Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization

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

Alessandro Marchetti
CELCT
Trento, Italy
amarchetti@celct.it

Yashar Mehdad
FBK-irst
Trento, Italy
mehdad@fbk.eu

Luisa Bentivogli
FBK-irst
Trento, Italy
bentivo@fbk.eu

Danilo Giampiccolo
CELCT
Trento, Italy
giampiccolo@celct.it

Abstract

This paper presents the first round of the task on *Cross-lingual Textual Entailment for Content Synchronization*, organized within SemEval-2012. The task was designed to promote research on semantic inference over texts written in different languages, targeting at the same time a real application scenario. Participants were presented with datasets for different language pairs, where multi-directional entailment relations ("forward", "backward", "bidirectional", "no_entailment") had to be identified. We report on the training and test data used for evaluation, the process of their creation, the participating systems (10 teams, 92 runs), the approaches adopted and the results achieved.

1 Introduction

The cross-lingual textual entailment task (Mehdad et al., 2010) addresses textual entailment (TE) recognition (Dagan and Glickman, 2004) under the new dimension of cross-linguality, and within the new challenging application scenario of content synchronization.

Cross-linguality represents a dimension of the TE recognition problem that has been so far only partially investigated. The great potential for integrating monolingual TE recognition components into NLP architectures has been reported in several areas, including question answering, information retrieval, information extraction, and document summarization. However, mainly due to the absence of cross-lingual textual entailment (CLTE) recognition components, similar improvements have not been achieved yet in any cross-lingual application. The CLTE task aims at prompting research to fill this gap. Along such direction, research can now benefit from recent advances in other fields, especially machine translation (MT), and the availability of: *i)* large amounts of parallel and comparable corpora in many languages, *ii)* open source software to compute word-alignments from parallel corpora, and *iii)* open source software to set up MT systems. We believe that all these resources can positively contribute to develop inference mechanisms for multilingual data.

Content synchronization represents a challenging application scenario to test the capabilities of advanced NLP systems. Given two documents about the same topic written in different languages (*e.g.* Wiki pages), the task consists of automatically detecting and resolving differences in the information they provide, in order to produce aligned, mutually enriched versions of the two documents. Towards this objective, a crucial requirement is to identify the information in one page that is either equivalent or novel (more informative) with respect to the content of the other. The task can be naturally cast as an entailment recognition problem, where bidirectional and unidirectional entailment judgments for two text fragments are respectively mapped into judgments about semantic equivalence and novelty. Alternatively, the task can be seen as a machine translation evaluation problem, where judgments about semantic equivalence and novelty depend on the possibility to fully or partially translate a text fragment into the other.

```
<entailment-corpus languages="spa-eng">
 <pair id="1" entailment="bidirectional">
   <t1>Mozart nació en la ciudad de Salzburgo</t1>
   <t2>Mozart was born in Salzburg</t2>
 </pair>
 <pair id="2" entailment="forward">
   <t1>Mozart nació en la ciudad de Salzburgo</t1>
   <t2>Mozart was born on the 27th January 1756 in Salzburg</t2>
 </pair>
 <pair id="3" entailment="backward">
   <t1>Mozart nació el 27 de enero de 1756 en Salzburgo</t1>
   <t2> Mozart was born in 1756 in the city of Salzburg</t2>
 </pair>
 <pair id="4" entailment="no_entailment">
   <t1>Mozart nació el 27 de enero de 1756 en Salzburgo</t1>
   <t2>Mozart was born to Leopold and Anna Maria Pertl Mozart</t2>
 </pair>
</entailment-corpus>
```

Figure 1: "bidirectional", "forward", "backward" and "no_entailment" judgments for SP/EN CLTE pairs.

The recent advances on monolingual TE on the one hand, and the methodologies used in Statistical Machine Translation (SMT) on the other, offer promising solutions to approach the CLTE task. In line with a number of systems that model the RTE task as a similarity problem (*i.e.* handling similarity scores between T and H as useful evidence to draw entailment decisions), the standard sentence and word alignment programs used in SMT offer a strong baseline for CLTE. However, although representing a solid starting point to approach the problem, similarity-based techniques are just approximations, open to significant improvements coming from semantic inference at the multilingual level (*e.g.* cross-lingual entailment rules such as "perro"→"animal"). Taken in isolation, similarity-based techniques clearly fall short of providing an effective solution to the problem of assigning directions to the entailment relations (especially in the complex CLTE scenario, where entailment relations are multi-directional). Thanks to the contiguity between CLTE, TE and SMT, the proposed task provides an interesting scenario to approach the issues outlined above from different perspectives, and large room for mutual improvement.

2 The task

Given a pair of topically related text fragments (*T1* and *T2*) in different languages, the CLTE task consists of automatically annotating it with one of the following entailment judgments (see Figure 1 for Spanish/English examples of each judgment):

- **bidirectional** (*T1*→*T2* & *T1*←*T2*): the two fragments entail each other (semantic equivalence);
- **forward** (*T1*→*T2* & *T1*↮*T2*): unidirectional entailment from *T1* to *T2*;
- **backward** (*T1*↛*T2* & *T1*←*T2*): unidirectional entailment from *T2* to *T1*;
- **no_entailment** (*T1*↛*T2* & *T1*↮*T2*): there is no entailment between *T1* and *T2* in both directions;

In this task, both *T1* and *T2* are assumed to be true statements. Although contradiction is relevant from an application-oriented perspective, contradictory pairs are not present in the dataset created for the first round of the task.

3 Dataset description

Four CLTE corpora have been created for the following language combinations: Spanish/English (SP-EN), Italian/English (IT-EN), French/English (FR-EN), German/English (DE-EN). The datasets are released in the XML format shown in Figure 1.

3.1 Data collection and annotation

The dataset was created following the crowdsourcing methodology proposed in (Negri et al., 2011), which consists of the following steps:

1. First, English sentences were manually extracted from copyright-free sources (Wikipedia and Wikinews). The selected sentences represent one of the elements (*T1*) of each entailment pair;
2. Next, each *T1* was modified through crowdsourcing in various ways in order to obtain a corresponding *T2* (*e.g.* introducing meaning-preserving lexical and syntactic changes, adding and removing portions of text);
3. Each *T2* was then paired to the original *T1*, and the resulting pairs were annotated with one of the four entailment judgments. In order to reduce the correlation between the difference in sentences' length and entailment judgments,

only the pairs where the difference between the number of words in *T1* and *T2* (*length_diff*) was below a fixed threshold (10 words) were retained.[1] The final result is a monolingual English dataset annotated with multi-directional entailment judgments, which are well distributed over *length_diff* values ranging from 0 to 9;

4. In order to create the cross-lingual datasets, each English *T1* was manually translated into four different languages (*i.e.* Spanish, German, Italian and French) by expert translators;

5. By pairing the translated *T1* with the corresponding *T2* in English, four cross-lingual datasets were obtained.

To ensure the good quality of the datasets, all the collected pairs were manually checked and corrected when necessary. Only pairs with agreement between two expert annotators were retained. The final result is a multilingual parallel entailment corpus, where *T1*s are in 5 different languages (*i.e.* English, Spanish, German, Italian, and French), and *T2*s are in English. It's worth mentioning that the monolingual English corpus, a by-product of our data collection methodology, will be publicly released as a further contribution to the research community.[2]

3.2 Dataset statistics

Each dataset consists of 1,000 pairs (500 for training and 500 for test), balanced across the four entailment judgments (bidirectional, forward, backward, and no_entailment).

For each language combination, the distribution of the four entailment judgments according to *length_diff* is shown in Figure 2. Vertical bars represent, for each *length_diff* value, the proportion of pairs belonging to the four entailment classes. As can be seen, the *length_diff* constraint applied to the length difference in the monolingual English pairs (step 3 of the creation process) is substantially reflected in the cross-lingual datasets for all language combinations. In fact, as shown in Table 1, the majority of the pairs is always included in the same *length_diff* range (approximately [-5,+5]) and, within this range, the distribution of the four classes is substantially uniform. Our assumption is that such data distribution makes entailment judgments based on mere surface features such as sentence length ineffective, thus encouraging the development of alternative, deeper processing strategies.

	SP-EN	IT-EN	FR-EN	DE-EN
Forward	104	132	121	179
Backward	202	182	191	123
No entailment	163	173	169	174
Bidirectional	175	199	193	209
ALL	644	686	674	685

Table 1: CLTE pairs distribution within the -5/+5 *length_diff* range.

4 Evaluation metrics and baselines

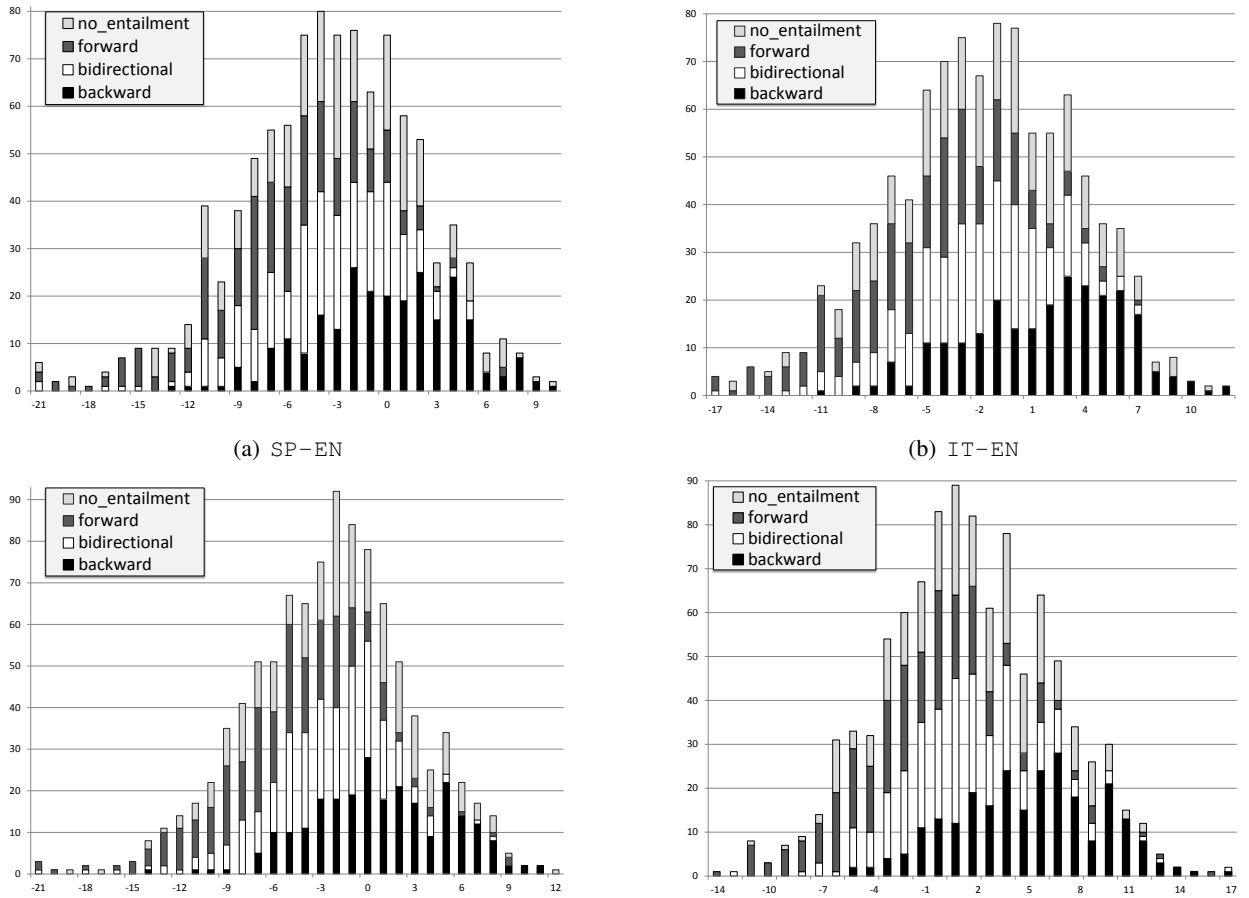
Evaluation results have been automatically computed by comparing the entailment judgments returned by each system with those manually assigned by human annotators. The metric used for systems' ranking is accuracy over the whole test set, *i.e.* the number of correct judgments out of the total number of judgments in the test set. Additionally, we calculated precision, recall, and F1 measures for each of the four entailment judgment categories taken separately. These scores aim at giving participants the possibility to gain clearer insights into their system's behavior on the entailment phenomena relevant to the task.

For each language combination, two baselines considering the length difference between *T1* and *T2* have been calculated (besides the trivial 0.25 accuracy score obtained by assigning each test pair in the balanced dataset to one of the four classes):

- **Composition of binary judgments (Binary).** To calculate this baseline an SVM classifier is trained to take binary entailment decisions ("YES", "NO"). The classifier uses *length(T1)/length(T2)* as a single feature to check for entailment from *T1* to *T2*, and *length(T2)/length(T1)* for the opposite direction. For each test pair, the unidirectional

[1] Such constraint has been applied in order to focus as much as possible on semantic aspects of the problem, by reducing the applicability of simple association rules such as *IF length(T1)>length(T2) THEN T1→T2*.

[2] The cross-lingual datasets are already available for research purposes at `http://www.celct.it/resourcesList.php`. The monolingual English dataset will be publicly released to non participants in July 2012.



(a) SP−EN (b) IT−EN

(c) FR−EN (d) DE−EN

Figure 2: CLTE pairs distribution for different *length_diff* values across all datasets.

judgments returned by the two classifiers are composed into a single multi-directional judgment ("YES-YES"="bidirectional", "YES-NO"="forward", "NO-YES"="backward", "NO-NO"="no_entailment");

- **Multi-class classification (Multi-class).** A single SVM classifier is trained with the same features to directly assign to each pair one of the four entailment judgments.

Both the baselines have been calculated with the LIBSVM package (Chang and Lin, 2011), using a linear kernel with default parameters. Baseline results are reported in Table 2.

Although the four CLTE datasets are derived from the same monolingual EN-EN corpus, baseline results present slight differences due to the effect of translation into different languages.

	SP-EN	**IT-EN**	**FR-EN**	**DE-EN**
1-class	0.25	0.25	0.25	0.25
Binary	0.34	0.39	0.39	0.40
Multi-class	**0.43**	**0.44**	**0.42**	**0.42**

Table 2: Baseline accuracy results.

5 Submitted runs and results

Participants were allowed to submit up to five runs for each language combination. A total of 17 teams registered to participate in the task and downloaded the training set. Out of them, 12 downloaded the test set and 10 (including one of the task organizers) submitted valid runs. Eight teams produced submissions for all the language combinations, while two teams participated only in the SP-EN task. In total, 92 runs have been submitted and evaluated (29 for SP-EN, and 21 for each of the other language pairs).

Despite the novelty and the difficulty of the problem, these numbers demonstrate the interest raised by the task, and the overall success of the initiative.

System_name	SP-EN	IT-EN	FR-EN	DE-EN
BUAP_run1	0.350	0.336	0.334	**0.330**
BUAP_run2	**0.366**	**0.344**	**0.342**	0.268
celi_run1	0.276	0.278	0.278	0.280
celi_run2	**0.336**	**0.338**	**0.300**	**0.352**
celi_run3	0.322	0.334	0.298	0.350
celi_run4	0.268	0.280	0.280	0.274
DirRelCond3_run1	0.300	0.280	0.362	0.336
DirRelCond3_run2	0.300	0.284	0.360	0.336
DirRelCond3_run3	0.300	**0.338**	**0.384**	0.364
DirRelCond3_run4	**0.344**	0.316	**0.384**	**0.374**
FBK_run1*	0.502	-	-	-
FBK_run2*	0.490	-	-	-
FBK_run3*	**0.504**	-	-	-
FBK_run4*	0.500	-	-	-
HDU_run1	0.630	0.554	0.564	**0.558**
HDU_run2	**0.632**	**0.562**	**0.570**	0.552
ICT_run1	**0.448**	**0.454**	**0.456**	**0.460**
JU-CSE-NLP_run1	**0.274**	0.316	0.288	0.262
JU-CSE-NLP_run2	0.266	**0.326**	0.294	**0.296**
JU-CSE-NLP_run3	0.272	0.314	**0.296**	0.264
Sagan_run1	0.342	0.352	**0.346**	**0.342**
Sagan_run2	0.328	0.352	0.336	0.310
Sagan_run3	**0.346**	**0.356**	0.330	0.332
Sagan_run4	0.340	0.330	0.310	0.310
SoftCard_run1	**0.552**	**0.566**	**0.570**	**0.550**
UAlacant_run1_LATE	**0.598**	-	-	-
UAlacant_run2	0.582	-	-	-
UAlacant_run3_LATE	0.510	-	-	-
UAlacant_run4	0.514	-	-	-
Highest	0.632	0.566	0.570	0.558
Average	0.440	0.411	0.408	0.408
Median	0.407	0.350	0.365	0.363
Lowest	0.274	0.326	0.296	0.296

Table 3: Accuracy results (92 runs) over the 4 language combinations. Highest, average, median and lowest scores are calculated considering the best run for each team (*task organizers' system).

Accuracy results are reported in Table 3. As can be seen from the table, overall accuracy scores are quite different across language pairs, with the highest result on SP-EN (0.632), which is considerably higher than the highest score on DE-EN (0.558). This might be due to the fact that most of the participating systems rely on a "pivoting" approach that addresses CLTE by automatically translating *T1* in the same language of *T2* (see Section 6). Regarding the DE-EN dataset, pivoting methods might be penalized by the lower quality of MT output when German *T1*s are translated into English.

The comparison with baselines results leads to interesting observations. First of all, while all systems significantly outperform the lowest 1-class baseline (0.25), both other baselines are surprisingly hard to beat. This shows that, despite the effort in keeping the distribution of the entailment classes uniform across different *length_diff* values, eliminating the correlation between sentences' length and correct entailment decisions is difficult. As a consequence, although disregarding semantic aspects of the problem, features considering such information are quite effective.

In general, systems performed better on the SP-EN dataset, with most results above the binary baseline (8 out of 10), and half of the systems above the multi-class baseline. For the other language pairs the results are lower, with only 3 out of 8 participants above the two baselines in all datasets. Average results reflect this situation: the average scores are always above the binary baseline, whereas only the SP-EN average result is higher than the multi-class baseline(0.44 vs. 0.43).

To better understand the behaviour of each system (also in relation to the different language combinations), Table 4 provides separate precision, recall, and F1 scores for each entailment judgment, calculated over the best runs of each participating team. Overall, the results suggest that the "bidirectional" and "no_entailment" categories are more problematic than "forward" and "backward" judgments. For most datasets, in fact, systems' performance on "bidirectional" and "no_entailment" is significantly lower, typically on recall. Except for the DE-EN dataset (more problematic on "forward"), also average F1 results on these judgments are lower. This might be due to the fact that, for all datasets, the vast majority of "bidirectional" and "no_entailment" judgments falls in a *length_diff* range where the distribution of the four classes is more uniform (see Figure 2).

Similar reasons can justify the fact that "backward" entailment results are consistently higher on all datasets. Compared with "forward" entailment, these judgments are in fact less scattered across the entire *length_diff* range (*i.e.* less intermingled with the other classes).

6 Approaches

A rough classification of the approaches adopted by participants can be made along two orthogonal dimensions, namely:

- **Pivoting vs. Cross-lingual.** Pivoting methods rely on the automatic translation of one of the two texts (either single words or the entire sentence) into the language of the other (typically English) in order perform monolingual TE recognition. Cross-lingual methods assign entailment judgments without preliminary translation.

- **Composition of binary judgments vs. Multi-class classification.** Compositional approaches map unidirectional entailment decisions taken separately into single judgments (similar to the *Binary* baseline in Section 4). Methods based on multi-class classification directly assign one of the four entailment judgments to each test pair (similar to our *Multi-class* baseline).

Concerning the former dimension, most of the systems (6 out of 10) adopted a pivoting approach, relying on Google Translate (4 systems), Microsoft Bing Translator (1), or a combination of Google, Bing, and other MT systems (1) to produce English *T2*s. Regarding the latter dimension, the compositional approach was preferred to multi-class classification (6 out of 10). The best performing system relies on a "hybrid" approach (combining monolingual and cross-lingual alignments) and a compositional strategy. Besides the frequent recourse to MT tools, other resources used by participants include: on-line dictionaries for the translation of single words, word alignment tools, part-of-speech taggers, NP chunkers, named entity recognizers, stemmers, stopwords lists, and Wikipedia as an external multilingual corpus. More in detail:

BUAP [pivoting, compositional] (Vilariño et al., 2012) adopts a pivoting method based on translating *T1* into the language of *T2* and vice versa (Google Translate[3] and the OpenOffice Thesaurus[4]). Similarity measures (*e.g.* Jaccard index) and rules are respectively used to annotate the two resulting sentence pairs with entailment judgments and combine them in a single decision.

CELI [cross_lingual, compositional & multi-class] (Kouylekov, 2012) uses dictionaries for word matching, and a multilingual corpus extracted from Wikipedia for term weighting. Word overlap and similarity measures are then used in different approaches to the task. In one run (Run_1), they are used to train a classifier that assigns separate entailment judgments for each direction. Such judgments are finally composed into a single one for each pair. In the other runs, the same features are used for multi-class classification.

DirRelCond3 [cross_lingual, compositional] (Perini, 2012) uses bilingual dictionaries (Freedict[5] and WordReference[6]) to translate content words into English. Then, entailment decisions are taken combining directional relatedness scores between words in both directions (Perini, 2011).

FBK [cross_lingual, compositional & multi-class] (Mehdad et al., 2012a) uses cross-lingual matching features extracted from lexical phrase tables, semantic phrase tables, and dependency relations (Mehdad et al., 2011; Mehdad et al., 2012b; Mehdad et al., 2012c). The features are used for multi-class and binary classification using SVMs.

HDU [hybrid, compositional] (Wäschle and Fendrich, 2012) uses a combination of binary classifiers for each entailment direction. The classifiers use both monolingual alignment features based on METEOR (Banerjee and Lavie, 2005) alignments (translations obtained from Google Translate), and cross-lingual alignment features based on GIZA++ (Och and Ney, 2000) (word alignments learned on Europarl).

ICT [pivoting, compositional] (Meng et al., 2012) adopts a pivoting method (using Google Translate and an in-house hierarchical MT system), and the open source EDITS system (Kouylekov and Negri, 2010) to calculate similarity scores between monolingual English pairs. Separate unidirectional entailment judgments obtained from binary classifier are combined to return one of the four valid CLTE judgments.

[3] http://translate.google.com/

[4] http://extensions.services.openoffice.org/en/taxonomy/term/233

[5] http://www.freedict.com/

[6] http://www.wordreference.com/

	SP-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
BUAP_spa-eng_run2	0,337	0,664	0,447	0,406	0,568	0,473	0,333	0,088	0,139	0,391	0,144	0,211
celi_spa-eng_run2	0,324	0,368	0,345	0,411	0,368	0,388	0,306	0,296	0,301	0,312	0,312	0,312
DirRelCond3_spa-eng_run4	0,358	0,608	0,451	0,444	0,448	0,446	0,286	0,032	0,058	0,243	0,288	0,264
FBK_spa-eng_run3	0,515	**0,704**	0,595	0,546	0,568	0,557	0,447	0,304	0,362	0,482	0,440	0,460
HDU_spa-eng_run2	0,607	0,656	**0,631**	**0,677**	0,704	**0,690**	**0,602**	**0,592**	**0,597**	**0,643**	**0,576**	**0,608**
ICT_spa-eng_run1	**0,750**	0,240	0,364	0,440	0,472	0,456	0,395	0,560	0,464	0,436	0,520	0,474
JU-CSE-NLP_spa-eng_run1	0,211	0,288	0,243	0,272	0,296	0,284	0,354	0,232	0,280	0,315	0,280	0,297
Sagan_spa-eng_run3	0,225	0,200	0,212	0,269	0,224	0,245	0,418	0,448	0,432	0,424	0,512	0,464
SoftCard_spa-eng_run1	0,602	0,616	0,609	0,650	0,624	0,637	0,471	0,448	0,459	0,489	0,520	0,504
UAlacant_spa-eng_run1_LATE	0,689	0,568	0,623	0,645	**0,728**	0,684	0,507	0,544	0,525	0,566	0,552	0,559
AVG.	*0,462*	*0,491*	*0,452*	*0,476*	*0,5*	*0,486*	*0,412*	*0,354*	*0,362*	*0,43*	*0,414*	*0,415*

	IT-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
BUAP_ita-eng_run2	0,324	0,456	0,379	0,327	0,672	0,440	0,538	0,056	0,101	0,444	0,192	0,268
celi_ita-eng_run2	0,349	0,360	0,354	0,455	0,36	0,402	0,294	0,320	0,307	0,287	0,312	0,299
DirRelCond3_ita-eng_run3	0,323	0,488	0,389	0,480	0,288	0,360	0,331	0,368	0,348	0,268	0,208	0,234
HDU_ita-eng_run2	0,564	**0,600**	0,581	**0,628**	0,648	0,638	0,551	**0,520**	**0,535**	**0,500**	0,480	0,490
ICT_ita-eng_run1	**0,661**	0,296	0,409	0,554	0,368	0,442	0,427	0,448	0,438	0,383	**0,704**	**0,496**
JU-CSE-NLP_ita-eng_run2	0,240	0,280	0,258	0,339	0,480	0,397	0,412	0,280	0,333	0,359	0,264	0,304
Sagan_ita-eng_run3	0,306	0,296	0,301	0,252	0,216	0,233	0,395	0,512	0,446	0,455	0,400	0,426
SoftCard_ita-eng_run1	0,602	0,616	**0,609**	0,617	**0,696**	**0,654**	**0,560**	0,448	0,498	0,481	0,504	0,492
AVG.	*0,421*	*0,424*	*0,410*	*0,457*	*0,466*	*0,446*	*0,439*	*0,369*	*0,376*	*0,397*	*0,383*	*0,376*

	FR-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
BUAP_fra-eng_run2	0,447	0,272	0,338	0,291	**0,760**	0,420	0,250	0,016	0,030	0,449	0,320	0,374
celi_fra-eng_run2	0,316	0,296	0,306	0,378	0,360	0,369	0,270	0,296	0,282	0,244	0,248	0,246
DirRelCond3_fra-eng_run3	0,393	0,576	0,468	0,441	0,512	0,474	0,387	0,232	0,290	0,278	0,216	0,243
HDU_fra-eng_run2	0,564	**0,672**	**0,613**	0,582	0,736	0,650	**0,676**	0,384	0,490	0,500	**0,488**	0,494
ICT_fra-eng_run1	**0,750**	0,192	0,306	0,517	0,496	0,506	0,385	**0,656**	0,485	0,444	0,480	0,462
JU-CSE-NLP_fra-eng_run3	0,215	0,208	0,211	0,289	0,296	0,292	0,341	0,496	0,404	0,333	0,184	0,237
Sagan_fra-eng_run1	0,244	0,168	0,199	0,297	0,344	0,319	0,394	0,568	0,466	0,427	0,304	0,355
SoftCard_fra-eng_run1	0,551	0,608	0,578	**0,649**	0,696	**0,672**	0,560	0,488	**0,521**	**0,513**	**0,488**	**0,500**
AVG.	*0,435*	*0,374*	*0,377*	*0,431*	*0,525*	*0,463*	*0,408*	*0,392*	*0,371*	*0,399*	*0,341*	0,364

	DE-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
BUAP_deu-eng_run1	0,395	0,120	0,184	0,248	0,224	0,235	0,344	**0,688**	0,459	0,364	0,288	0,321
celi_deu-eng_run2	0,347	0,416	0,378	0,402	0,392	0,397	0,339	0,312	0,325	0,319	0,288	0,303
DirRelCond3_deu-eng_run4	0,429	0,312	0,361	0,408	0,552	0,469	0,367	0,320	0,342	0,298	0,312	0,305
HDU_deu-eng_run1	0,559	0,528	0,543	0,600	**0,696**	**0,644**	**0,540**	0,488	**0,513**	**0,524**	0,520	**0,522**
ICT_deu-eng_run1	**0,718**	0,224	0,341	0,493	0,552	0,521	0,390	0,512	0,443	0,439	**0,552**	0,489
JU-CSE-NLP_deu-eng_run2	0,182	0,048	0,076	0,307	0,496	0,379	0,315	0,560	0,403	0,233	0,080	0,119
Sagan_deu-eng_run1	0,250	0,168	0,201	0,239	0,256	0,247	0,405	0,600	0,484	0,443	0,344	0,387
SoftCard_deu-eng_run1	0,568	**0,568**	**0,568**	**0,611**	0,640	0,625	0,521	0,488	0,504	0,496	0,504	0,500
AVG.	*0,431*	*0,298*	*0,332*	*0,414*	*0,476*	*0,440*	*0,403*	*0,496*	*0,434*	*0,390*	*0,361*	*0,368*

Table 4: precision, recall and F1 scores, calculated for each team's best run for all the language combinations.

JU-CSE-NLP [pivoting, compositional] (Neogi et al., 2012) uses Microsoft Bing translator[7] to produce monolingual English pairs. Separate lexical mapping scores are calculated (from *T1* to *T2* and vice-versa) considering different types of information and similarity metrics. Binary entailment decisions are then heuristically combined into single decisions.

Sagan [pivoting, multi-class] (Castillo and Cardenas, 2012) adopts a pivoting method using Google Translate, and trains a monolingual system based on a SVM multi-class classifier. A CLTE corpus derived from the RTE-3 dataset is also used as a source of additional training material.

[7] http://www.microsofttranslator.com/

SoftCard [pivoting, multi-class] (Jimenez et al., 2012) after automatic translation with Google Translate, uses SVMs to learn entailment decisions based on information about the cardinality of: *T1*, *T2*, their intersection and their union. Cardinalities are computed in different ways, considering tokens in *T1* and *T2*, their IDF, and their similarity (computed with edit-distance)

UAlacant [pivoting, multi-class] (Esplà-Gomis et al., 2012) exploits translations obtained from Google Translate, Microsoft Bing translator, and the Apertium open-source MT platform (Forcada et al., 2011).[8] Then, a multi-class SVM classifier is used to take entailment decisions using information about overlapping sub-segments as features.

7 Conclusion

Despite the novelty of the problem and the difficulty to capture multi-directional entailment relations across languages, the first round of the *Cross-lingual Textual Entailment for Content Synchronization* task organized within SemEval-2012 was a successful experience. This year a new interesting challenge has been proposed, a benchmark for four language combinations has been released, baseline results have been proposed for comparison, and a monolingual English dataset has been produced as a by-product which can be useful for monolingual TE research. The interest shown by participants was encouraging: 10 teams submitted a total of 92 runs for all the language pairs proposed. Overall, the results achieved on all datasets are encouraging, with best systems significantly outperforming the proposed baselines. It is worth observing that the nature of the task, which lies between semantics and machine translation, led to the participation of teams coming from both these communities, showing interesting opportunities for integration and mutual improvement. The proposed approaches reflect this situation, with teams traditionally working on MT now dealing with entailment, and teams traditionally participating in the RTE challenges now dealing with cross-lingual alignment techniques. Our ambition, for the future editions of the CLTE task, is to further consolidate the bridge between the semantics and MT communities.

[8] http://www.apertium.org/

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853). The authors would also like to acknowledge Giovanni Moretti from CELCT for evaluation scripts and technical assistance, and the volunteer translators that contributed to the creation of the dataset: María Sol Accossato, Laura Barthélémy, Claudia Biacchi, Jane Brendler, Amandine Chantrel, Hanna Cheda Patete, Ellen Clancy, Rodrigo Damian Tejeda, Daniela Dold, Valentina Frattini, Debora Hedy Amato, Geniz Hernandez, Bénédicte Jeannequin, Beate Jones, Anne Kauffman, Marcia Laura Zanoli, Jasmin Lewis, Alicia López, Domenico Loseto, Sabrina Luján Sánchez, Julie Mailfait, Gabriele Mark, Nunzio Pruiti, Lourdes Rey Cascallar, Sylvie Martlew, Aleane Salas Velez, Monica Scalici, Andreas Schwab, Marianna Sicuranza, Chiara Sisler, Stefano Tordazzi, Yvonne.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Julio Castillo and Marina Cardenas. 2012. Sagan: A Cross Lingual Textual Entailment system based on Machine Traslation. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Ido Dagan and Oren Glickman. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Proceedings of the PASCAL Workshop of Learning Methods for Text Understanding and Mining*.

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012. UAlacant: Using Online Machine Translation for Cross-Lingual Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Mikel L. Forcada, Ginestí-Rosell Mireia, Nordfalk Jacob, O'Regan Jim, Ortiz-Rojas Sergio, Pérez-Ortiz Juan A., Sánchez-Martínez Felipe, Ramírez-Sánchez Gema,

and Tyers Francis M. 2011. Apertium: a Free/Open-Source Platform for Rule-Based Machine Translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality + ML: Learning Adaptive Similarity Functions for Cross-lingual Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Milen Kouylekov and Matteo Negri. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*.

Milen Kouylekov. 2012. CELI: An Experiment with Cross Language Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

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

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*.

Yashar Mehdad, Matteo Negri, and José G. C. de Souza. 2012a. FBK: Cross-Lingual Textual Entailment Without Translation. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012b. Detecting Semantic Equivalence and Information Disparity in Cross-lingual Documents. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012c. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT 2012)*.

Fandong Meng, Hao Xiong, and Qun Liu. 2012. ICT: A Translation based Cross-lingual Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Matto Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*.

Snehasis Neogi, Partha Pakray, Sivaji Bandyopadhyay, and Alexander Gelbukh. 2012. JU-CSE-NLP: Language Independent Cross-lingual Textual Entailment System. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Franz J. Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*.

Alpár Perini. 2011. Detecting textual entailment with conditions on directional text relatedness scores. *Studia Universitatis Babes-Bolyai Series Informatica*, LVI(2):13–18.

Alpár Perini. 2012. DirRelCond3: Detecting Textual Entailment Across Languages With Conditions On Directional Text Relatedness Scores. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Darnes Vilariño, David Pinto, Mireya Tovar, Saul León, and Esteban Castillo. 2012. BUAP: Lexical and Semantic Similarity for Cross-lingual Textual Entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Katharina Wäschle and Sascha Fendrich. 2012. HDU: Cross-lingual Textual Entailment with SMT Features. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.