Measuring the behavioral impact of machine translation quality improvements with A/B testing

Benjamin Russell and **Duncan Gillespie**
Etsy
{brussell, dgillespie}@etsy.com

Abstract

In this paper we discuss a process for quantifying the behavioral impact of a domain-customized machine translation system deployed on a large-scale e-commerce platform. We discuss several machine translation systems that we trained using aligned text from product listing descriptions written in multiple languages. We document the quality improvements of these systems as measured through automated quality measures and crowdsourced human quality assessments. We then measure the effect of these quality improvements on user behavior using an automated A/B testing framework. Through testing we observed an increase in key e-commerce metrics, including a significant increase in purchases.

1 Introduction

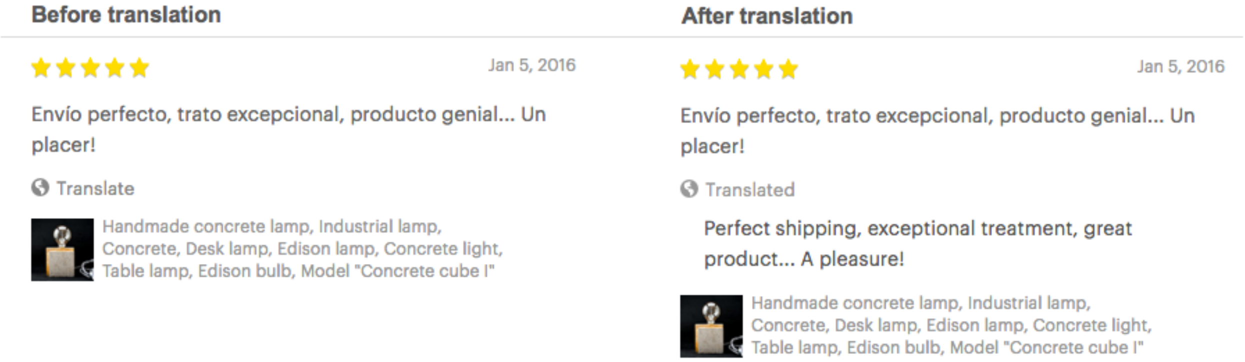
Quality evaluation is an essential task when training a machine translation (MT) system. While automatic evaluation methods like BLEU (Papineni et al., 2002) can be useful for estimating translation quality, a higher score is no guarantee of quality improvement (Callison-Burch et al., 2006). Previous studies (e.g. Coughlin, 2003) have compared human evaluations of MT to metrics like BLEU and found close correspondence between the two. Koehn (2004) argued that relatively small differences in BLEU can indicate significant MT quality differences and suggested that human evaluation, the traditional alternative to automated metrics like BLEU, is therefore unnecessarily time-consuming and costly. Callison-Burch (2009) explored the use of crowdsourcing platforms for evaluating MT quality, with good results. However, we are not aware of any research that investigates the effect of improved MT on human behavior. In a commercial application, like an e-commerce platform, it is desirable to have a high degree of confidence in the material effect of MT quality differences: any MT system change should positively impact user experiences.

Etsy is an online marketplace for handmade and vintage items, with over 40 million active listings and a community of buyers and sellers located around the world. Visitors can use MT to translate the text of product descriptions, product reviews, and private messages, making it possible for members to communicate effectively with one another, even when they speak different languages. These multilingual interactions facilitated by MT, such as reading nonnative listing descriptions or conversing with a foreign seller, are integral to the user experience.

However, due to the unique nature of the products available in the marketplace, a generic third party MT system[1] often falls short when translating user-generated content. One challenging lexical item is "clutch." A generic engine, trained on commonly available parallel text, translates clutch as an "automotive clutch." In this marketplace, however, clutch almost always means "purse." A mistake like this is problematic: a user who sees this incorrect machine translation may lose confidence in that listing and possibly in the marketplace as a whole.

[1] We use Microsoft's Bing Translator for our machine translations.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2295–2299,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics



Before translation

Jan 5, 2016

Envío perfecto, trato excepcional, producto genial... Un placer!

Translate

Handmade concrete lamp, Industrial lamp, Concrete, Desk lamp, Edison lamp, Concrete light, Table lamp, Edison bulb, Model "Concrete cube I"

After translation

Jan 5, 2016

Envío perfecto, trato excepcional, producto genial... Un placer!

Translated

Perfect shipping, exceptional treatment, great product... A pleasure!

Handmade concrete lamp, Industrial lamp, Concrete, Desk lamp, Edison lamp, Concrete light, Table lamp, Edison bulb, Model "Concrete cube I"

Figure 1: An example review translation on the website.

To improve the translation quality for terms like clutch, we used an interface provided by a third party machine translation service[2] to train a custom MT engine for English to French translations. To validate that the retrained MT systems were materially improved, we used a two step validation process, first using crowd-sourced evaluations with Amazon's Mechanical Turk, and secondly using A/B testing, a way of conducting randomized experiments on web sites, to measure the effect of the trained system on user behavior.

2 Data Collection

Our online marketplace contains millions of listing descriptions posted by tens of thousands of multilingual sellers. We conducted an MT system training using aligned texts from these product listings. We used our third-party translation service's automated retraining framework to train multiple MT systems that were specifically tuned to the marketplace's corpus. To gather this data, we used a Hadoop job to parse through 130 million active and expired product listings to find listing descriptions that were written in both English and French. Once we found these listing descriptions, we tokenized the text on sentence boundary. We removed any descriptions where there was a mismatch in the number of sentences between the source and target descriptions.

Next, we used a language detection service to ensure the source and target strings were the correct languages (source: English, target: French). After language detection, we removed all sentences where the ratio of alphabetic characters to total characters was below 70%. This 70% threshold was determined through manual assessment of the result set, and was used to eliminate strings with low numbers of alphabetic characters, such as "25.5 in x 35.5 in".

After these preliminary filtering steps, our training set consisted of 885,732 aligned sentences. To supplement the aligned text, we also collected 2,625,162 monolingual French sentences for the training. The monolingual text was parsed and cleaned in the same manner as the aligned sentences.

The commercial MT system's automatic training framework provides tools for the upload of bilingual and monolingual training data, tuning data, and testing data for customization of the underlying statistical MT system. Bilingual training data is used to modify the base translation model; monolingual data customizes the language model; the system is optimized for the tuning data; and the testing data is used to calculate a BLEU score. We trained over a dozen systems with a variety of datasets and selected the three systems that had the highest BLEU scores.

System 1 was trained using the aligned sentences, along with the 2.6 million monolingual sentences. The system was tuned using 2,500 sentences automatically separated from the training sentences by the third party's training system, and used an additional 2,500 automatically separated sentences for testing.

For System 2, we used a variation of the Gale-Church alignment algorithm (1993) to remove sentences predicted to be misaligned based on their length differences. The subject of sentence alignment in parallel texts has been researched exten-

[2] http://hub.microsofttranslator.com

Training Data	Sys. 1	Sys. 2	Sys. 3
886K aligned sentences	x		
766K aligned sentences after Gale-Church applied		x	x
2.6M monolingual segments	x	x	x
Auto tuning*	x	x	
Tuning with 2K in-domain sentences			x

Table 1: Data sets used for three MT system retrainings. *The third party's training platform automatically sets aside data to use for the parameter tuning.

sively (e.g. Brown et al., 1991; Gale and Church, 1993). Although more sophisticated methods exist (e.g. Chen, 1993; Wu, 1994; Melamed, 1996; Munteanu and Marcu, 2005), we used Gale-Church due to its relatively high accuracy and low implementation overhead. Misalignment between the same listing descriptions written in multiple languages could be caused by several factors, the most common problem being that sellers do not translate descriptions sentence for sentence from one language to the next. We detected possible misalignments in 13.5% of the original 886K aligned sentences, leaving 776K sentences to use for training System 2. We used auto-tuning and auto-testing for this engine, as we did for System 1.

System 3 was trained using the same training data as the second engine, but was tuned using 2,000 professionally-translated sentences taken from listing descriptions. Two hundred of these sentences were drawn semi-randomly to represent a general sample of listing description text; the remaining 1,800 contained terms, like "clutch," that were being mistranslated by the generic system. This system used the same automatically-generated testing data as the other two to calculate a BLEU score. Table 1 shows the training and tuning data used for the three systems.

3 Crowdsourced Evaluation

For evaluation of the trained translation systems, we generated translations of sentences drawn randomly from our monolingual English corpus (product listings that sellers had not translated into languages other than English). We excluded segments that were translated the same by both the trained and generic systems. (For System 1, 48 of 2,000 test sentences had the same translation as the generic system, for System 2 that number was 42, and for System 3 that number was 148.)

	BLEU Score	BLEU Score Improvement Over Generic System
System 1	48.16	+9.82
System 2	50.36	+12.02
System 3	46.85	+8.51

Table 2: BLEU score improvements for three translation systems over a baseline BLEU for the generic system of 38.34.

To obtain judgments about the quality of these translations, we used Mechanical Turk to obtain human evaluations of our candidate translation systems (Callison-Burch, 2009). To recruit Mechanical Turk workers with bilingual competence, we required workers to achieve at least 80% accuracy in a binary translation judgment task (workers were asked to judge whether each of 20 translations was "Good" or "Bad"; their answers were compared with those of professional translators).

Qualified workers completed a survey indicating their preference for the translation of a particular trained system compared to the generic commercial translation system. Translation pairs were presented in random order with no indication of whether a translation was produced by a human, a generic translation system, or an untrained translation system. Workers were asked to choose the better of the two translations or to indicate, "Neither is better". Workers were offered $2.00 to complete a 50-question survey. Each survey contained five hidden questions with known answers (translation pairs judged by professional translators) for quality control (we excluded responses from workers who did not answer the hidden questions with at least 80% accuracy).

4 Results

4.1 BLEU Evaluation

We used the automated BLEU calculation provided by the third-party translation service to obtain scores for each of the three translation systems.

All three systems had significant BLEU improvements after retraining, as shown in Table 2. We be-

	Trained	**Generic**	**Neither**	**Ratio**
Sys. 1	129 (34%)	109 (29%)	138 (36%)	1.18
Sys. 2	71 (25%)	85 (31%)	123 (44%)	0.84
Sys. 3	203 (36%)	150 (27%)	205 (37%)	1.35

Table 3: Results from crowdsourced evaluations of three translation systems. Columns labeled **Trained**, **Generic**, and **Neither** include the number of responses and percentage of total responses for each response type. The **Ratio** column shows the number of responses that favored the trained system to the number of responses that favored the generic system.

lieve System 3 has a lower BLEU score than the others because it was tuned on a different data set: the professionally-translated, in-domain sentences from product listing descriptions. This made the system's output less like the automatically-selected test set than the others, but closer, presumably, to the high-quality, low-noise tuning translations sourced from professional translators.

4.2 Crowdsourced evaluation

The crowdsourced evaluation of the three systems favored System 3. Table 3 provides a summary of the results. Neither System 1 nor System 2 showed a significant difference between selection of translations provided by the trained or untrained system: chi-squared tests did not detect a significant difference between number of responses favoring the trained system and number of responses favoring the generic system ($p = 0.1948$ and $p = 0.26$, respectively, for the two systems). However, a chi-squared test indicated a significant preference for System 3, which was chosen 35% more often than the generic system ($p = 0.0048$). Based on the crowd-sourced results, we proceeded to A/B test System 3 against the generic translation system baseline.

The lack of improvements for System 1 and System 2 detected using the crowd-sourcing methods was somewhat surprising, given the large BLEU score improvements observed for all three systems. We believe this lends further support to Callison-Burch, et al.'s (2006) critiques of BLEU as a stand-alone machine translation quality metric. In this case, it is possible that Systems 1 and 2 achieved high BLEU improvements due to over-fitting the training data from which the test set was drawn. We might speculate that this is due to the presence of low-quality translations from limited-bilingual sellers, or the presence of MT generated by a different online tool in some sellers' translations. By tuning the system using a high-quality, professionally-translated test set, we reduced overall BLEU but increased quality as judged by bilingual evaluators.

4.3 A/B testing

A/B testing is a strategy for comparing two different versions of a website to see which one performs better. Traditionally, one of these experiences is the existing, A, control experience, and the other experience is a new, B, variant experience. By randomly grouping users into one of the two experiences, and measuring the on-site behavior (e.g., clicks on a listing or items purchased) of each group, we can make data-driven decisions about whether new experiences are actually an improvement for our users. For our use case, the control experience is showing users content machine translated with the generic engine, and the variant experience is showing content translated with the retrained engine. A/B testing allows us to answer the following question: will users who read a product description translated by a domain-customized translation engine be more or less likely to purchase a product?

To test the effects of the quality improvement obtained, we used our in-house automated A/B testing framework to compare the behavioral effects on users who translated text using the generic engine and those who translated using System 3. Visitors to the online marketplace were randomly ”bucketed” into an experimental group or a control group. Random bucketing was achieved via a hash of a user's browser ID, which allows users who return to the site during the experimental period to be bucketed consistently across visits. For visitors who requested translations from English into French, the generic system's translations were displayed to visitors in the control group, and System 3 translations were displayed to visitors in the experimental group.

The experiment ran for 66 days for a total of 88,106 visitors (43,306 control and 44,800 experimental). The key metrics tracked were pages per visit (the number of pages seen in one user session), conversion rate (the percent of visits that include at least one purchase), and add-to-cart rate (the percent of visits in which a user adds an item to their shopping cart). We observed a significant positive

Metric	Trained engine
Conversion rate	+8.72%
Visit add-to-cart rate	+2.92%
Pages per visit	+3.37%

Table 4: The trained translation system's (System 3) improvement over the generic engine on key business metrics. All differences are statistically significant ($p < 0.05$). Base rates are omitted for data privacy reasons.

effect of the trained system on all three metrics, as shown in Table 4: a 3.37% increase in pages per visit ($p = 0.00153$ 95% CI $[1.29, 5.46]$), an 8.72% increase in purchase rate ($p = 0.00513$ 95% CI $[2.61, 14.82]$), and a 2.92% increase in add-to-cart ($p = 0.04689$ 95% CI $[0.04, 5.8]$).

5 Conclusion

Numerous studies have shown that automatic machine translation quality estimates, such as BLEU, are correlated with human evaluations of translation quality. Our work shows that those improvements in translation quality can have a positive effect on user behavior in a commercial setting, as measured through conversion rate. These considerations suggest that, in domains where machine translation conveys information upon which individuals base decisions, the effort needed to gather and process data to customize a machine translation system can be worthwhile. Additionally, our experiments show A/B testing can be a valuable tool to evaluate machine translation quality. A/B testing goes beyond measuring the quality of translation improvements: it allows us to see the positive impact that quality improvements are having on users' purchase behavior in a measurable way.

References

Peter F. Brown, Jennifer C. Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, pages 169–176.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*, volume 6, pages 249–256.

Chris Callison-Burch. 2009. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1, pages 286–295.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 9–16.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

I Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

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 on Association for Computational Linguistics*, pages 311–318.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 80–87.