauri Karttunen. 2006. *Computing relative polarity for textual inference.* Inference in Computational Semantics (ICoS-5). April 20-21. Buxton, UK.

Aurelien Max, Guillaume Wisniewski. 2010. *Mining naturally-occurring corrections and paraphrases from wikipedia's revision history*. Proceedings of the Seventh conference on International Language Resources and Evaluation. May 19-21. Valletta, Malta.

Yashar Mehdad, Matteo Negri, Elena Cabrio, Milen Kouylekov, Bernardo Magnini. 2009. *Using Lexical Resources in a Distance-Based Approach to RTE*. Proceedings of the TAC 2009 Workshop on TE. November 17. Gaithersburg, Maryland.

Shachar Mirkin, Roy Bar-Haim, Jonathan Beran, Ido Dagan, Eyal Shnarch, Asher Stern, Idan Szpektor. 2009. *Addressing Discourse and Document Structure in the RTE Search Task*. Proceedings of the TAC 2009 Workshop on TE. November 17. Gaithersburg, Maryland.

Rani Nelken, Elif Yamangil. 2008. *Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms*. Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence. July 13-14, Chicago, Illinois.

Satoshi Sekine. 2005. *Automatic Paraphrase Discovery based on Context and Kwywords between NE Pairs.* Proceedings of the International Workshop on Paraphrasing (IWP-05). October 14. Jeju Island, South Korea.

Idan Szpektor, Hristo Tanev, Ido Dagan, Bonaventura Coppola. 2004. *Scaling Web-based Acquisition of Entailment Relations.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. July 25-26. Barcelona, Spain.

Idan Szpektor, Ido Dagan. 2008. *Learning Entailment Rules for Unary Templates.* Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). August 18-22. Manchester, UK.

Idan Szpektor, Eyal Shnarch, Ido Dagan. 2007. *Instance-based Evaluation of Entailment Rule Acquisition.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. June 23-30. Prague, Czech Republic.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, Hassan Sawaf. 1997. *Accelerated DP based search for statistical translation.* Proceedings

of the European Conf. on Speech Communication and Technology, pages 26672670. September. Rhodes, Greece.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, Lillian Lee. 2010. *For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia.* Proceedings of the NAACL, pp. 365-368, 2010. Short paper. June 1-6. Los Angeles, USA.

Fabio Massimo Zanzotto, Marco Pennacchiotti. 2010. *Expanding textual entailment corpora from Wikipedia using co-training*. Proceedings of the COLING-Workshop on The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources. August 28. Beijing, China.

Shiqi Zhao, Haifeng Wang, Ting Liu, Sheng Li. 2009. *Extracting Paraphrase Patterns from Bilingual Parallel Corpora*. Journal of Natural Language Engineering, 15 (4): 503:526.