Getting Expert Quality from the Crowd for Machine Translation Evaluation

Luisa Bentivogli and **Marcello Federico** and **Giovanni Moretti** and **Michael Paul**

FBK-irst   CELCT   NICT

Via Sommarive, 18   Via alla Cascata, 56c   Hikaridai 3-5

38123 Povo-Trento, Italy   38123 Povo-Trento, Italy   619-0289 Kyoto, Japan

`{bentivo,federico}@fbk.eu`   `moretti@celct.it` `michael.paul@nict.go.jp`

Abstract

This paper addresses the manual evaluation of Machine Translation (MT) quality by means of crowdsourcing. To this purpose, we replicated the ranking evaluation of the Arabic-English BTEC task proposed at the IWSLT 2010 Workshop by hiring non-experts through the CrowdFlower interface to Amazon's Mechanical Turk. In particular, we investigated the effectiveness of "gold units" offered by CrowdFlower as the main quality control mechanism. The analysis of the collected data shows that agreement rates for non-experts are comparable to those obtained for experts, and that the crowd-based system ranking has a very strong correlation with expert-based ranking. Our results confirm that crowdsourcing is an effective way to reduce the costs of MT evaluation without sacrificing quality, and demonstrate that just exploiting the Crowd-Flower control mechanism is enough to approximate expert-level data quality.

1 Introduction

The evaluation of Machine Translation quality is a difficult task because there may exist many possible ways to translate a given source sentence. Moreover, the usability of a given translation depends on numerous factors like the intended use of the translation, the characteristics of the MT software, and the nature of the translation process. Early attempts tried to manually produce numerical judgements of MT quality with respect to a set of reference translations (White et al., 1994). Recently, human assessment of MT quality has been carried out by either assigning a single grade on a scale of 5 or 7 specifying the fluency or adequacy of a given translation (Przybocki et al., 2008), or by relatively ranking to each other multiple translations of the same input (Callison-Burch et al., 2007).

Although human evaluation of MT output provides the most direct and reliable assessment, it is time consuming, costly and subjective, i.e., evaluation results might vary from person to person due to different backgrounds, bilingual experience, and inconsistent judgements caused by the high complexity of the multi-class grading task.

These drawbacks to human assessment schemes have encouraged many researchers to seek reliable methods for estimating such measures automatically. Various automatic evaluation measures have been proposed to make the evaluation of MT outputs cheaper and faster. However, automatic metrics have not yet proved able to consistently predict the usefulness of MT technologies. Each automatic metric focuses on different aspects of the translation output and its correlation with human judges depends largely on the type of human assessment.

In order to minimize inconsistent judgements, recent evaluation campaigns like IWSLT (Paul et al., 2010) have employed paid expert graders. These graders are bilingual judges that must take part in dry-run evaluation exercises prior to the shared task evaluation and prove highly consistent judgements for the given translation task. However, resorting to expert annotators is particularly expensive, especially in the case of MT evaluation campaigns which offer many different tasks and count a high number of participants.

To counter the high costs for human assessment of MT outputs, new possibilities are offered by the advent of crowdsourcing services such as Amazon's Mechanical Turk and CrowdFlower, which in recent years have attracted a lot of attention both from industry and academia as a means to collect data for human language technologies at low cost.

Amazon's Mechanical Turk[1] (MTurk) is one of the leading on-line work marketplaces, where peo-

---

[1]http://www.mturk.com/

ple are paid small sums of money to work on Human Intelligence Tasks (HITs), i.e. tasks that machines have hard time doing. MTurk allows anyone to work on available HITs, but in order to be a requester it is necessary to own a US billing address. The CrowdFlower[2] (CF) platform works across multiple crowdsourcing services[3], including MTurk. CF gives unrestricted access to all the offered channels, making it possible for non US-based requesters to place HITs on MTurk.

This paper investigates crowdsourcing as a method to reduce evaluation costs by using non-expert graders for the human assessment of machine translation quality. To this purpose, the official ranking evaluation of the Arabic-English BTEC task proposed at the IWSLT 2010 Workshop was replicated by hiring MTurk workforce through the CrowdFlower service. The aim of this experiment is to determine the quality of non-expert judgements by comparing them to the expert judgements available from the evaluation campaign.

The task of replicating existing evaluation settings in order to control the quality of non-expert data has already been addressed in previous works (Callison-Burch, 2009; Callison-Burch et al., 2010), which report on non-expert data gathered upon direct access to MTurk. However, since MTurk and CF offer different data quality control mechanisms, differences could emerge in the quality of the data depending on whether they are collected from the MTurk workforce directly or through the CF platform.

Our paper seeks to investigate the effectiveness of the main quality control mechanism offered by CF, i.e. the use of gold units to filter out bad workers, in terms of *(i)* effort required to use it in our task and *(ii)* actual quality of the collected data. We present a simple and cost-effective methodology to use the CF gold-based control mechanism in the ranking evaluation task and analyse the quality of the collected assessments focusing on (1) intra-/inter-annotator agreement of the non-expert graders, (2) the differences in the ranking evaluation carried out by expert graders and non-expert graders, and (3) the correlation of non-expert rankings with expert rankings. Finally, we compare this correlation against the cor-

relations between experts and common automatic evaluation metrics.

The comparison of expert-based and crowd-based assessments leads to two main findings. First, in line with previous work, we demonstrate that crowdsourcing is an effective way of reducing the costs of MT evaluation without sacrificing quality. Furthermore, we show that the CF gold-based control mechanism is enough to achieve expert-level data quality. This suggests that other quality control procedures, such as collecting a high number of redundant non-expert judgements or resorting to further a posteriori data filtering, are not strictly necessary.

2 Related Work

A rapidly growing number of recent studies have shown that MTurk can be a source of high quality and low cost annotated data for a wide range of research fields, including Information Retrieval, Speech, Vision, and Natural Language Processing tasks such as relation extraction, word sense disambiguation, textual entailment, named entity annotation, and natural language generation. Machine Translation is one of the fields where research on crowdsourcing is most active. The feasibility of collecting good quality crowdsourced data has been explored for many different MT tasks, including the creation of parallel corpora, the word-level alignment of parallel sentences, the creation of paraphrases of existing reference translations, and the creation of translation lexica for low resource languages (Callison-Burch and Dredze, 2010).

Annotated data is also crucial for evaluation purposes. Very recently, crowdsourced data has started being officially used in international evaluation campaigns. In the CLEF 2010 Web People Search Clustering Task (Artiles et al., 2010) and the SemEval-2010 Task of Noun Compounds Interpretation Using Paraphrasing Verbs and Prepositions (Butnariu et al., 2010), for example, MTurk was used for the annotation of the training/test data sets.

For MT evaluation, a number of studies have been carried out with the aim of understanding the feasibility of substituting expert data with non-expert data for different types of human evaluation tasks. In (Callison-Burch, 2009), it is shown that MTurk can be effectively used to collect relative rankings, to perform human-mediated translation edit rate

(HTER), and to carry out evaluation through reading comprehension questions. In (Denkowski and Lavie, 2010) MTurk was used to obtain translation adequacy assessments. Finally, in (Callison-Burch et al., 2010) a subset of the official WMT10 relative ranking evaluation was reproduced with non-expert judges and various methods to improve the quality of the collected data are presented. All these MT evaluation experiments have been conducted using MTurk directly, as have most of the available studies on the effectiveness of crowdsourcing. Up to now, only a few papers have reported on the use of CF as an interface to MTurk (Wang and Callison-Burch, 2010; Finin et al., 2010; Negri and Mehdad, 2010), none of them addressing the task of MT evaluation.

3 IWSLT 2010 BTEC Task Evaluation

The *International Workshop on Spoken Language Translation* (IWSLT) is a yearly, open evaluation campaign for spoken language translation. IWSLT's evaluations are not competition-oriented; their goal is to foster cooperative work and scientific exchange. In this respect, IWSLT proposes challenging research tasks and an open experimental infrastructure for the scientific community working on spoken and written language translation.

In the IWSLT 2010 campaign (Paul et al., 2010), three different shared tasks addressing various source and target languages were proposed. Among them, a translation task focusing on frequently used utterances in the domain of travel conversations was provided for the translation of Arabic (A) spoken language text into English (E). This translation task was carried out using the *Basic Travel Expression Corpus* (BTEC), a multilingual speech corpus containing tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad (Kikui et al., 2006). Twelve participants took part in the Arabic-English BTEC task (BTEC-AE). In addition, the organizers used the output of an online MT server, resulting in a total of 13 MT systems.

The translation quality of the submissions was evaluated using both manual evaluation and automatic metrics. Manual evaluation was considered primary as it was used both to officially rank the system submissions and to assess how well automatic metrics correlate with human judgements.

Human evaluation was carried out by three paid expert judges. The official measure chosen to evaluate translation quality was *Ranking*. In order to carry out the Ranking evaluation, the expert judges were asked to *"rank translations from Best to Worst relative to the other choices (ties are allowed)"* (Callison-Burch et al., 2007). The unit of evaluation was the *ranking set*, which is composed of a source sentence, a reference human translation, and up to five machine translations to be ordered by assigning a single grade to each of them. The evaluation was carried out using a web-browser interface where experts were shown screens containing three different ranking sets to be evaluated.

The manual evaluation of the BTEC-AE task was carried out on 392 test sentences. For each test sentence, the set of 13 MT outputs was randomly split into three ranking sets so that two of them contain four MT outputs, and the third contains five MT outputs [4]. The ranking sets were randomly created three times, one for each annotator. Moreover, around 100 ranking sets were repeated for each annotator in order to calculate intra-annotator agreement. Therefore, a total of 3,828 ranking sets[5] were evaluated by the three expert judges.

The time needed by the experts to complete the evaluation task[6] was around 13 working days, for a total cost of $3121. In addition, around two working days were necessary to prepare the evaluation data sets and setup the evaluation interface by the IWSLT organizers.

4 Crowdsourcing the BTEC Task Evaluation to Non-Expert Judges

The crowdsourcing experiments reported in this paper reproduce the official BTEC-AE Ranking evaluation by posting the task on MTurk through the CF platform. In Section 4.1, we will introduce the default quality control mechanisms offered by MTurk and CF. In Section 4.2, we present the whole

---

[4] Due to the high evaluation costs, not all the possible pairwise comparisons could be evaluated. However, by creating three ranking sets for each test sentence, it was assured that all the 13 MT outputs were evaluated by a human grader.

[5] (1,176 original ranking sets + 100 repetitions) * 3 graders

[6] The completion time includes a dry-run evaluation period of three working days that each expert grader was required to carry out in order to get used to the evaluation specifications and the evaluation interface.

data collection process, describing the design of the Ranking task using the CF interface, the gold units created for quality control, and giving details about the amount of collected human assessments as well as the cost and the time needed to collect them.

4.1 Data Quality Control

One of the most crucial issues to consider when collecting crowdsourced data is how to ensure their quality. Both MTurk and CF provide requesters with quality control mechanisms; some quality check options are offered by both, while others are specific to each service. Consequently, there can be differences in the data collected from the MTurk workforce directly or through CF.

Both services offer the "locale qualification" option (to restrict workers by country) and the "flag worker" option (to disable specific workers).

MTurk allows setting preliminary qualifications for workers, such as (i) a past HIT Approval Rate higher than a given threshold, and (ii) a minimum number of previously approved submissions. Moreover, workers can be required to complete a specific qualification test before working on the HIT.

CF does not prevent anyone from accepting a HIT (unless restricted via locale qualification or flag worker options) since its basic quality control mechanism consists of an on-the-fly verification of the workers' reliability. Workers who are not performing well on the accepted task are filtered out, and consequently poor quality data is removed before returning the results for a given job to the requester. To this purpose, the HIT design interface provided by CF allows including gold units, i.e. items with known labels, along with the other units composing the required HIT. These control units allow distinguishing between trusted workers (those who correctly replicate the gold units) and untrusted workers (those who fail the gold units). In order to be considered trusted in a job, by default, workers are required to judge a minimum of four gold units and to be above an accuracy threshold of 70%. Untrusted workers are automatically blocked[7] and not paid, and their labels are filtered out from the final data

---

[7]The HITs generated by CF also display the accuracy score to the MTurk workers while they are working on the HIT, in order to give them feedback on their performance.
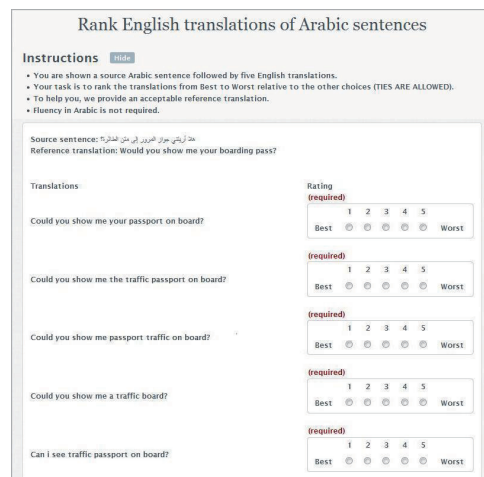

Figure 1: The CrowdFlower Ranking Interface.

set. As a further control, CF automatically pauses a job if workers are failing too many gold units.

Gold units are randomly mixed with the other units by CF when it creates the worker assignments, and the suggested amount of gold units to be provided is around 10% of the requested units[8].

4.2 Data Collection through CrowdFlower

In order to recreate the existing set of expert judgements for the BTEC-AE task, the official experiment settings were exactly replicated. However, instead of the expert graders, the judgements were collected from MTurk workers. CF provides an interface for designing HITs that we exploited to reproduce the official ranking web interface used by experts. A snapshot of the ranking interface presented to MTurk workers is given in Figure 1.

We posted all the 3,528 official ranking sets on MTurk, providing workers with the same task instructions given to experts. When posting the job, CF gives the possibility to choose how many times each single unit[9] has to be completed by different workers. This can be used to collect agreement in-

---

[8]In principle, nothing prevents MTurk requesters from inserting gold units in a HIT. However, the difference with respect to CF-generated HITs is that with MTurk, the requester has to do an a-posteriori filter on the obtained (and paid) data, whereas in CF, the mechanism is built-in, and the requester does not have to take care of it. Moreover, CF keeps a record of each worker, and uses the workers' history to apply confidence scores to their annotations.

[9]A unit is each single work item composing the HIT. In our task, the unit corresponds to one ranking set.

formation or to compute label aggregation by applying majority voting schemes. However, to conform to the BTEC-AE evaluation setting, we required each unit to be judged only once. In order to ensure enough data to measure agreement, we also required five redundant judgements from different workers for a subset of the units (80 ranking sets chosen randomly). Moreover, following the official IWSLT evaluation where each interface screen displayed three ranking sets to be assessed, MTurk workers were assigned three units each, meaning that they had to judge three ranking sets at a time.

In creating the gold units, we followed the "Reference Preference" assumption, in which a reference translation wins (or ties in) a comparison when it appears in one ranking set together with MT outputs (Callison-Burch, 2009). To this purpose, 360 ranking sets, i.e. 10% of the total units, were selected ensuring that different test sentences were represented and, for each of them, one MT output was automatically substituted with a reference translation. Gold units were manually checked to verify that the references were actually better then the MT outputs. It is worth noticing that as MT systems performed in general quite well, the BTEC-AE task was particularly suited to creating good gold units, since in most of the investigated ranking sets, the reference translation was not immediately recognizable as the best one, thus making the gold units not easy to judge.

In order to differentiate the MTurk costs according to the work required, we subdivided the data into jobs with ranking sets composed of either four or five system outputs. The 4-output (5-output) jobs paid $0.04 ($0.05) per assignment, respectively. The jobs were posted to MTurk through CF with *(i)* locale qualifications[10] and *(ii)* gold units required for each assignment, i.e., among the three ranking sets presented at a time to workers, one was gold.

The work was carried out by 52 "trusted" graders, who returned a total of 3,964 ranking sets[11]. An-

other 705 ranking sets from "untrusted" workers, corresponding to the 15.1% of the units judged by all workers, were filtered out of the results.

The time needed by the MTurk workforce to complete the jobs was around 6 days, for a total cost of $126. In addition, around two working days were necessary to prepare the gold units[12].

It is worth noting that gold units can be reused for later MT ranking tasks, provided that the domain and the languages addressed remain the same. Given the possibility of creating only a small number of gold units, the related cost for reference translations (which remains hidden in our experiment as we already had them available) is not particularly relevant. On top of this, it has already been demonstrated in a number of previous works that crowdsourcing translations is also a feasible and cheap task. Moreover, the cost of creating one reference translation for each gold unit would not be higher than that required for an automatic evaluation.

Up to now we have shown the cost-effectiveness of our crowdsourcing approach, which provides a cheap and fast way to collect data. In the next section we analyse the quality of the collected data by comparing them with the expert assessments used for the official IWSLT task evaluation.

5 Expert vs. Non-Expert Evaluation

In this section, we compare non-expert and expert assessments, focusing on *(i)* intra-annotator and inter-annotator agreement rates and on *(ii)* the resulting ranking of the systems participating in the task.

5.1 Quality of Data

Some details about the characteristics of the ranking tasks carried out by experts and non-experts are presented in Table 1. Information about grading time[13] can be useful to determine the quality of judgements, as it gives an indication of how carefully the workers completed their task. We can see that non-experts were slightly faster than experts, but the difference is not particularly relevant. Concerning the evalua-

---

[10]It is known by the literature that locale qualifications are definitely necessary to increase the quality of the data. This was further demonstrated by our experience with CF. In fact, jobs containing gold units and posted without locale qualifications were always paused due to an excessive gold failure rate.

[11]We required a total of 3,928 judged units (3,528 official ranking sets plus 80 repetitions with 5 judgments each) but usually jobs return a slightly larger number than required, due to the labour distribution mechanism internal to MTurk.

[12]Notice that this effort can be further reduced by splitting big jobs into smaller jobs, so that not only are fewer gold units required for each job, but also the same gold units can be reused in all the smaller jobs.

[13]The CF output record contains various worker information, including the time it took them to complete each assignment.

	Non-Experts	Experts
# of graders	52	3
# of ranking sets	3,964	3,816
grading time	36:26 h	36:30 h
evaluation period	6 days	13 days
evaluation setup	2 days	2 days

Table 1: Non-Expert and Expert Assessments

	Non-Experts			Experts		
	comp	Pr(a)	K	comp	Pr(a)	K
Intra-Agreement	547	0.7751	0.6627	1,686	0.8463	0.7695
Inter-Agreement	10,093	0.6335	0.4502	6,210	0.6878	0.5317

Table 2: Intra- and Inter-Annotator Agreement

tion period, 10 days for BTEC-AE evaluations and 3 days for dry-run evaluations were allocated for the expert grading task to allow for a careful completion of the task, whereas the CF data collection lasted 6 days. Other interesting information given by CF, and related to the gold-based control mechanism, is the "trust level" of the workers, i.e. the average accuracy obtained by trusted workers on the gold units. While the minimum threshold to be "trusted" is 70%, the average trust level of the "trusted" workers was much higher, amounting to 89%.

The most informative indicator of the quality of a dataset is given by the agreement rate, or grading consistency, both between different judges and within the same judge. To this purpose, inter- and intra-annotator agreement were calculated on the MTurk data and compared to the agreement rates obtained by expert judges. Agreement rates are calculated using the *Fleiss' kappa coefficient* κ (Landis and Koch, 1977):

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where Pr(a) is the relative observed agreement among graders, and Pr(e) is the hypothetical probability of chance agreement. In our task, Pr(a) is given by the proportion of times that two judges (or the same judge for intra-annotator agreement) assessing the same pair of systems on the same source sentence agree that A>B, A=B, or A<B. Pr(e) is 0.33 as the possible decision classes are three.

Table 2 shows the κ values for intra-annotator and inter-annotator agreement for both non-expert and expert judges. For computing non-expert intra-annotator agreement, we were able to collect only a small number of comparisons, because it is not possible to ensure that the same worker completes the same unit more than once. On the other hand, gathering data for inter-annotator agreement is easy, because CF allows requiring that different workers judge the same unit. In our experiment, it turned out that only 9 graders out of 52 judged the same

pair of systems on the same test sentence, for a total 547 comparisons. On the contrary, a lot of comparisons were available for inter-annotator agreement, both because the HITs were judged by 52 annotators (and not only 3, as for the experts) and because we required 5 redundant judgements from different workers for several HITs.

Concerning the κ values obtained, it is very interesting to note that, even though - as expected - agreement rates for experts are all higher than those for non-experts, the differences among them are not particularly remarkable: namely, a κ value that is around 0.11 less for intra-annotator and 0.08 for inter-annotator agreement. Furthermore, according to the standard interpretation of the κ values, both expert and non-expert agreement fall in the same range, i.e. *substantial* for intra-annotator agreement and *moderate* for inter-annotator agreement.

This finding is particularly relevant when compared to a similar evaluation carried out in (Callison-Burch et al., 2010). Their experiment was carried out on the WMT 2010 data, and ranking sets were posted directly on MTurk requiring all the MTurk preliminary qualifications. Even though a direct comparison is not possible due to different data sets, the agreement rates obtained are informative. Unlike our experiment, the intra-/inter-annotator agreement rates for non-experts were found to be markedly lower than those of experts, with a gap in the κ values of 0.29 for intra-annotator and 0.28 for inter-annotator agreement. Even after applying posterior data filtering techniques like removing bad workers, the gaps reduced only to a value near 0.15. These results show that the CF gold unit control mechanism is very effective, making it possible to obtain good quality data in a very simple way.

5.2 Ranking Results

The primary evaluation metric used for the IWSLT 2010 BTEC-AE Task was the manual ranking of the system outputs. The *Ranking* scores were obtained

MT Systems Ranking	Experts	Non-Experts	Automatic Metrics						
			BLEU	METEOR	WER	PER	TER	GTM	NIST
1	0.4863	0.4795	33.85	67.75	48.91	42.47	43.54	68.53	6.666
2	0.4485	0.4073	42.96	72.88	40.72	35.46	35.10	71.41	7.285
3	0.4396	0.4166	46.73	73.22	37.51	32.72	32.30	72.97	7.345
4	0.4020	0.3855	43.76	71.48	39.95	35.22	34.93	73.29	7.248
5	0.3991	0.3711	41.55	70.84	42.27	36.76	36.31	70.25	7.042
6	0.3889	0.3626	43.47	71.69	40.31	35.73	34.68	71.73	7.123
7	0.3438	0.3081	40.57	69.23	42.39	36.71	36.25	70.12	6.734
8	0.3300	0.2794	35.15	66.13	47.61	41.45	41.05	68.65	6.522
9	0.2967	0.2329	33.62	68.37	49.26	41.69	41.71	68.16	6.586
10	0.2588	0.1980	29.04	64.14	50.70	45.17	43.13	63.23	5.857
11	0.2535	0.1953	20.11	57.60	58.97	52.59	50.18	56.46	4.978
12	0.2529	0.2199	27.04	56.88	53.79	48.13	45.75	59.84	4.602
13	0.2249	0.1837	35.92	65.95	46.64	40.33	40.32	66.75	6.482
Spearman Rank Correlation ρ	0.9780		0.6483	0.7802	-0.6483	-0.6154	-0.5330	0.7582	0.8316

Table 3: Ranking Results for Experts, Non-Experts, and Automatic Metrics and Correlations with Expert Ranking

as the average number of times that a system was judged better than any other system.

The IWSLT 2010 translation results were also evaluated using a variety of standard automatic evaluation metrics including BLEU, NIST, METEOR, GTM, WER, PER, TER (Paul et al., 2010). A total of 7 reference translations were made available. The automatic evaluation specifications for the BTEC task were defined as case-sensitive with punctuation. Tokenization scripts were applied automatically to all run submissions prior to evaluation.

The correlations of automatic metrics with expert ranking were calculated using the *Spearman rank correlation coefficient* ρ[14]. In order to verify the feasibility of using non-expert assessments for the *Ranking* evaluation of MT systems, the rankings obtained according to non-expert data are compared to those based on expert data.

Table 3 shows all the ranking results for experts, non-experts, and automatic metrics. These include: *(i)* the official BTEC-AE ranking of MT systems - including the online MT system - according to experts; *(ii)* the scores assigned to each of the 13 MT systems, obtained applying the *Ranking* metric respectively on the expert data, the non-expert data, and the 7 automatic evaluation metrics; *(iii)* the correlations of non-expert and automatic metrics rankings with expert ranking calculated using the Spearman rank correlation coefficient. It can be observed that the non-expert ranking is more similar to the expert-based ranking than those obtained by the automatic evaluation metrics. This result is clearly

shown by the correlation figures, from which we can see that the ranking produced using non-experts has a much stronger correlation with the BTEC-AE expert ranking than all the automatic metrics.

6 Conclusions

In this paper we addressed crowdsourcing as a method to reduce MT human evaluation costs without sacrificing quality. In particular, we have investigated the use of the CF platform and the effectiveness of its gold-based data quality control mechanism. Whereas MTurk controls the workers by relying on the quality of their previous work, the gold-based mechanism allows it to directly evaluate workers on the given task by using the workers' own data while they are in the process of creating it.

The results obtained in our experiment demonstrate the effectiveness of CF, both in terms of effort required to use it for our task and of the actual quality of the collected data. In fact, CF provides a cheap and fast way to collect data in a number of respects. The Ranking task was quick and easy to design, as the CF HIT design interface provides a lot of functionalities which allow the creation of screens to be displayed to MTurk workers with no or little programming effort. As for gold units, only limited effort is required to create them, the gold-based mechanism has already been implemented and does not require any work to manage it and, most important, it is very effective for controlling data quality.

Concerning the quality of the collected data, we found that the agreement rates for non-experts are comparable to those obtained for experts, unlike

[14]http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

similar experiments which used the MTurk work-force directly. Moreover, non-expert rankings are more similar to and correlate better with expert rankings than automatic metrics.

The presented experiments show that (1) crowd-sourcing is an effective way to reduce MT evaluation costs in terms of time as well as money, achieving a cost reduction of 47% in time (13 days of evaluation + 2 days of setup for experts *vs.* 6 + 2 days for non-experts) and of 96% in cash ($3,121 for experts *vs.* $126 for non-experts), and that (2) exploiting only the gold-based mechanism is enough to approximate expert-level data quality (Spearman rank correlation coefficient $\rho = 0.9780$) without the need of additional quality control procedures, such as collecting a high number of redundant non-expert judgements or resorting to a posteriori data filtering.

A problematic issue related to crowdsourcing in general should be taken into account when examining the possibility of substituting expert data, and that is the lack of continuity in workers. Due to the fact that workers change over time, results obtained in one experiment may not be replicable in others.

As future work, we are planning to investigate the applicability of the presented approach to more complex translation tasks, including tasks where the target language is not English. Moreover, we want to investigate how much applying other ways of filtering data besides the use of gold units can help further improve quality. Finally, we will try other CF channels (e.g. those rated highest in terms of work quality, such as SamaSource) in order to see if differences arise in the resulting data quality.

References

J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In *Notebook Papers of the CLEF 2010 LABs*, Padova, Italy.

C. Butnariu, S.N. Kim, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale. 2010. SemEval-2 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proc. of the 5th International Workshop on Semantic Evaluation*, pages 39–44, Uppsala, Sweden.

C. Callison-Burch and M. Dredze. 2010. Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, USA.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proc. of the 2nd Workshop on SMT*, pages 136–158, Prague, Czech Republic.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the Workshop on SMT and Metrics for Machine Translation. In *Proc. of the Joint 5th Workshop on SMT and MetricsMATR*, pages 17–53, Uppsala, Sweden.

C. Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluation Translation Quality Using Amazon's Mechanical Turk. In *Proc. of the EMNLP*, pages 286–295, Singapore.

M. Denkowski and A. Lavie. 2010. Exploring Normalization Techniques for Human Judgments of MT Adequacy Collected Using Amazon Mechanical Turk. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 57–61, Los Angeles, USA.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles, USA.

G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1674–1682.

J.R. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174.

M. Negri and Y. Mehdad. 2010. Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 212–216, Los Angeles, USA.

M. Paul, M. Federico, and S. Stücker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proc. of IWSLT*, pages 3–27, Paris, France.

M. Przybocki, K. Peterson, and S. Bronsart. 2008. Metrics for MAchine TRanslation Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results.

R. Wang and C. Callison-Burch. 2010. Cheap facts and counter-facts. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 163–167, Los Angeles, USA.

J.S. White, T. O'Connell, and F. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proc of the AMTA*, pages 193–205.