Review and Analysis of China Workshop on Machine Translation 2013 Evaluation

Sitong Yang[1,2] yangsitong@ict.ac.cn
Heng Yu[1] yuheng@ict.ac.cn
Hongmei Zhao[1] zhaohongmei@ict.ac.cn
Qun Liu[1,3] qliu@computing.dcu.ie
Yajuan lü[1] lvyajuan@ict.ac.cn
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] CNGL, School of Computing, Dublin City University

Abstract

This paper gives a general review and detailed analysis of China Workshop on Machine Translation (CWMT) Evaluation. Compared with the past CWMT evaluation campaigns, CWMT2013 evaluation is characterized as follows: first, adopting gray-box evaluation which makes the results more replicable and controllable; second, adding one rule-based system as a counterpart; third, carrying out manual evaluations on some specific tasks to give a more comprehensive analysis of the translation errors. Boosted by those new features, our analysis and case study on the evaluation results shows the pros and cons of both rule-based and statistical systems, and reveals some interesting correlations bewteen automatic and manual evaluation metrics on different translation systems.

1 Introduction

The China Workshop on Machine Translation has always been focusing on catching the latest development of Machine Translation (MT) and promoting the communication between related organizations in China. By convention, we organized a unified machine translation evaluation in 2013, sponsored by Chinese Information Processing Society of China and held by the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS).

Compared with the previous evaluation [Zhao et al., 2009], the main improvements of CWMT2013 are as follows: First, we follow the "Gray-Box Evaluation" mode, which not only requires the participants to submit the final translation results but also

results of some key intermediate procedures as gray-box files, such as alignment results, k-best translation, etc. This mechanism makes the results more replicable and controllable, at the same time it enables the participants to identify the weak link in their system pipeline, and make targeted adjustment to improve translation quality. Second, we adopt one rule-based system along with its statistical counterparts to make a more comprehensive comparison between different kinds of MT systems. To increase the diversity of evaluation method, two additional automatic evaluation metrics are also introduced: METEOR [Banerjee and Lavie, 2005] and TER [Snover et al., 2006]. Finally, Besides the automatic evaluation, manual evaluation is also involved in this evaluation. It provides not only the fidelity score and fluency score but also the error types of the translation. This will help us to identify the advantage for each system and the distribution of error types.

Boosted by the above new features, our analysis and case study first shows that rule-based and statistical systems have very different error distributions, and the distributions also vary with different domains. Second we find a serious discrepancy between the automatic and manual evaluation results of the rule-based MT systems, and a detailed study in this problem reveals that automatic evaluation metrics such as BLEU-SBP [Chiang et al., 2008] and METEOR have some bias against rule-based systems. And we also find some correlations between other automatic evaluation metrics.

The rest of the paper is arranged as follows: in the next section, we give an overall introduction to the CWMT2013 evaluation. Section 3 presents manual evaluation results. Section 4 shows the analysis of correlations between several automatic evaluation metrics. In Section 5, we present a case-study on the mismatch between manual and automatic evaluation results. Finally we draw the conclusion and future work in Section 6.

2 Overall Introduction to CWMT2013 Evaluation

2.1 Evaluation Tracks

There are six tracks in CWMT2013 evaluation, covering 5 different language pairs and 4 domains: news domain for Chinese-to-English direction (CE), news domain for English-to-Chinese direction (EC_n), scientific domains for English-to-Chinese direction (EC_s), and three Chinese minority language tasks including, daily expression domain for Mongolian-to-Chinese (MC), government-doc domain for Tibetan-to-Chinese (TC) and news domain for Uighur-to-Chinese (UC), shown in Table 1.

Task Code	Domain	Language Pair	# of test-set	Pr.	Pc.
CE	News	CH-EN	1	7	10
EC_n	News	EN-CH	2	9	15
EC_s	Scientific	EN-CH	2	7	8
MC	Daily-expression	MO-CH	2	6	6
TC	Goverment-doc	TI-CH	2	6	8
UC	News	UI-CH	2	9	9
Total	4	5	11	44	56

Table 1: Track and system information for CWMT2013 Evaluation Tasks. The last two columns present the number of participanting systems in each task, where Pr. for Primary systems, Pc. for Constrast systems.

2.2 Participants and Systems

There are 16 participants, most of which are institutes and universities such as Chinese Academy of Sciences and Harbin Institute of Technology. Besides, we also have one industrial participant and one foreign participant. 183 translation results of both primary and contrast systems are submitted in this evaluation. The so-called primary system is the main system of each participant in this evaluation and its training data must within the range that the evaluation organizer specified. Contrast system refers to the system that participants use to produce comparative results and its training data is not restricted. We further categorize contrast systems into restricted/non-restricted systems by whether external data is used. Table 1 shows the number of the participants and their systems in each evaluation task.

2.3 Evaluation Data for MT Tracks

The evaluation corpus contains five language directions (Chinese-to-English, English-to-Chinese, Mongolian-to-Chinese, Uighur-to-Chinese, and Tibetan-to-Chinese) and four domains (news, scientific, daily expressions, and government-doc). The input and the output files in the evaluation are encoded in UTF-8 (with BOM) and in strict XML format. All development sets and test sets contain an original text and 4 references. All 4 references are translated from the original text independently by four professional translators. The test-set includes the current test-set of CWMT2013 for ranking and the progress test-set from previous CWMT evaluations to investigate the improvement of each participanting system.

The evaluation data inherit all the data in previous CWMT evaluation [Zhao et al., 2009]. Further more, we add new test sets in 4 tasks (EC_s, MC, TC, UC) and update a number of training corpus in Chinese minority language-to-Chinese

Task Code	Training-set	Dev-set	Progress test-set		Current test-set
			$cwmt_{09}$	$cwmt_{11}$	
CE	5.84M	1,006	1,003	–	–
EC_n	5,84M	1,000	1,002	1,001	–
EC_s	0.9M	1,116	–	1,497	1,000
MC	0.11M	1,000	–	400	1,005
TC	0.12M	650	–	286	1,000
UC	0.11M	700	–	574	1,000

Table 2: Number of sentences in the data-sets for CWMT 2013

tasks. The statistics of evaluation data are shown in Table 2.

2.4 Gray-Box Evaluation

In order to get a deeper understanding of each translation system, we adopt "Gray-box testing" mode for the first time in our evaluation. It requires participants submit not only the final translation files, but also result files of several key intermediate procedures as gray-box files. Specifically as follows:

Gray-box files for statistical machine translation system includes: Source language preprocessing results of the training corpus; Target language preprocessing results of the training corpus; Word alignment results of the training corpus; Translation rule table filtered by the development set and test set; Preprocessing result of monolingual corpus for language model (LM) training; Language model documentation (instructing LM toolkit, commands and parameters used for LM training); Development set preprocessing results; Decoder configuration file; Test set preprocessing results; Decoder output; Final translation results.

Gray-box files for rule-based machine translation system includes: Test set preprocessing results; Decoder output; Final translation results; Translation rules used for translating test set sentences (optional)

After the evaluation, the organizer shares all the gray-box files of primary systems and baseline systems with the participants, so they could identify the weak link in their translation pipeline and make adjustments accordingly.

2.5 Baseline System

This evaluation provides one or more baseline systems for each evaluation task, including source code and corresponding gray-box files. Participants can build their own machine translation systems by optimizing the given baseline system, or they can use

Task code	Systems	Providers
CE/EC_n	Moses	Harbin Institute of Technology
CE/EC_n	Niu-Trans	Northeastern University
EC_s	Moses	Institute of Scientific and Technical Information of China
MC	Moses	Institue of Computing Technology, CAS.
TC	Moses	Xiamen University
UC	Moses	Institute of Automation, CAS.

Table 3: CWMT2013 Evaluation Baseline Systems.

their own systems. The data and translation result provided by baseline system could also be used by participants for research purpose. The evaluations baseline systems are mainly based on two open source systems: Moses [Koehn et al., 2007] and NiuTrans [Xiao et al., 2012]. The corresponding gray-box files are provided by six domestic participants. We show all baseline systems and their providers in Table 3.

2.6 Performance Measurement

In this evaluation we use a variety of automatic evaluation metrics. The main evaluation metric is BLEU-SBP for its decomposability at sentence level. Other automatic evaluation metrics include: BLEU [Papineni et al., 2002], NIST [Doddington, 2002], GTM [Turian et al., 2006], mWER [Nießen et al., 2000], mPER [Gregor Leusch, 2003], ICT (a metric developed by the Institute of Computing Technology, CAS.), METEOR and TER. In Chinese-to-English direction we also introduce Woodpecker Methodology [Bo et al., 2013], since it could utilize rich linguistic knowledge by setting checkpoints in evaluation.

We adopt two new automatic evaluation metrics METEOR and TER based on the following considerations: BLEU metric is based on n-gram precision, without considering the syntax structure, synonyms, and paraphrase. To solve these problems, recently researchers put forward a variety of new evaluation methods. Among them, the automatic evaluation metric METEOR has been widely accepted. It uses stemming match, synonyms match as well as the exact literal match and considers not only precision but also recall. TER is a classic metric in machine translation [Snover et al., 2006], we use it by calculating the minimum editing distance between translation and reference to

Loyalty		Fluency
Score	Criteria	Criteria
0	No translation at all	Completely incomprehensible
1	Only a few individual words are translated.	Only individual phrases or grammatical components are understandable
2	A few phrases or grammatical components are translated	40% of text is translated fluently, a few grammatical components are understandable
3	60% of text is correctly translated, or SVO of the translation is correct	60% of text is translated fluently
4	80% of text is correctly translated	80% of text is translated fluently, or SVO of the text is basically fluent.
5	All text is correctly translated	Translation is fluent.

Table 4: Scoring Criteria for Manual Evaluation

ease the shortcoming of exact literal match.

All metrics (including WoodPecker) are case-sensitive, the evaluation of Chinese is based on Chinese character instead of word. We do significant test [Collins et al., 2005] on the BLEU-SBP results of each primary system. Specifically, for each primary system we test the significant degree of the differences between its translation results and all other primary systems, constructing the significance of difference matrix of all primary systems.

Besides the above automatic evaluation metrics, we carry out manual evaluation on EC_n task and UC task. The manual evaluation data of EC_n task comes from EC_n task in CWMT2011 and manual evaluation data of UC task comes from UC task in CWMT2013. We select 500 sentences from each test set as the manual evaluation corpus.

Manual evaluation focus on the loyalty and fluency of translation results, and these evaluation criteria refer to *the Language Norms Based Assessment Specifications of Machine Translation Systems(draft)* released by State Language Affairs Commission and the Ministry of Education of People's Republic of China. Taking practical operability into account, we made some minor modifications. The scoring criteria are shown in Table 4.

Translation results of each participating system were manually evaluated by three native speakers. Then, we take the arithmetic mean of all loyalty/fluency scores of each system as their final loyalty/fluency evaluation scores. During manual evaluation, in addition to evaluating loyalty and fluency, evaluators also need to give translation results a brief analysis of error types pre-set by evaluation organizer including:

- a:translation and original text have opposite meanings
- b:lack of content words in translation
- c:word order error
- d:named entity problems
- e:quantifier / temporal words problems
- f:word selection error
- o:other errors

2.7 Official Evaluation Results

The official evaluation results are released online[1]. In the following section we will do some meaningful comparison among the participating systems and give a detailed analysis on these results.

3 Analysis on Manual Evaluation Result

3.1 Error Type Analysis

By analyzing the error types of manual evaluation results in EC_n task and UC task, we find out that in EC_n task, The most frequent errors are "f: word selection error in translation", "b: lack of content words in translation", and "c: word order error". This validates the common wisdom that English and Chinese have very different structures resulting in a lot of long-distance-reordering which the current system could't handle. It also reveals that the current system is prone to omit content words, which is mainly caused by alignment errors. In UC tasks, however, the frequency of "c: word order error" is much lower than that of EC_n task, while the frequency of "b: lack of content words in translation" is much higher. This indicates that Uighur and Chinese have a more similar structure, but due to the rich morphology of Uighur, there are more alignment error and quantifier/temporal errors.

We show the distribution of the error types of the two systems in Figure 1 and Figure 2.

3.2 Statistical MT System vs. Rule-based MT System

In recent years, along with the success of the statistical MT system, rule-based MT system has been gradually fading away from the translation community. In this evaluation, we did a detailed comparision between this two kind of systems and the result is shown in Table 5.

[1] http://nlp.ict.ac.cn/Admin/kindeditor/attached/file/20140310/20140310173732_36859.pdf

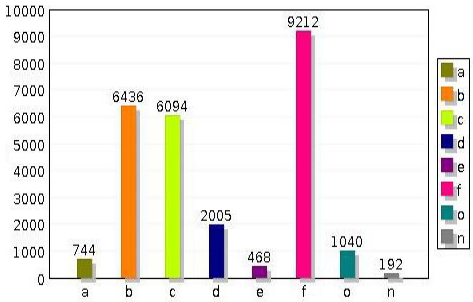

Figure 1: Distribution of Overall Error Type of EC_n Task. n means no error.

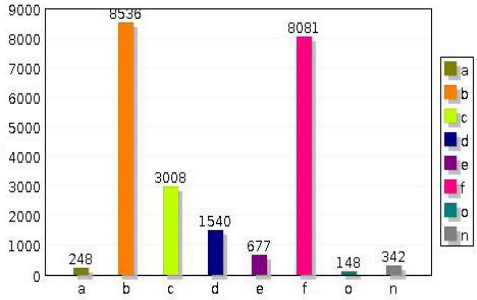

Figure 2: Distribution of Overall Error Type of UC Task.

System	Loyalty	Fluency	BLEU-SBP
RB	**3.27**	**3.00**	0.22
SB	2.93	2.76	0.34
SB+SC	3.10	2.97	**0.35**

Table 5: Manual and Automatic evaluation results of three systems in EC_n task. RB denotes a rule-based system. SB is a statistical system, and SB+SC means statistical system with system combination technology.

The first and second column shows the manual evaluation results, we can see that rule-based system still have some advantage over statistical systems. We further analyze the result and plot the distributions of error types of the RB and SB systems in Figure 3 and Figure 4. We can see that rule-based system has a clear advantage in translating content words, resulting in a more complete translation and a higher manual evaluation score, while statistical system is trained to optimize BLEU score and makes less word selection errors.

Another intersting finding is that the system combination technology for statistical MT system brings a positive impact on both manual and automatic evaluation. The 4th row in Table 5 shows the performance of statistical system with system combination technology. We can see that both manual and automatic evaluation scores get a big boost with 1 BLEU-SBP point and about 0.2 points in Loyalty and Fluency scores.

4 Correlations between Automatic Evaluation Metrics

In this evaluation we use a variety of automatic evaluation metrics to evaluate all the systems and produce a large amount of evaluation scores, which enables us to further study the correlations between those automatic evaluation metrics. Eleven evaluation metrics are involved in most tasks including: 5-gram BLEU-SBP, METEOR, TER, 5-gram BLEU, 6-gram BLEU, 6-gram NIST, 7-gram NIST, GTM, mWER, mPER, and ICT. For each task, we calculate the Spearman Rank Correlation Coefficient (SRCC)

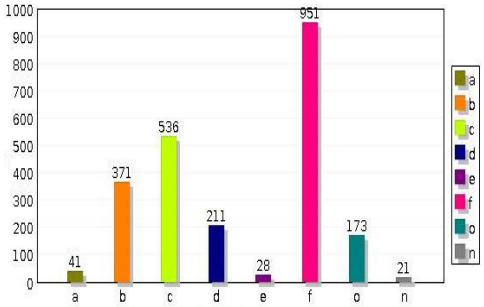

Figure 3: Distribution of One Rule-based MT Systems Error Type of EC_n Task.

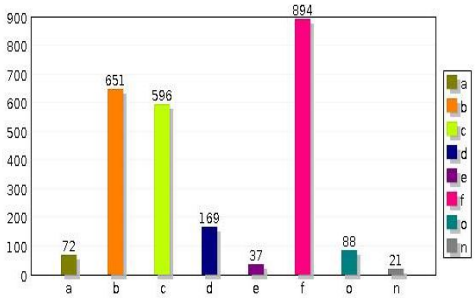

Figure 4: Distribution of One Statistical MT Systems Error Type of EC_nTask.

[Pirie, 1988] between the evaluation scores of two different metrics. The results of EC_n and UC task are shown in Figure 5 and Figure 6. Each node denotes one or more metrics and the distant between them is based on their SRCC score. The orange double arrow connects the metrics with a higher SRCC score and the blue dotted line connects the metrics with relatively lower SRCC scores. Noted that if the SRCC score of two metrics is greater than 0.99, we merge them as one node in the figure. In EC_n task, we can find that:

- Same metrics with different n-gram settings always have the highest correlation with each other, such as 5-gram and 6-gram BLEU, 6-gram and 7-gram NIST.
- BLEU-SBP has a very high correlation with BLEU.
- NIST, GTM, and mPER have a high correlation with each other.
- TER, mWER, and ICT have a low correlation with NIST and GTM.

Most of these findings are in accord with the common wisdom: metrics based on n-gram precision such as BLEU and BLEU-SBP have a high correlation. And metrics mainly based on edit-distance such as TER, mWER and ICT are much similar with each other. It's also interesting to find that METEOR is kind of at the middle ground of all automatic metrics, since it incorporates a wide variety of linguistic knowledges. In UC task, the results is similar with the EC_n tasks except that GTM has the highest correlation with NIST. And TER, mWER and ICT have a low correction with NIST and GTM.

5 Case Study: Automatic Evaluation vs. Manual Evaluation

In our analysis of the correlation between automatic and manual evaluation scores, we find an inconsistent case for rule-based MT system: Unlike statistical MT systems, the rule-based MT system has very different performances in automatic evaluation and manual evaluation. We show the SRCC between automatic evaluation metrics and manual evaluation metric with/without rule-based MT system in Table 6. We can see that rule-based system caused a drastic jump in SRCC, this denotes a obvious conflict between automatic and manual evaluation in rule-based system.

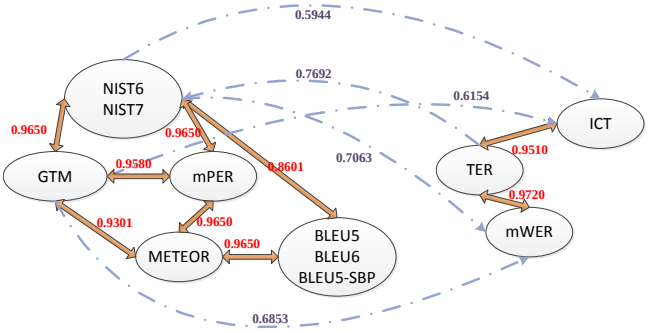

Figure 5: Correlations between Automatic Evaluation Metrics of EC_n task.

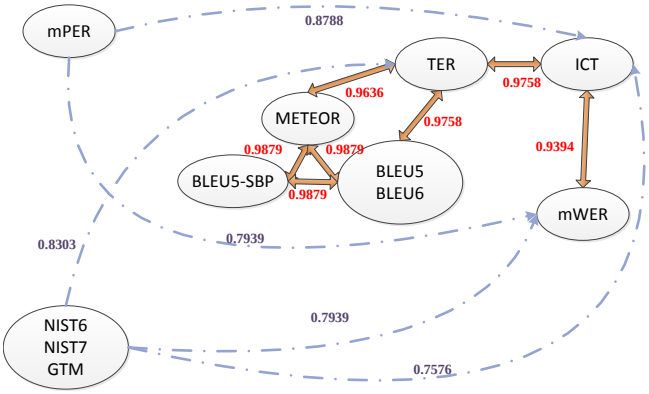

Figure 6: Correlations between Automatic Evaluation Metrics of UC task.

	Loyalty	Fluency
BLEU5-SBP	0.33	0.35
METEOR	0.37	0.33
TER	0.33	0.37

	Loyalty	Fluency
BLEU5-SBP	0.91	0.93
METEOR	0.95	0.91
TER	0.91	0.95

Table 6: SRCC between the Automatic Evaluation scores and Manual Evaluation scores in EC_n Task. The score in the left table is calculated based on results from all participating systems, whereas in the right we exclude the rule-based system to show its significant effect on SRCC.

One possible reason is the translation format of rule-based MT system participated in this evaluation caused this great performance difference: since the output of rule-based MT system sometimes contains optional words in parentheses and multiple choices of words in brackets, shown as "Org" in Figure 7. And this format will affect the n-gram precision in automatic evaluation.

Org：超人(已经)[起动;开始]一挑起一条重大提议他打算在一个举措里放弃他的美国国籍旨在把更(多)全球的[影响;敲击]和威望给他.

Pos：超人已经起动一挑起一条重大提议他打算在一个举措里放弃他的美国国籍旨在把更多全球的影响和威望给他.

Figure 7: Output sample of the rule-based MT system. "Org" denotes the original output of the system. "Pos" denotes the post-processed results.

To exclude the above side-effect, we carry out an additional experiment: we turn the output of rule-based system into standard translation format by removing redundant words, and evaluate the post-processed results (shown as "Pos" in Figure 7). The evaluation results are shown in Figure 8, where S_1 is the original rule-based system and S_1^* is the same system with post-processing. We can see that format problem indeed causes a little drop in automatic evaluation score(about 0.5 points in BLEU-SBP). However, it doesn't change the overall trend that rule-based system has very different performances in automatic and manual evaluation. This suggests that automatic evaluation metrics such as BLEU-SBP and METEOR have some bias against rule-based system which may result in a unilateral evaluation. And this mismatch further indicates that the current automatic evaluation metrics are still not good enough to reflect the real quality of the translation. We need to explore better automatic evaluation metrics which has a better correlation with manual evalution metrics.

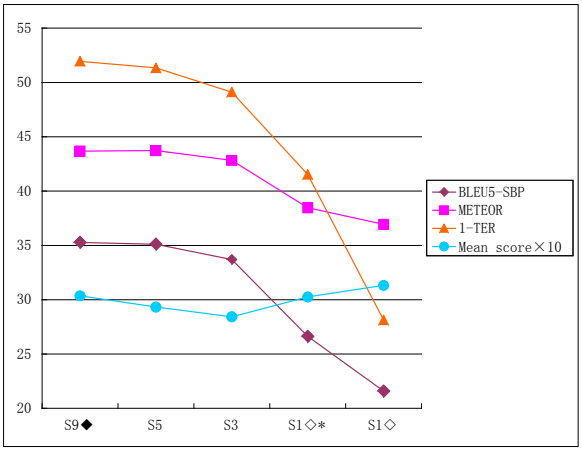

Figure 8: Evaluation Scores of different Systems in CWMT2013 EC_n Task. S_1 = Rule-based system, S_1^* = Rule-based system with Post-processing, S_3,S_5,S_9 are statistical systems.

6 Conclusions and Future Work

In this paper, we gave a detailed description of the CWMT2013 evaluation. Our analysis revealed some interesting correlations between different evaluation metrics. And the case study on rule-based system showed that automatic evaluation metrics such as BLEU-SBP and METEOR have some bias against rule-based system, causing the conflict in automatic and manual evaluation results. In the future evaluation, we will continue to explore better evaluation metrics and add more tasks on Chinese minority languages to promote the research in related fields.

Acknowledgement

We thank the three anonymous reviewers for helpful suggestions. The authors were supported by CAS Action Plan for the Development of Western China (No. KGZD-EW-501) and National Natural Science Foundation of China (Contract 61379086). Liu's work was partially supported by the Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the CNGL at Dublin City University. The views and findings in this paper are those of the authors and are not endorsed by the Chinese governments.

References

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72.

[Bo et al., 2013] Bo, W., Zhou, M., Liu, S., Li, M., and Zhang, D. (2013). Woodpecker: An automatic methodology for machine translation diagnosis with rich linguistic knowledge. *Journal of information science and engineering*.

[Chiang et al., 2008] Chiang, D., DeNeefe, S., Chan, Y. S., and Ng, H. T. (2008). Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 610–619. Association for Computational Linguistics.

[Collins et al., 2005] Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

[Doddington, 2002] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

[Gregor Leusch, 2003] Gregor Leusch, Nicola Ueffing, H. N. (2003). A novel string-to-string distance measure with applications to machine translation evaluation. In *In Proceedings of MT Summit IX, New Orleans, U.S.A.*

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL: Demonstrations*.

[Nießen et al., 2000] Nießen, S., Och, F. J., Leusch, G., Ney, H., et al. (2000). An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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. Association for Computational Linguistics.

[Pirie, 1988] Pirie, W. (1988). Spearman rank correlation coefficient. *Encyclopedia of statistical sciences*.

[Snover et al., 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

[Turian et al., 2006] Turian, J. P., Shea, L., and Melamed, I. D. (2006). Evaluation of machine translation and its evaluation. Technical report, DTIC Document.

[Xiao et al., 2012] Xiao, T., Zhu, J., Zhang, H., and Li, Q. (2012). Niutrans: an open source toolkit for phrase-based and syntax-based machine translation. In *Proceedings of the ACL 2012 System Demonstrations*, pages 19–24. Association for Computational Linguistics.

[Zhao et al., 2009] Zhao, H., Xie, J., Liu, Q., Lü, Y., Zhang, D., and Li, M. (2009). Introduction to china's cwmt2008 machine translation evaluation. *Proceedings of the twelfth Machine Translation Summit, Ottawa, Canada*.