Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment

Yashar Mehdad
FBK - irst and Uni. of Trento
Povo (Trento), Italy
mehdad@fbk.eu

Matteo Negri
FBK - irst
Povo (Trento), Italy
negri@fbk.eu

Marcello Federico
FBK - irst
Povo (Trento), Italy
federico@fbk.eu

Abstract

This paper explores the use of bilingual parallel corpora as a source of lexical knowledge for cross-lingual textual entailment. We claim that, in spite of the inherent difficulties of the task, phrase tables extracted from parallel data allow to capture both lexical relations between single words, and contextual information useful for inference. We experiment with a phrasal matching method in order to: *i)* build a system portable across languages, and *ii)* evaluate the contribution of lexical knowledge in isolation, without interaction with other inference mechanisms. Results achieved on an English-Spanish corpus obtained from the RTE3 dataset support our claim, with an overall accuracy above average scores reported by RTE participants on monolingual data. Finally, we show that using parallel corpora to extract paraphrase tables reveals their potential also in the monolingual setting, improving the results achieved with other sources of lexical knowledge.

1 Introduction

Cross-lingual Textual Entailment (CLTE) has been proposed by (Mehdad et al., 2010) as an extension of Textual Entailment (Dagan and Glickman, 2004) that consists in deciding, given two texts T and H *in different languages*, if the meaning of H can be inferred from the meaning of T. The task is inherently difficult, as it adds issues related to the multilingual dimension to the complexity of semantic inference at the textual level. For instance, the reliance of current monolingual TE systems on lexical resources (*e.g.* WordNet, VerbOcean, FrameNet) and deep processing components (*e.g.* syntactic and semantic parsers, co-reference resolution tools, temporal expressions recognizers and normalizers) has to confront, at the cross-lingual level, with the limited availability of lexical/semantic resources covering multiple languages, the limited coverage of the existing ones, and the burden of integrating language-specific components into the same cross-lingual architecture.

As a first step to overcome these problems, (Mehdad et al., 2010) proposes a "basic solution", that brings CLTE back to the monolingual scenario by translating H into the language of T. Despite the advantages in terms of modularity and portability of the architecture, and the promising experimental results, this approach suffers from one main limitation which motivates the investigation on alternative solutions. Decoupling machine translation (MT) and TE, in fact, ties CLTE performance to the availability of MT components, and to the quality of the translations. As a consequence, on one side translation errors propagate to the TE engine hampering the entailment decision process. On the other side such unpredictable errors reduce the possibility to control the behaviour of the engine, and devise *ad-hoc* solutions to specific entailment problems.

This paper investigates the idea, still unexplored, of a tighter integration of MT and TE algorithms and techniques. Our aim is to embed cross-lingual processing techniques inside the TE recognition process in order to avoid any dependency on external MT components, and eventually gain full control of the system's behaviour. Along this direction, we

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1336–1345,
Portland, Oregon, June 19-24, 2011.

start from the acquisition and use of lexical knowledge, which represents the basic building block of any TE system. Using the basic solution proposed by (Mehdad et al., 2010) as a term of comparison, we experiment with different sources of multilingual lexical knowledge to address the following questions:

(1) What is the potential of the existing multilingual lexical resources to approach CLTE? To answer this question we experiment with lexical knowledge extracted from bilingual dictionaries, and from a multilingual lexical database. Such experiments show two main limitations of these resources, namely: *i)* their limited coverage, and *ii)* the difficulty to capture contextual information when only associations between single words (or at most named entities and multiword expressions) are used to support inference.

(2) Does MT provide useful resources or techniques to overcome the limitations of existing resources? We envisage several directions in which inputs from MT research may enable or improve CLTE. As regards the resources, phrase and paraphrase tables extracted from bilingual parallel corpora can be exploited as an effective way to capture both lexical relations between single words, and contextual information useful for inference. As regards the algorithms, statistical models based on co-occurrence observations, similar to those used in MT to estimate translation probabilities, may contribute to estimate entailment probabilities in CLTE. Focusing on the resources direction, the main contribution of this paper is to show that the lexical knowledge extracted from parallel corpora allows to significantly improve the results achieved with other multilingual resources.

(3) In the cross-lingual scenario, can we achieve results comparable to those obtained in monolingual TE? Our experiments show that, although CLTE seems intrinsically more difficult, the results obtained using phrase and paraphrase tables are better than those achieved by average systems on monolingual datasets. We argue that this is due to the fact that parallel corpora are a rich source of cross-lingual paraphrases with no equivalents in monolingual TE.

(4) Can parallel corpora be useful also for monolingual TE? To answer this question, we experiment on monolingual RTE datasets using paraphrase tables extracted from bilingual parallel corpora. Our results improve those achieved with the most widely used resources in monolingual TE, namely WordNet, Verbocean, and Wikipedia.

The remainder of this paper is structured as follows. Section 2 shortly overviews the role of lexical knowledge in textual entailment, highlighting a gap between TE and CLTE in terms of available knowledge sources. Sections 3 and 4 address the first three questions, giving motivations for the use of bilingual parallel corpora in CLTE, and showing the results of our experiments. Section 5 addresses the last question, reporting on our experiments with paraphrase tables extracted from phrase tables on the monolingual RTE datasets. Section 6 concludes the paper, and outlines the directions of our future research.

2 Lexical resources for TE and CLTE

All current approaches to monolingual TE, either syntactically oriented (Rus et al., 2005), or applying logical inference (Tatu and Moldovan, 2005), or adopting transformation-based techniques (Kouleykov and Magnini, 2005; Bar-Haim et al., 2008), incorporate different types of lexical knowledge to support textual inference. Such information ranges from *i)* lexical paraphrases (textual equivalences between terms) to *ii)* lexical relations preserving entailment between words, and *iii)* word-level similarity/relatedness scores. WordNet, the most widely used resource in TE, provides all the three types of information. Synonymy relations can be used to extract lexical paraphrases indicating that words from the text and the hypothesis entail each other, thus being interchangeable. Hypernymy/hyponymy chains can provide entailment-preserving relations between concepts, indicating that a word in the hypothesis can be replaced by a word from the text. Paths between concepts and glosses can be used to calculate similarity/relatedness scores between single words, that contribute to the computation of the overall similarity between the text and the hypothesis.

Besides WordNet, the RTE literature documents the use of a variety of lexical information sources (Bentivogli et al., 2010; Dagan et al., 2009). These include, just to mention the most popular

ones, DIRT (Lin and Pantel, 2001), VerbOcean (Chklovski and Pantel, 2004), FrameNet (Baker et al., 1998), and Wikipedia (Mehdad et al., 2010; Kouylekov et al., 2009). DIRT is a collection of statistically learned inference rules, that is often integrated as a source of lexical paraphrases and entailment rules. VerbOcean is a graph of fine-grained semantic relations between verbs, which are frequently used as a source of precise entailment rules between predicates. FrameNet is a knowledge-base of frames describing prototypical situations, and the role of the participants they involve. It can be used as an alternative source of entailment rules, or to determine the semantic overlap between texts and hypotheses. Wikipedia is often used to extract probabilistic entailment rules based word similarity/relatedness scores.

Despite the consensus on the usefulness of lexical knowledge for textual inference, determining the actual impact of these resources is not straightforward, as they always represent one component in complex architectures that may use them in different ways. As emerges from the ablation tests reported in (Bentivogli et al., 2010), even the most common resources proved to have a positive impact on some systems and a negative impact on others. Some previous works (Bannard and Callison-Burch, 2005; Zhao et al., 2009; Kouylekov et al., 2009) indicate, as main limitations of the mentioned resources, their limited coverage, their low precision, and the fact that they are mostly suitable to capture relations mainly between single words.

Addressing CLTE we have to face additional and more problematic issues related to: *i)* the stronger need of lexical knowledge, and *ii)* the limited availability of multilingual lexical resources. As regards the first issue, it's worth noting that in the monolingual scenario simple "bag of words" (or "bag of n-grams") approaches are *per se* sufficient to achieve results above baseline. In contrast, their application in the cross-lingual setting is not a viable solution due to the impossibility to perform direct lexical matches between texts and hypotheses in different languages. This situation makes the availability of multilingual lexical knowledge a necessary condition to bridge the language gap. However, with the only exceptions represented by WordNet and Wikipedia, most of the aforementioned resources are available only for English. Multilingual lexical databases aligned with the English WordNet (*e.g.* MultiWordNet (Pianta et al., 2002)) have been created for several languages, with different degrees of coverage. As an example, the 57,424 synsets of the Spanish section of MultiWordNet aligned to English cover just around 50% of the WordNet's synsets, thus making the coverage issue even more problematic than for TE. As regards Wikipedia, the cross-lingual links between pages in different languages offer a possibility to extract lexical knowledge useful for CLTE. However, due to their relatively small number (especially for some languages), bilingual lexicons extracted from Wikipedia are still inadequate to provide acceptable coverage. In addition, featuring a bias towards named entities, the information acquired through cross-lingual links can at most complement the lexical knowledge extracted from more generic multilingual resources (*e.g* bilingual dictionaries).

3 Using Parallel Corpora for CLTE

Bilingual parallel corpora represent a possible solution to overcome the inadequacy of the existing resources, and to implement a portable approach for CLTE. To this aim, we exploit parallel data to: *i)* learn alignment criteria between phrasal elements in different languages, *ii)* use them to automatically extract lexical knowledge in the form of *phrase tables*, and *iii)* use the obtained phrase tables to create monolingual *paraphrase tables*.

Given a cross-lingual T/H pair (with the text in l_1 and the hypothesis in l_2), our approach leverages the vast amount of lexical knowledge provided by phrase and paraphrase tables to map H into T. We perform such mapping with two different methods. The **first method** uses a single phrase table to directly map phrases extracted from the hypothesis to phrases in the text. In order to improve our system's generalization capabilities and increase the coverage, the **second method** combines the phrase table with two monolingual paraphrase tables (one in l_1, and one in l_2). This allows to:

1. use the paraphrase table in l_2 to find paraphrases of phrases extracted from H;
2. map them to entries in the phrase table, and extract their equivalents in l_1;

3. use the paraphrase table in l_1 to find paraphrases of the extracted fragments in l_1;

4. map such paraphrases to phrases in T.

With the second method, phrasal matches between the text and the hypothesis are indirectly performed through paraphrases of the phrase table entries.

The final entailment decision for a T/H pair is assigned considering a model learned from the similarity scores based on the identified phrasal matches. In particular, "YES" and "NO" judgements are assigned considering the proportion of words in the hypothesis that are found also in the text. This way to approximate entailment reflects the intuition that, as a directional relation between the text and the hypothesis, the full content of H has to be found in T.

3.1 Extracting Phrase and Paraphrase Tables

Phrase tables (PHT) contain pairs of corresponding phrases in two languages, together with association probabilities. They are widely used in MT as a way to figure out how to translate input in one language into output in another language (Koehn et al., 2003). There are several methods to build phrase tables. The one adopted in this work consists in learning phrase alignments from a word-aligned bilingual corpus. In order to build English-Spanish phrase tables for our experiments, we used the freely available Europarl V.4, News Commentary and United Nations Spanish-English parallel corpora released for the WMT10[1]. We run TreeTagger (Schmid, 1994) for tokenization, and used the Giza++ (Och and Ney, 2003) to align the tokenized corpora at the word level. Subsequently, we extracted the bilingual phrase table from the aligned corpora using the Moses toolkit (Koehn et al., 2007). Since the resulting phrase table was very large, we eliminated all the entries with identical content in the two languages, and the ones containing phrases longer than 5 words in one of the two sides. In addition, in order to experiment with different phrase tables providing different degrees of coverage and precision, we extracted 7 phrase tables by pruning the initial one on the direct phrase translation probabilities of $0.01, 0.05, 0.1, 0.2, 0.3, 0.4$ and 0.5. The resulting phrase tables range from 76 to 48 million entries, with an average of 3.9 words per phrase.

Paraphrase tables (PPHT) contain pairs of corresponding phrases in the same language, possibly associated with probabilities. They proved to be useful in a number of NLP applications such as natural language generation (Iordanskaja et al., 1991), multidocument summarization (McKeown et al., 2002), automatic evaluation of MT (Denkowski and Lavie, 2010), and TE (Dinu and Wang, 2009).

One of the proposed methods to extract paraphrases relies on a pivot-based approach using phrase alignments in a bilingual parallel corpus (Bannard and Callison-Burch, 2005). With this method, all the different phrases in one language that are aligned with the same phrase in the other language are extracted as paraphrases. After the extraction, pruning techniques (Snover et al., 2009) can be applied to increase the precision of the extracted paraphrases.

In our work we used available[2] paraphrase databases for English and Spanish which have been extracted using the method previously outlined. Moreover, in order to experiment with different paraphrase sets providing different degrees of coverage and precision, we pruned the main paraphrase table based on the probabilities, associated to its entries, of $0.1, 0.2$ and 0.3. The number of phrase pairs extracted varies from 6 million to about 80000, with an average of 3.2 words per phrase.

3.2 Phrasal Matching Method

In order to maximize the usage of lexical knowledge, our entailment decision criterion is based on similarity scores calculated with a phrase-to-phrase matching process.

A phrase in our approach is an *n-gram* composed of up to 5 consecutive words, excluding punctuation. Entailment decisions are estimated by combining phrasal matching scores ($Score_n$) calculated for each level of *n-grams* , which is the number of *1-grams*, *2-grams*,..., *5-grams* extracted from H that match with *n-grams* in T. Phrasal matches are performed either at the level of tokens, lemmas, or stems, can be of two types:

[1] http://www.statmt.org/wmt10/

[2] http://www.cs.cmu.edu/ alavie/METEOR

1. **Exact**: in the case that two phrases are identical at one of the three levels (token, lemma, stem);

2. **Lexical**: in the case that two different phrases can be mapped through entries of the resources used to bridge T and H (*i.e.* phrase tables, paraphrases tables, dictionaries or any other source of lexical knowledge).

For each phrase in H, we first search for exact matches at the level of token with phrases in T. If no match is found at a token level, the other levels (lemma and stem) are attempted. Then, in case of failure with exact matching, lexical matching is performed at the same three levels. To reduce redundant matches, the lexical matches between pairs of phrases which have already been identified as exact matches are not considered.

Once matching for each *n-gram* level has been concluded, the number of matches (M_n) and the number of phrases in the hypothesis (Nn) are used to estimate the portion of phrases in H that are matched at each level (*n*). The phrasal matching score for each *n-gram* level is calculated as follows:

$$Score_n = \frac{M_n}{Nn}$$

To combine the phrasal matching scores obtained at each *n-gram* level, and optimize their relative weights, we trained a Support Vector Machine classifier, SVMlight (Joachims, 1999), using each score as a feature.

4 Experiments on CLTE

To address the first two questions outlined in Section 1, we experimented with the phrase matching method previously described, contrasting the effectiveness of lexical information extracted from parallel corpora with the knowledge provided by other resources used in the same way.

4.1 Dataset

The dataset used for our experiments is an English-Spanish entailment corpus obtained from the original RTE3 dataset by translating the English hypothesis into Spanish. It consists of 1600 pairs derived from the RTE3 development and test sets (800+800). Translations have been generated by the CrowdFlower[3] channel to Amazon Mechanical Turk[4] (MTurk), adopting the methodology proposed by (Negri and Mehdad, 2010). The method relies on translation-validation cycles, defined as separate jobs routed to MTurk's workforce. Translation jobs return one Spanish version for each hypothesis. Validation jobs ask multiple workers to check the correctness of each translation using the original English sentence as reference. At each cycle, the translated hypothesis accepted by the majority of trustful validators[5] are stored in the CLTE corpus, while wrong translations are sent back to workers in a new translation job. Although the quality of the results is enhanced by the possibility to automatically weed out untrusted workers using gold units, we performed a manual quality check on a subset of the acquired CLTE corpus. The validation, carried out by a Spanish native speaker on 100 randomly selected pairs after two translation-validation cycles, showed the good quality of the collected material, with only 3 minor "errors" consisting in controversial but substantially acceptable translations reflecting regional Spanish variations.

The T-H pairs in the collected English-Spanish entailment corpus were annotated using TreeTagger (Schmid, 1994) and the Snowball stemmer[6] with token, lemma, and stem information.

4.2 Knowledge sources

For comparison with the extracted phrase and paraphrase tables, we use a large bilingual dictionary and MultiWordNet as alternative sources of lexical knowledge.

Bilingual dictionaries (DIC) allow for precise mappings between words in H and T. To create a large bilingual English-Spanish dictionary we processed and combined the following dictionaries and bilingual resources:
- XDXF Dictionaries[7]: 22,486 entries.

[3] http://crowdflower.com/

[4] https://www.mturk.com/mturk/

[5] Workers' trustworthiness can be automatically determined by means of hidden gold units randomly inserted into jobs.

[6] http://snowball.tartarus.org/

[7] http://xdxf.revdanica.com/

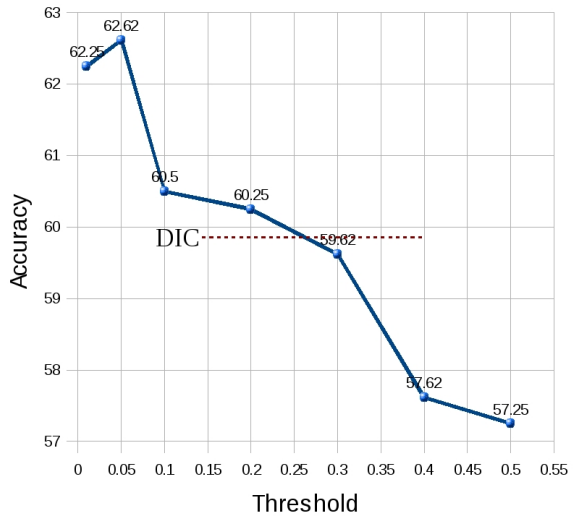

Figure 1: Accuracy on CLTE by pruning the phrase table with different thresholds.

- Universal dictionary database[8]: 9,944 entries.
- Wiktionary database[9]: 5,866 entries.
- Omegawiki database[10]: 8,237 entries.
- Wikipedia interlanguage links[11]: 7,425 entries.

The resulting dictionary features 53,958 entries, with an average length of 1.2 words.

MultiWordNet (MWN) allows to extract mappings between English and Spanish words connected by entailment-preserving semantic relations. The extraction process is dataset-dependent, as it checks for synonymy and hyponymy relations only between terms found in the dataset. The resulting collection of cross-lingual words associations contains 36,794 pairs of lemmas.

4.3 Results and Discussion

Our results are calculated over 800 test pairs of our CLTE corpus, after training the SVM classifier over 800 development pairs. This section reports the percentage of correct entailment assignments (accuracy), comparing the use of different sources of lexical knowledge.

Initially, in order to find a reasonable trade-off between precision and coverage, we used the 7 phrase tables extracted with different pruning thresholds (see Section 3.1). Figure 1 shows that with the pruning threshold set to 0.05, we obtain the highest result of 62.62% on the test set. The curve demonstrates that, although with higher pruning thresholds we retain more reliable phrase pairs, their smaller number provides limited coverage leading to lower results. In contrast, the large coverage obtained with the pruning threshold set to 0.01 leads to a slight performance decrease due to probably less precise phrase pairs.

MWN	DIC	PHT	PPHT	Acc.	δ
x				55.00	0.00
	x			59.88	+4.88
		x		**62.62**	+7.62
		x	x	**62.88**	+7.88

Table 1: Accuracy results on CLTE using different lexical resources.

Once the threshold has been set, in order to prove the effectiveness of information extracted from bilingual corpora, we conducted a series of experiments using the different resources mentioned in Section 4.2.

As it can be observed in Table 1, the highest results are achieved using the phrase table, both alone and in combination with paraphrase tables (62.62% and 62.88% respectively). These results suggest that, with appropriate pruning thresholds, the large number and the longer entries contained in the phrase and paraphrase tables represent an effective way to: *i)* obtain high coverage, and *ii)* capture cross-lingual associations between multiple lexical elements. This allows to overcome the bias towards single words featured by dictionaries and lexical databases.

As regards the other resources used for comparison, the results show that dictionaries substantially outperform MWN. This can be explained by the low coverage of MWN, whose entries also represent weaker semantic relations (preserving entailment, but with a lower probability to be applied) than the direct translations between terms contained in the dictionary.

Overall, our results suggest that the lexical knowledge extracted from parallel data can be successfully used to approach the CLTE task.

[8] http://www.dicts.info/
[9] http://en.wiktionary.org/
[10] http://www.omegawiki.org/
[11] http://www.wikipedia.org/

Dataset	WN	VO	WIKI	PPHT	PPHT 0.1	PPHT 0.2	PPHT 0.3	AVG
RTE3	61.88	62.00	61.75	62.88	**63.38**	**63.50**	63.00	62.37
RTE5	62.17	61.67	60.00	61.33	**62.50**	**62.67**	62.33	61.41
RTE3-G	62.62	61.5	60.5	62.88	**63.50**	62.00	61.5	-

Table 2: Accuracy results on monolingual RTE using different lexical resources.

5 Using parallel corpora for TE

This section addresses the third and the fourth research questions outlined in Section 1. Building on the positive results achieved on the cross-lingual scenario, we investigate the possibility to exploit bilingual parallel corpora in the traditional monolingual scenario. Using the same approach discussed in Section 4, we compare the results achieved with English paraphrase tables with those obtained with other widely used monolingual knowledge resources over two RTE datasets.

For the sake of completeness, we report in this section also the results obtained adopting the "basic solution" proposed by (Mehdad et al., 2010). Although it was presented as an approach to CLTE, the proposed method brings the problem back to the monolingual case by translating H into the language of T. The comparison with this method aims at verifying the real potential of parallel corpora against the use of a competitive MT system (Google Translate) in the same scenario.

5.1 Dataset

We experiment with the original RTE3 and RTE5 datasets, annotated with token, lemma, and stem information using the TreeTagger and the Snowball stemmer.

In addition to confront our method with the solution proposed by (Mehdad et al., 2010) we translated the Spanish hypotheses of our CLTE dataset into English using Google Translate. The resulting dataset was annotated in the same way.

5.2 Knowledge sources

We compared the results achieved with paraphrase tables (extracted with different pruning thresholds[12]) with those obtained using the three most widely used English resources for Textual Entailment (Bentivogli et al., 2010), namely:

WordNet (WN). WordNet 3.0 has been used to extract a set of 5396 pairs of words connected by the hyponymy and synonymy relations.

VerbOcean (VO). VerbOcean has been used to extract 18232 pairs of verbs connected by the "stronger-than" relation (*e.g. "kill" stronger-than "injure"*).

Wikipedia (WIKI). We performed Latent Semantic Analysis (LSA) over Wikipedia using the jLSI tool (Giuliano, 2007) to measure the relatedness between words in the dataset. Then, we filtered all the pairs with similarity lower than 0.7 as proposed by (Kouylekov et al., 2009). In this way we obtained 13760 word pairs.

5.3 Results and Discussion

Table 2 shows the accuracy results calculated over the original RTE3 and RTE5 test sets, training our classifier over the corresponding development sets.

The first two rows of the table show that pruned paraphrase tables always outperform the other lexical resources used for comparison, with an accuracy increase up to 3%. In particular, we observe that using 0.2 as a pruning threshold provides a good trade-off between coverage and precision, leading to our best results on both datasets (63.50% for RTE3, and 62.67% for RTE5). It's worth noting that these results, compared with the average scores reported by participants in the two editions of the RTE Challenge (AVG column), represent an accuracy improvement of more than 1%. Overall, these results confirm our claim that increasing the coverage using context sensitive phrase pairs obtained from large parallel corpora, results in better performance not only in CLTE,

[12]We pruned the paraphrase table (PPHT), with probabilities set to 0.1 (PPHT 0.1), 0.2 (PPHT 0.2), and 0.3 (PPHT 0.3)

but also in the monolingual scenario.

The comparison with the results achieved on monolingual data obtained by automatically translating the Spanish hypotheses (RTE3-G row in Table 2) leads to four main observations. First, we notice that dealing with MT-derived inputs, the optimal pruning threshold changes from 0.2 to 0.1, leading to the highest accuracy of 63.50%. This suggests that the noise introduced by incorrect translations can be tackled by increasing the coverage of the paraphrase table. Second, in line with the findings of (Mehdad et al., 2010), the results obtained over the MT-derived corpus are equal to those we achieve over the original RTE3 dataset (*i.e.* 63.50%). Third, the accuracy obtained over the CLTE corpus using combined phrase and paraphrase tables (62.88%, as reported in Table 1) is comparable to the best result gained over the automatically translated dataset (63.50%). In all the other cases, the use of phrase and paraphrase tables on CLTE data outperforms the results achieved on the same data after translation. Finally, it's worth remarking that applying our phrase matching method on the translated dataset without any additional source of knowledge would result in an overall accuracy of 62.12%, which is lower than the result obtained using only phrase tables on cross-lingual data (62.62%). This demonstrates that phrase tables can successfully replace MT systems in the CLTE task.

In light of this, we suggest that extracting lexical knowledge from parallel corpora is a preferable solution to approach CLTE. One of the main reasons is that placing a black-box MT system at the front-end of the entailment process reduces the possibility to cope with wrong translations. Furthermore, the access to MT components is not easy (*e.g.* Google Translate limits the number and the size of queries, while open source MT tools cover few language pairs). Moreover, the task of developing a full-fledged MT system often requires the availability of parallel corpora, and is much more complex than extracting lexical knowledge from them.

6 Conclusion and Future Work

In this paper we approached the cross-lingual Textual Entailment task focusing on the role of lexical knowledge extracted from bilingual parallel corpora. One of the main difficulties in CLTE raises from the lack of adequate knowledge resources to bridge the lexical gap between texts and hypotheses in different languages. Our approach builds on the intuition that the vast amount of knowledge that can be extracted from parallel data (in the form of phrase and paraphrase tables) offers a possible solution to the problem. To check the validity of our assumptions we carried out several experiments on an English-Spanish corpus derived from the RTE3 dataset, using phrasal matches as a criterion to approximate entailment. Our results show that phrase and paraphrase tables allow to: *i)* outperform the results achieved with the few multilingual lexical resources available, and *ii)* reach performance levels above the average scores obtained by participants in the monolingual RTE3 challenge. These improvements can be explained by the fact that the lexical knowledge extracted from parallel data provides good coverage both at the level of single words, and at the level of phrases.

As a further contribution, we explored the application of paraphrase tables extracted from parallel data in the traditional monolingual scenario. Contrasting results with those obtained with the most widely used resources in TE, we demonstrated the effectiveness of paraphrase tables as a mean to overcome the bias towards single words featured by the existing resources.

Our future work will address both the extraction of lexical information from bilingual parallel corpora, and its use for TE and CLTE. On one side, we plan to explore alternative ways to build phrase and paraphrase tables. One possible direction is to consider linguistically motivated approaches, such as the extraction of syntactic phrase tables as proposed by (Yamada and Knight, 2001). Another interesting direction is to investigate the potential of paraphrase patterns (*i.e.* patterns including part-of-speech slots), extracted from bilingual parallel corpora with the method proposed by (Zhao et al., 2009). On the other side we will investigate more sophisticated methods to exploit the acquired lexical knowledge. As a first step, the probability scores assigned to phrasal entries will be considered to perform weighted phrase matching as an improved criterion to approximate entailment.

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of COLING-ACL.*

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).*

Roy Bar-haim , Jonathan Berant , Ido Dagan , Iddo Greental , Shachar Mirkin , Eyal Shnarch , and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. *Proceedings of the TAC 2008 Workshop on Textual Entailment*.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. *Proceedings of the the Text Analysis Conference (TAC 2010)*.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Proceedings of the PASCAL Workshop of Learning Methods for Text Understanding and Mining*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering* , Volume 15, Special Issue 04, pp i-xvii.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. *Proceedings of Human Language Technologies (HLT-NAACL 2010).*

Georgiana Dinu and Rui Wang. 2009. Inference Rules and their Application to Recognizing Textual Entailment. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009).*

Claudio Giuliano. 2007. jLSI a tool for latent semantic indexing. *Software available at http://tcc.itc.it/research/textec/tools-resources/jLSI.html.*

Lidija Iordanskaja, Richard Kittredge, and Alain Polg re.. 1991. Lexical selection and paraphrase in a meaning text generation model. *Natural Language Generation in Articial Intelligence and Computational Linguistics*.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical.

Philipp Koehn, Franz Josef Och, and Daniel Marcu 2003. Statistical Phrase-Based Translation. *Proceedings of HLT/NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Milen Kouleykov and Bernardo Magnini. 2005. Tree edit distance for textual entailment. *Proceedings of RALNP-2005, International Conference on Recent Advances in Natural Language Processing*.

Milen Kouylekov, Yashar Mehdad, and Matteo Negri. 2010. Mining Wikipedia for Large-Scale Repositories of Context-Sensitive Entailment Rules. *Proceedings of the Language Resources and Evaluation Conference (LREC 2010)*.

Yashar Mehdad, Alessandro Moschitti and Fabio Massimo Zanzotto. 2010. Syntactic/semantic structures for textual entailment recognition. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*.

Dekang Lin and Patrick Pantel. 2001. DIRT - Discovery of Inference Rules from Text.. *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbias Newsblaster. *Proceedings of the Human Language Technology Conference.*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual Entailment. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010).*

Dan Moldovan and Adrian Novischi. 2002. Lexical chains for question answering. *Proceedings of COLING*.

Matteo Negri and Yashar Mehdad. 2010. Creating a Bilingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush. *Proceedings of the NAACL 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk* .

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):1951.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing and Aligned Multilingual Database. *Proceedings of the First International Conference on Global WordNet*.

Vasile Rus, Art Graesser, and Kirtan Desai 2005. Lexico-Syntactic Subsumption for Textual Entailment. *Proceedings of RANLP 2005*.

Helmut Schmid 2005. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*.

Marta Tatu andDan Moldovan. 2005. A semantic approach to recognizing textual entailment. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of WMT09.*

Rui Wang and Yi Zhang,. 2009. Recognizing Textual Relatedness with Predicate-Argument Structures. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*.

Kenji Yamada and Kevin Knight 2001. A Syntax-Based Statistical Translation Model. *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2009. Extracting Paraphrase Patterns from Bilingual Parallel Corpora. *Journal of Natural Language Engineering* , Volume 15, Special Issue 04, pp 503-526.