RC25031 (W1008-004) August 3, 2010
Computer Science

IBM Research Report

Approaches to Automatic Quality Estimation of Manual Translations in Crowdsourcing Parallel Corpora Development: A Quality Equivalence and Cohort-Consensus Approach

Juan M. Huerta
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Approaches to Automatic Quality Estimation of Manual Translations in Crowdsourcing Parallel Corpora Development: A Quality Equivalence and Cohort-Consensus Approach

Juan M. Huerta
IBM T.J. Watson Research Center

1. Overview.

We address the topic of metrics and approaches to automatic quality analysis and validation of sentence translations when manually developing a parallel corpus of translations. We focus specifically on the crowdsourcing-centered approach. We propose a set of metrics which provide the corpus developers with translation quality estimates. These estimates are particularly necessary when, due to the particular circumstances of the data collection, the quality of the translation provided is expected to vary significantly from person to person as well as from sentence to sentence. Our approach is based on the concept of quality equivalence and cohort-consensus. We also describe our experience and results using our proposed metrics when developing a large parallel corpus in a crowdsourcing approach.

2. Background

Parallel textual corpora play a vital role in contemporary research and development of machine translation technologies. Parallel text corpora (corpora of sentences for which every sentence has an associated translation into other language) are typically developed through the manual translation of documents by a team of professional or at least experienced translators and corpus developers. The assurance of the translation quality typically involves an additional step in which mistakes and inconsistencies are corrected by an editor. Overall, these are costly and labor intensive processes.

The emergence of the crowdsourcing model provides corpus and language resource developers with opportunities to lower the cost of corpus development [2,3,4]. In the case of parallel corpus development, the kernel task consists of providing small groups of sentences to individuals in the crowd which in turn translate these sentences in exchange for some incentive (i.e., remuneration, raffle entry, charity donation etc.). This approach is gaining popularity in the corpus development community.

While the cost of development is typically reduced when doing crowdsourcing, the resulting quality can be inconsistent: the quality uniformity and homogeneity of the resulting translations is expected to vary significantly due not only to the style of the individual participants but also due to their skills. While ideally, participants should be fluent in the source language and native in the target language, this is a condition that is hard to evaluate and enforce.

Thus, crowdsourcing presents us with the need for a more stringent and systematic quality estimation process than in the traditional approach.

3. Issues related to Crowdsourcing Parallel Corpora Development

There are two main issues affecting the quality of the translated sentences and our ability to estimate it. The first relates to the natural variability in the distribution of translation quality. The second is the extent to which redundancy (or translation overlap) in the work is desired or exists.

With respect to the variability in the quality, several factors influence it. The combination of multiple factors like language skill of each participant, nature of the sentences, attention or effort devoted etc., results in a non-uniform translation quality distribution across the various translators. Even within a single translator's work the attention and effort might not be homogeneous and will result in within translator variability . However, our assumption is that in the long run, as more translations become available the *distribution* of the quality of these translations will provide us with a glimpse about the skill of the translator.

With respect to redundancy, there is a tradeoff between aiming for collecting multiple-translations in each sentence (i.e., sentence overlap or redundant translation [2]) and coverage (i.e., maximizing the number of different sentences translated). If no overlap or redundancy exists then each sentence will be translated only once thus maximizing the coverage. The risk of maximizing coverage is that if for a given sentence a defective or of low quality translation has been entered, not only we will not have alternative translations to replace the defective translation but we will have no *direct* way of knowing that this particular translation is of low quality.

4. Estimating Quality of Human Translations Using Automatic Machine Translation: the Quality Equivalence Condition

Given that we assume that we have no professionally-translated references (i.e., "gold-standard references") to score the translator's data against, the only additional resource we can use for this purpose is the machine translation output of the desired sentences.

The justification of using MT as pseudo-references is as follows: let us assume that we know the "true" distribution of BLEU scores of a translation engine given the corpus to be translated. This distribution of scores reflects the distance between MT output and a gold-standard set of reference human translations (the empirical value of BLEU corresponds to the Maximum Likelihood estimate of the expected value of this distribution). Lacking this gold-standard but having MT and crowdsourcing translations, we focus on the question: *how likely it is that the crowd's translations can take the place of reference translations?* A *necessary* condition for this *quality equivalence* is that the distribution of BLEU scores between the gold-standard references and MT output has to be similar to the distribution between the crowdsourcing references and the MT output. We refer to this necessary condition as the Quality Equivalence Condition. We leverage this Condition in the next sections and we propose metrics based on the machine generated translation as reference.

5. Metrics for Quality Assurance and Monitoring

We now describe two proposed sets of metrics for automatic quality estimation: the first set is usable when there is no overlap in the translation set and is based on of the quality equivalence condition and the second set is for when there is at least some overlap.

5.1 Metrics with No Existing Sentence Overlap: Figure 1 show below a summary of the 4 measurements proposed that focus on the distance between 2 sentences s1 and s2 (let s1 be the translator's sentence and s2 be the Machine translation for that sentence).

- F1: Length ratio: consists of the ratio of the absolute difference between the lengths of the two sentences smoothed by an exponent over the length of the MT sentence. The idea is that the length of two similar sentences should not diverge too much from each other if they are coarsely equivalent.
- F2: Core Language words ratio or symmetry: This metric is similar to F1 except that instead of considering the length of each sentence, we consider the length in terms of function or grammatical words (e.g., prepositions pronouns, conjunctions etc) the rationale is that the proportion of function words in between two equivalent sentences should not vary too much if they are coarsely equivalent.
- F3: BLEU: This is the BLEU distortion (1 – BLEU similarity) between the MT sentence and the translated sentence. The rationale of applying BLEU to this task is based directly on the quality equivalence condition described in section 4.
- F4: String Edit Distance: measures the similarity between two sentences and is more stringent than BLEU.

F1: Length ratio	$f1(s_1,s_2)=\frac{\vert length(s_1)-length(s_2)\vert^{0.3}}{length(s_2)}$
F2: Language words ratio	$f2(s_1,s_2)=\frac{\vert language\_length(s_1)-language\_length(s_2)\vert^{0.6}}{language\_length(s_2)}$
F3: 1-BLEU	$f3(s_1,s_2)=1-BLEU(s_1,s_2)$
F4: String Edit	$f4(s_1,s_2)=edit(s_1,s_2)$

FIGURE 1. Summary of proposed metrics

5.2 Sentence with Overlap

In this case we assume overlap. We assume that agreement (cohort consensus) in the multiple translation instances coming from crowdsourcing for each sentence (what we call a *cohort*) reflects features that are likely to exist in a gold standard reference.
We can leverage cross-sentence agreement through the concept of the Cohort Consensus as follows:

1. Consensus Divergence: It refers to the cumulative *edit distance* score between a sentence and the set of translations provided by others (the cohort set). This essentially tells us how different the features of this sentence (n-grams) are from those of the rest of the cohort.

$$ConsensusDivergence(s_i)=\sum_{\forall j\in corpus} edit(s_i,s_j)$$

2. Cohort Rank Distribution: We rank the participants in a cohort based on increasing participant's sentence consensus divergence and look at the participant's rank in the cohort list. This tells us how different this participant is from the pack on average. A low rank will indicate that the sentence is close to being the most popular translation of the cohort while high rank (being deep in the cohort list) means that the translations provided by the participant are unconventional. The distribution of ranks of a given participant can provide insight on systematic quality issues.

6. Summary of Crowdsourcing Results

We now describe our experience in a large crowdsourcing activity and the application of the metrics and approaches introduced in this paper.

- Description of Crowdsourcing Activity: We carried out a large crowdsourcing activity aimed at translating as many sentences as possible without overlap (except in a small optional subset of benchmarking sentences). We had a total of 1700 unique participants and collected data in 11 languages plus English, resulting in a total of 22 language pairs. We collected 55k sentences. Participation is highly skewed as described in figure 2: We can see that 1% of the participants produced 38% of the translations, and so on. The MT output was made available to the translators because we assumed that it could save time to the translators to edit the MT output rather than entering each translation from scratch.

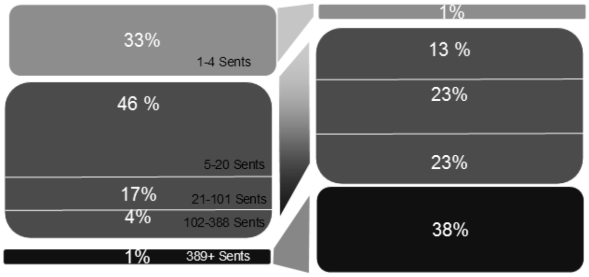

FIGURE 2. Participation distribution

- Analysis of Translations for multiple translators in non-overlapping set: Figure 3 below the scatter plot of F3 vs F4 for all the sentences translated by 4 users (user green, black, red, and blue). This plot reflects the nature of the changes introduced in the sentences: in general F4 > F3 and to the extent to which this is true reflects the nature of the translator's effort. In figure 3 We can see that the patterns of the green, black and blue translator's overlap, while the pattern of red falls on top of a line that do not extend far from the origin. This reflects an agreement between the SMT and the translator which is unusually large and that there is very little non-linear editing (reflected in F4). This translator, in other words, disagreed very little with the machine translator output much more than the other 3 translators and thus can be considered to be providing little new information in his/her translations. Under the quality equivalence assumption, translators blue, black and green have consistent score distributions and satisfy the necessary condition for quality. The red translator is inconsistent and these translations should be further inspected.

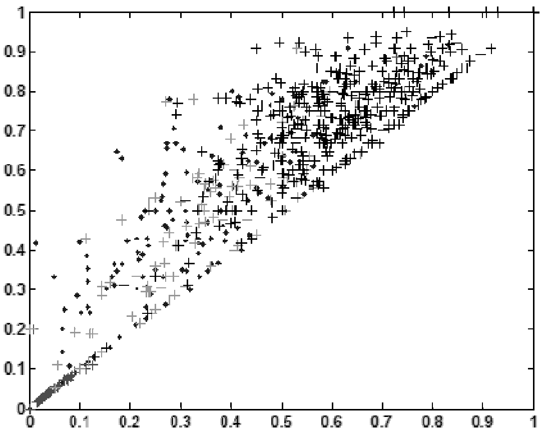

FIGURE 3 (a and b). Distribution of Quality Plots

- Analysis of the overlapping set: We Analysis of cross-translator agreement using the overlapping set: 846 Sentences having 3213 translations across a subset of the language pair set (9 pairs). These sentences map to 110 unique "originating" sentences. Figure 5 shows a scatter plot of the cohort rank vs. cohort depth for two translators (green and red). Each sentence is represented by a point in the scatter plot. We can see that there are clear regions or patterns wit green translator showing consistently lower (better translations) ranks than red translator. The work of the red translator should be further inspected for systematic problems.

7. Conclusions

We have presented approaches based on quality equivalence assumption and the cohort rank distribution that can be of value in assessing the quality of sentences in a corpus translated by the crowd. As more and more data are generated through crowdsourcing we believe this approach can be of great value.

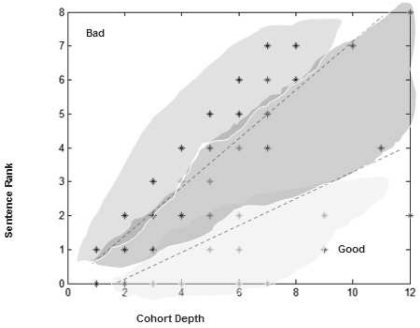

FIGURE 4. Cohort rank-distribution for 2 annotators

References

[1] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318

[2] V. Sheng, F. Provost and P. Ipeirotis , (2008) “Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers”. In Proc. KDD, 2008.

[3] R. Snow, B. O’Connor, D. Jurafsky, A. Y. Ng (2008) “Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Nat. Lang. Tasks” In Proc EMNLP 2008

[4] Stewart, O., Huerta, J. M., and Sader, M. 2009. Designing crowdsourcing community for the enterprise. In Proceedings of the ACM SIGKDD Workshop on Human Computation (Paris, France, June 28 - 28, 2009).