SAGAN: A Machine Translation Approach for Cross-Lingual Textual Entailment

Julio Castillo[1,2] **and Marina Cardenas**[2]
[1]UNC-FaMAF, Argentina
[2]UTN-FRC, Argentina
{jotacastillo, ing.marinacardenas}@gmail.com

Abstract

This paper describes our participation in the task denominated Cross-Lingual Textual Entailment (CLTE) for content synchronization. We represent an approach to CLTE using machine translation to tackle the problem of multilinguality. Our system resides on machine learning and in the use of WordNet as semantic source knowledge. Results are very promising always achieving results above mean score.

1 Introduction

This paper describes the participation of Sagan, a TE and CLTE system, in the new task of Cross Lingual Textual Entailment for Content Synchronization.

The objective of the Recognizing Textual Entailment (RTE) task (Dagan et al., 2006) is determining whether the meaning of a text fragment that we call hypothesis H can be inferred from another text fragment T. In this manner, we say that T entails H, if a person reading T would infer that H is most likely true. Thus, this definition assumes common human understanding of language and common background knowledge.

In that context, Cross-Lingual Textual Entailment addresses textual entailment recognition in the challenging application scenario of content synchronization. Thus, CLTE constitutes a generalization of Textual Entailment task (also Monolingual Textual Entailment) , but envisioning a larger number of application areas in NLP, including question answering, information retrieval, information extraction, and document summarization, across different languages.

Content synchronization could be used to keep consistence among documents written in different languages. For example, a CLTE system can be used in Wikipedia articles to inform lectors which information is absent or inconsistent in comparison to other page in a different language.

This new task has to face more additional issues than monolingual TE. Among them, we emphasize the ambiguity, polysemy, and coverage of the resources. Another additional problem is the necessity for semantic inference across languages, and the limited availability of multilingual knowledge resources.

The CLTE for content synchronization specifically consist on determining the entailment relationship between two text fragment T1 and T2 which are assumed belong a related topic.

Four alternatives are possible in this relationship:

- Bidirectional : It is a semantic equivalence between T1 and T2.
- Forward : It is an unidirectional entailment from T1 to T2.
- Backward: It is an unidirectional entailment from T2 to T1.
- No Entailment: It means that there is no entailment between T1 and T2.

The paper is organized as follows: Section 2 describes the relevant work done on cross-lingual textual entailment and related tasks, Section 3 describes the architecture of the system, then Section 4 shows experiments and results; and finally Sec-

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 721–726,
Montréal, Canada, June 7-8, 2012.

tion 5 summarize some conclusions and future work.

2 Related work

In this section we briefly describe two tasks that are closely related to CLTE.

2.1 Textual Entailment

The objective of the recognizing textual entailment (RTE) task (Dagan et al., 2006) is determining whether or not the meaning of a ‘‘hypothesis’’ (H) can be inferred from a ‘‘text’’ (T).

The two-way RTE task consists of deciding whether: T entails H, in which case the pair will be marked as ‘‘Entailment’’, otherwise the pair will be marked as ‘‘No Entailment’’. This definition of entailment is based on (and assumes) average human understanding of language as well as average background knowledge.

Recently the RTE4 Challenge has changed to a three-way task (Bentivogli et al, 2009) that consists in distinguishing among ‘‘Entailment’’, ‘‘Contradiction’’ and ‘‘Unknown’’ when there is no information to accept or reject the hypothesis.

The RTE challenge has mutated over the years, aiming at accomplishing more accurate and specific solutions; in 2009 the organizers proposed a pilot task, the Textual Entailment Search (Bentivogli et al, 2009), consisting in finding all the sentences in a set of documents that entail a given Hypothesis and since 2010 there is a Novelty Detection Task, which means that RTE systems are required to judge whether the information contained in each H is novel with respect to (i.e., not entailed by) the information contained in the corpus.

Thus, the new CLTE task can be thought as a generalized problem of RTE, which has to face new challenges as scarcity of resources to multilingual scenario, among others issues.

2.2 Semantic Textual Similarity

The pilot task STS was recently defined in Semeval 2012 (Aguirre et al., 2012) and has as main objective measuring the degree of semantic equivalence between two text fragments. STS is related to both Recognizing Textual Entailment (RTE) and Paraphrase Recognition, but has the advantage of being a more suitable model for multiple NLP applications.

As mentioned before, the goal of the RTE task (Bentivogli et al, 2009) is determining whether the meaning of a hypothesis H can be inferred from a text T. The main difference with STS is that STS consists in determining how similar two text fragments are, in a range from 5 (total semantic equivalence) to 0 (no relation). Thus, STS mainly differs from TE and Paraphrasing in that the classification is graded instead of binary and also STS assumes bidirectional equivalence but in TE the equivalence is only directional. In this manner, STS is filling the gap between TE and Paraphrase.

2.3 Cross-Lingual Textual Entailment

There are a few previous works on CLTE, the first one was the definition of this new task (Mehdad et al., 2010). Afterwards, the creation of CLTE corpus by using Mechanical Turk is described on (Negri et al., 2011) and a corpus freely available for CLTE is published (Castillo, 2011).

To our knowledge, two approach are proposed to address this new challenging task, one consist of using machine translation to move on towards monolingual textual entailment scenario and then apply classic techniques for RTE (Mehdad et al., 2010; Castillo and Cardenas, 2011), and the other is based on exploit databases of paraphrases (Mehdad et al., 2011). Both techniques obtained similar results and the accuracy achieved by them is not a statically significant difference.

In previous work (Castillo, 2010; Castillo and Cardenas, 2011) we addressed the CLTE focusing on English-Spanish language pair and released a bilingual textual entailment corpus. This paper is based on that work in order to tackling the problem across different language pairs Spanish-English (SPA-ENG), Italian-English (ITA-ENG), French-English (FRA-ENG) and German-English (GER-ENG) and we also used an approach based on machine translation.

3 System architecture

Sagan is a CLTE system (Castillo and Cardenas, 2010) which has taken part of several challenges, including the Textual Analysis Conference 2009 and TAC 2010, and the Semantic Textual Similari-

ty Semeval 2012 (Aguirre et al., 2012; Castillo and Estrella, 2012) and Cross Lingual Textual Entailment for content synchronization as part of the Semeval 2012 (Negri et al., 2012).

The system is based on a machine learning approach and it utilizes eight WordNet-based (Fellbaum, 1998) similarity measures with the purpose of obtaining the maximum similarity between two concepts. We used SVM as classifier with polynomial kernel. The system determines the entailment based on the semantic similarity of two texts (T,H) viewed as a function of the semantic similarity of the constituent words of both phrases. Thereby, we expect that combining word to word similarity metrics to text level would be a good indicator of text to text similarity.

These text-to-text similarity measures are based on the following word-to-word similarity metrics: (Resnik, 1995), (Lin, 1997), (Jiang and Conrath, 1997), (Pirrò and Seco, 2008), (Wu and Palmer, 1994), Path Metric, (Leacock and Chodorow, 1998), and a semantic similarity to sentence level named SemSim (Castillo and Cardenas, 2010).

Additional information about how to produce feature vectors as well as each word- and sentence-level metric can be found in (Castillo, 2011). The architecture of the system is shown in Figure 1.

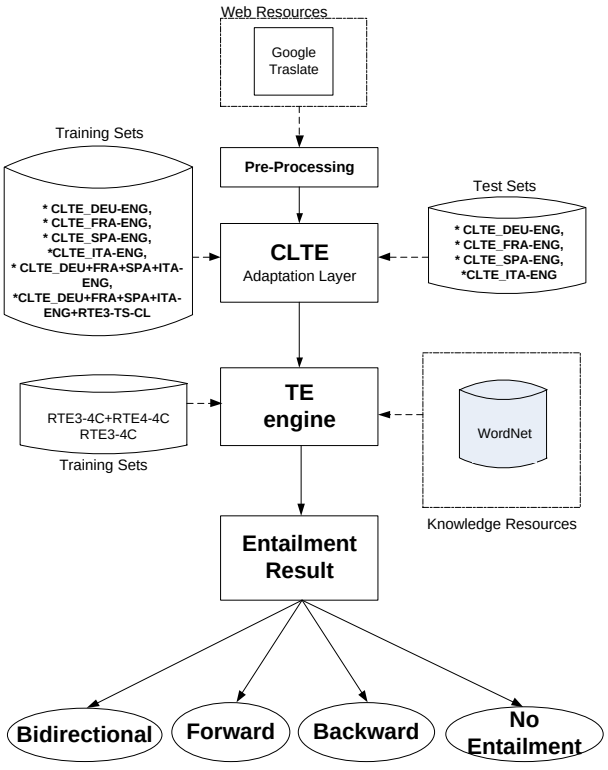

Fig.1. System architecture

In the preprocessing module we performed string normalization across different languages by using a lookup table for lexical entries, and then date and time normalization is carried out.

CLTE adaption layer is composed by four machine translation sub-modules that bring back each $<T_i, H>$ pair into the monolingual case ENG-ENG. Where T_i can be given in Spanish, German, Italian or French.

The training set used to the submitted runs are whose provided by the organizers of the CLTE for Content Synchronization Task and a combination of RTE datasets, such as it is described in the Section Experiments and Results.

4 Experiments and Results

The dataset provided by the organizers consists of 500 CLTE pairs translated to four languages following the crowdsourcing-based methodology proposed in (Negri et al., 2011). Also, for test purpose additional 500 pairs are provided. Both datasets are balanced with respect to the four entailment judgments (bidirectional, forward, backward, and no entailment).

We also performed experiments using traditional RTE datasets. Because of the RTE datasets are binary classified as NO (no-entailment) and YES (entailment), then we assumed that NO class is "no-entailment" and YES class is "forward" in the CLTE task. Certainly, the corpus tagged in this way will have contradictory information, since several pairs classified as forward should be classified as bidirectional, and also several pairs classified as no-entailment could be backwards, but the objective is experimenting whether we can gain accuracy in our RTE system despite of these (few) contradictory cases.

Additionally, in our experiments we used an algorithm (Castillo,2010) to generate additional training data, in other words to expand a data set. It is based on a Double Translation Process (dtp) or round-trip translation. Double translation process can be defined as the process of starting with an S (String in English), translating it to a foreign language F(S), for example Spanish, and finally back into the English source language F-1(S).

We applied the algorithm starting with RTE3 and RTE4 datasets. Thus, the augmented corpus is denoted RTE3-4C which is tagged according to the three-way task composed of: 340 pairs Contradic-

tion, 1520 pairs Yes, and 1114 pairs Unknown. In the case of the two-way task, it is composed by 1454 pairs No, and 1520 pairs Yes.

The other dataset augmented is denoted RTE4-4C, and has the following composition: 546 pairs Contradiction, 1812 pairs Entailment, and 1272 pairs Unknown. Therefore, in the two-way task, there are 1818 pairs No (No Entailment), and 1812 pairs Yes (Entailment) in this data set.

The idea behind using RTE3-4C and RTE3-4C is providing to our system an increased dataset aiming to acquire more semantic variability.

In our system submission we report the experiments performed with the test sets provided by CLTE organizers which is composed by four datasets of 500 pairs each one.

4.1 Submission to the CLTE shared task

With the aims of applying the monolingual textual entailment techniques, in the CLTE domain, we utilized the Google translate as MT system to bring back the <T,H> pairs into the monolingual case.

Then we generated a feature vector for every <T,H> pair with both training and test sets, and used monolingual textual entailment engine to classify the pairs. First we described the dataset used and then explain each submitted run.

The datasets used are listed below:

- CLTE_Esp+Fra+Ita+Ger: dataset composed by all language pairs.
- RTE3-TS-CL: a ENG-SPA cross lingual textual entailment corpus (Castillo,2011) composed by 200 pairs (108 Entailment, 32 Contradiction, 60 Unknown).
- RTE3-4C: an augmented dataset based on RTE3.
- RTE4-4C: an augmented dataset based on RTE4.

Our participation in the shared task consisted of four different runs produced with the same feature set, and the main differences rely on the amount and type of training data. Each run is described below:

- RUN1: system trained on CLTE_Esp+Fra+Ger+Ita corpus in addition to the RTE3-TS-CL dataset.
- RUN2: system trained on CLTE_Esp, CLTE_Fra, CLTE_Ger and CLTE_Ita corpus. At testing phase, the system chooses the right dataset according to the language that it is processing.
- RUN3: system trained using all training data that came from different language pairs.

We remark that we can combine the training data because of we used a machine translation submodule that bring back each <T,H> pair into the monolingual case ENG-ENG.

- RUN4: In RUN4 the training set is composed by all pairs of CLTE_Esp+Fra+Ita+Ger and RTE3-4C+ RTE4-4C datasets.

Ten teams participated in this CLTE task, eight submitting runs to all language pairs. For Spanish 28 runs were submitted and 20 runs were submitted for the other languages. The results achieved by our system is showed in Table 1.

Team id	Team system id	Score (Accuracy)				Run Rank			
		SPA-ENG	*ITA-ENG*	*FRA-ENG*	*DEU-ENG*	*SPA*	*ITA*	*FRA*	*DEU*
Sagan	*run1*	0.342	0.352	**0.346**	**0.342**	16	6	**9**	**9**
Sagan	*run2*	0.328	0.352	0.336	0.310	19	7	11	13
Sagan	*run3*	**0.346**	**0.356**	0.330	0.332	**14**	**5**	12	12
Sagan	*run4*	0.340	0.330	0.310	0.310	17	12	13	14
System Rank		7	4	5	6				

The results reported show that our best run is ranking above the average for all languages. The same situation occurs when ranking the systems, except for Spanish where the system is placed on 7th among 10 teams.

We achieved the highest result of 0.356 with RUN3 in the pair ITA-ENG which is placed fourth among participating systems.

We also note that, in general, training the system with the pairs of all datasets achieved better results than training separately for each dataset. Furthermore, if we analyze RUN4 vs. RUN3 we can conclude that incorporating additional RTE dataset produces a very unbalanced dataset resulting in a decrease in performance. In (Castillo, 2011) we experimented with these expanded datasets over monolingual RTE and CLTE tasks and we showed gain in performance, thus we suspect that the decrease is more due to unbalanced dataset than to noise introduced by the double translation process.

Interesting, the Corpus RTE3-TS-CL dataset utilized in the RUN1 helps to improve the results in FRA-ENG and DEU-ENG language pairs.

The Table 2 shows that our system predict with high F-measure to *bidirectional* and *no-entailment* entailment judgments in all language pairs, but has problems to distinguish the *forward* and *backward* entailment judgments.

It is probably due to our systems is based on semantic overlap between T and H, resulting the backwards particularly difficult to predict to our system.

Run id	Language pair	Precision				Recall				F-measure				Score (Accuracy)	Mean Score-all runs
		F	*B*	*NE*	*BI*	*F*	*B*	*NE*	*BI*	*F*	*B*	*NE*	*BI*		
Run3	SPA-ENG	0.23	0.27	**0.42**	0.42	0.20	0.22	0.45	0.51	0.21	0.25	0.43	0.46	**0.346**	**0.346**
Run3	ITA-ENG	**0.31**	0.25	0.40	**0.46**	0.30	0.22	0.51	0.40	0.30	0.23	0.45	0.43	**0.356**	**0.336**
Run1	FRA-ENG	0.24	**0.30**	0.39	0.43	0.17	0.34	0.57	0.30	0.20	0.32	0.47	0.36	**0.346**	**0.336**
Run1	DEU-ENG	0.25	0.23	0.41	0.44	0.17	0.26	0.60	0.34	0.20	0.25	0.49	0.39	**0.342**	**0.336**

Table 2. Official results for Precision, Recall and F-measure

5 Conclusions and future work

In this paper we explained our participation in the new challenging task of Cross-Lingual Textual Entailment (CLTE) for Content Synchronization. This task also could presents benefit as a metric for machine translation evaluation, as reported in (Castillo and Estrella, 2012).

This work focuses on CLTE based on Machine translation to bring back the problem into the monolingual Textual Entailment (TE) scenario. This decoupled approach between Textual Entailment and Machine Translation has several advantages, such as taking benefits of the most recent advances in machine translation, the ability to test the efficiency of different MT systems, as well as the ability to scale the system easily to any language pair.

Results achieved are promising and additional work is needed in order to address the problem of distinguish among *forward*, *backward* and *bidirectional* entailment judgments.

Future work will be oriented to tackle the problem with backwards. Finally, we remark the necessity of bigger corpus tagged in four-way classification, for all language pairs.

References

Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. *The PASCAL Recognising Textual Entailment Challenge*. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.) Machine Learning Challenges. Lecture Notes in Computer Science , Vol. 3944, pp. 177-190, Springer.

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. 2012. *Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization.* In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012).

L. Bentivogli, P. Clark, I. Dagan, H. T. Dang, and D. Giampiccolo. 2010. *The Sixth PASCAL Recognizing Textual Entailment Challenge*. In TAC 2010 Workshop Proceedings, NIST, Gaithersburg, MD, USA.

Y. Mehdad, M. Negri, and M. Federico. 2010. *Towards Cross-Lingual Textual Entailment*. In Proceedings of NAACL-HLT 2010.

Eneko Agirre, Daniel Cer, Mona Diab and Aitor Gonzalez-Agirre. 2012. *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*. In Proceedings of the 6th International Workshop on Semantic Evalua-tion (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Bentivogli, Luisa, Dagan Ido, Dang Hoa, Giampiccolo, Danilo, Magnini Bernardo.2009.*The Fifth PASCAL RTE Challenge*. In: Proceedings of the Text Analysis Conference.

Fellbaum C. 1998. *WordNet: An Electronic Lexical Database*, volume 1. MIT Press.

Castillo Julio. 2011. *A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment.* International Journal of Machine Learning and Cybernetics - Springer, Volume 2, Number 3.

Castillo Julio and Cardenas Marina. 2010. *Using sentence semantic similarity based onWordNet in recognizing textual entailment*. Iberamia 2010. In LNCS, vol 6433. Springer, Heidelberg, pp 366–375.

Castillo Julio. 2010. *A semantic oriented approach to textual entailment using WordNet-based measures*. MICAI 2010. LNCS, vol 6437. Springer, Heidelberg, pp 44–55.

Castillo Julio. 2010. *Using machine translation systems to expand a corpus in textual entailment*. In: Proceedings of the Icetal 2010. LNCS, vol 6233, pp 97–102.

M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. 2011. *Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textu-*

al Entailment Corpora. In Proceedings of the Conference on Empirical Methods in Natural. EMNLP 2011.

Resnik P. 1995. *Information content to evaluate semantic similarity in a taxonomy*. In: Proceedings of IJCAI 1995, pp 448–453.

Castillo Julio, Cardenas Marina. 2011. *An Approach to Cross-Lingual Textual Entailment using Online Machine Translation Systems*. Polibits Journal. Vol 44.

Castillo Julio and Estrella Paula. 2012. *Semantic Textual Similarity for MT evaluation*. NAACL 2012 Seventh Workshop on Statistical Machine Translation. WMT 2012, Montreal, Canada.

Lin D. 1997.*An information-theoretic definition of similarity*. In: Proceedings of Conference on Machine Learning, pp 296–304.

Jiang J, Conrath D.1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In: Proceedings of the ROCLINGX.

Pirro G., Seco N. 2008. *Design, implementation and evaluation of a new similarity metric combining feature and intrinsic information content*. In: ODBASE 2008, Springer LNCS.

Wu Z, Palmer M. 1994. *Verb semantics and lexical selection*. In: Proceedings of the 32nd ACL 916.

Leacock C, Chodorow M. 1998. *Combining local context and WordNet similarity for word sense identification*. MIT Press, pp 265–283.

Hirst G, St-Onge D . 1998. *Lexical chains as representations of context for the detection and correction of malapropisms*. MIT Press, pp 305–332.

Banerjee S, Pedersen T. 2002. *An adapted lesk algorithm for word sense disambiguation using WordNet*. In: Proceeding of CICLING-02.

William B. Dolan and Chris Brockett.2005. *Automatically Constructing a Corpus of Sentential Paraphrases*. Third International Workshop on Paraphrasing (IWP2005). Asia Federation of Natural Language Processing.

Castillo Julio and Estrella Paula. 2012. *SAGAN: An approach to Semantic Textual Similarity based on Textual Entailment.* In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Mehdad Y., M. Negri, and M. Federico. 2011. *Using Parallel Corpora for Cross-lingual Textual Entailment*. In Proceedings of ACL-HLT 2011.