Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush

Matteo Negri[1] and **Yashar Mehdad**[1,2]
FBK-Irst[1], University of Trento[2]
Trento, Italy
{negri,mehdad}@fbk.eu

Abstract

This paper reports on experiments in the creation of a bi-lingual Textual Entailment corpus, using non-experts' workforce under strict cost and time limitations ($100, 10 days). To this aim workers have been hired for translation and validation tasks, through the CrowdFlower channel to Amazon Mechanical Turk. As a result, an accurate and reliable corpus of 426 English/Spanish entailment pairs has been produced in a more cost-effective way compared to other methods for the acquisition of translations based on crowdsourcing. Focusing on two orthogonal dimensions (*i.e. reliability* of annotations made by non experts, and overall corpus creation *costs*), we summarize the methodology we adopted, the achieved results, the main problems encountered, and the lessons learned.

1 Introduction

Textual Entailment (TE) (Dagan and Glickman, 2004) has been proposed as a generic framework for modelling language variability. Given a *text* T and an *hypothesis* H, the task consists in deciding if the meaning of H can be inferred from the meaning of T. At the monolingual level, the great potential of integrating TE recognition (RTE) components into NLP architectures has been demonstrated in several areas, including question answering, information retrieval, information extraction, and document summarization. In contrast, mainly due to the absence of cross-lingual TE (CLTE) recognition components, similar improvements have not been achieved yet in any cross-lingual application. Along such direction, focusing on feasibility and architectural issues, (Mehdad et al., 2010) recently proposed baseline results demonstrating the potential of a simple approach that integrates Machine Translation and monolingual TE components.

As a complementary research problem, this paper addresses the data collection issue, focusing on the definition of a fast, cheap, and reliable methodology to create CLTE corpora. The main motivation is that, as in many other NLP areas, the availability of large quantities of annotated data represents a critical bottleneck in the systems' development/evaluation cycle. Our first step in this direction takes advantage of an already available monolingual corpus, casting the problem as a translation one. The challenge consists in taking a publicly available RTE dataset of English T-H pairs (*i.e.* the PASCAL-RTE3 dataset[1]), and create its English-Spanish CLTE equivalent by translating the hypotheses into Spanish. To this aim non-expert workers have been hired through the CrowdFlower[2] channel to Amazon Mechanical Turk[3] (MTurk), a crowdsourcing marketplace recently used with success for a variety of NLP tasks (Snow et al., 2008; Callison-Burch, 2009; Mihalcea and Strapparava, 2009; Marge et al., 2010; Ambati et al., 2010).

The following sections overview our experiments, carried out under strict time (10 days) and cost ($100) limitations. In particular, Section 2 describes our data acquisition process; Section 3 summarizes

[1] Available at: http://www.nist.gov/tac/data/RTE/index.html

[2] http://crowdflower.com/

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

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 212–216,
Los Angeles, California, June 2010.

the successive approximations that led to the definition of our methodology, and the lessons learned at each step; Section 4 concludes the paper and provides directions for future work.

2 Corpus creation cycles

Starting from the RTE3 Development set (800 English T-H pairs), our corpus creation process has been organized in sentence *translation-validation* cycles, defined as separate "jobs" routed to CrowdFower's workforce. At the first stage of each cycle, the original English hypotheses are used to create a *translation* job for collecting their Spanish equivalents. At the second stage, the collected translations are used to create a *validation* job, where multiple judges are asked to check the correctness of each translation, given the English source. Translated hypotheses that are positively evaluated by the majority of trustful validators (*i.e.* those judged correct with a confidence above 0.8) are retained, and directly stored in our CLTE corpus together with the corresponding English texts. The remaining ones are used to create a new translation job. The procedure is iterated until substantial agreement for each translated hypothesis is reached.

As regards the first phase of the cycle, we defined our **translation HIT** as follows:

In this task you are asked to:

- *First, judge if the Spanish sentence is a correct translation of the English sentence. If the English sentence and its Spanish translation are blank (marked as -), you can skip this step.*
- *Then, translate the English sentence above the text box into Spanish.*

Please make sure that your translation is:

1. *Faithful to the original phrase in both meaning and style.*
2. *Grammatically correct.*
3. *Free of spelling errors and typos.*

Don't use any automatic (machine) translation tool! You can have a look at any on-line dictionary or reference for the meaning of a word.

This HIT asks workers to first check the quality of an English-Spanish translation (used as a gold unit), and then write the Spanish translation of a new English sentence. The quality check allows to collect accurate translations, by filtering out judgments made by workers missing more than 20% of the gold units.

As regards the second phase of the cycle, our **validation HIT** has been defined as follows:

Su tarea es verificar si la traducción dada de una frase del Inglés al espaol es correcta o no. La traducción es correcta si:

1. *El estilo y sentido de la frase son fieles a los de la original.*
2. *Es gramaticalmente correcta.*
3. *Carece de errores ortográficos y tipográficos.*

Nota: el uso de herramientas de traducción automática (máquina) no está permitido!

This HIT asks workers to take binary decisions (Yes/No) for a set of English-Spanish translations including gold units. The title and the description are written in Spanish in order to weed out untrusted workers (*i.e.* those speaking only English), and attract the attention of Spanish speakers.

In our experiments, both the translation and validation jobs have been defined in several ways, trying to explore different strategies to quickly collect reliable data in a cost effective way. Such cost reduction effort led to the following differences between our work and similar related approaches documented in literature (Callison-Burch, 2009; Snow et al., 2008):

- Previous works built on redundancy of the collected translations (up to 5 for each source sentence), thus resulting in more costly jobs. For instance, adopting a redundancy-based approach to collect 5 translations per sentence at the cost of $0.01 each, and 5 validations per translation at the cost of $0.002 each, would result in $80 for 800 sentences.

  Assuming that the translation process is complex and expensive, our cycle-based technique builds on simple and cheap validation mechanisms that drastically reduce the amount of translations required. In our case, 1 translation per sentence at the cost of $0.01, and 5 validations per translation at the cost of $0.002 each, would result in \$32 for 800 sentences, making a conservative assumption of up to 8 iterations with 50% wrong translations at each cycle (*i.e.* 800 sentences in the first cycle, 400 in the second, 200 in the third, etc.).

- Previous works involving validation of the collected data are based on ranking/voting mechanisms, where workers are asked to order a number of translations, or select the best one given the source. Our approach to validation is based on asking workers to take binary decisions over source-target pairs. This results in an easier, faster, and eventually cheaper task.

- Previous works did not use any specific method to qualify the workers' knowledge, apart from *post-hoc* agreement computation. Our approach systematically includes gold units to filter out untrusted workers during the process. As a result we pay only for qualified judgments.

3 Experiments and lessons learned

The overall methodology, and the definition of the HITs described in Section 2, are the result of successive approximations that took into account two correlated aspects: the quality of the collected translations, and the current limitations of the CrowdFlower service. On one side, simpler, cheaper, and faster jobs launched in the beginning of our experiments had to be refined to improve the quality of the retained translations. On the other side, *ad-hoc* solutions had to be found to cope with the limited quality control functionalities provided by CrowdFlower. In particular, the lack of regional qualifications of the workers, and of any qualification tests mechanism (useful features of MTurk) raised the need of defining more controlled, but also more expensive jobs.

Table 1 and the rest of this section summarize the progress of our work in defining the methodology adopted, the main improvements experimented at each step, the overall costs, and the lessons learned.

Step 1: a naïve approach. Initially, translation/validation jobs were defined without using qualification mechanisms, giving permission to any worker to complete our HITs. In this phase, our goal was to estimate the trade-off between the required development time, the overall costs, and the quality of translations collected in the most naïve conditions.

As expected, the job accomplishment time was negligible, and the overall cost very low. More specifically, it took about 1 hour for translating the 800 hypotheses at the cost of \$12, and less than 6 hours to obtain 5 validations per each translation at the same cost of \$12.

Nevertheless, as revealed by further experiments with the introduction of gold units, the quality of the collected translations was poor. In particular, 61% of them should have been rejected, often due to gross mistakes. As an example, among the collected material several translations in languages other than English revealed a massive and defective use of on-line translation tools by untrusted workers, as also observed by (Callison-Burch, 2009).

Step 2: reducing *validation* errors. A first improvement addressed the validation phase, where we introduced *gold units* as a mechanism to qualify the workers, and consequently prune the untrusted ones. To this aim, we launched the validation HIT described in Section 2, adding around 50 English-Spanish control pairs. The pairs (equally distributed into positive and negative samples) have been extracted from the collected data, and manually checked by a Spanish native speaker.

The positive effect of using gold units has been verified in two ways. First, we checked the quality of the translations collected in the first naïve translation job, by counting the number of rejections (61%) after running the improved validation job. Then, we manually checked the quality of the translations retained with the new job. A manual check on 20% of the retained translations was carried out by a Spanish native speaker, resulting in 97% Accuracy. The 3% errors encountered are equally divided into minor translation errors, and controversial (but substantially acceptable) cases due to regional Spanish variations.

The considerable quality improvement observed has been obtained with a small increase of 25% in the cost (less than \$3). However, as regards the accomplishment time, adding the gold units to qualify workers led to a considerable increase in duration (about 4 days for the first iteration). This is mainly due to the high number of automatically rejected judgments, obtained from untrusted workers missing the gold units. Because of the discrepancy between trusted and untrusted judgments, we faced another limitation of the CrowdFlower service, which further delayed our experiments. Often, in fact, the rapid growth of untrusted judgments activates automatic pausing mechanisms, based on the assumption that gold units are not accurate. This, however, is a strong assumption which does not take into account the huge amount of non-qualified workers accepting (or even just playing with) the HITs. For instance, in our case the vast majority of errors came from workers located in specific regions where the native language is not Spanish nor English.

Step 3: reducing *translation* errors. The observed improvement obtained by introducing gold units in the validation phase, led us to the definition of a new translation task, also involving a similar qualification mechanism. To this aim, due to language variability, it was clearly impossible to use reference translations as gold units. Taking into account the limitations of the CrowdFlower interface, which does not allow to set qualification tests or split the jobs into sequential subtasks (other effective and widely used features of MTurk), we solved the problem by defining the translation HITs as described in Section 2. This solution combines a validity check and a translation task, and proved to be effective with a decrease in the translations eventually rejected (45%).

Step 4: reducing *time*. Considering the extra time required by using gold units, we decided to spend more money on each HIT to boost the speed of our jobs. In addition, to overcome the delays caused by the automatic pausing mechanism, we obtained from CrowdFlower the possibility to pose regional qualification, as commonly used in MTurk.

As expected, both solutions proved to be effective, and contributed to the final definition of our methodology. On one side, doubling the payment for each task (from $0.01 to $0.02 for each translation and from from $0.002 to $0.005 for each validation), we halved the required time to finish each job. On the other side, by imposing the regional qualification, we eventually avoided unexpected automatic pauses.

4 Conclusion and future work

We presented a set of experiments targeting the creation of bi-lingual Textual Entailment corpora by means of non experts' workforce (*i.e.* the CrowdFlower channel to Amazon Mechanical Turk).

As a first step in this direction, we took advantage of an already existing monolingual English RTE corpus, casting the problem as a translation task where Spanish translations of the hypotheses are collected and validated by the workers. Strict time and cost limitations on one side, and the current limitations of the CrowdFlower service on the other side, led us to the definition of an effective corpus creation methodology. As a result, less than $100 were spent in 10 days to define such methodology, leading to collect 426 pairs as a by-product. However, it's worth remarking that applying this technique to create the full corpus would cost about $30.

The limited costs, together with the short time required to acquire reliable results, demonstrate the effectiveness of crowdsourcing services for simple sentence translation tasks. However, while MTurk is already a well tested, stable, and rich of functionalities platform, some limitations emerged during our experience with the more recent CrowdFlower service (currently the only one accessible to non-US citizens). Some of these limitations, such as the regional qualification mechanism, have been overcome right after the end of our experimentation with the introduction of new functionalities provided as "Advanced Options". Others (such as the lack of other qualification mechanisms, and the automatic pausing of the HITs in case of high workers' error rates on the gold units) at the moment still represent a possible complication, and have to be carefully considered when designing experiments and interpreting the results[4].

In light of this positive experience, next steps in our research will further explore crowdsourcing-based data acquisition methods to address the complementary problem of collecting new entailment pairs from scratch. This will allow to drastically reduce data collection bottlenecks, and boost research both on cross-lingual and mono-lingual Textual En-

[4] However, when asked through the provided support service, the CrowdFlower team proved to be quite reactive in providing *ad-hoc* solutions to specific problems.

Elapsed time	Running cost	Focus	Lessons learned
1 day	$24	Approaching CrowdFlower, defining a naïve methodology	Need of qualification mechanism, task definition in Spanish.
7 days	$58	Improving validation	Qualification mechanisms (gold units and regional) are effective, need of payment increase to boost speed.
9 days	$99.75	Improving translation	Combined HIT for qualification, payment increase worked!
10 days	$99.75	Obtaining bi-lingual RTE corpus	Fast, cheap, and reliable method.

Table 1: $100 for a 10-day rush (summary and lessons learned)

tailment.

Acknowledgments

We would like to thank MTurk and CrowdFlower for providing the $100 credit used for the experiments, and our colleague Silvana Bernaola Biggio, who kindly accepted to validate our results.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248531.

References

V. Ambati, S. Vogel and J. Carbonell 2010. *Active Learning and Crowd-Sourcing for Machine Translation.* To appear in Proceedings of LREC 2010.

C. Callison-Burch 2009. *Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk.* In Proceedings of EMNLP 2009.

I. Dagan and O. 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.

M. Marge, S. Banerjee and A. Rudnicky 2010. *Using the Amazon Mechanical Turk for Transcription of Spoken Language.* In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Spoken Language (ICASSP 2010).

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

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

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng 2008. *Cheap and Fast - but is it Good? Evaluating Non-expert Annotations for Natural Language Tasks.* In Proceedings of EMNLP 2008.