Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora

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

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

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

Danilo Giampiccolo
CELCT
Trento, Italy
giampiccolo@celct.it

Alessandro Marchetti
CELCT
Trento, Italy
amarchetti@celct.it

Abstract

We address the creation of cross-lingual textual entailment corpora by means of crowdsourcing. Our goal is to define a cheap and replicable data collection methodology that minimizes the manual work done by expert annotators, without resorting to preprocessing tools or already annotated monolingual datasets. In line with recent works emphasizing the need of large-scale annotation efforts for textual entailment, our work aims to: *i)* tackle the scarcity of data available to train and evaluate systems, and *ii)* promote the recourse to crowdsourcing as an effective way to reduce the costs of data collection without sacrificing quality. We show that a complex data creation task, for which even experts usually feature low agreement scores, can be effectively decomposed into simple subtasks assigned to non-expert annotators. The resulting dataset, obtained from a pipeline of different jobs routed to Amazon Mechanical Turk, contains more than 1,600 aligned pairs for each combination of texts-hypotheses in English, Italian and German.

1 Introduction

Cross-lingual Textual Entailment (CLTE) has been recently proposed by (Mehdad et al., 2010; Mehdad et al., 2011) as an extension of Textual Entailment (Dagan and Glickman, 2004). The task consists of deciding, given a text (T) and an hypothesis (H) *in different languages*, if the meaning of H can be inferred from the meaning of T. As in other NLP applications, both for monolingual and cross-lingual TE, the availability of large quantities of annotated data is an enabling factor for systems development and evaluation. Until now, however, the scarcity of such data on the one hand, and the costs of creating new datasets of reasonable size on the other, have represented a bottleneck for a steady advancement of the state of the art.

In the last few years, monolingual TE corpora for English and other European languages have been created and distributed in the framework of several evaluation campaigns, including the RTE Challenge[1], the Answer Validation Exercise at CLEF[2], and the Textual Entailment task at EVALITA[3]. Despite the differences in the design of the tasks, all the released datasets were collected through similar procedures, always involving expensive manual work done by expert annotators. Moreover, in the data creation process, large amounts of hand-crafted T-H pairs often have to be discarded in order to retain only those featuring full agreement, in terms of the assigned entailment judgements, among multiple annotators. The amount of discarded pairs is usually high, contributing to increase the costs of creating textual entailment datasets[4].

The issues related to the shortage of datasets and the high costs for their creation are more evident

[1] http://www.nist.gov/tac/2011/RTE/

[2] http://nlp.uned.es/clef-qa/ave/

[3] http://www.evalita.it/2009/tasks/te

[4] For instance, in the first five RTE Challenges, the average effort needed to create 1,000 pairs featuring full agreement among 3 annotators was around 2.5 person-months. Typically, around 25% of the original pairs had to be discarded during the process, due to low inter-annotator agreement (Bentivogli et al., 2009).

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 670–679,
Edinburgh, Scotland, UK, July 27–31, 2011.

in the CLTE scenario, where: *i)* the only dataset currently available is an English-Spanish corpus obtained by translating the RTE-3 corpus (Negri and Mehdad, 2010), and *ii)* the application of the standard methods adopted to build RTE pairs requires proficiency in multiple languages, thus significantly increasing the costs of the data creation process.

To address these issues, in this paper we devise a cost-effective methodology to create cross-lingual textual entailment corpora. In particular, we focus on the following problems:

(1) Is it possible to collect T-H pairs minimizing the intervention of expert annotators? To address this question, we explore the feasibility of crowdsourcing the corpus creation process. As a contribution beyond the few works on TE/CLTE data acquisition, we define an effective methodology that: *i)* does not involve experts in the most complex (and costly) stages of the process, *ii)* does not require preprocessing tools, and *iii)* does not rely on the availability of already annotated RTE corpora.

(2) How can we guarantee good quality of the collected data at a low cost? We address the quality control issue through the decomposition of a complex task (*i.e.* creating and annotating entailment pairs) into smaller sub-tasks. Complex tasks are usually hard to explain in a simple way understandable to non-experts, difficult to accomplish, and not suitable for the application of the quality-check mechanisms provided by current crowdsourcing services. Our "divide and conquer" solution represents the first attempt to address a complex task involving content *generation* and *labelling* through the definition of a cheap and reliable pipeline of simple tasks which are easy to define, accomplish, and control.

(3) Can we adapt such methodology to collect cross-lingual T-H pairs? We tackle this question by separating the problem of creating and annotating TE pairs from the issues related to the multilingual dimension. Our solution builds on the assumption that entailment annotations can be projected across aligned T-H pairs in different languages. In this case, a complex multilingual task is reduced to a sequence of simpler subtasks where the most difficult one, the generation of entailment pairs, is entirely monolingual. Besides ensuring cost-effectiveness, our solution allows us to overcome the problem of finding workers that are proficient in multiple languages. Moreover, since the core monolingual tasks of the process are carried out by manipulating English texts, we are able to address the very large community of English speaking workers, with a considerable reduction of costs and execution time. Finally, as a by-product of our method, the acquired pairs are fully aligned for all language combinations, thus enabling meaningful comparisons between scenarios of different complexity (monolingual TE, and CLTE between close or distant languages).

We believe that, in the same spirit of recent works promoting large-scale annotation efforts around entailment corpora (Sammons et al., 2010; Bentivogli et al., 2010), the proposed approach and the resulting dataset[5] will contribute to meeting the strong need for resources to develop and evaluate novel solutions for textual entailment.

2 Related Works

Crowdsourcing services, such as Amazon Mechanical Turk[6] (MTurk) and CrowdFlower[7], have been recently used with success for a variety of NLP applications (Callison-Burch and Dredze, 2010). The idea is that the acquisition and annotation of large amounts of data needed to train and evaluate NLP tools can be carried out in a cost-effective manner by defining simple Human Intelligence Tasks (HITs) routed to a crowd of non-expert workers (aka "Turkers") hired through on-line marketplaces.

As regards textual entailment, the first work exploring the use of crowdsourcing services for data *annotation* is described in (Snow et al., 2008), which shows high agreement between non-expert annotations of the RTE-1 dataset and existing gold standard labels assigned by expert labellers.

Focusing on the actual *generation* of monolingual entailment pairs, (Wang and Callison-Burch, 2010) experiments the use of MTurk to collect facts and counter facts related to texts extracted from an existing RTE corpus annotated with named entities. Taking a step beyond the task of annotating exist-

[5]The CLTE corpora described in this paper will be made freely available for research purposes through the website of the funding EU Project CoSyne (http://www.cosyne.eu/).

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

[7]Although MTurk is directly accessible only to US citizens, the CrowdFlower service (http://crowdflower.com/) provides an interface to MTurk for non-US citizens.

ing datasets, and showing the feasibility of involving non-experts also in the generation of TE pairs, this approach is more relevant to our objectives. However, at least two major differences with our work have to be remarked. First, they still use available RTE data to obtain a monolingual TE corpus, whereas we pursue the more ambitious goal of generating from scratch aligned CLTE corpora for different language combinations. To this aim, we do not resort to already annotated data, nor language-specific preprocessing tools. Second, their approach involves qualitative analysis of the collected data only *a posteriori*, after manual removal of invalid and trivial generated hypotheses. In contrast, our approach integrates quality control mechanisms at all stages of the data collection/annotation process, thus minimizing the recourse to experts to check the quality of the collected material.

Related research in the CLTE direction is reported in (Negri and Mehdad, 2010), which describes the creation of an English-Spanish corpus obtained from the RTE-3 dataset by translating the English hypotheses into Spanish. Translations have been crowdsourced adopting a methodology based on translation-validation cycles, defined as separate HITs. Although simplifying the CLTE corpus creation problem, which is recast as the task of translating already available annotated data, this solution is relevant to our work for the idea of combining gold standard units and "validation HITS" as a way to control the quality of the collected data at runtime.

3 Quality Control of Crowdsourced Data

The design of data acquisition HITs has to take into account several factors, each having a considerable impact on the difficulty of instructing the workers, the quality and quantity of the collected data, the time and overall costs of the acquisition. A major distinction has to be made between jobs requiring data *annotation*, and those involving content *generation*. In the former case, Turkers are presented with the task of labelling input data referring to a fixed set of possible values (*e.g.* making a choice between multiple alternatives, assigning numerical scores to rank the given data). In the latter case, Turkers are faced with creative tasks consisting in the production of textual material (*e.g.* writing a correct translation, or a summary of a given text).

The ease of controlling the quality of the acquired data depends on the nature of the job. For annotation jobs, quality control mechanisms can be easily set up by calculating Turkers' agreement, by applying voting schemes, or by adding hidden gold units to the data to be annotated[8]. In contrast, the quality of the results of content generation jobs is harder to assess, due to the fact that multiple valid results are acceptable (*e.g.* the same content can be expressed, translated, or summarized in different ways). In such situations the standard quality control mechanisms are not directly applicable, and the detection of errors requires either costly manual verification at the end of the acquisition process, or more complex and creative solutions integrating HITs for quality check.

Most of the approaches to content generation proposed so far rely on *post hoc* verification to filter out undesired low-quality data (Mrozinski et al., 2008; Mihalcea and Strapparava, 2009; Wang and Callison-Burch, 2010). The few solutions integrating validation HITs address the translation of single sentences, a task that is substantially different from ours (Negri and Mehdad, 2010; Bloodgood and Callison-Burch, 2010). Compared to sentence translation, the task of creating CLTE pairs is both harder to explain without recurring to notions that are difficult to understand to non-experts (*e.g.* "semantic equivalence", "unidirectional entailment"), and harder to execute without mastering these notions. To tackle these issues the "divide and conquer" approach described in the next section consists in the decomposition of a difficult *content generation* job into easier subtasks that are: *i)* self-contained and easy to explain, *ii)* easy to execute without any NLP expertise, and *iii)* suitable for the integration of a variety of runtime control mechanisms (regional qualifications, gold units, "validation HITs") able to ensure a good quality of the collected material.

[8]Both MTurk and CrowdFlower provide means to check workers' reliability, and weed out untrusted ones without money waste. These include different types of qualification mechanisms, the possibility of giving work only to known trusted Turkers (only with MTurk), and the possibility of adding hidden gold standard units in the data to be annotated (offered as a built-in mechanism only by CrowdFlower).

4 CLTE Corpus Creation Methodology

Our approach builds on a pipeline of HITs routed to MTurk's workforce through the CrowdFlower interface. The objective is to collect aligned T-H pairs for different language combinations, reproducing an RTE-like annotation style. However, our annotation is not limited to the standard RTE framework, where only unidirectional entailment from T to H is considered. As a useful extension, we annotate any possible entailment relation between the two text fragments, including: *i)* bidirectional entailment (*i.e.* semantic equivalence between T and H), *ii)* unidirectional entailment from T to H, and *iii)* unidirectional entailment from H to T. The resulting pairs can be easily used to generate not only standard RTE datasets[9], but also general-purpose collections featuring multi-directional entailment relations.

4.1 Data Acquisition and Annotation

We collect large amounts of CLTE pairs carrying out the most difficult part of the process (the creation of entailment-annotated pairs) at a monolingual level. Starting from a set of parallel sentences in *n* languages, (*e.g.* L1, L2, L3), *n* entailment corpora are created: *one* monolingual (L1/L1), and *n-1* cross-lingual (L1/L2, and L1/L3).

The monolingual corpus is obtained by modifying the sentences only in one language (L1). Original and modified sentences are then paired and annotated to form an entailment dataset for L1. The CLTE corpora are obtained by combining the modified sentences in L1 with the original sentences in L2 and L3, and projecting to the multilingual pairs the annotations assigned to the monolingual pairs.

In principle, only two stages of the process require crowdsourcing multilingual tasks, but do not concern entailment annotations. The first one, at the beginning of the process, aims to obtain a set of parallel sentences to start with, and can be done in different ways (*e.g.* crowdsourcing the translation of a set of sentences). The second one, at the end of the process, consists of translating the modified L1 sentences into other languages (*e.g.* L2) in order to extend the corpus to cover new language combinations (*e.g.* L2/L2, L2/L3).

The execution of the two "multilingual" stages is not strictly necessary but depends on: *i)* the availability of parallel sentences to start the process, and *ii)* the actual objectives in terms of language combinations to be covered[10].

As regards the first stage, in this work we started from a set of 467 English/Italian/German aligned sentences extracted from parallel documents downloaded from the Cafebabel European Magazine[11]. Concerning the second multilingual stage, we performed only one round of translations from English to Italian to extend the *3* combinations obtained without translations (ENG/ENG, ENG/ITA, and ENG/GER) with the new language combinations ITA/ITA, ITA/ENG, and ITA/GER.

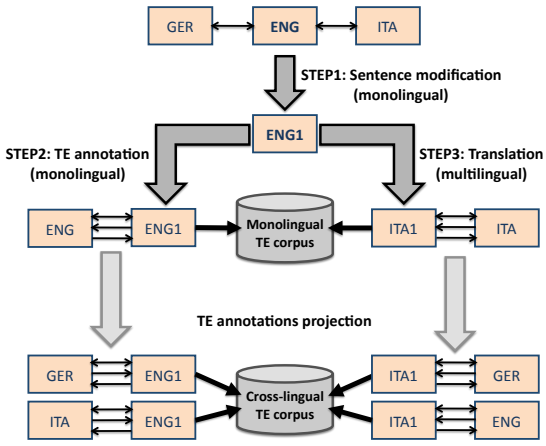

Figure 1: Corpus creation process.

The main steps of our corpus creation process, depicted in Figure 1, can be summarized as follows:

Step1: Sentence modification. The original English sentences (ENG) are modified through (monolingual) *generation* HITs asking Turkers to: *i)* preserve the meaning of the original sentences using different surface forms, or *ii)* slightly change their meaning by adding or removing content. Our assumption, in line with (Bos et al., 2009), is that

[9] With the positive examples drawn from bidirectional and unidirectional entailments from T to H, and the negative ones drawn from unidirectional entailments from H to T.

[10] Starting from parallel sentences in *n* languages, the *n* corpora obtained without recurring to translations can be augmented, by means of translation HITs, to create the full set of language combinations. Each round of translation adds 1 monolingual corpus, and *n-1* CLTE corpora.

[11] http://www.cafebabel.com/

another way to think about entailment is to consider whether one text *T1* adds new information to the content of another text *T*: if so, then *T* is entailed by *T1*.

The result of this phase is a set of texts (ENG1) that can be of three types:

1. Paraphrases of the original ENG texts, that will be used to create bidirectional entailment pairs (ENG↔ENG1);

2. More specific sentences (the outcome of content addition operations), used to create ENG←ENG1 unidirectional entailment pairs;

3. More general sentences (the outcome of content removal operations), used to create ENG→ENG1 unidirectional entailment pairs.

Step2: TE Annotation. Entailment pairs composed of the original sentences (ENG) and the modified ones (ENG1) are used as input of (monolingual) *annotation* HITs asking Turkers to decide which of the two texts contains more information. As a result, each ENG/ENG1 pair is annotated as an example of uni-/bidirectional entailment, and stored in the monolingual English corpus. Since the original ENG texts are aligned with the ITA and GER texts, the entailment annotations of ENG/ENG1 pairs can be projected to the other language pairs and the ITA/ENG1 and GER/ENG1 pairs are stored in the CLTE corpus. The possibility of projecting TE annotations is based on the assumption that the semantic information is mostly preserved during the translation process. This particularly holds at the denotative level (i.e. regarding the truth values of the sentence) which is crucial to semantic inference. At other levels (e.g. lexical) there might be slight semantic variations which, however, are very unlikely to play a crucial role in determining entailment relations.

Step3: Translation. The modified sentences (ENG1) are translated into Italian (ITA1) through (multilingual) *generation* HITs reproducing the approach described in (Negri and Mehdad, 2010). As a result, three new datasets are produced by automatically projecting annotations: the monolingual ITA/ITA1, and the cross-lingual ENG/ITA1 and GER/ITA1.

Since the solution adopted for sentence translation does not present novelty factors, the remainder of this paper will omit further details on it. Instead, the following sections will focus on the more challenging tasks of sentence modification and TE annotation.

4.2 Crowdsourcing Sentence Modification and TE Annotation

Sentence modification and TE annotation have been decomposed into a pipeline of simpler monolingual English sub-tasks. Such pipeline, depicted in Figure 2, involves several types of generation/annotation HITs designed to be easily understandable to non-experts. Each HIT consists of: *i)* a set of instructions for a specific task (*e.g.* paraphrasing a text), *ii)* the data to be manipulated (*e.g.* an English sentence), and *iii)* a test to check workers' reliability. To cope with the quality control issues discussed in Section 3, such tests are realized using gold standard units, either hidden in the data to be annotated (annotation HITs) or defined as test questions that workers must correctly answer (generation HITs). Moreover, regional qualifications are applied to all HITs. As a further quality check, all the annotation HITs consider Turkers' agreement as a way to filter out low quality results (only annotations featuring agreement among 4 out of 5 workers are retained). The six HITs defined for each subtask can be described as follows:

1. Paraphrase (generation). Modify an English text (ENG), in order to produce a semantically equivalent variant (ENG1). As a reliability test, before creating the paraphrase workers are asked to judge if two English sentences contain the same information.

2. Grammaticality (annotation). Decide if an English sentence is grammatically correct. This validation HIT represents a quality check of the output of each generation task (*i.e.* paraphrasing, and add/remove information HITs).

3. Bidirectional Entailment (annotation). Decide whether two English sentences, the original ENG and the modified ENG1, contain the same information (*i.e.* are semantically equivalent).

4a. Add Information (generation). Modify an English text to create a more specific one by adding content. As a reliability test, before generating the

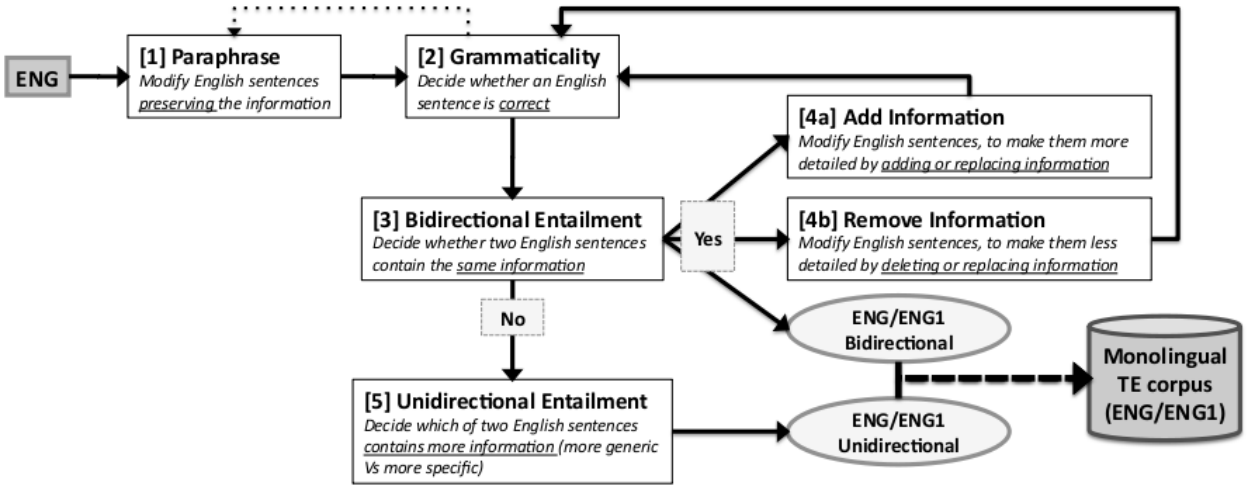

Figure 2: Sentence modification and TE annotation pipeline.

new sentence workers are asked to judge which of two given English sentences is more detailed.

4b. Remove Information (generation). Modify an English text to create a more general one by removing part of its content. As a reliability test, before generating the new sentence workers are asked to judge which of two given English sentences is less detailed.

5. Unidirectional Entailment (annotation). Decide which of two English sentences (the original ENG, and a modified ENG1) provides more information.

These HITs are combined in an iterative process that alternates text generation, grammaticality check, and entailment annotation steps. As a result, for each original ENG text we obtain multiple ENG1 variants of the three types (paraphrases, more general texts, and more specific texts) and, in turn, a set of annotated monolingual (ENG/ENG1) TE pairs.

As described in Section 4.1, the resulting monolingual English TE corpus (ENG/ENG1) is used to create the following mono/cross-lingual TE corpora:

- ITA/ENG1, and GER/ENG1 (by projecting TE annotations)
- ITA/ITA1, GER/ITA1, and ENG/ITA1 (by translating the ENG1 texts into Italian, and projecting TE annotations)

5 The Resulting CLTE Corpora

This section provides a quantitative and qualitative analysis of the results of our corpus creation methodology, focusing on the collected ENG-ENG1 monolingual dataset. It has to be remarked that, as an effect of the adopted methodology, all the observations and the conclusions drawn hold for the collected CLTE corpora as well.

5.1 Quantitative Analysis

Table 1 provides some details about each step of the pipeline shown in Figure 2. For each HIT the table presents: *i)* the number of items (sentences, or pairs of sentences) given in input, *ii)* the number of items (sentences or annotations) produced as output, *iii)* the number of items discarded when the agreement threshold was not reached, *iv)* the number of entailment pairs added to the corpus, *v)* the time (days and hours) required by the MTurk workforce to complete the job, and *vi)* the cost of the job.

In **HIT-1** (Paraphrase) 1,414 paraphrases were collected asking three different meaning-preserving modifications of each of the 467 original sentences[12]. From a practical point of view, such redundancy aims to ensure a sufficient number of grammatically correct and semantically equivalent modified sentences. From a theoretical point of view,

[12]Often, crowdsourced jobs return a number of output items that is slightly larger than required, due to the labour distribution mechanism internal to MTurk.

HIT	# Input items	# Output items	# Discarded items	# Pairs to corpus	MTurk time	Cost ($)
1. Paraphrase	467	1,414			5d+10.5h	45.48
2. Grammaticality	1,414	1,326	88 (6.22%)		1d+15h	56.88
3. Bidirectional Ent.	1,326	1,213 (yes=1,205 no=8)	113 (8.52%)	301	3d+2h	53.47
4a. Add Info	452	916			3d	37.02
4b. Remove Info	452	923			2d+22h	29.73
2. Grammaticality	1,839	1,749	90 (4.89%)		2d+5h	64.37
3. Bidirectional Ent.	1,749	1,438 (yes=148 no=1,290)	311 (17.78%)	148	3d+20.5h	70.52
5. Unidirectional Ent.	1,298	1,171	127 (9.78%)	1,171 (491 + 680)	8.5h	78.24
TOTAL			721	**1,620**	**22d+11h**	**435.71**

Table 1: The monolingual dataset creation pipeline.

collecting many variants of a small pool of original sentences aims to create pairs featuring different entailment relations with similar superficial forms. This, in principle, should allow to obtain a dataset which requires TE systems to focus more on deeper semantic phenomena than on the surface realization of the pairs.

The collected paraphrases were sent as input to **HIT-2** (Grammaticality). After this validation HIT, the number of acceptable paraphrases was reduced to 1,326 (with 88 discarded sentences, corresponding to 6.22% of the total).

The retained paraphrases were paired with their corresponding original sentences, and sent to **HIT-3** (Bidirectional Entailment) to be judged for semantic equivalence. The pairs marked as bidirectional entailments (1,205) were divided in three groups: 25% of the pairs (301) were directly stored in the final corpus, while the ENG1 paraphrases of the remaining 75% (904) were equally distributed to the next modification steps.

In both **HIT-4a** (Add Information) and **HIT-4b** (Remove information) two new modified sentences were asked for each of the 452 paraphrases received as input. The sentences collected in these generation tasks were respectively 916 and 923.

The new modified sentences were sent back to **HIT-2** (Grammaticality) and **HIT-3** (Bidirectional Entailment). As a result 1,438 new pairs were created; out of these, 148 resulted to be bidirectional entailments and were stored in the corpus.

Finally, the 1,298 entailment pairs judged as non-bidirectional in the two previously completed HIT-3 (8+1,290) were given as input to **HIT-5** (Unidirectional Entailment). The pairs which passed the agreement threshold were classified according to the judgement received, and stored in the corpus as unidirectional entailment pairs.

The analysis of Table 1 allows to formulate some considerations. First, the percentage of discarded items confirms the effectiveness of decomposing complex generation tasks into simpler subtasks that integrate validation HITs and quality checks based on non-experts' agreement. In fact, on average, around 9.5% of the generated items were discarded without experts' intervention[13]. Second, the amount of discarded items gives evidence about the relative difficulty of each HIT. As expected, we observe lower rejection rates, corresponding to higher inter-annotator agreement, for grammaticality HITs (5.55% on average) than for more complex entailment-related tasks (12.02% on average).

Looking at costs and execution time, it is hard to draw definite conclusions due to several factors that influence the progress of the crowdsourced jobs (*e.g.* the fluctuations of Turkers' performances, the time of the day at which jobs are posted, the difficulty to set the optimal cost for a given HIT[14]). On the one hand, as expected, the more creative "Add Info" task proved to be more demanding than the "Remove Info": even though it was paid more,

[13]Moreover, it is worthwhile noticing that around 20% of the collected items were automatically rejected (and not paid) due to failures on the gold standard controls created both for generation and annotation tasks.

[14]The payment for each HIT was set on the basis of a previous feasibility study aimed at determining the best trade-off between cost and execution time. However, replicating our approach would not necessarily result in the same costs.

it still took little more time to be completed. On the other hand, although the "Unidirectional Entailment" task was expected to be more difficult and thus rewarded more than the "Bidirectional Entailment" one, in the end it took notably less time to be completed. Nevertheless, the overall figures (435 USD, and about 22.5 days of MTurk work to complete the process)[15] clearly demonstrate the effectiveness of the approach. Even considering the time needed for an expert to manage the pipeline (*i.e.* one week to prepare gold units, and to handle the I/O of each HIT), these figures show that our methodology provides a cheaper and faster way to collect entailment data in comparison with the RTE average costs reported in Section 1.

As regards the amount of data collected, the resulting corpus contains 1,620 pairs with the following distribution of entailment relations: *i)* 449 bidirectional entailments, *ii)* 491 ENG→ENG1 unidirectional entailments, and *iii)* 680 ENG←ENG1 unidirectional entailments.

It must be noted that our methodology does not lead to the creation of pairs where some information is provided in one text and not in the other, and viceversa, as Example 1 shows:

Example 1.
ENG: New theories were emerging in the field of psychology.
ENG1: New theories were rising, which announced a kind of veiled racism.

These negative examples in both directions represent a natural extension of the dataset, relevant also for specific application-oriented scenarios, and their creation will be addressed in future work.

Besides the achievement of our primary objectives, the adopted approach led to some interesting by-products. First, the generated corpora are perfectly suitable to produce entailment datasets similar to those used in the traditional RTE evaluation framework. In particular, considering any possible entailment relation between two text fragments, our annotation subsumes the one proposed in RTE campaigns. This allows for the cost-effective generation of RTE-like annotations from the acquired corpora by combining ENG↔ENG1 and ENG→ENG1 pairs to form 940 positive examples (449+491), keeping the 680 ENG←ENG1 as negative examples. Moreover, by swapping ENG and ENG1 in the unidirectional entailment pairs, 491 additional negative examples and 680 positive examples can be easily obtained.

Finally, the output of HITs 1-2-3 in Table 1 represents *per se* a valuable collection of 1,205 paraphrases. This suggests the great potential of crowdsourcing for paraphrase acquisition.

5.2 Qualitative Analysis

Through manual verification of more than 50% of the corpus (900 pairs), a total number of 53 pairs (5.9%) were found incorrect. The different errors were classified as follows:

Type 1: Sentence modification errors. Generation HITs are a minor source of errors, being responsible for 10 problematic pairs. These errors are either introduced by generating a false statement (Example 2), or by forming a not fully understandable, awkward, or non-natural sentence (Example 3).

Example 2.
ENG: Kosovo was the subject of major riots in 1989.
ENG1: The Russian city of Kosovo was the subject of ...

Example 3.
ENG: Balat is the Kurdish-Armenian district of Instanbul.
ENG1: Balat is a place, which is the Kurdish-Armenian ...

Type 2: TE annotation errors. The notion of containing more/less information, used in the "Unidirectional Entailment" HIT, can mostly be applied straightforwardly to the entailment definition. However, the concept of "more/less detailed", which generally works for factual statements, in some cases is not applicable. In fact, the MTurk workers have regularly interpreted the instructions about the amount of information as concerning the quantity of concepts contained in a sentence. This is not always corresponding to the actual entailment relation between the sentences. As a consequence, 43 pairs featuring wrong entailment annotations were encountered. These errors can be classified as follows:

a) 13 pairs, where the added/removed information changes the meaning of the sentence. In these cases, the modified sentence was judged more/less specific

[15] Although by projecting annotations the ENG1/ITA and ENG1/GER CLTE corpora came for free, the ITA1/ITA, ITA1/ENG, and ITA1/GER combinations created by crowdsourcing translations added 45 USD and approximately 5 days to these figures.

than the original one, leading to unidirectional entailment annotation. On the contrary, in terms of the standard entailment definition, the correct annotation is "no entailment" (as in Example 4, which was annotated as ENG→ENG1):

Example 4.
ENG: If you decide to live in Bulgaria, you have to like difficulties because they are not difficulties, they are challenges.
ENG1: You have to like difficulties as they are not difficulties, they are challenges.

b) 10 pairs where the incorrect annotation is due to a coreference problem, as in:

Example 5.
ENG: John Smith is the new CEO of the company.
ENG1: He is the new CEO of the company.

These pairs were labelled as unidirectional entailments (in the example above ENG→ENG1), under the assumption that a proper name is more specific and informative than a pronoun. However, adhering to the TE definition, co-referring expressions are equivalent, and their realization does not play any role in the entailment decision. This implies that the correct entailment annotation is "bidirectional".

c) 9 pairs where the sentences are semantically equivalent, but contain a piece of information which is explicit in one sentence, and implicit in the other. In these cases, Turkers judged the sentence containing the explicit mention as more specific, and thus the pair was annotated as unidirectional entailment.

Example 6.
ENG: I hear the click of the trigger and the burst of bullets reach me immediately.
ENG1: I hear the trigger and the burst of bullets reach me instantly.

In Example 6, the expression *"the trigger"* in ENG1 implicitly means *"the click of the trigger"*, making the two sentences equivalent, and the entailment bidirectional (instead of ENG→ENG1).

d) 7 pairs where the information removed from or added to the sentence is not relevant to the entailment relation. In these cases, the modified sentence was judged less/more specific than the original one (and thus considered as unidirectional entailment), even though the correct judgement is "bidirectional", as in:

Example 7.
ENG: At the same time, AKP is struggling with its approach to the EU.
ENG1: AKP is struggling with its approach to the European Union.

e) 4 pairs where the added/removed information concerns universally quantified general statements, about which the interpretation of "more/less specific" given by Turkers resulted in the wrong annotation.

Example 8.
ENG: I think the success of multicultural couples depends on the size of the cultural gap between the two partners
ENG1: I believe the success of the couples depends on the size of the cultural gap between the 2 partners.

In Example 8, the additional information (*"multicultural"*) restricts the set to which it refers (*"couples"*) making ENG entailed by ENG1, and not vice versa as resulted from Turkers' annotation.

In light of this analysis, we conclude that the sentence modification methodology proved to be successful, as the low number of Type 1 errors shows. Considering that the most expensive phase in the creation of a TE dataset is the generation of the pairs, this is a significant achievement. Differently, the entailment assessment phase appears to be more problematic, accounting for the majority of errors. As shown by Type 2 errors, this is due to a partial misalignment between the instructions given in our HITs, and the formal definition of textual entailment. For this reason, further experimentation will explore different ways to instruct workers (*e.g.* asking to consider proper names and pronouns as equivalent) in order to reduce the amount of errors produced. As a final remark, considering that in the creation of a TE dataset the manual check of the annotated pairs represents a minor cost, even the involvement of experts to filter out wrong annotations would not decrease the cost-effectiveness of the proposed methodology.

6 Conclusions

There is an increasing need of annotated data to develop new solutions to the Textual Entailment problem, explore new entailment-related tasks, and set up experimental frameworks targeting real-world applications. Following the recent trends promoting annotation efforts that go beyond the established RTE Challenge framework (unidirectional entailment between monolingual T-H pairs), in this

paper we addressed the multilingual dimension of the problem. Our primary goal was the creation of large-scale collections of entailment pairs for different language combinations. Besides that, we considered cost effectiveness and replicability as additional requirements. To achieve our objectives, we developed a "divide and conquer" methodology based on crowdsourcing. Our approach presents several key innovations with respect to the related works on TE data acquisition. These include the decomposition of a complex content generation task in a pipeline of simpler subtasks accessible to a large crowd of non-experts, and the integration of quality control mechanisms at each stage of the process. The result of our work is the first large-scale dataset containing both monolingual and cross-lingual corpora for several combinations of texts-hypotheses in English, Italian, and German. Among the advantages of our method it is worth mentioning: *i)* the full alignment between the created corpora, *ii)* the possibility to easily extend the dataset to new languages, and *iii)* the feasibility of creating general-purpose corpora, featuring multi-directional entailment relations, that subsume the traditional RTE-like annotation.

Acknowledgments

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853). The authors would like to thank Emanuele Pianta for the helpful discussions, and Giovanni Moretti for the valuable support in the creation of the CLTE dataset.

References

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *Proceedings of TAC 2009*.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. *Proceedings of LREC 2010*.

Michael Bloodgood and Chris Callison-Burch. 2010. Using Mechanical Turk to Build Machine Translation Evaluation Sets. *Proceedings of the NAACL 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*.

Johan Bos, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual Entailment at EVALITA 2009. *Proceedings of EVALITA 2009*.

Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data With Amazons Mechanical Turk. *Proceedings NAACL-2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*.

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

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual Entailment. *Proceedings of NAACL-HLT 2010*.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment. *Proceedings of ACL-HLT 2011*.

Rada Mihalcea and Carlo Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. *Proceedings of ACL 2009*.

Joanna Mrozinski, Edward Whittaker, and Sadaoki Furui. 2008. Collecting a Why-Question Corpus for Development and Evaluation of an Automatic QA-System. *Proceedings of ACL 2008*.

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

Mark Sammons, V. G. Vinod Vydiswaran, and Dan Roth. 2010. Ask Not What Textual Entailment Can Do for You... *Proceedings of ACL 2010*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of EMNLP 2008*.

Rui Wang and Chris Callison-Burch. 2010. Cheap Facts and Counter-Facts. *Proceedings of the NAACL 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*.