Semeval-2013 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
UBC
Vancouver, Canada
mehdad@cs.ubc.ca

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

Danilo Giampiccolo
CELCT
Trento, Italy
giampiccolo@celct.it

Abstract

This paper presents the second round of the task on *Cross-lingual Textual Entailment for Content Synchronization*, organized within SemEval-2013. 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 (six teams, 61 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. Given two texts in different languages, the cross-lingual textual entailment (CLTE) task consists of deciding if the meaning of one text can be inferred from the meaning of the other text. Cross-linguality represents an interesting direction for research on recognizing textual entailment (RTE), especially due to its possible application in a variety of tasks. Among others (*e.g.* question answering, information retrieval, information extraction, and document summarization), multilingual content synchronization represents a challenging application scenario to evaluate CLTE recognition components geared to the identification of sentence-level semantic relations.

Given two documents about the same topic written in different languages (*e.g.* Wikipedia pages), the content synchronization 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 (Monz et al., 2011; Bronner et al., 2012). 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 judgements for two text fragments are respectively mapped into judgements about semantic equivalence and novelty. The task can also be seen as a machine translation evaluation problem, where judgements about semantic equivalence and novelty depend on the possibility to fully or partially translate a text fragment into the other.

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 features contributing to the entailment decision), the standard sentence and word alignment programs used in SMT offer a strong baseline for CLTE (Mehdad et al., 2011;

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, June 14-15, 2013.

```
<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 el 27 de enero de 1756 en Salzburgo</t1>
   <t2> Mozart was born in 1756 in the city of Salzburg</t2>
 </pair>
<pair id="3" entailment="backward">
   <t1>Mozart nació en la ciudad de Salzburgo</t1>
   <t2>Mozart was born on the 27th January 1756 in 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: Example of SP-EN CLTE pairs.

Mehdad et al., 2012). 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 offers large room for mutual improvement.

Building on the success of the first CLTE evaluation organized within SemEval-2012 (Negri et al., 2012a), the remainder of this paper describes the second evaluation round organized within SemEval-2013. The following sections provide an overview of the datasets used, the participating systems, the approaches adopted, the achieved results, and the lessons learned.

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 judgements (see Figure 1 for Spanish/English examples of each judgement):

- **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 either direction;

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.

3 Dataset description

The CLTE-2013 dataset is composed of four CLTE corpora created for the following language combinations: Spanish/English (SP-EN), Italian/English (IT-EN), French/English (FR-EN), German/English (DE-EN). Each corpus consists of 1,500 sentence pairs (1,000 for training and 500 for test), balanced across the four entailment judgements.

In this year's evaluation, as training set we used the CLTE-2012 corpus[1] that was created for the SemEval-2012 evaluation exercise[2] (including both training and test sets). The CLTE-2013 test set was created from scratch, following the methodology described in the next section.

3.1 Data collection and annotation

To collect the entailment pairs for the 2013 test set we adopted a slightly modified version of the crowdsourcing methodology followed to create the CLTE-2012 corpus (Negri et al., 2011). The main difference with last year's procedure is that we did not take advantage of crowdsourcing for the whole data collection process, but only for part of it.

As for CLTE-2012, the collection and annotation process consists of the following steps:

1. First, English sentences were manually extracted from Wikipedia and Wikinews. The selected sentences represent one of the elements (*T1*) of each entailment pair;

[1] http://www.celct.it/resources.php?id_page=CLTE
[2] http://www.cs.york.ac.uk/semeval-2012/task8/

2. Next, each *T1* was modified in various ways in order to obtain a corresponding *T2*. While in the CLTE-2012 dataset the whole *T2* creation process was carried out through crowdsourcing, for the CLTE-2013 test set we crowdsourced only the first phase of *T1* modification, namely the creation of paraphrases. Focusing on the creation of high quality paraphrases, we followed the crowdsourcing methodology experimented in Negri et al. (2012b), in which a paraphrase is obtained through an iterative modification process of an original sentence, by asking workers to introduce meaning-preserving lexical and syntactic changes. At each round of the iteration, new workers are presented with the output of the previous iteration in order to increase divergence from the original sentence. At the end of the process, only the more divergent paraphrases according to the Lesk score (Lesk, 1986) are selected. As for the second phase of *T2* creation process, this year it was carried out by expert annotators, who followed the same criteria used last year for the crowdsourced tasks, i.e. *i)* remove information from the input (paraphrased) sentence and *ii)* add information from sentences surrounding *T1* in the source article;

3. Each *T2* was then paired to the original *T1*, and the resulting pairs were annotated with one of the four entailment judgements. In order to reduce the correlation between the difference in sentences' length and entailment judgements, 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.[3] The final result is a monolingual English dataset annotated with multi-directional entailment judgements, 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 cross-annotated and filtered to retain only those pairs with full agreement in the entailment judgement between two expert annotators. 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 is 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.

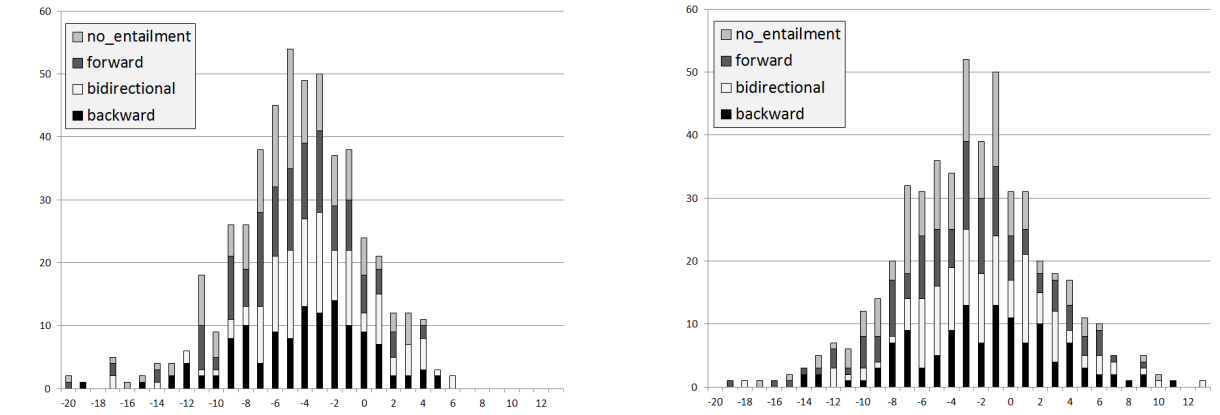
3.2 Dataset statistics

As described in section 3.1, the methodology followed to create the training and test sets was the same except for the crowdsourced tasks. This allowed us to obtain two datasets with the same balance across the entailment judgements, and to keep under control the distribution of the pairs for different *length_diff* values in each language combination.

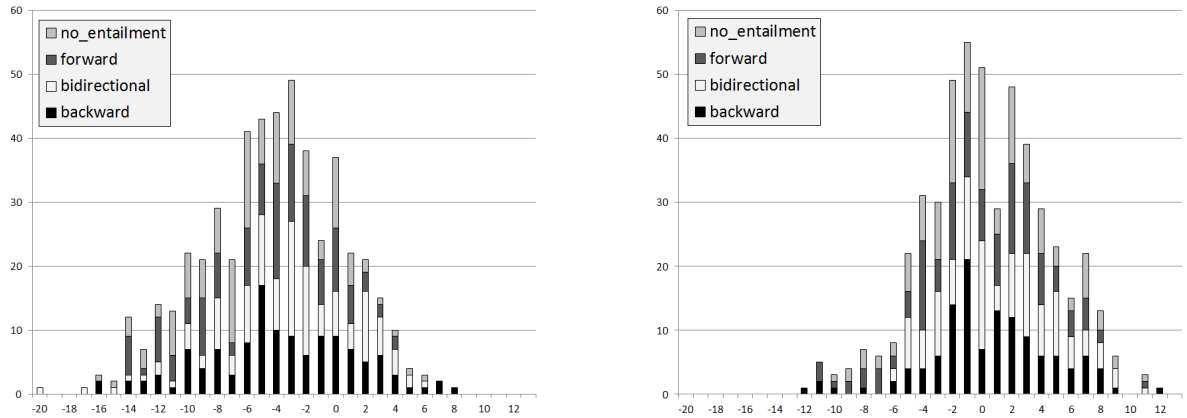
Training Set. The training set is composed of 1,000 CLTE pairs for each language combination, balanced across the four entailment judgements (bidirectional, forward, backward, and no_entailment). As shown in Table 1, our data collection procedure led to a dataset where the majority of the pairs falls in the +5 -5 *length_diff* range for each language pair (67.2% on average across the four language pairs). This characteristic is particularly relevant as our assumption is that such data distribution makes entailment judgements based on mere surface features such as sentence length ineffective, thus encouraging the development of alternative, deeper processing strategies.

Test Set. The test set is composed of 500 entailment pairs for each language combination, balanced across the four entailment judgements. As shown in Table 2, also in this dataset the majority of the collected entailment pairs is uniformly distributed

[3] 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*.



(a) SP−EN (b) IT−EN



(c) FR−EN (d) DE−EN

Figure 2: Pair distribution in the 2013 test set: total number of pairs (y-axis) for different *length_diff* values (x-axis).

	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
% (out of 1,000)	64.4	68.6	67.4	68.5

Table 1: Training set pair distribution within the -5/+5 *length_diff* range.

	SP-EN	IT-EN	FR-EN	DE-EN
backward	82	89	82	102
bidirectional	89	92	90	106
forward	69	78	76	98
no_entailment	71	80	59	100
ALL	311	339	307	406
% (out of 500)	62.2	67.8	61.4	81.2

Table 2: Test set pair distribution within the -5/+5 *length_diff* range.

in the [-5,+5] *length_diff* range (68.1% on average across the four language pairs).

However, comparing training and test set for each language pair, it can be seen that while the Spanish-English and Italian-English datasets are homogeneous with respect to the *length_diff* feature, the French-English and German-English datasets present noticeable differences between training and test set. These figures show that, despite the considerable effort spent to produce comparable training and test sets, the ideal objective of a full homogeneity between the datasets for these two languages was difficult to reach.

Complete details about the distribution of the pairs in terms of *length_diff* for the four cross-lingual corpora in the test set are provided in Figure 2. Vertical bars represent, for each *length_diff* value, the proportion of pairs belonging to the four entailment classes.

4 Evaluation metrics and baselines

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

To allow comparison with the CLTE-2012 results, the same three baselines were calculated on the CLTE-2013 test set for each language combination. The first one is the 0.25 accuracy score obtained by assigning each test pair in the balanced dataset to one of the four classes. The other two baselines consider the length difference between *T1* and *T2*:

- **Composition of binary judgements (Binary).** To calculate this baseline an SVM classifier is trained to take binary entailment decisions ("YES", "NO"). The classifier uses *length(T1)/length(T2)* and *length(T2)/length(T1)* as features respectively to check for entailment from *T1* to *T2* and vice-versa. For each test pair, the unidirectional judgements returned by the two classifiers are composed into a single multi-directional judgement ("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 judgements.

Both the baselines have been calculated with the LIBSVM package (Chang and Lin, 2011), using default parameters. Baseline results are reported in Table 3.

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. With respect to last year's evaluation, we can observe a slight drop in the binary classification baseline results. This might be due to the fact that the length distribution of examples is slightly different this year. However, there are no significant differences between the multi-class baseline results of this year in comparison with the previous round results. This might suggest that multi-class classification is a more robust approach for recognizing multi-directional entailment relations. Moreover, both baselines failed in capturing the *"no-entailment"* examples in all datasets ($F1_{no-entailment} = 0$).

	SP-EN	**IT-EN**	**FR-EN**	**DE-EN**
1-class	0.25	0.25	0.25	0.25
Binary	0.35	0.39	0.37	0.39
Multi-class	**0.43**	**0.44**	**0.42**	**0.42**

Table 3: Baseline accuracy results.

5 Submitted runs and results

Like in the 2012 round of the CLTE task, participants were allowed to submit up to five runs for each language combination. A total of twelve teams registered for participation and downloaded the training set. Out of them, six[4] submitted valid runs. Five teams produced submissions for all the four language combinations, while one team participated only in the DE-EN task. In total, 61 runs have been submitted and evaluated (16 for DE-EN, and 15 for each of the other language pairs).

Accuracy results are reported in Table 4. As can be seen from the table, the performance of the best systems is quite similar across the four language combinations, with the best submissions achieving results in the 43.4-45.8% accuracy interval. Similarly, also average and median results are close to each other, with a small drop on DE-EN. This drop might be explained by the difference between the training and test set with respect to the *length_diff* feature. Moreover, the performance of DE-EN automatic translation might affect approaches based on "pivoting", (*i.e.* addressing CLTE by automatically translating *T1* in the same language of *T2*, as described in Section 6).

[4]Including the task organizers.

System_name	SP-EN	IT-EN	FR-EN	DE-EN
altn_run1*	**0.428**	**0.432**	**0.420**	**0.388**
BUAP_run1	0.364	0.358	0.368	0.322
BUAP_run2	0.374	0.358	0.364	0.318
BUAP_run3	0.380	0.358	0.362	0.316
BUAP_run4	0.364	**0.388**	**0.392**	**0.350**
BUAP_run5	**0.386**	0.360	0.372	0.318
celi_run1	0.340	**0.324**	0.334	**0.342**
celi_run2	**0.342**	**0.324**	**0.340**	**0.342**
ECNUCS_run1	**0.428**	**0.426**	0.438	0.422
ECNUCS_run2	0.404	0.420	0.450	0.436
ECNUCS_run3	0.408	**0.426**	**0.458**	0.432
ECNUCS_run4	0.422	0.416	0.436	**0.452**
ECNUCS_run5	0.392	0.402	0.442	0.426
SoftCard_run1	**0.434**	**0.454**	0.416	**0.414**
SoftCard_run2	0.432	0.448	**0.426**	0.402
umelb_run1	–	–	–	**0.324**
Highest	0.434	0.454	0.458	0.452
Average	0.404	0.404	0.401	0.378
Median	0.428	0.426	0.420	0.369
Lowest	0.342	0.324	0.340	0.324

Table 4: CLTE-2013 accuracy results (61 runs) over the 4 language combinations. Highest, average, median and lowest scores are calculated considering only the best run for each team (*task organizers' system).

Compared to the results achieved last year, shown in Table 5, a sensible decrease in the highest scores can be observed. While in CLTE-2012 the top systems achieved an accuracy well above 0.5 (with a maximum of 0.632 in SP-EN), the results for this year are far below such level (the peak is now at 45,8% for FR-EN). A slight decrease with respect to 2012 can also be noted for average performances. However, it's worth remarking the general increase of the lowest and median scores, which are less sensitive to isolate outstanding results achieved by single teams. This indicates that a progress in CLTE research has been made building on the lessons learned after the first round of the initiative.

To better understand the behaviour of each system, Table 6 provides separate precision, recall, and F1 scores for each entailment judgement, calculated over the best runs of each participating team. In contrast to CLTE-2012, where the "bidirectional" and "no entailment" categories consistently proved to be more problematic than "forward" and "backward" judgements, this year's results are more homogeneous across the different classes. Nevertheless, on average, the classification of "bidirectional" pairs is still worse for three language pairs (SP-EN, IT-EN and FR-EN), and results for "no entailment" are lower for two of them (SP-EN and DE-EN).

	SP-EN	IT-EN	FR-EN	DE-EN
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 5: CLTE-2012 accuracy results. Highest, average, median and lowest scores are calculated considering only the best run for each team.

As regards the comparison with the baselines, this year's results confirm that the *length_diff*-based baselines are hard to beat. More specifically, most of the systems are slightly above the binary classification baseline (with the exception of the DE-EN dataset where only two systems out of six outperformed it), whereas for all the language combinations the multi-class baseline was beaten only by the best participating system.

This shows that, despite the effort in keeping the distribution of the entailment classes uniform across different *length_diff* values, eliminating the correlation between sentence length and correct entailment decisions is difficult. As a consequence, although disregarding semantic aspects of the problem, features considering length information are quite effective in terms of overall accuracy. Such features, however, perform rather poorly when dealing with challenging cases (*e.g. "no-entailment"*), which are better handled by participating systems.

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 judgements without preliminary translation.

- **Composition of binary judgements vs. Multi-class classification.** Compositional approaches map unidirectional ("YES"/"NO")

	SP-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
altn_full_spa-eng	**0.509**	**0.464**	**0.485**	0.440	0.264	0.330	**0.464**	0.416	0.439	0.357	**0.568**	0.438
BUAP_spa-eng_run5	0.446	0.360	0.398	0.521	0.296	0.378	0.385	0.456	0.418	0.300	0.432	0.354
celi_spa-eng_run2	0.396	0.352	0.373	0.431	0.400	0.415	0.325	0.328	0.327	0.245	0.288	0.265
ECNUCS_spa-eng_run1	0.458	0.432	0.444	0.533	0.320	0.400	0.406	0.416	0.411	**0.380**	0.544	**0.447**
SoftCard_spa-eng_run1	0.462	0.344	0.394	**0.619**	**0.480**	**0.541**	0.418	**0.472**	**0.444**	0.325	0.440	0.374
AVG.	0.454	0.390	0.419	0.509	0.352	0.413	0.400	0.418	0.408	0.321	0.454	0.376

	IT-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
altn_full_ita-eng	0.448	0.376	0.409	0.417	0.344	0.377	**0.512**	**0.496**	**0.504**	**0.374**	**0.512**	**0.432**
BUAP_ita-eng_run4	0.418	0.328	0.368	0.462	0.384	0.419	0.379	0.440	0.407	0.327	0.400	0.360
celi_ita-eng_run1	0.288	0.256	0.271	0.395	0.408	0.402	0.336	0.304	0.319	0.279	0.328	0.301
ECNUCS_ita-eng_run1	0.422	**0.456**	0.438	0.592	0.336	0.429	0.440	0.440	0.440	0.349	0.472	0.401
SoftCard_ita-eng_run1	**0.514**	**0.456**	**0.483**	**0.612**	**0.480**	**0.538**	0.392	0.464	0.425	0.364	0.416	0.388
AVG.	0.418	0.374	0.394	0.496	0.390	0.433	0.412	0.429	0.419	0.339	0.426	0.376

	FR-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
altn_full_fra-eng	0.405	0.392	0.398	0.420	0.296	0.347	0.500	0.440	0.468	0.381	0.552	0.451
BUAP_fra-eng_run4	0.407	**0.472**	0.437	0.431	0.376	0.402	0.379	0.376	0.378	0.352	0.344	0.348
celi_fra-eng_run2	0.394	0.344	0.368	0.364	0.376	0.370	0.352	0.352	0.352	0.263	0.288	0.275
ECNUCS_fra-eng_run3	0.422	0.432	0.427	**0.667**	0.352	0.461	**0.514**	**0.432**	**0.470**	**0.383**	**0.616**	**0.472**
SoftCard_fra-eng_run2	**0.477**	0.416	**0.444**	0.556	**0.400**	**0.465**	0.412	**0.432**	0.422	0.335	0.456	0.386
AVG.	0.421	0.411	0.415	0.488	0.360	0.409	0.431	0.406	0.418	0.343	0.451	0.386

	DE-EN											
	Forward			Backward			No entailment			Bidirectional		
System name	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	F1
altn_full_deu-eng	0.432	0.408	0.420	0.378	0.272	0.316	**0.445**	0.392	**0.417**	0.330	0.480	0.391
BUAP_deu-eng_run4	0.364	0.344	0.354	0.389	0.280	0.326	0.352	0.352	0.352	0.317	0.424	0.363
celi_deu-eng_run1	0.346	0.352	0.349	0.414	**0.424**	0.419	0.351	0.264	0.301	0.272	0.328	0.297
ECNUCS_deu-eng_run4	0.429	**0.432**	**0.430**	**0.611**	0.352	**0.447**	0.415	0.392	0.403	**0.429**	**0.632**	**0.511**
SoftCard_deu-eng_run1	**0.511**	0.368	0.428	0.527	0.384	0.444	0.417	**0.400**	0.408	0.317	0.504	0.389
umelb_deu-eng_run1	0.323	0.320	0.321	0.240	0.184	0.208	0.362	0.376	0.369	0.347	0.416	0.378
AVG.	0.401	0.371	0.384	0.426	0.316	0.360	0.390	0.363	0.375	0.335	0.464	0.389

Table 6: Precision, recall and F1 scores, calculated for each team's best run for all the language combinations.

entailment decisions taken separately into single judgements (similar to the *Binary* baseline in Section 4). Methods based on multi-class classification directly assign one of the four entailment judgements to each test pair (similar to our *Multi-class* baseline).

In contrast with CLTE-2012, where the combination of pivoting and compositional methods was the option adopted by the majority of the approaches, this year's solutions do not show a clear trend. Concerning the former dimension, participating systems are equally distributed in cross-lingual and pivoting methods relying on external automatic translation tools. Regarding the latter dimension, in addition to compositional and multi-class strategies, also alternative solutions that leverage more sophisticated meta-classification strategies have been proposed. Besides the recourse to MT tools (*e.g.* Google Translate), other tools and resources used by participants include: WordNet, word alignment tools (*e.g.* Giza++), part-of-speech taggers (*e.g.* Stanford POS Tagger), stemmers (*e.g.* Snowball), machine learning libraries (*e.g.* Weka, SVMlight), parallel corpora (*e.g.* Europarl), and stopword lists. More in detail:

ALTN [cross-lingual, compositional] (Turchi and Negri, 2013) adopts a supervised learning method based on features that consider word alignments between the two sentences obtained with GIZA++ (Och et al., 2003). Binary entailment judgements are taken separately, and combined into final CLTE decisions.

BUAP [pivoting, multi-class and meta-classifier] (Vilariño et al., 2013) adopts a pivoting method based on translating *T1* into the language of

T2 and vice versa (using Google Translate[5]). Similarity measures (*e.g.* Jaccard index) and features based on n-gram overlap, computed at the level of words and part of speech categories, are used (either alone or in combination) by different classification strategies including: multi-class, a meta-classifier (*i.e.* combining the output of 2/3/4-class classifiers), and majority voting.

CELI [cross-lingual, meta-classifier] (Kouylekov, 2013) uses dictionaries for word matching, and a multilingual corpus extracted from Wikipedia for term weighting. A variety of distance measures implemented in the RTE system EDITS (Kouylekov and Negri, 2010; Negri et al., 2009) are used to extract features to train a meta-classifier. Such classifier combines binary decisions ("YES"/"NO") taken separately for each of the four CLTE judgements.

ECNUCS [pivoting, multi-class] (Jiang and Man, 2013) uses Google Translate to obtain the English translation of each T1. After a pre-processing step aimed at maximizing the commonalities between the two sentences (*e.g.* abbreviation replacement), a number of features is extracted to train a multi-class SVM classifier. Such features consider information about sentence length, text similarity/difference measures, and syntactic information.

SoftCard [pivoting, multi-class] (Jimenez et al., 2013) 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.

Umelb [cross-lingual, pivoting, compositional] (Graham et al., 2013) adopts both pivoting and cross-lingual approaches. For the latter, GIZA++ was used to compute word alignments between the input sentences. Word alignment features are used to train binary SVM classifiers whose decisions are eventually composed into CLTE judgements.

7 Conclusion

Following the success of the first round of the *Cross-lingual Textual Entailment for Content Synchronization* task organized within SemEval-2012, a second evaluation task has been organized within SemEval-2013. Despite the decrease in the number of participants (six teams - four less than in the first round - submitted a total of 61 runs) the new experience is still positive. In terms of data, a new test set has been released, extending the old one with 500 new CLTE pairs. The resulting 1,500 cross-lingual pairs, aligned over four language combinations (in addition to the monolingual English version), and annotated with multiple entailment relations, represent a significant contribution to the research community and a solid starting point for further developments.[6] In terms of results, in spite of a significant decrease of the top scores, the increase of both median and lower results demonstrates some encouraging progress in CLTE research. Such progress is also demonstrated by the variety of the approaches proposed. While in the first round most of the teams adopted more intuitive and "simpler" solutions based on pivoting (*i.e.* translation of *T1* and *T2* in the same language) and compositional entailment decision strategies, this year new ideas and more complex solutions have emerged. Pivoting and cross-lingual approaches are equally distributed, and new classification methods have been proposed. Our hope is that the large room for improvement, the increase of available data, and the potential of CLTE as a way to address complex NLP tasks and applications will motivate further research on the proposed problem.

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-248531). The authors would also like to acknowledge Pamela Forner and Giovanni Moretti from CELCT, and the volunteer translators that contributed to the creation of the dataset: Giusi Calo, Victoria Díaz, Bianca Jeremias, Anne Kauffman, Laura López Ortiz, Julie Mailfait, Laura Morán Iglesias, Andreas Schwab.

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

[6] Together with the datasets derived from translation of the RTE data (Negri and Mehdad, 2010), this is the only material currently available to train and evaluate CLTE systems.

References

Amit Bronner, Matteo Negri, Yashar Mehdad, Angela Fahrni, and Christof Monz. 2012. Cosyne: Synchronizing multilingual wiki content. In *Proceedings of WikiSym 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*.

Yvette Graham, Bahar Salehi, and Tim Baldwin. 2013. Unimelb: Cross-lingual Textual Entailment with Word Alignment and String Similarity Features. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Zhao Jiang and Lan Man. 2013. ECNUCS: Recognizing Cross-lingual Textual Entailment Using Multiple Feature Types. . In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2013. Soft Cardinality-CLTE: Learning to Identify Directional Cross-Lingual Entailmens from Cardinalities and SMT. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

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

Milen Kouylekov. 2013. Celi: EDITS and Generic Text Pair Classification. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Michael Lesk. 1986. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th annual international conference on Systems documentation (SIGDOC86)*.

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 Marcello Federico. 2012. 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)*.

Christof Monz, Vivi Nastase, Matteo Negri, Angela Fahrni, Yashar Mehdad, and Michael Strube. 2011. Cosyne: a framework for multilingual content synchronization of wikis. In *Proceedings of WikiSym 2011*.

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

Matteo Negri, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. 2009. Towards extensible textual entailment engines: the edits package. In *AI* IA 2009: Emergent Perspectives in Artificial Intelligence*, pages 314–323. Springer.

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)*.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012a. Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*.

Matteo Negri, Yashar Mehdad, Alessandro Marchetti, Danilo Giampiccolo, and Luisa Bentivogli. 2012b. Chinese Whispers: Cooperative Paraphrase Acquisition. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC12)*, volume 2, pages 2659–2665.

F. Och, H. Ney, F. Josef, and O. H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Marco Turchi and Matteo Negri. 2013. ALTN: Word Alignment Features for Cross-Lingual Textual Entailment. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Darnes Vilariño, David Pinto, Saul León, Yuridiana Alemán, and Helena Gómez-Adorno. 2013. BUAP: N-gram based Feature Evaluation for the Cross-Lingual Textual Entailment Task. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.