Building English-Chinese and Chinese-English MT engines for the computer software domain

Maxim Khalilov
TAUS Labs
Oudeschans 85-3
Amsterdam, 1011KW
The Netherlands
maxim@tauslabs.com

Rahzeb Choudhury
TAUS
Oudeschans 85-3
Amsterdam, 1011KW
The Netherlands
rahzeb@translationautomation.com

Abstract

In this paper we present two sets of English-Chinese and Chinese-English machine translation trials conducted by TAUS Labs on computer software content. The goal of this study is twofold: (1) to share our experience on training and optimizing of Moses-based engines driven by translation memories provided by industrial users and (2) to give to the users the idea of results, cost and effort associated with training of MT engines.

1 Introduction: goals and approach

We describe a series of English-Chinese and Chinese-English machine translation trials conducted by TAUS Labs[1] on computer software content. Statistical MT engines were trained and tested on the basis of open-source software using Amazon Elastic Cloud[2] as a remote server. Parallel corpora were downloaded from the TAUS Data Association repository[3].

In this study we focused on the following particular questions that MT users are interested in:

- How well do statistical customizable MT engines based on Moses perform in comparison with Google Translate?
- Which Chinese word segmentation and reordering strategies improve translation performance?
- How expensive (in terms of time and money) is the process of MT engine training?
- How well do the automatic evaluation metrics BLEU, TER and GTM correlate with each other? What are the results of human evaluation?

While in the majority of experiments published in academic conferences tend to use only free publicly available corpora to feed MT engines, we trained our systems on the data provided by 10 industrial publishers.

2 Data

Experiments were conducted using different variations of the Chinese-English training corpus, built on a basis of translation memories coming from the software industry. Test and development datasets were provided by EMC[4].

The training dataset contains around 22 million words on the English side and around 23 millions on the Chinese side. The development set was 500 lines long (7,000 on the Chinese side), while translation systems were tested on the corpus of around 15,000 words.

3 Baseline and experiments

The SMT system used in the experiments was implemented within the open-source MOSES toolkit (Koehn et al., 2007). Training and tuning procedures are detailed on the MOSES web page[5].

Word alignment was estimated with GIZA++ tool[6] (Och, 2003), coupled with mkcls[7] (Och, 1999), which allows for statistical word clustering for better generalization. A 3-gram target lan-

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

[2] http://aws.amazon.com/ec2

[3] http://www.tausdata.org

[4] http://www.emc.com/

[5] http://www.statmt.org/moses/

[6] code.google.com/p/giza-pp/

[7] http://www.fjoch.com/mkcls.html

guage model was estimated using the SRI LM toolkit (Stolcke, 2002).

In the writing system of Chinese, texts are not segmented by words, but Moses operates with words (tokens) rather than with unbroken strings. We used two alternative segmenters for Chinese in the pre-processing step: the Stanford Chinese segmenter[8] (Tseng et al., 2005) and the Simplified Chinese segmenter (the Peterson segmenter[9]) with the goal to determine which segmentation strategy leads to better MT system performance.

Two reordering methods are widely used along with Moses-based MT systems:

Distance-based reordering (Koehn et al., 2003): a simple distance-based reordering model default for Moses system.

MSD (Tillman, 2004): a lexicalized data-driven reordering model. The MSD model is used together with a distance-based reordering.

4 Evaluation methodology

Automatic evaluation. In English-Chinese experiments, Chinese reference translation was pre-segmented using one of the two segmentation tools (the Peterson or the Stanford segmenter) in order to make the evaluation as fair as possible. The reason for that was an intention to minimize the segmentation effect for Chinese portion of the data and focus the evaluation on the correctness of lexical choice and word order.

In Chinese-English trials, all the automatically generated translation hypotheses and reference translation were detokenized using *detokenizer.pl* script distributed as a Moses package.

We used three evaluation metrics to measure translation quality in a resource-light way:

- GTM (Turian et al., 2003), a precision-recall metric measuring similarity between MT output and reference translation. It takes into account the number of adjacent words shared by translation hypothesis and reference.

- TER (Snover et al., 2006), a metric based on the counting transformations required to reconstruct the reference translation from the MT output, while preserving the content of the source. TER estimates the number of edits required to change a system output into one of the references.

- BLEU (Papineni et al., 2002), a simple evaluation metric that performs better on capturing fluency rather than adequacy of the translation. BLEU shows how many words are shared between MT output and human-made reference, benefiting sequential words.

BLEU is still a de-facto standard evaluation tool for academic research on MT, despite its obvious disadvantages. BLEU tends to give a very high score with a short output, so long as all its n-grams are present as a reference. Besides, BLEU is mostly a precision metric, taking recall into account in a very simple way by considering only the measure for sentence length.

While BLEU is criticized within academic and industrial MT communities because in many cases it does not show good correlation with human judgment (Callison-Burch, 2006), GTM is reported to be a more reliable way to measure translation quality, at least for certain domains (O'Brien, 2011). Due to this reason and since it has the strong correlation with post-editing effort (Tatsumi, 2009) GTM was selected as the primary indicator of translation quality.

TER is currently considered more reliable metric than BLEU for some of the most popular translation applications since it gives a better indication of the post-editing effort compared to BLEU (O'Brien, 2011).

A comparison with free online engine was completed for informative purposes. Since Google is not a member of TAUS Data Association, it does not have access to the parallel corpus that was used to train the Moses systems.

The evaluation conditions for English were case-sensitive and included punctuation marks. The Chinese translation generated by Google Translate was re-segmented to preserve the consistency of evaluation.

Human evaluation. The native speaker evaluator was provided with the original text in source language and the outputs of the four translation systems. They were asked to assess the quality of 100 lines from the test corpus using the following unique scale to measure the acceptability of the output at the segment level .

Using the methodology described in Roturier (2009) and Specia (2011) we apply a 4-level scale to measure output acceptability:

[8] http://nlp.stanford.edu/software/segmenter.shtml

[9] http://www.mandarintools.com/segmenter.html

- Excellent (E): no post-editing required;
- Good (G): only minor post-editing is required;
- Medium (M): significant post-editing is required;
- Poor (P): it would be better to manually retranslate from scratch (post-editing is not worthwhile).

We also used the aggregated score following a simple approach to assign a certain weight to each category, multiply the number of occurrence by those weights, sum them up and normalize:

$$K = \frac{\sum\limits_{i \in P,M,G,E} w_i * N_i}{N} \qquad (1)$$

where $N = \sum\limits_{i \in P,M,G,E} N_i$, $w_P = 1$, $w_M = 2$, $w_G = 3$ and $w_E = 4$.

5 Results

5.1 Automatic scores

We contrast the results shown by 4 translation systems per direction with the performance delivered by Google Translate (Table 1 and Figure 1).

5.2 Human evaluation

Some of the systems under consideration were analyzed by human judges following the strategy described in Section 4. Early results can be found in Table 2.

SID	Segment.	Reord.	GMT	TER	BLEU
		Chinese-English			
1	Peters.	MSD	67.95	36.51	49.41
2	Peters.	Dist.	67.22	37.81	48.46
3	Stanf.	MSD	64.99	40.32	45.16
4	Stanf.	Dist.	64.32	40.55	44.52
G	Google	N/A	62.95	63.40	24.78
		English-Chinese			
1	Peters.	MSD	76.75	39.35	36.51
2	Peters.	Dist.	76.63	39.62	34.29
3	Stanf.	Dist.	76.57	40.95	32.44
4	Stanf.	MSD	76.54	40.82	33.69
G	Google	N/A	60.81	56.99	9.40

Table 1: Automatic scores.

SID	Segment.	Reord.	P	M	G	E	**K**
			Chinese-English				
1	Peters.	MSD	11	18	43	28	**2.88**
4	Stanf.	Dist.	11	17	46	26	**2.87**
G	Google	N/A	13	26	40	21	**2.69**
			English-Chinese				
1	Peters.	MSD	65	10	10	11	**1.59**
4	Stanf.	MSD	66	12	10	11	**1.64**
G	Google	N/A	64	17	11	8	**1.63**

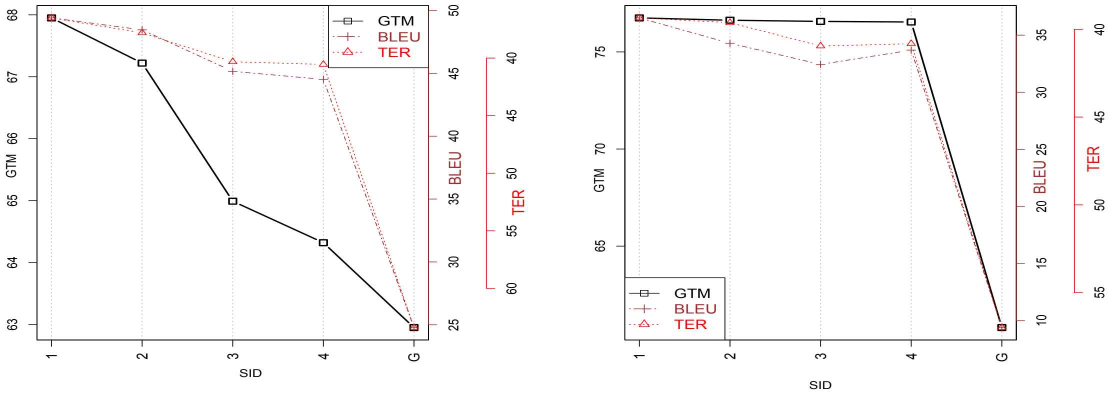
Table 2: Human evaluation results.

5.3 Correlation of automatic scores

The experiments gave us an opportunity to check how well GTM, TER and BLEU correlate for a single-reference evaluation task.

Trial	BLEU-TER	BLEU-GTM	TER-GTM
Ch.-En.	-0.84	-0.60	0.66
En.-Ch.	-0.99	-0.98	0.45

Table 3: BLEU, TER and GTM correlation.



(a) Chinese-English experiments.

(b) English-Chinese experiments.

Figure 1: GTM, TER and BLEU scores.

The Pearson correlation coefficient shows strong dependence between BLEU and TER and, to a lesser extent, between BLEU and GTM metrics, which is significantly stronger for Chinese-English translation.

The results of manual judgement for Chinese-English confirm the results of automatic evaluation in grosso mode. However, there is a significant discrepancy in BLEU/TER/GTM scores and human results for English-Chinese: while according to the automatic scores Moses translations are much better than Google Translate, human evaluation shows that customized Moses and Google Translate perform virtually indistinguishable from each other. We explain it by an effect of non-ideal target-side segmentation that affects automatic scores, but is disregarded by the evaluator[10].

6 Cost and effort

Data processing has been done on a local machine (regular laptop). We assume that the cost associated with its usage is virtually zero.

Cost associated with AWS usage: the total technology costs for these trials were around 28 euros per direction (4 MT engines, 2 different datasets).

Human and time resources: data preparation was done in parallel for both translation directions. While the most time-consuming part, which is training corpus processing, was shared for both trials, development and test corpora were cleaned, tokenized and segmented independently. The data preparation process took around 16 hours.

MT engine training, system optimization, and backing-up amounted to 60 hours, equally distributed between 2 master engines. Around 30% (18 hours) of that lapsed time required human resources (mostly, on the data preparation step).

7 Findings and future work

We operate with an open-source Moses toolkit that implements the entirely data-driven approach to MT. The corpus-based nature of this software implies high dependence on parallel data that is fed to the MT engine.

- Unsurprisingly, we find that in domain texts are translated much better by Moses MT solutions trained on specific material than the high-quality, but general-purpose Google Translate tool. The best Moses-based system performs two times better than Google Translate in terms of BLEU score (+7% in terms of GMT) for Chinese-English translation. Moses-based solutions outperform the online Google solution by almost four times (BLEU) and +20 % (GMT) when translating from English into Chinese.

- Access to the right data, which is a core element of MT customization is the key aspect in getting competitive translation performance. This should be taken into account by decision makers when adopting or integrating MT.

- The choice of word segmentation strategy for Chinese can have significant impact on the delivered translation. Segmentation of Chinese portion of parallel corpora (training, development and test) with the use of a rather simple, but efficient Peterson segmenter leads to a better performance than segmentation done using Stanford segmenter based on pattern recognition algorithm.

- Notable finding of this study is that some of the evaluation metrics based on different principles are well correlated. BLEU (the metric that estimates the number of n-grams shared by translation hypothesis and human reference) and TER (the metric based on counting of number of text transformations) report high correlation for both directions ($|r|$=0.84 for English-Chinese and $|r|$=-0.99 for Chinese-English). GTM is well correlated with BLEU. Correlation for English-Chinese translation is much weaker than for Chinese-English.

- The results of human evaluation confirm the scores shown by automatic metrics for Chinese-English trials, but do not verify the huge degradation in Google Translate performance shown by BLEU, TER and GTM scores for English-Chinese.

[10] A calculation of *r* correlation between automatic scores and human evaluation results is not presented in this paper due to a low number of manually evaluated systems.

References

Koehn, Ph., F. Och, and D. Marcu. 2003. Statistical phrase-based machine translation. In *Proceedings of the HLT-NAACL'03*, pages 48–54.

Koehn, Ph., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open-source toolkit for statistical machine translation. In *Proceedings of ACL'07*, pages 177–180, Prague, Czech Republic.

O'Brien, Sh. 2011. Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215.

Och, F. 1999. An efficient method for determining bilingual word classes. In *Proceedings of ACL'99*, pages 71–76, Maryland, MD, USA.

Och, F. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL'03*, pages 160–167, Sapporo, Japan.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318, Philadelphia, PA, USA.

Roturier, J. 2009. Deploying novel mt technology to raise the bar for quality: a review of key advantages and challenges. In *Proceedings of the MT Summit XII*, pages 1–8, Ottawa, Canada, August.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Specia, L. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of EAMT 2011*.

Stolcke, A. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of SLP'02*, pages 901–904.

Tatsumi, M. 2009. Correlation between automatic evaluation scores, post-editing speed and some other factors. In *Proceedings of MT Summit XII*, pages 332–339, Ottawa, Canada, August.

Tillman, C. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*, pages 101–104, Boston, MA, USA.

Tseng, H., P. Chang, G. Andrew, D. Jurafsky, and Ch. Manning. 2005. A conditional random field word segmenter. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Turian, J., L. Shen, and I.D. Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX*, pages 386–393, New Orleans, USA, September.