Expanding textual entailment corpora from Wikipedia using co-training

Fabio Massimo Zanzotto
University of Rome "Tor Vergata"
Rome, Italy
zanzotto@info.uniroma2.it

Marco Pennacchiotti
Yahoo! Lab
Sunnyvale, CA, 94089
pennac@yahoo-inc.com

Abstract

In this paper we propose a novel method to automatically extract large textual entailment datasets homogeneous to existing ones. The key idea is the combination of two intuitions: (1) the use of Wikipedia to extract a large set of textual entailment pairs; (2) the application of semi-supervised machine learning methods to make the extracted dataset homogeneous to the existing ones. We report empirical evidence that our method successfully expands existing textual entailment corpora.

1 Introduction

Despite the growing success of the Recognizing Textual Entailment (RTE) challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007), the accuracy of most textual entailment recognition systems are still below 60%. An intuitive way to improve performance is to provide systems with larger annotated datasets. This is especially true for machine learning systems, where the size of the training corpus is an important factor. As a consequence, several attempts have been made to train systems using larger datasets obtained by merging RTE corpora of different challenges. Unfortunately, experimental results show a significant decrease in accuracy (de Marneffe et al., 2006). There are two major reasons for this counter-intuitive result:

Homogeneity. As indicated by many studies (e.g. (Siefkes, 2008)), homogeneity of the training corpus is an important factor for the applicability of supervised machine learning models, since examples with similar properties often imply more effective models. Unfortunately, the corpora of the four RTE challenges are not homogenous. Indeed, they model different properties of the textual entailment phenomenon, as they have been created using slightly (but significantly) different methodologies. For example, part of the RTE-1 dataset (Dagan et al., 2006) was created using comparable documents, where positive entailments have a lexical overlap higher than negative ones (Nicholson et al., 2006; Dagan et al., 2006). Comparable documents have not been used as a source of later RTE corpora, making RTE-1 odd with respect to other datasets.

Corpus size. RTE corpora are relatively small in size (typically 800 pairs). The increase in size obtained by merging corpora from different challenges is not a viable solution. Much larger datasets, of one or more order of magnitude, are needed to capture the complex properties characterizing entailment.

A key issue for the future development of RTE is then the creation of datasets fulfilling two properties: (1) large size; (2) homogeneity wrt. existing RTE corpora. The task of creating large datasets is unfeasible for human annotators. Collaborative annotation environments such as the Amazon Mechanical Turk[1] can help to annotate pairs of sentences in positive or negative entailment (Zaenen, submitted; Snow et al., 2008). Yet, these environments can hardly solve the problem of finding relevant pairs of sentences. Completely automatic processes of dataset creation have been proposed (Burger and Ferro, 2005; Hickl et al., 2006). Unfortunately, these datasets are not homogeneous wrt. to the RTE datasets, as they are

[1] http://mturk.com

created using different methodologies. In this paper we propose a novel method to automatically extract entailment datasets which are guaranteed to be large and homogeneous to RTE ones. The key idea is the combination of two factors: (1) the use of Wikipedia as source of a large set of textual entailment pairs; (2) the application of semi-supervised machine learning methods, namely co-training, to make corpora homogeneous to RTE.

The paper is organized as follows. In Section 2 we report on previous attempts in automatically creating RTE corpora. In Section 3 we outline important properties that these corpora should have, and introduce our methodology to extract an RTE corpus from Wikipedia (the *WIKI corpus*), conforming to these properties. In Section 4 we describe how co-training techniques can be leveraged to make the WIKI corpus homogeneous to existing RTE corpora. In Section 5 we report empirical evidence that the combination of the WIKI corpus and co-training is successful. Finally, in Section 6 we draw final conclusions and outline future work.

2 Related Work

The first attempt to automatically create large RTE corpora was proposed by Burger and Ferro (Burger and Ferro, 2005), with the *MITRE corpus*, a corpus of positive entailment examples extracted from the XIE section of the Gigaword news collection (Graff, 2003). The idea of the approach is that the headline and the first paragraph of a news article should be (near-)paraphrase. Authors then collect paragraph-headline pairs as Text (T) - Hypothesis (H) examples, where the headlines plays the role of H. The final corpus consists of 100,000 pairs, with an estimated accuracy of 70% – i.e. two annotators checked a sample of about 500 pairs, and verified that 30% of these were either false entailments or noisy pairs. The major limitation of the Burger and Ferro (Burger and Ferro, 2005)'s approach is that the final corpus consist only of positive examples. Because of this imbalance, the corpus cannot be positively used by RTE learning systems.

Hickl et al. (2006) propose a solution to the problem, providing a methodology to extract both positive and negative pairs (the *LCC corpus*). A positive corpus consisting of 101,000 pairs is extracted similarly to (Burger and Ferro, 2005). Corpus accuracy is estimated on a sample of 2,500 examples, achieving 92% (i.e. almost all examples are positives), 22 points higher than Burger and Ferro. A negative corpus of 119,000 is extracted either: (1) selecting sequential sentences including mentions of a same named entity (98.000 pairs); (2) selecting pairs of sentences connected by words such as *even though*, *although*, *otherwise*, *but* (21,000 pairs). Estimated accuracy for the two techniques is respectively 97% and 94%.

Hickl and colleagues show that expanding the RTE-2 training set with the LCC corpus (the expansion factor is 125), their RTE system improves 10% accuracy. This suggests that by expanding with a large and balanced corpus, entailment recognition performance drastically improves. This intuition is later contradicted in a second experiment by Hickl and Bensley (2007). Authors use the LCC corpus with the RTE-3 training set to train a new RTE system, showing an improvement in accuracy of less than 1% wrt. the RTE-3 training alone.

Overall, evidence suggests that automatic expansion of the RTE corpora do not always lead to performance improvement. This highly depends on how balanced the corpus is, on the RTE system adopted, and on the specific RTE dataset that is expanded.

3 Extracting the WIKI corpus

In this section we outline some of the properties that a reliable corpus for RTE should have (Section 3.1), and show that a corpus extracted from Wikipedia conforms to these properties (Section 3.2).

3.1 Good practices in building RTE corpora

Previous work in Section 2 and the vast literature on RTE suggest that a "reliable" corpus for RTE should have, among others, the following properties:

(1) Not artificial. Textual entailment is a complex phenomenon which encompasses different linguistic levels. Entailment types range from very simple polarity mismatches and syntactic alternations, to very complex semantic and knowledge-

S_1'	*In this regard, some have charged the New World Translation Committee with being inconsistent.*
S_2'	*In this regard, some have charged the New World Translation Committee with not be consistent.*
S_1''	*The 'Stockholm Network' is Europe's only dedicated service organisation for market-oriented think tanks and thinkers.*
S_2''	*The 'Stockholm Network' is, according to its own site, Europe's only dedicated service organisation for market-oriented think tanks and thinkers.*

Figure 1: Sentence pairs from the Wikipedia revision corpus

based inferences. These different types of entailments are naturally distributed in texts, such as news and every day conversations. A reliable RTE corpus should preserve this important property, i.e. it should be rich in entailment types whose distribution in the corpus is similar to that in real texts; and should not include unrepresentative hand-crafted prototypical examples.

(2) Balanced and consistent. A reliable corpus should be *balanced*, i.e. composed by an equal or comparable number of positive and negative examples. This is particularly critical for RTE systems based on machine learning: highly imbalanced class distributions often result in poor learning performance (Japkowicz and Stephen, 2002; Kubat and Matwin, 1997). Also, the positive and negative subsets of the corpus should be *consistent*, i.e. created using the same methodology. If this property is not preserved, the risk is a learning system building a model which separates positive and negatives according to the properties characterizing the two methodologies, instead of those of the entailment phenomenon.

(3) Not biased on lexical overlap. A major criticism on the RTE-1 dataset was that it contained too many positive examples with high lexical overlap wrt. negative examples (Nicholson et al., 2006). Glickman et al. (2005) show that an RTE system using word overlap to decide entailment, surprisingly achieves an accuracy of 0.57 on RTE-1 test set. These performances are comparable to those obtained on the same dataset by more sophisticated and principled systems. Learning from this experience, a good corpus for RTE should avoid imbalances on lexical overlap.

(4) Homogeneous to existing RTE corpora. Corpus homogeneity is a key property for any machine learning approach (Siefkes, 2008). A new corpus for RTE should then model the same or similar entailments types of the reliable existing ones (e.g., those of the RTE challenges). If this is not the case, RTE system will be unable to learn a coherent model, thus resulting in a decrease in performance.

The MITRE corpus satisfies property (1), but does not (2) and (3), as it is highly imbalanced (it contains mostly positive examples), and is fairly biased on lexical overlap, as most examples of headline-paragraph pairs have many words in common. The LCC corpus suffers the problem of inconsistency, as positive and negative examples are derived with radically different methodologies. Both the MITRE and the LCC corpora are difficult to merge with the RTE challenge datasets, as they are not homogeneous – i.e. they have been built using very different methodologies.

3.2 Extracting the corpus from Wikipedia revisions

Our main intuition in using Wikipedia to build an entailment corpus is that the wiki framework should provide a natural source of non-artificial examples of true and false entailments, through its revision system. Wikipedia is an open encyclopedia, where every person can behave as an author, inserting new entries or modifying existing ones. We call *original entry* S_1 a piece of text in Wikipedia before it is modified by an author, and *revision* S_2 the modified text. The primary concern of Wikipedia authors is to reshape a document according to their intent, by adding or replacing pieces of text. Excluding vandalism, there are several reasons for making a revision: missing information, misspelling, syntactic errors, and, more importantly, disagreement on the content. For example, in Fig. 1, S_1'' is revised to S_2'', as the author disagrees on the content of S_1''.

Our hypothesis is that (S_1, S_2) pairs represent good candidates of both true and false entailment pairs (T, H), as they represent semantically close

pieces of texts. Also, Wikipedia pairs conform to the properties listed in the previous section, as described in the following.

(S_1, S_2) pairs are *not artificial*, as we extract them from pieces of original texts, without any modification or post-processing. Also, pairs are rich of different entailment types, whose distribution is a reliable sample of language in use[2]. As shown later in the paper, a collection of (S_1, S_2) pairs is likely *balanced* on positive and negative examples, as authors either contradict the content of the original entry (false entailment) or add new information to the existing content (true entailment). Positive and negative pairs are guaranteed to be *consistent*, as they are drawn from the same Wikipedia source. Finally, the Wikipedia is *not biased in lexical overlap*: A sentence S_2 replacing S_1, usually changes only a few words. Yet, the meaning of S_2 may or may not change wrt. the meaning of S_1 – i.e. the lexical overlap of the two sentences is very high, but the entailment relation between S_1 and S_2 may be either positive or negative. For example, in Fig. 1 both pairs have high overlap, but the first is a positive entailment ($S'_1 \rightarrow S'_2$), while the second is negative ($S''_1 \rightarrow S''_2$).

An additional interesting property of Wikipedia revisions is that the transition from S_1 to S_2 is commented by the author. The *comment* is a piece of text where authors explain and motivate the change (e.g. "general cleanup of spelling and grammar", "revision: Eysenck died in 1997!!"). Even if very small, the comment can be used to determine if S_1 and S_2 are in entailment or not. In the following section we show how we leverage comments to make the WIKI corpus *homogeneous* to those of the RTE challenges.

4 Expanding the RTE corpus with WIKI using co-training

Unlike the LCC corpus where negative and positive examples are clearly separated, the WIKI corpus mixes the two sets – i.e. it is unlabelled. In order to exploit the WIKI corpus in the RTE task, one should either manually annotate the corpus, or find an alternative strategy to leverage the corpus even if unlabelled. As manual annotation is unfeasible, we choose the second solution. The goal is then to expand a *labelled* RTE challenge training set with the *unlabelled* WIKI, so that the performance of an RTE system can increase over an RTE test set.

[2]It has been shown that web documents (as Wikipedia) are reliable samples of language (Keller and Lapata, 2003).

CO-TRAINING_ALGORITHM(L,U,k)
returns h_1,h_2,L_1,L_2

set $L_1 = L_2 = L$

while stopping condition is not met

- learn h_1 on F_1 from L_1, and learn h_2 on F_1 from L_2,
- classify U with h_1 obtaining U_1, and classify U with h_2 obtaining U_2
- select and remove k-best classified examples u_1 and u_2 from respectively U_1 and U_2
- add u_1 to L_2 and u_2 to L_1

Figure 2: General co-training algorithm

In the literature, several techniques have been proposed to use unlabelled data to expand a training labelled corpus, e.g. Expectation-Maximization (Dempster et al., 1977). We here apply the co-training technique, first proposed by (Blum and Mitchell, 1998) and then successfully leveraged and analyzed in different settings (Abney, 2002). Co-training can be applied when the unlabelled dataset allows two independent views on its instances (*applicability condition*).

In this section, we first provide a short description of the co-training algorithm (Section 4.1). We then investigate if different RTE corpora conform to the applicability condition (Section 4.2). Finally, we show that our WIKI corpus conforms to the condition, and then apply co-training by creating two independent views (Section 4.3).

4.1 Co-training

The co-training algorithm uses unlabelled data to increase classification performance, and to indirectly increasing the size of labelled corpora. The algorithm can be applied only under a specific applicability condition: corpus' instances must have two *independent views*, i.e. they can be modeled by two independent feature sets.

We here adopt a slightly modified version of the

cotraining algorithm, as described in Fig.2. Under the applicability condition, instances are modeled on a feature space $F = F_1 \times F_2 \times C$, where F_1 and F_2 are the two independent views and C is the set of the target classes (in our case, true and false entailment). The algorithm starts with an initial set of training labelled examples L and a set of unlabelled examples U. The set L is copied in two sets L_1 and L_2, used to train two different classifiers h_1 and h_2, respectively using views F_1 and F_2. The two classifiers are used to classify the unlabelled set U, obtaining two different classifications, U_1 and U_2. Then comes the *co-training step*: the k-best classified instances in U_1 are added to L_2 and feed the learning of a new classifier h_2 on the feature space F_2. Similarly, the k-best instances in U_2 are added to L_1 and train a new classifier h_1 on F_1.

The procedure repeats until a stopping condition is met. This can be either a fixed number of added unlabelled examples (Blum and Mitchell, 1998), the performance drop on a control set of labelled instances, or a filter on the disagreement of h_1 and h_2 in classifying U (Collins and Singer, 1999). The final outcome of co-training is the new set of labelled examples $L_1 \cup L_2$ and the two classifier h_1 and h_2, obtained from the last iteration.

4.2 Applicability condition on RTE corpora

In order to leverage co-training for homogeneously expanding an RTE corpus, it is necessary to have a large unlabelled corpus which satisfies the applicability condition. Unfortunately, existing methodologies cannot guarantee the condition.

For example, the corpora from which the datasets of the RTE challenges were derived, were created from the output of applications performing specific tasks (e.g., Question&Answering, Information Extraction, Machine Translation, etc.). These corpora do not offer the possibility to create two completely independent views. Indeed, each extracted pair is composed only by the textual fragments of T and H, i.e. the only information available are the two pieces of texts, from which it is difficult to extract completely independent sets of features, as linguistic features tend to be dependent.

The MITRE corpus is extracted using two subsequent sentences, the title and the first paragraph. The LCC negative corpus is extracted using two correlated sentences or subsentences. Also in these two cases, it is very hard to find a view that is independent from the space of the sentence pairs.

None of the existing RTE corpora can then be used for co-training. In the next section we show that this is not the case for the WIKI corpus.

4.3 Creating independent views on the WIKI corpus

The WIKI corpus is naturally suited for co-training, as for each (S_1, S_2) pair, it is possible to clearly define two independent views:

- *content-pair view*: a set of features modeling the actual textual content of S_1 and S_2. This view is typically available also in any other RTE corpus.
- *comment view*: a set of features regarding the revision comment inserted by an author. This view represents "external" information (wrt. to the text fragments) which are peculiar of the WIKI corpus.

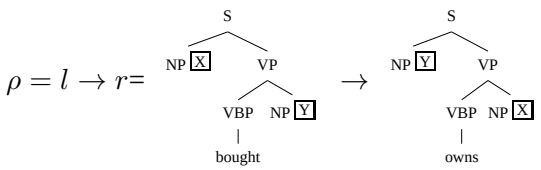
These two views are most likely independent. Indeed, the content-pair view deals with the content of the Wikipedia revision, while the comment view describes the reason why a revision has been made. This setting is very similar to the original one proposed for co-training by Blum and Mitchell (Blum and Mitchell, 1998), where the target problem was the classification of web pages, and the two independent views on a page were (1) its content and (2) its hyperlinks.

In the rest of this section we describe the feature spaces we adopt for the two independent views.

4.3.1 Content-pair view

The content-pair view is the classical view used in RTE. The original entry S_1 represents the Text T, while the revision S_2 is the Hypothesis H. Any feature space of those reported in the textual entailment literature could be applied. We here adopt the space that represents first-order syntactic rewrite rules (FOSR), as described in (Zanzotto and Moschitti, 2006). In this feature space, each feature represents a syntactic first-order or

grounded rewrite rule. For example, the rule:



is represented by the feature $< l, r >$. A (T, H) pair activates a feature if it unifies with the related rule. A detailed discussion of the FOSR feature space is given in (Zanzotto et al., 2009) and efficient algorithms for the computation of the related kernel functions can be found in (Moschitti and Zanzotto, 2007; Zanzotto and Dell'Arciprete, 2009).

4.4 Comment view

A review comment is typically a textual fragment describing the reason why an author has decided to make a revision. In most cases the comment is not a well-formed sentence, as authors tend to use informal slang expressions and abbreviations (e.g. "details: Trelew Massacre; cat: Dirty War, copyedit", "removed a POV vandalism by Spylab", "dab ba:clean up using Project:AWB"). In these cases, where syntactic analysis would mostly fail, it is advisable to use simpler surface approaches to build the feature space. We then use a standard bag-of-word space, combined with a bag-of-2-grams space. For the first space we keep only meaningful content words, by using a standard stop-list including articles, prepositions, and very frequent words such as *be* and *have*. The second space should help in capturing small text fragments containing functional words: we then keep all words without using any stop-list.

5 Experiments

The goals of our experiments are the following: (1) check the quality of the WIKI corpus, i.e. if positive and negative examples well represent the entailment phenomenon; (2) check if WIKI contains examples similar to those of the RTE challenges, i.e. if the corpus is homogeneous to RTE; (3) check if the WIKI corpus improves classification performance when used to expand the RTE datasets using the co-training technique described in Section 4.

5.1 Experimental Setup

In order to check the above claims, we need to experiment with both manually labelled and unlabelled corpora. As unlabelled corpora we adopt:

wiki_unlabelled: An unlabelled WIKI corpus of about 3,000 examples. The corpus has been built by downloading 40,000 Wikipedia pages dealing with 800 entries about politics, scientific theories, and religion issues. We extracted original entries and revisions from the XML and wiki code, collecting an overall corpus of 20,000 (S_1, S_2) pairs. We then randomly selected the final 3,000 pairs.

news: A corpus of 1,600 examples obtained using the methods adopted for the LCC corpus, both for negative and positive examples (Hickl et al., 2006).[3] We randomly divided the corpus in two parts: 800 training and 800 testing examples. Each set contains an equal number of 400 positive and negative pairs.

As labelled corpora we use:

RTE-1,RTE-2, and **RTE-3**: The corpora from the first three RTE challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007). We use the standard split between training and testing.

wiki: A manually annotated corpus of 2,000 examples from the WIKI corpus. Pairs have been annotated considering the original entry as the H and the revision as T. Noisy pairs containing vandalism or grammatical errors were removed (these accounts for about 19% of the examples). In all, the annotation produced 945 positive examples (strict entailments and paraphrases) and 669 negative examples (reverse strict entailments and contradictions). The annotation was carried out by two experienced researchers, each one annotating half of the corpus. Annotation guidelines follow those used for the RTE challenges.[4]

[3] For negative examples, we adopt the headline - first paragraph extraction methodology.

[4] Annotators were initially trained on a small development corpus of 200 pairs. The inter-annotator agreement on this set, computed using the Kappa-statistics (Siegel and Castellan, 1988), was 0.60 corresponding to *substantial agreement*,

The corpus has been randomly split in three equally numerous parts: development, training, and testing. We kept aside the development to design the features, while we used training and testing for the experiments.

We use the Charniak Parser (Charniak, 2000) for parsing sentences, and SVM-light (Joachims, 1999) extended with the syntactic first-order rule kernels described in (Zanzotto and Moschitti, 2006; Moschitti and Zanzotto, 2007) for creating the FOSR feature space.

5.2 Experimental Results

The first experiment aims at checking the quality of the WIKI corpus, by comparing the performance obtained by a standard RTE system over the corpus in exam with those obtained over any RTE challenge corpus. The hypothesis is that if performance is comparable, then the corpus in exam has the same complexity (and quality) as the RTE challenge corpora. We then independently experiment with the $wiki$ and the $news$ corpora with the training-test splits reported in Section 5.1. As RTE system we adopt an SVM model learnt on the FOSR feature space described in Section 4.3.1.

The accuracies of the system on the $wiki$ and $news$ corpora are respectively 70.73% and 94.87%. The performance of the system on the $wiki$ corpus are in line with those obtained over the RTE-2 dataset (60.62%). This suggests that the WIKI corpus is at least as complex as the RTE corpora (i.e. positive and negatives are not trivially separable). On the contrary, the $news$ corpus is much easier to separate. Pilot experiments show that increasing the size of the $news$ corpus, accuracy reaches nearly 100%. This indicates that positive and negative examples in the $news$ corpus are extremely different. Indeed, as mentioned in Section 3.1, $news$ is not consistent – i.e. the extraction methods for the positives and the negatives are so different that the examples can be easily recognized using evidence not representative of the entailment phenomenon (e.g. for negative examples, the lexical overlap is extremely low wrt. positives).

in line with the RTE challenge annotation efforts.

Training Corpus	*Accuracy*
RTE-2	60.62
RTE-1	51.25
RTE-3	57.25
wiki	56.00
news	53.25
RTE-2+RTE-1	58.5
RTE-2+RTE-3	59.62
RTE-2+news	56.75
RTE-2+wiki	59.25
RTE-1+wiki	53.37
RTE-3+wiki	59.00

Table 1: Accuracy of different training corpora over RTE-2 test.

In a second experiment we aim at checking if WIKI is homogeneous to the RTE challenge corpora – i.e. if it contains (T, H) pairs similar to those of the RTE corpora. If this holds, we would expect the performance of the RTE system to improve (or at least not decrease) when expanding a given RTE challenge corpus with WIKI. de Marneffe et al. (2006) already showed in their experiment that it is extremely difficult to obtain better performance by simply expanding an RTE challenge training corpus with corpora of other challenges, since different corpora are usually not homogeneous.

We here repeat a similar experiment: we experiment with different combinations of training sets, over the same test set (namely, RTE-2 test). Results are reported in Table 1. The higher performance is the one of the system when trained on RTE-2 training set (second row) – i.e. a corpus completely homogeneous to RTE-2 would produce the same performance as RTE-2 training.

As expected, the models learnt on RTE-1 and RTE-3 perform worse (third and fourth rows): in particular, RTE-1 seems extremely different from RTE-2, as results show. The $wiki$ corpus is more similar to RTE-2 than the $news$ corpus, i.e. performance are higher. Yet, it is quite surprising that the $news$ corpus yields to a performance drop as in (Hickl et al., 2006) it shows a high performance increase.

The expansion of RTE-2 with the above corpora (seventh-tenth rows) lead to a drop in performance, suggesting that none of the corpora is completely homogeneous to RTE-2. Yet, the performance drop of the $wiki$ corpus (*RTE-2 +*

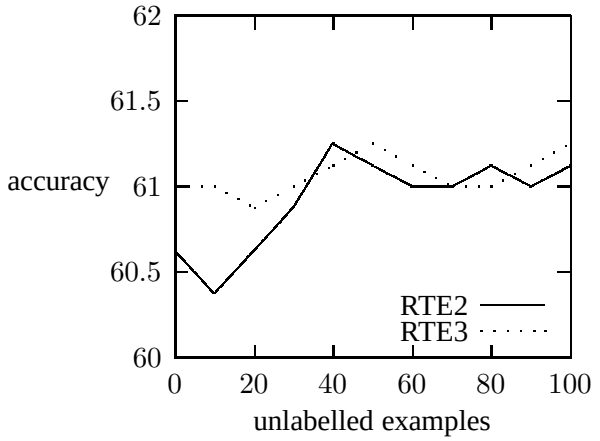


Figure 3: Co-training accuracy curve on the two corpora.

wiki) is comparable to the performance drop obtained using the other two RTE corpora (*RTE-2 + RTE-1* and *RTE-2 + RTE-3*). This indicates that $wiki$ is more homogeneous to RTE than $news$ – i.e. it contains (T, H) pairs that are similar to the RTE examples. Interestingly, $wiki$ combined with other RTE corpora (*RTE-1 + wiki* and *RTE-3 + wiki*) increases performance wrt. the models obtained with RTE-1 and RTE-3 alone (last two rows).

In a final experiment, we check if the WIKI corpus improves the performance when combined with the RTE-2 training in a co-training setting, as described in Section 4. This would confirm that WIKI is homogeneous to the RTE-2 corpus, and could then be successfully adopted in future RTE competitions. As test sets, we experiment both with RTE-2 and RTE-3 test. In the co-training, we use the RTE-2 training set as initial set L, and $wiki\_unlabelled$ as the unlabelled set U.[5]

Figure 3 reports the accuracy curves obtained by the classifier h_1 learnt on the content view, at each co-training iteration, both on the RTE-2 and RTE-3 test sets. As the comment view is not available in the RTE sets, the comment-view classifier become active only after the first 10 examples are fed as training from the content view classifier. As expected, performance increase for some steps and then become stable for RTE-3 and decrease for RTE-2. This is the only case in which we verified an increase in performance using corpora other than the official ones from RTE challenges. This result suggests that the WIKI corpus can successfully contribute to learn better textual entailment models for RTE.

[5]Note that only $wiki\_unlabelled$ allows both views described in Section 4.3.

6 Conclusions

In this paper we proposed a method for expanding existing textual entailment corpora that leverages Wikipedia. The method is extremely promising as it allows building corpora homogeneous to existing ones. The model we have presented is not strictly related to the RTE corpora. This method can then be used to expand corpora such as the Fracas test-suite (Cooper et al., 1996) which is more oriented to specific semantic phenomena.

Even if the performance increase of the completely unsupervised cotraining method is not extremely high, this model can be used to semi-automatically expanding corpora by using active learning techniques (Cohn et al., 1996). The initial increase of performances is an interesting starting point.

In the future, we aim at releasing the annotated portion of the WIKI corpus to the community; we will also carry out further experiments and refine the feature spaces. Finally, as Wikipedia is a multilingual resource, we will use the WIKI methodology to semi-automatically build RTE corpora for other languages.

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Idan Magnini, Bernardo Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Conference on Computational Learning Theory*. Morgan Kaufmann.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of the 1st NAACL*, pages 132–139, Seattle, Washington.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. Technical Report LRE 62-051 D-16, The FraCaS Consortium. Technical report.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Quionero-Candela et al., editor, *LNAI 3944: MLCW 2005*, pages 177–190, Milan, Italy. Springer-Verlag.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. 2006. Learning to distinguish valid textual entailments. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June. Association for Computational Linguistics.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, Southampton, UK.

David Graff. 2003. English gigaword.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176, Prague, June. ACL.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with LCCs GROUNDHOG system. In *Proceedings of the 2nd PASCAL Challenge Workshop on RTE*, Venice, Italy.

N. Japkowicz and S. Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5).

Thorsten Joachims. 1999. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3), September.

M. Kubat and S. Matwin. 1997. Addressing the curse of inbalanced data sets: One-side sampleing. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

Alessandro Moschitti and Fabio Massimo Zanzotto. 2007. Fast and effective kernels for relational learning from texts. In *Proceedings of the International Conference of Machine Learning (ICML)*, Corvallis, Oregon.

Jeremy Nicholson, Nicola Stokes, and Timothy Baldwin. 2006. Detecting entailment using an extended implementation of the basic elements overlap metric. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Christian Siefkes. 2008. *An Incrementally Trainable Statistical Approach to Information Extraction*. VDM Verlag, Saarbrucken, Germany.

S. Siegel and Jr. N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on EmNLP*, pages 254–263, Honolulu, Hawaii. ACL.

Annie Zaenen. submitted. Do give a penny for their thoughts. *Journal of Natural Language Engineering*.

Fabio Massimo Zanzotto and Lorenzo Dell'Arciprete. 2009. Efficient kernels for sentence pair classification. In *Conference on Empirical Methods on Natural Language Processing*, pages 91–100, 6-7 August.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st Coling and 44th ACL*, pages 401–408, Sydney, Australia, July.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2009. A machine learning approach to textual entailment recognition. *NATURAL LANGUAGE ENGINEERING*, 15-04:551–582. Accepted for pubblication.