Systematic Comparison of Professional and Crowdsourced Reference Translations for Machine Translation

Rabih Zbib, Gretchen Markiewicz, Spyros Matsoukas,
Richard Schwartz, John Makhoul
Raytheon BBN Technologies
Cambridge, MA 02138, USA
{rzbib,gmarkiew,smatsouk,schwartz,makhoul}@bbn.com

Abstract

We present a systematic study of the effect of crowdsourced translations on Machine Translation performance. We compare Machine Translation systems trained on the same data but with translations obtained using Amazon's Mechanical Turk vs. professional translations, and show that the same performance is obtained from Mechanical Turk translations at 1/5th the cost. We also show that adding a Mechanical Turk reference translation of the development set improves parameter tuning and output evaluation.

1 Introduction

Online crowdsourcing services have been shown to be a cheap and effective data annotation resource for various Natural Language Processing (NLP) tasks (Callison-Burch and Dredze, 2010; Zaidan and Callison-Burch, 2011a; Zaidan and Callison-Burch, 2011b). The resulting quality of annotations is high enough to be used for training statistical NLP models, with a saving in cost and time of up to an order of magnitude. Statistical Machine Translation (SMT) is one of the NLP tasks that can benefit from crowdsourced annotations. With appropriate quality control mechanisms, reference translations collected by crowdsourcing have been successfully used for training and evaluating SMT systems (Zbib et al., 2012; Zaidan and Callison-Burch, 2011b).

In this work, we used Amazon's Mechanical Turk (MTurk) to obtain alternative reference translations of four Arabic-English parallel corpora previously released by the Linguistic Data Consortium (LDC) for the DARPA BOLT program. This data, totaling over 500K Arabic tokens, was originally collected from web discussion forums and translated professionally to English. We used alternative MTurk translations of the same data to train and evaluation MT systems; and conducted the first systematic study that quantifies the effect of the reference translation process on MT output. We found that:

- Mechanical Turk can be used to translate enough data for training an MT system at 1/10th the price of professional translation, and at a much faster rate.
- Training MT systems on MTurk reference translations gives the same performance as training with professional translations at 20% of the cost.
- A second translation of the development set obtained via MTurk improves parameter tuning and output evaluation.

2 Previous Work

There have been several publications on crowdsourcing data annotation for NLP. Callison-Burch and Dredze (2010) give an overview of the NAACL-2010 Workshop on using Mechanical Turk for data annotation. They describe tasks for which MTurk can be used, and summarize a set of best practices. They also include references to the workshop contributions.

Zaidan and Callison-Burch (2011a) created a monolingual Arabic data set rich in dialectal content from user commentaries on newspaper websites. They hired native Arabic speakers on MTurk

Proceedings of NAACL-HLT 2013, pages 612–616,
Atlanta, Georgia, 9–14 June 2013.

to identify the dialect level and used the collected labels to train automatic dialect identification systems. They did not translate the collected data, however. Zaidan and Callison-Burch (2011b) obtained multiple translations of the NIST 2009 Urdu-English evaluation set using MTurk. They trained a statistical model on a set of features to select among the multiple translations. They showed that the MTurk translations selected by their model approached the range of quality of professional translations, and that the selected MTurk translations can be used reliably to score the outputs of different MT systems submitted to the NIST evaluation. Unlike our work, they did not investigate the use of crowdsourced translations for training or parameter tuning. Zbib et al. (2012) trained a Dialectal Arabic to English MT system using Mechanical Turk translations. But the data they translated on MTurk does not have professional translations to conduct the systematic comparison we do in this paper.

It is well known that scoring MT output against multiple references improves MT scores such as BLEU significantly, since it increases the chance of matching *n-grams* between the MT output and the references. Tuning system parameter with multiple references also improves machine translation for the same reason Madnani et al. (2007) and Madnani et al. (2008) showed that tuning on additional references obtained by automatic paraphrasing helps when only few tuning references are available.

3 Data Translation

The data we used are Arabic-English parallel corpora released by the LDC for the DARPA BOLT Phase 1 program[1]. The data was collected from Egyptian online discussion forums, and consists of separate discussion threads, each composed of an initial user posting and multiple reply postings. The data tends to be bimodal: the first posting in the thread is often formal and expressed in Modern Standard Arabic, while the subsequent threads use a less formal style, and contain colloquial Egyptian dialect. The data was manually segmented into sentence units, and translated professionally.

We used non-professional translators hired on MTurk to get second translations. We used several measures to control the quality of translations and detect cheaters. Those include the rendering of Arabic sentences as images, comparing the output to Google Translate and Bing Translator, and other automatic checks. The quality of individual worker's translations was quantified by asking a native Arabic speaker judge to score a sample of the Turker's translations. The translation task unit (aka Human Intelligence Task or HIT) consisted of a sequence of contiguous sentences from a discussion thread amounting to between 40 and 60 words. The instructions were simply to translate the Arabic source fully and accurately, and to take surrounding sentence segments into account to help resolve ambiguities. The HIT rewards were set to 2.5¢ per word.

At the end of the effort, we had 26 different workers translate 567K Arabic tokens in 4 weeks. The resulting translations were less fluent than their professional counterparts, and 10% shorter on average. The following section presents results of MT experiments using the MTurk translations.

4 MT Experiments

The MT system used is based on a string-to-dependency-tree hierarchical model of Shen et al. (2008). Sentence alignment was done using GIZA++ (Och and Ney, 2003). Decoder features include translation probabilities, smoothed lexical probabilities, and a dependency tree language model. Additionally, we used 50,000 sparse, binary-valued source and target features based on Chiang et al. (2009). The English language model was trained on 7 billion words from the LDC Gigaword corpus and from a web crawl. We used expected BLEU maximization (Devlin, 2009) to tune feature weights.

We defined a tuning set (3581 segments, 43.3K tokens) and a test set (4166 segments, 47.7K tokens) using LDC2012E30, the corpus designated as a development set by the LDC, augmented with around 50K Words held out from LDC2012E15 and LDC2012E19, to make a development set large enough to tune the large number of feature weights[2]. The remaining data was used for training. We defined three nested training sets containing 100K, 200K and 400K Arabic tokens respectively, with

[1]Corpora: LDC2012E15, LDC2012E19, LDC2012E55

[2]Only full forum threads were held out

Training	Web-forum Only			Newswire(10MW)+Web-forum			
	100KW	200KW	400KW	0KW	100KW	200KW	400KW
Prof. refs	17.71	20.23	22.61	22.82	24.05	24.85	25.19
MTurk refs	16.41	18.43	20.08	22.82	23.79	24.20	24.51
Two Training refs	19.03	21.19	23.06	22.82	24.26	25.19	25.38
Add'l Training data	-	19.80	21.53	22.82	-	24.31	25.16

Table 1: Comparison of the effect of web forum training data when using professional and MTurk reference translations. All results use professional references for the tuning and test sets.

two versions of each set: one with the professional reference translations for the target, and the other with the same source data, but the MTurk translations. We defined two versions of the test and tuning sets similarly. We report translation results in terms of lower-case BLEU scores (Papineni et al., 2002).

4.1 Training Data References

We first study the effect of training data references, varying the amount of training data and type of translations, while using the same professional translation references for tuning and scoring. The first set of baseline experiments were trained on web forum data only, using professional translations. The first line of Table 1 shows that doubling of the training data adds 2.5 then 2.3 BLEU points. We repeated the experiments, but with MTurk training references, and saw that the scores are lower by 1.3-2.5 BLEU points, depending on the size of training data, and that the gain obtained from doubling the training data decreases to 2.0 and 1.6 BLEU points.

The lower MT scores and slower learning curve of the MTurk systems are both due to the lower quality of the translations, and to the mismatch with the professional development set translations (we discuss this issue further in §4.3). However, by interpolation of the MT scores, we find that the same MT performance can be obtained by using twice the amount of MTurk translated data as professional data. Considering that the MTurk translations is 10 times cheaper than professional translations (2.5¢ versus 25-30¢), this constitutes a cost ratio of 5x.

We repeated the above experiments, but this time added 10 million words of parallel data from the NIST MT 2012 corpora (mostly news) for training. We weighted the web forum part of the training data by a factor of 5. Note from the results in the right half of Table 1 that the newswire data improves the BLEU score by 2.5 to 6.3 BLEU points, depending on the size of the web forum data. This significant improvement is because some of the web forum user postings are formal and written in MSA (§3). More relevant to our aims is the comparison when we vary the web forum training references in the presence of the newswire training. The difference between the MTurk translation systems and the professional translation drops to 0.26-0.68 points. We conclude that in a domain adaptation scenario, where out-of-domain training data (i.e. newswire) already exists, crowdsourced translations for the in-domain (i.e. web forum) training data can be used with little to no loss in MT performance.

4.2 More Data vs. Multiple Translations

To our knowledge no previous work has compared using multiple reference translations for training data versus using additional training data of the same size. We studied this question by using both translations on the target side of the training data. Using the MTurk translations in addition to the professional translations in training gave a gain of 0.4 to 1.3 BLEU points (bottom half of Table 1). The gain was smaller in the presence of the GALE newswire data. When we compared with using the same amount of different training data instead of multiple references, we saw that training on new data with crowdsourced translations is better: training on two translations of 100KW gives 19.03, compared to 19.80 when training on a single translation of 200KW. The advantage of different-source data drops to 0.34 points when we start with 200KW. With a larger initial corpus, the additional source coverage of new data is not as critical, and the advantage of more variety on the target-side of the extracted translation rules becomes more competitive. This coverage is even less critical in the presence of the news data, where the ad-

Training	Tuning	Test	Training Data Size			
			100KW	200KW	400KW	400KW(no_lex)
Prof.	**Prof.**	**Prof.**	17.71	20.23	22.61	20.01
Prof.	**Prof.**	**Prof.+MTurk**	22.53	25.75	28.38	25.42
Prof.	**Prof. (len=0.95)**	**Prof.+MTurk**	23.63	26.84	**29.54**	26.17
Prof.	**Prof.+MTurk**	**Prof.+MTurk**	25.26	28.44	**30.94**	27.22
MTurk	**MTurk**	**MTurk**	16.66	18.47	20.35	17.75
MTurk	**MTurk**	**Prof.+MTurk**	23.83	26.45	28.66	25.44
MTurk	**MTurk (len=1.05)**	**Prof.+MTurk**	23.73	26.19	**28.74**	25.87
MTurk	**Prof.+MTurk**	**Prof.+MTurk**	24.91	27.66	**29.78**	26.45

Table 2: Effect of Tuning and Scoring References on MT.

vantage of new web forum source data disappears (lower-right quadrant of Table 1).

4.3 Development Data References

So far, we have focused on varying training data conditions, and kept the tuning and evaluation conditions fixed. But since we have re-translated the tuning and test sets on MTurk as well, we can study the effect of their reference translations on MT. As Table 2 shows, scoring the MT output using both reference translations, the BLEU scores increase by over 5 points (and more for the MTurk-trained system). This increase by itself is not remarkable. What is important to note is that the gain obtained by doubling the amount of training data is larger when measured using the multiple reference test set. We also ran experiments with 400KW training data, but with the lexical smoothing features (Koehn et al., 2003; Devlin, 2009) turned off. The bigger gains show that improvements in the MT output (from additional training or new features) can be better measured using a second MTurk reference of the test set.

Finally, we study the effect of tuning the system parameters using both translation references. Looking at the system trained on the professional translations, we see a gain of 2.5 to 2.7 BLEU points from adding the MTurk references to the tuning set. But as we mentioned earlier, the MTurk translations are shorter than the professional translations by around 10% on average. Tuning on both references, therefore, shortens the system output by around 5%. To neutralize the effect of length mismatch, we compared to a fairer baseline tuned on the professional references only, but we tuned the output-to-reference length ratio to be 0.95 (thus producing a shorter output). In this case, we see a gain of 1.4 points from adding the MTurk references to the tuning set.

We also used the multiple-reference tuning set to retune the systems trained on MTurk translations. Comparing that to a baseline that is tuned and scored using MTurk references only, we see a gain of around 1%. Note, however, that in this case the length mismach is reversed, and the output of the multiple-reference system is around 5% longer than that of the baseline. If we compare with a baseline that is tuned with a length ratio of 1.05 (to produce a longer output), we see the gain shrink only slightly.

To sum up this section, a second set of reference translations obtained via MTurk makes measurements of improvement on the test set more reliable. Also, a second set of references for tuning improves the output of the MT systems trained on either professional or MTurk references.

5 Conclusion

We compared professional and crowdsourced translations of the same data for training, tuning and scoring Arabic-English SMT systems. We showed that the crowdsourced translations yield the same MT performance as professional translations for as little as 20% of the cost. We also showed that a second crowsourced reference translation of the development set allows for a more accurate evaluation of MT output.

Acknowledgments

This work was supported in part by DARPA/IPTO Contract No. HR0011-12-C-0014 under the BOLT

Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Distribution Statement A (Approved for Public Release, Distribution Unlimited).

References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *NAACL '09: Proceedings of the 2009 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado.

Jacob Devlin. 2009. Lexical features for statistical machine translation. Master's thesis, University of Maryland, December.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton, Canada.

Nitin Madnani, Necip Fazil, Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Nitin Madnani, Philip Resnik, Bonnie Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the 8th Conf. of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawaii, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 577–585, Columbus, Ohio.

Omar F. Zaidan and Chris Callison-Burch. 2011a. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, June.

Omar F. Zaidan and Chris Callison-Burch. 2011b. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, June.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, June. Association for Computational Linguistics.