An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation

Liling Tan
Universität des Saarland
Campus, Saarbrücken, Germany
liling.tan@uni-saarland.de

Jon Dehdari, Josef van Genabith
Deutsches Forschungszentrum für Künstliche Intelligenz / Saarbrücken, Germany
{first.last_name}@dfki.de

Abstract

Automatic evaluation of machine translation (MT) quality is essential in developing high quality MT systems. Despite previous criticisms, BLEU remains the most popular machine translation metric. Previous studies on the schism between BLEU and manual evaluation highlighted the poor correlation between MT systems with low BLEU scores and high manual evaluation scores. Alternatively, the RIBES metric—which is more sensitive to reordering—has shown to have better correlations with human judgements, but in our experiments it also fails to correlate with human judgements. In this paper we demonstrate, via our submission to the Workshop on Asian Translation 2015 (WAT 2015), a patent translation system with very high BLEU and RIBES scores and very poor human judgement scores.

1 Introduction

Automatic Machine Translation (MT) evaluation metrics have been criticized for a variety of reasons (Babych and Hartley, 2004; Callison-Burch et al., 2006). However, the relatively consistent correlation of higher BLEU scores (Papineni et al., 2002) and better human judgements in major machine translation shared tasks has led to the conventional wisdom that translations with significantly higher BLEU scores generally suggests a better translation than its lower scoring counterparts (Bojar et al., 2014; Bojar et al., 2015; Nakazawa et al., 2014; Cettolo et al., 2014).

Callison-Burch et al. (2006) has anecdotally presented possible failures of BLEU by showing examples of translations with the same BLEU score but of different translation quality. Through meta-evaluation[1] of BLEU scores and human judgements scores of the 2005 NIST MT Evaluation exercise, they have also showed high correlations of R^2 = 0.87 (for adequacy) and R^2 = 0.74 (for fluency) when an outlier rule-based machine translation system with poor BLEU score and high human score is excluded; when included the correlations drops to 0.14 for adequacy and 0.74 for fluency.

Despite showing the poor correlation between BLEU and human scores, Callison-Burch et al. (2006) had only empirically meta-evaluated a scenario where low BLEU score does not necessary result in a poor human judgement score. In this paper, we demonstrate a real-world example of machine translation that yielded high automatic evaluation scores but failed to obtain a good score on manual evaluation in an MT shared task submission.

2 BLEU

Papineni et al. (2002) originally define BLEU n-gram precision p_n by summing the n-gram matches for every hypothesis sentence S in the test corpus C:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)} \tag{1}$$

BLEU is a precision based metric; to emulate recall, the brevity penalty (BP) is introduced to compensate for the possibility of high precision translation that are too short. The BP is calculated as:

[1]Meta-evaluation refers to the measurement of the Pearson correlation R^2 between an automatic evaluation metrics and human judgment scores. More recently, meta-evaluation involves the calculation using other correlation measures such as the Spearman's rank correlation ρ (Callison-Burch et al., 2007) or the Kendall's Tau τ (Stanojević et al., 2015; Graham et al., 2015)

Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 74–81,
Kyoto, Japan, 16th October 2015.

$$BP = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases} \tag{2}$$

where c and r respectively refers to the length of the hypothesis translations and the reference translations. The resulting system BLEU score is calculated as follows:

$$\text{BLEU} = \text{BP} \times \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{3}$$

where n refers to the orders of n-gram considered for p_n and w_n refers to the weights assigned for the n-gram precisions; in practice, the weights are uniformly distributed.

A BLEU score can range from 0 to 1 and the closer to 1 indicates that a hypothesis translation is closer to the reference translation[2].

Traditionally, BLEU scores has showed high correlation with human judgements and is still used as the *de facto* standard automatic evaluation metric for major machine translation shared tasks. And BLEU continues to show high correlations primarily for *n*-gram-based machine translation systems (Bojar et al., 2015; Nakazawa et al., 2014).

However, the fallacy of BLEU-human correlations can be easily highlighted with the following example:

Source:
이러한작용을발휘하기위해서는, 각각 0.005% 이상함유하는것이바람직하다.

Hypothesis:
このような作用を発揮するためには、夫々０．００５％以上含有することが好ましい。

Baseline:
このような作用を発揮するためには、それぞれ ０．００５％以上含有することが好ましい。

Reference:
このような作用を発揮させるためには、夫々０．００５％以上含有させることが好ましい。

Source/Reference English Gloss:
"So as to achieve the reaction, it is preferable that it contains more 0.005% of each [chemical]"

The unigram, bigram, trigrams and fourgrams (p_1, p_2, p_3, p_4) precision of the hypothesis translation are 90.0, 78.9, 66.7 and 52.9 respectively. The p_n score for the hypothesis sentence precision score for the reference is 70.75. When considering the brevity penalty of 0.905, the overall BLEU is 64.03. Comparatively, the *n*-gram precisions for the baseline translations are p_1=84.2, p_2=66.7, p_3=47.1 and p_4=25.0 and the overall BLEU is 43.29 with a BP of 0.854. In this respect, one would consider the baseline translation inferior to the hypothesis with a >10 BLEU difference. However, there is only a subtle difference between the hypothesis and the baseline translation (それぞれvs 夫々).

This is an actual example from the 2nd Workshop on Asian Translation (WAT 2015) MT shared task evaluation, and five crowd-sourced evaluators consider the baseline translation a better translation. For this particular example, the human evaluators preferred the natural translation from Korean 각각 *gaggag* to Japanese それぞれ *sorezore* instead of the patent document usage of 夫々 *sorezore*, both それぞれ and 夫々 can be loosely translated as '*respectively*' or '*(for) each*' in English.

The big difference in BLEU for a single lexical difference in translation is due to the geometric averaged scores for the individual *n*-gram precisions. It assumes the independence of *n*-gram precisions and accentuates the precision disparity by involving the single lexical difference in all possible *n*-grams that capture the particular position in the sentence. This is clearly indicated by the growing precision difference in the higher order *n*-grams.

2.1 RIBES

Another failure of BLEU is the lack of explicit consideration for reordering. Callison-Burch et al. (2006) highlighted that since BLEU only takes reordering into account by rewarding the higher *n*-gram orders, freely permuted unigrams and bigrams matches are able to sustain a high BLEU score with little penalty caused by tri/fourgram mismatches. To overcome reordering, the RIBES

[2] Alternatively, researchers would choose to inflate the BLEU score to a range between 0 to 100 to improve readability of the scores without the decimal prefix.

score was introduced by adding a rank correlation coefficient[3] prior to unigram matches without the need for higher order *n*-gram matches (Isozaki et al., 2010).

Let us consider another example:

Source:
T용융(DSC) = 89.9℃; T결정화(DSC) = 72℃(
5℃/ 분에서DSC 로측정) .

Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．9結晶化度（Ｔ）。

Baseline:
Ｔ溶融（ＤＳＣ）＝８９．９℃；Ｔ結晶化（ＤＳＣ）＝７２℃（５℃／分でＤＳＣで測定）。

Reference:
Ｔｍｅｌｔ（ＤＳＣ）＝８９．９℃；Ｔｃｒｙｓｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣを用いて測定）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)

The example above shows the marginal effectiveness of RIBES when penalizing wrongly ordered phrases in the hypothesis. The baseline translation accurately translates the meaning of the sentence with a minor partial translation of the technical variables (i.e. *Tmelt* -> Ｔ溶融 and T결정화 -> Ｔ結晶化. However, the hypothesis translation made serious adequacy errors when inverting the values of the technical variables but the hypothesis translation was minimally penalized in RIBES and also BLEU.

The RIBES score for the hypothesis and baseline translations are 94.04 and 86.33 respectively whereas their BLEU scores are 53.3 and 58.8. In the WAT 2015 evaluation, five evaluators unanimously voted in favor for the baseline translation. Although the RIBES score presents a wider difference between the hypothesis and baseline translation than BLEU, it is insufficient to account for the arrant error that the hypothesis translation made.

[3]normalized Kendall τ of all *n*-gram pairs between the hypothesis and reference translations

2.2 Other Shades of BLEU / RIBES

It is worth noting that there are other automatic MT evaluation metrics that depend on the same precision-based score with primary differences in how the $Count_{match}(ngram)$ is measured; Giménez and Màrquez (2007) described other linguistics features that one could match in place of surface *n*-grams, such as lexicalized syntactic parse features, semantic entities and roles annotations, etc. As such, the modified BLEU-like metrics can present other aspects of syntactic fluency and semantic adequacy complementary to the string-based BLEU.

A different approach to improve upon the BLEU scores is to allow paraphrases or gappy variants and replace the proportion of $Count_{match}(ngram)$ / Count(ngram) by a lexical similarity measure. Banerjee and Lavie (2005) introduced the METEOR metric that allows hypotheses' *n*-grams to match paraphrases and stems instead of just the surface strings. Lin and Och (2004) presented the ROUGE-S metrics that uses skip-grams matches. More recently, pre-trained regression models based on semantic textual similarity and neural network-based similarity measures trained on skip-grams are applied to replace the *n*-gram matching (Vela and Tan, 2015; Gupta et al., 2015).

While enriching the surface *n*-gram matching allows the automatic evaluation metric to handle variant translations, it does not resolves the "prominent crudeness" of BLEU (Callison-Burch, 2006) involving (i) the omission of content-bearing materials not being penalized, and (ii) the inability to calculate recall despite the brevity penalty.

3 Experimental Setup

We describe our system submission[4] to the WAT 2015 shared task (Nakazawa et al., 2015) for Korean to Japanese patent translation.[5].

The Japan Patent Office (JPO) Patent Corpus is the official resource provided for the shared task. The training dataset is made up of 1 million sentences (250k each from the chemistry, electricity, mechanical engineering and physics do-

[4]Our Team ID in WAT 2015 is `Sense`

[5]Although, we have also participated in the English-Japanese-Chinese scientific text translation subtask using the ASPEC corpus, our results have been presented in Tan and Bond (2014) and Tan et al. (2015)

Parameters	Organizers	Ours
Input document length	40	80
Korean tokenizer	MeCab	KoNLPy
Japanese tokenizer	Juman	MeCab
LM n-gram order	5	5
Distortion limit	0	20
Quantized & binarized LM	no	yes
`devtest.txt` in LM	no	yes
Binarized phrase tables	no	yes
MERT runs	1	2

Table 1: Differences between Organizer's and our Phrase-based SMT system

mains). Two development datasets[6] and one test set each comprises 2000 sentences with 500 sentences from each of the training domains. The Korean and Japanese texts were tokenized using KoNLPy (Park and Cho, 2014) and MeCab (Kudo et al., 2004) respectively.

We used the phrase-based SMT implemented in the Moses toolkit (Koehn et al., 2003; Koehn et al., 2007) with the following vanilla Moses experimental settings:

- `MGIZA++` implementation of IBM word alignment model 4 with grow-diagonal-final-and heuristics for word alignment and phrase-extraction (Och and Ney, 2003; Koehn et al., 2003; Gao and Vogel, 2008)(Koehn et al., 2003; Och and Ney, 2003; Gao and Vogel, 2008)
- Bi-directional lexicalized reordering model that considers monotone, swap and discontinuous orientations (Koehn, 2005; Galley and Manning, 2008)
- To minimize the computing load on the translation model, we compressed the phrase-table and lexical reordering model (Junczys-Dowmunt, 2012)
- Language modeling is trained using KenLM using 5-grams, with modified Kneser-Ney smoothing (Heafield, 2011; Kneser and Ney, 1995; Chen and Goodman, 1998). The language model is quantized to reduce filesize and improve querying speed (Heafield et al., 2013; Whittaker and Raj, 2001).
- Minimum Error Rate Training (MERT) (Och, 2003) to tune the decoding parameters.

[6] `dev.txt` and `devtest.txt`

3.1 Human Evaluation

The human judgment scores for the WAT evaluations were acquired using the Lancers crowdsourcing platform (WAT, 2014). Human evaluators were randomly assigned documents from the test set. They were shown the source document, the hypothesis translation and a baseline translation generated by the baseline phrase-based MT system.

3.1.1 Baseline System

Human evaluations were conducted as pairwise comparisons between translations from our system and the WAT organizers' phrase-based statistical MT baseline system. Table 1 highlights the parameter differences between the organizers and our phrase-based SMT system.

3.1.2 Pairwise Comparison

The human judgment scores for the WAT evaluations were acquired using the Lancers crowdsourcing platform. Human evaluators were randomly assigned documents from the test set. They were shown the source document, the hypothesis translation and a baseline translation generated by the phrase-based MT system. Five evaluators were asked to judge each document.

The crowdsourced evaluators were non-experts, thus their judgements were not necessary precise, especially for patent translations. The evaluators were asked to judge whether the hypothesis or the baseline translation was better, or they were tied. The translation that was judged better constituted a *win* and the other a *loss*. For each, the majority vote between the five evaluators for the hypothesis decided whether the hypothesis *won*, *lost* or *tied* the baseline. The final human judgment score, *HUMAN*, is calculated as follows:

$$\text{HUMAN} = 100 \times \frac{W - L}{W + L + T} \quad (4)$$

By definition, the *HUMAN* score ranges from -100 to $+100$, where higher is better.

4 Results

Moses' default parameter tuning method, MERT, is non-deterministic, and hence it is advisable to tune the phrase-based model more than once (Clark et al. 2011). We repeated the tuning step and submitted the system translations that achieved the higher BLEU score for manual evaluation.

As a sanity check we also replicated the organizers' baseline system and submitted it for manual evaluation. We expect this system to score close to zero. We submitted a total of three sets of output to the WAT 2015 shared task, two of which underwent manual evaluation.

Systems	**RIBES**	**BLEU**	**HUMAN**
Organizers' PBMT baseline	94.13	69.22	0.0
Our replica baseline	94.29	70.23	**+3.50**
Ours (MERT 1)	95.03	84.26	-
Ours (MERT 2)	**95.15**	**85.23**	-17.75

Table 2: BLEU and HUMAN scores for WAT 2015

Table 2 presents the BLEU scores achieved by our phrase-based MT system in contrast to the organizers' baseline phrase-based system. The difference in BLEU between the organizers' system and ours may be due to our inclusion of the second development set in building our language model and the inclusion of more training data by allowing a maximum of 80 tokens per document as compared to 40 (see Table 1).

Another major difference is the high distortion limit we have set as compared to the organizers' monotonic system, it is possible that the high distortion limit compensates for the long distance word alignments that might have been penalized by the phrasal and reordering probabilities which results in the higher RIBES and BLEU score.[7]

However, the puzzling fact is that our system being 15 BLEU points better than the organizers' baseline begets a terribly low human judgement score. We discuss this next.

5 Segment Level Meta-Evaluation

We perform a segment level meta-evaluation by calculating the BLEU and RIBES score difference for each hypothesis-baseline translation. Figures 1 and 2 show the correlations of the BLEU and RIBES score difference against the positive and negative human judgements score for every sentence.

Figure 1 presents the considerable incongruity between our system's high BLEU improvements (>+60 BLEU) being rated marginally better than the baseline translation, indicated by the orange and blue bubbles on the top right corner. There were even translations from our system with >+40 BLEU improvements that tied with the organizer's baseline translations, indicated by the grey bubbles at around the +40 BLEU and +5 RIBES region. Except for the a portion of segments that scored worse than the baseline system (lower right part of the graph where BLEU and RIBES falls below 0), the overall trend in Figure 1 presents the conventional wisdom that the BLEU improvements from our systems reflects positive human judgement scores.

However, Figure 2 presents the awkward disparity where many segments with BLEU improvements were rated strongly as poorer translations when compared against the baseline. Also, many segments with high BLEU improvements were tied with the baseline translations, indicated by the grey bubbles across the positive BLEU scores.

As shown in the examples in Section 2, a number of prominent factors contribute to these disparity in high BLEU / RIBES improvements and low HUMAN judgement scores:

- Minor lexical differences causing a huge difference in *n*-gram precision
- Crowd-sourced *vs.* expert preferences on terminology, especially for patents

[7] In our submission `Byte2String` refers to the encoding problem we encountered when tokenizing the Korean text with MeCab causing our system to read Korean byte-code instead of Unicode. But the decoder could still output Unicode since our Japanese data was successfully tokenized using MeCab, we submitted this output under the submission name `Byte2String`; the `Byte2String` submission is not reported in this paper. Later we rectified the encoding problem by using KoNLPy and re-ran the alignment, phrase extraction, MERT and decoding, hence the submission name, `Unicode2String`, i.e. the system reported in Table 2.

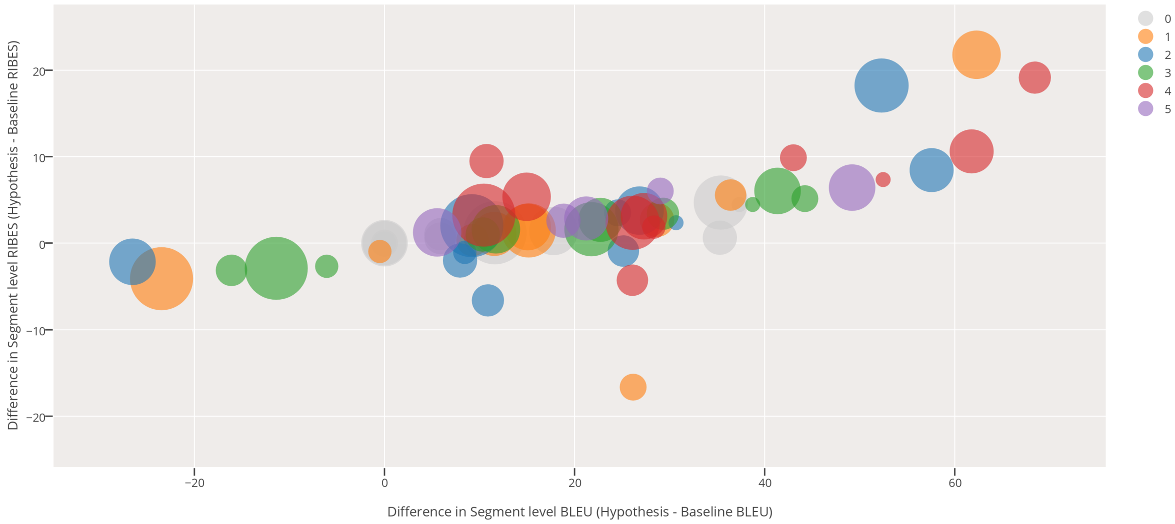

Figure 1: Correlation between BLEU, RIBES differences and *Positive* HUMAN Judgements (HUMAN Scores of 0, +1, +2, +3, +4 and +5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores)

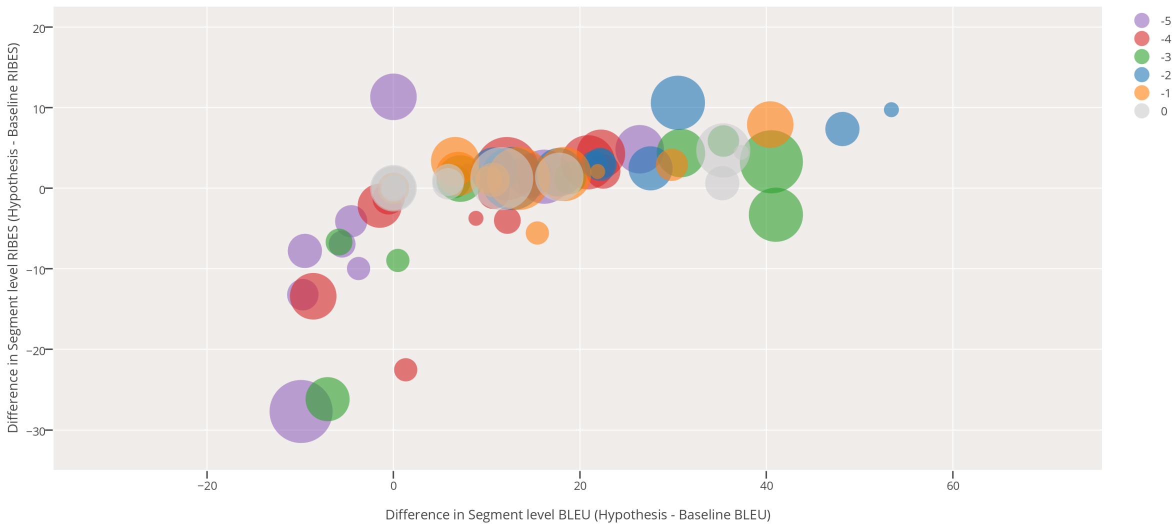

Figure 2: Correlation between BLEU, RIBES differences and *Negative* HUMAN Judgements (HUMAN Scores of 0, -1, -2, -3, -4 and -5 represented by the colored bubbles: *grey, orange, blue, green, red and purple*; larger area means more segments with the respective HUMAN Scores)

- Minor MT evaluation metric differences not reflecting major translation inadequacy

Each of these failures contributes to an increased amount of disparity between the automatic translation metric improvements and human judgement scores.

6 Conclusion

In this paper we have demonstrated a real-world case where high BLEU and RIBES scores do not correlate with better human judgement. Using our system's submission for the WAT 2015 patent shared task, we presented several factors that might contribute to the poor correlation, and also performed a segment level meta-evaluation to identify segments where our system's high BLEU / RIBES improvements were deemed substantially worse than the baseline translations. We hope our results and analysis will lead to improvements in automatic translation evaluation metrics.

Acknowledgements

We thank the Workshop on Asian Translation organizers for sharing the results of the manual annotations, the organization of the task and the helpful feedbacks to improve this paper.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

References

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 621.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

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

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, CA*, pages 2–17.

Stanley Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report 10-98, Harvard University.

Michel Galley and Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Stroudsburg, PA, USA.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on*

Empirical Methods in Natural Language Processing, pages 944–952.

Marcin Junczys-Dowmunt. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *The Prague Bulletin of Mathematical Linguistics*, 98:63–74.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit, vol. 5, pp. 79-86.*

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the first workshop on Asian translation. In *Proceedings of the First Workshop on Asian Translation (WAT2014)*, Tokyo, Japan.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd workshop on Asian translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, Kyoto, Japan.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167.

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.

Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*, Chuncheon, Korea.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal.

Liling Tan and Francis Bond. 2014. Manipulating input data in machine translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014).*

Liling Tan, Josef van Genabith, and Francis Bond. 2015. Passive and pervasive use of bilingual dictionary in statistical machine translation. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 30–34, Beijing.

Mihaela Vela and Liling Tan. 2015. Predicting machine translation adequacy with document embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 402–410, Lisbon, Portugal.

Edward W.D. Whittaker and Bhiksha Raj. 2001. Quantization-based language model compression. In *Proceedings of INTERSPEECH*, pages 33–36.