Automatic Evaluation of English-to-Korean and Korean-to-English Neural Machine Translation Systems by Linguistic Test Points

Sung-Kwon Choi, Gyu-Hyeun Choi and Youngkil Kim
Language Intelligence Research Group
Electronics and Telecommunications Research Institute, Daejeon, Korea
{choisk, choko93, kimyk}@etri.re.kr

Abstract

BLEU is the most well-known automatic evaluation technology in assessing the performance of machine translation systems. However, BLEU does not know which parts of the NMT translation results are good or bad. This paper describes the automatic evaluation approach of NMT systems by linguistic test points. This approach allows automatic evaluation of each linguistic test point not shown in BLEU and provides intuitive insight into the strengths and flaws of NMT systems in handling various important linguistic test points. The linguistic test points used for automatic evaluation were 58 and consisted of 630 sentences. We conducted the evaluation of two bidirectional English/Korean NMT systems. BLEUs of English-to-Korean NMT systems were 0.0898 and 0.2081 respectively, and their automatic evaluations by linguistic test points were 58.35% and 77.31%, respectively. BLEUs of Korean-to-English NMT systems were 0.3939 and 0.4512 respectively, and their automatic evaluations by linguistic test points were 33.10% and 40.47%, respectively. This means that the automatic evaluation approach by linguistic test points has similar results as BLEU assessment. According to automatic evaluation by linguistic test points, we know that both English-to-Korean NMT systems and Korean-to-English NMT systems have strengths in polysemy translations, but has flaws in style translations and translations of sentences with complex syntactic structures.

1 Introduction

Currently, the performance of NMT system is rapidly advancing. Its performance proved to be superior to that of SMT (Bojar et. al., 2016). The NMT is also applied to both English-to-Korean and Korean-to-English machine translation systems for commercial service. The most well-known machine translation evaluation technique in assessing the performance of NMT systems is BLEU (Papineni et. al., 2002). It is strength of BLEU that automatic scores for MT output can be provided in cases where there are existing reference translations by calculating similarity between the MT output and the references. Faults of BLEU are that it does not provide insight into the specific nature of problems encountered in the translation output and scores are tied to the particularities of the reference translations (Lommel et. al., 2014). By BLEU, developers and users cannot identify which part of the NMT translation result is vulnerable.

We propose an automatic evaluation approach of neural machine translation systems by linguistic test points. Instead of assigning a general score to an NMT system we conduct an automatic evaluation by each linguistic test point not shown in BLEU. This automatic evaluation approach of this paper can give developers an intuitive insight into the strengths and flaws of NMT systems. Also, the automatic evaluation method by linguistic test points, which is not like BLEU, may provide objective evaluation even without reference sentences.

Section 2 describes existing studies related to the automatic evaluation approach by linguistic test points. Section 3 introduces the design of test set

32nd Pacific Asia Conference on Language, Information and Computation
Hong Kong, 1-3 December 2018

including linguistic test points. In Section 4, we explain the results of automatic evaluation by linguistic test points and analyze the strengths and flaws of two bidirectional English/Korean NMT systems.

2 Related Work

Approaches for evaluating machine translation systems can be divided into automatic and manual assessment. An automatic evaluation is the automatic scoring of a machine translation result by calculating the similarity between the machine translation result and the reference. BLEU is representative of this automatic evaluation method. It has the strength of objectively assessing the results of machine translation. However, BLEU is dependent on reference sentences and cannot point to translation errors in machine translation results. Manual evaluation method is for human translators to assign scores to the results of machine translations according to the evaluation criteria. It has the strength of being able to assess the results of machine translations precisely, but depends on human evaluators and is costly and time-consuming.

In order to identify the strengths and weaknesses of the machine translation system, previous studies have introduced approaches of using linguistic test set for evaluation purposes. The process for constructing a test set including the linguistic test points can be described as following steps:

- Design taxonomy of linguistic test points of test set
- Collect a large amount of bilingual sentences from the web or book collections.
- For each category of test points, extract language expressions of the linguistic test points from the bilingual sentence pairs.
- Determine the references of each linguistic test point in source language.

A representative test set for evaluating machine translation systems was TSNLP (Test Suites for Natural Language Processing) (Balkan et al., 1994). Most of test sets were composed of language pairs in similar language family (Bentivogli et. al., 2016; Isabelle et. al., 2017). Among test sets with the language pairs in heterogeneous language family was Koh (2001), which was related to Korean, and was consisted of structure part and selection part. While these test sets have been manually constructed, Zhou (2008) has introduced how to automatically build the test set by parser and word aligner. Test set for grasping the strengths and faults of NMT systems started in 2016. Bentivogli (2016) used the English-German test set of IWSLT 2015 to compare PBMT with NMT. Isabelle (2017) constructed 108 English-French test sentences and evaluated them according to the Yes/No questions in order to identify strengths and weaknesses of English-to-French NMT systems. Guillou (2016) established a test set to conduct an assessment on the English-to-French machine translation of pronouns ‘it’ and ‘they’ in the DiscoMT 2015 shared task. However, the existing evaluation of machine translation system using test set was focused on three linguistic categories: morpho-syntactic divergences, lexico-syntactic divergences, and syntactic divergences. Unlike previous studies, this paper describes a test set and an evaluation method to evaluate various linguistic phenomena.

3 Construction of Test Set

3.1 Taxonomy of Linguistic Test Points

A linguistic test point is a linguistically motivated unit, which is pre-defined in test set for automatic evaluation. We attempted to collect a variety of linguistic test points that can target at identifying of the strengths and weaknesses of the neural machine translation systems. For this purpose, the linguistic test points related to part-of-speeches, syntactic structures, semantic relations, and target word selection were manually collected from the grammar books. The linguistic test points can be divided into the structure part related to source sentence of source language and the selection part related to target words of target language. They are subdivided into depth of 3. Table 1 shows the taxonomy of linguistic test points.

Currently, there are 58 linguistic test points. Each test point of the structure part consists of 10 sentences, and each test point of the selection part is composed of 20 sentences. As a result, the total number of sentences is 630. In practice, new linguistic test points can be easily added.

Category	Subcategory	Test Points	# of test points	# of sentences
Structure	Part-of-speech	Noun, Adverb, Verb, Conjunction, Symbol, Number, etc.	29	290
	Phrase and Clause	Gerund, Idiomatic expression, etc.	10	100
	Sentence	Negation, Passive, Mood, Ellipsis, Coordination, Tense, etc.	14	140
Selection	Ambiguity	Part-of-speech ambiguity, Structural ambiguity	2	40
	Collocation	Collocation	1	20
	Polysemy	Polysemy	1	20
	Style	Natural translation	1	20
2	7	58		630

Table 1: Taxonomy of Linguistic Test Points

3.2 Structure of Linguistic Test Points

A linguistic test point consists of <*source sentence*> <*source words*> and <*target words*>. The <*source sentence*> refers to sentence of source language that will be translated automatically. The <*source words*> are a check point words of <*source sentence*> for linguistic test points of the automatic evaluation. The <*target words*> indicate the reference of <*source words*>. The <*target words*> are extracted from the digital translation dictionary. The linguistic test point has a variety of structures, such as:

- <*source sentence*> <*source words*> <*target words*>

  The basic structure of a linguistic test point consists of one <*source sentence*>, one <*source words*>, and one <*target words*>.

  Ex.) <In these 3 years, 2.5 billion pencils were sold> <2.5 billion> <25 억>

- <*source sentence*> <*source words*> <*target words*$_1$>…<*target words*$_n$>

  A test point may consist of two or more <*target words*> corresponding to a <*source words*>.

  Ex.) <I could not but get angry> <could not but> <지 않고는 있을 수 없었> <지 않고 있을 수 없었> <지 않을 수 없었>

- <*source sentence*> <*source words*> <#*target words*$_1$>…<#*target words*$_n$>

  The <#*target words*> refers to <*target words*>, which can be regarded as a <*target words*> among the results of NMT.

  Ex.) <I could not but get angry.> <could not but> <지 않고는 있을 수 없었> <지 않고 있을 수 없었> <지 않을 수 없었><#수 밖에 없었>

- <*source sentence*> <*source words*> <*target words*$_{11}$//*target words*$_{12}$> … <*target words*$_{n1}$//*target words*$_{n2}$>

  <*target words*$_{11}$//*target words*$_{12}$> says that <*target words*> corresponding to <*source words*> can be separated into a <*target words*$_{11}$> and a <*target words*$_{12}$>.

  Ex.) <None of the books were interesting> <none of the books> <어느 책도//지 않>, <책들 중 어느 것도//지 않>

- <*source sentence*> <*source words*> <~*target words*>

  <~*target words*> means that <*target words*> corresponding to <*source words*> should not appear in results of NMT. In other words, '~' refers to a logical negation among logical connectives.

  Ex.) <It was September 17> <It> <~그것>

4 Evaluation Methodology

We have established an automatic evaluation program for automatic evaluation of NMT systems by linguistic test points. With the linguistic test point, the automatic evaluation of an MT system is performed with the following steps:

- For each test point, we calculate the number of matched references against the translated sentence of the NMT system.
- The score of a subcategory can be obtained by summing up the scores of all test points of this subcategory. Then the evaluation of an NMT system can be obtained by summing up the scores of all subcategories.
- Finally, scores of system, subcategories, and linguistic test points are all provided to the developers to evaluate the NMT system.

Figure 1 shows the overall construction of the automatic evaluation approach by linguistic test points.

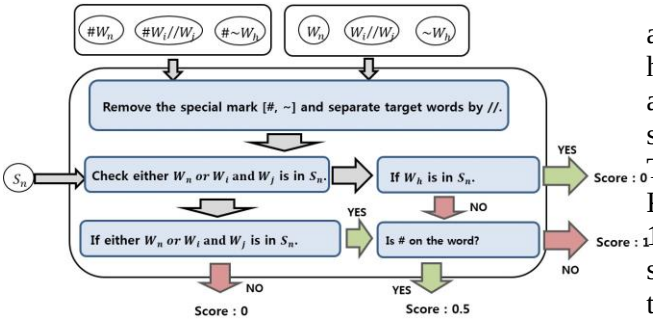

Figure 1: Automatic Evaluation Approach by Linguistic Test Points

In the Figure 1, S_n refers to the automatically translated result of <*source sentence*> and W_n, W_i, W_j *and* W_h mean <*target words*>. The program will determine the existence of a # and then check whether ~ and // are on the <*target words*>. If <*target words*> exist in S_n, give 1 point and otherwise 0 point. If <*#target words*> exists in S_n, give 0.5 point and otherwise 0 point. If <*~target words*> exists in S_n, give 0 point and otherwise 1 point. If <*target words*$_1$*//target words*$_2$> exists in S_n, give 1 point and otherwise 0 point.

5 Experiments

5.1 Automatic Evaluation by Linguistic Test Points of English-to-Korean NMT Systems

Two English-to-Korean NMT systems used in the experiment are well known for their automatic translation commercialization services in Korea. BLEU and the result of automatic evaluation by linguistic test points are as follows:

	# of sentences	NMT1	NMT2	NMT2 – NMT1
Total	630	58.35%	77.31%	16.41
Structure	530	58.21%	74.62%	21.50
Selection	100	58.50%	80.00%	18.96
BLEU		0.0898	0.2081	0.1183
NIST		3.7039	5.1132	1.4093

Table 2: Comparison of BLEU with Automatic Evaluation by Linguistic Test Points in English-to-Korean NMT Systems

According to Table 2, the results of automatic assessment of NMT2 are 16.41 % and 7.37% higher than those of NMT1, respectively. The automatic evaluation by linguistic test points shows the similar results in the BLEU and NIST. That is, BLEU and NIST of NMT2 in English-to-Korean NMT systems were higher 0.1183 and 1.4093 than those of NMT1, respectively. This shows that the automatic evaluation by linguistic test points is similar to the automatic translation metrics. Thus, we can identify the strengths and weaknesses of NMTs by building simple <*target words*> for the automatic evaluation by linguistic test points without having to create a lot of reference sentences. Table 3 indicate the strengths and flaws of NMT1 and NMT2 identified by subcategories:

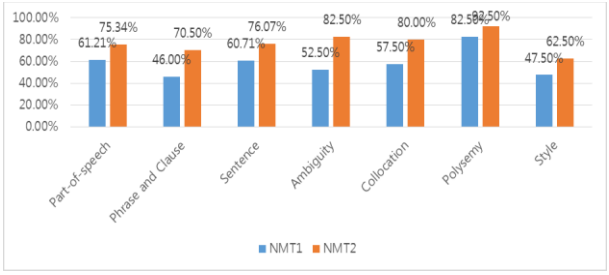

Table 3: Automatic Evaluation by Linguistic Test Points related to Subcategories in English-to-Korean NMT Systems

According to Table 3, the automatic evaluation by linguistic test points of NMT2 is better overall than that of NMT1. The biggest difference between NMT2 and NMT1 was ambiguity. Among ambiguities, structural ambiguity was the biggest difference. The linguistic test points in which NMT1 and NMT2 received low ratings were style, phrase and clause, part–of-speech, and sentence. In the NMT2, machine translation of style was weakest, followed by phrase and clause, part–of-speech, sentence and collocation. In the NMT1, machine translation of style was also weakest, followed by phrase and clause, ambiguity, collocation, sentence, and part–of-speech. Table 4 shows the detailed results of an automatic evaluation by linguistic test points.

Table 4 indicates that NMT1 needs to correct the linguistic test points such as number, numeric expression, participial construction, gerund, subjunctive, structural ambiguity, and natural translation. NMT2 needs to revise the linguistic test points such as symbol, participial construction and parenthesis. The following are examples of the automatic evaluation by linguistic test points about number and symbol.

- Number

<*source sentence*> They had three and a half dollars.
<*source words*> three and a half dollars
<*target words*> 3 달러 50 센트, 3.5 달러
<NMT1>그들에게는 3 달러 반이 있었습니다. // Evaluation: 0 point
<NMT2> 그들은 3 달러 50 센트를 가지고 있었다. // Evaluation: 1 point

- Symbol

<*source sentence*> It is difficult to talk about "typical" Americans.
<*source words*> "typical"
<*target words*> "전형적인", "일반적인"
<NMT1> "전형적인" 미국인에 대해서 이야기하는 것은 어렵습니다 // Evaluation: 1 point
<NMT2> 전형적인 미국인들에 대해 말하는 것은 어렵다. // Evaluation: 0 point

5.2 Automatic Evaluation by Linguistic Test Points of Korean-to-English NMT Systems

Table 5 indicates the result of automatic evaluation by linguistic test points of Korean-to-English NMT systems:

	# of sentences	NMT1	NMT2	NMT2 – NMT1
Total	630	33.10%	40.47%	7.37
Structure	530	33.21%	40.94%	7.73
S election	100	33.00%	40.00%	7
BLEU		0.3939	0.4512	0.0573
NIST		7.1893	7.7504	0.5611

Table 5: Comparison of BLEU with Automatic Evaluation by Linguistic Test Points in Korean-to-English NMT Systems

BLEU and NIST of NMT2 in Korean-to-English NMT systems are higher 0.0573 and 0.5611 than those of NMT1, respectively. As the evaluation in English-to-Korean NMT systems, the automatic evaluation by linguistic test points in Korean-to-English NMT shows the similar evaluation conducted by the automatic translation metrics. Table 6 indicate the strengths and flaws of NMT1 and NMT2 identified by subcategories:

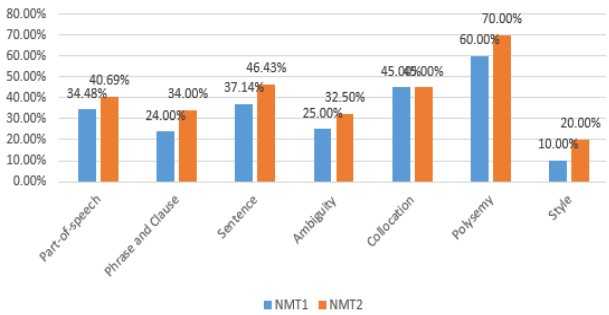

Table 6: Automatic Evaluation by Linguistic Test Points related to Subcategories in Korean-to-English NMT Systems

According to Table 6, the Korean-to-English NMT systems need to correct first the subcategories such as style, ambiguity and

phrase/clause. Table 7 shows the detailed results of an automatic evaluation by linguistic test points.

Table 7 indicates that NMT1 needs to correct the linguistic test points such as article, participial construction, parenthesis, structural ambiguity and natural translation. NMT2 needs to deal with the linguistic test points such as article, relative pronoun, participial construction, subjunctive, parenthesis, structural ambiguity and natural translation. Some examples of automatic evaluation about number and natural translation in Korean-to-English NMT Systems are as follows.

- Number

<*source sentence*> 집 계약금으로 1,000 만원을 지불했어요. // I paid ten million won down on the house.
<*source words*> 1,000 만원
<*target words*> 10 million won, ten million won
<NMT1> I paid 10 million won for the house deposit. // Evaluation: 1 point
<NMT2> I paid ten million won down on the house. // Evaluation: 1 point

- Natural translation

<*source sentence*> 그 사고는 그의 부주의로 일어났다. // His carelessness caused the accident.
<*source words*> 부주의로 일어났
<*target words*> carelessness caused
<NMT1> The accident happened due to his carelessness. // Evaluation: 0 point
<NMT2> The accident was brought about by his carelessness. // Evaluation: 0 point

6 Conclusion

In this paper, we introduced the automatic evaluation approach by linguistic test points for NMT systems. In the automatic evaluation by linguistic test points, the scoring method was as follows:

1) If a <*source sentence*> of a linguistic test point is automatically translated by an NMT system and <*target words*> of a linguistic test point are in the translation results of the NMT system, give 1 point.
2) If a <*source sentence*> of a linguistic test point is automatically translated by an NMT system and <*#target words*> of a linguistic test point are in the translation results of the NMT system, give 0.5 point.
3) If a <*source sentence*> of a linguistic test point is automatically translated by an NMT system and <*target words*> of a linguistic test point are not in the translation results of the NMT system, give 0 point.

In accordance with the automatic evaluation by linguistic test points, the strengths and weaknesses of two bidirectional English/Korean NMT systems were identified. According to the automatic evaluation by linguistic test points, the English-to-Korean NMT systems need to correct first the subcategories with linguistic test points such as style, phrase and clause, part–of-speech, and sentence. The Korean-to-English NMT systems need to deal with first the subcategories with linguistic test points such as style, ambiguity and phrase/clause.

According to the experiments, we could identify the strengths and weaknesses of NMT systems from the automatic evaluation by linguistic test points without having to create a lot of reference sentences.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement)

References

Balkan, L., Netter, K., Arnold, D. and Meijer, S. 1994. TSNLP - test suites for natural language processing, Proceedings of the Language Engineering Convention, ELSNET, Paris: 17-22.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. 2016. Neural versus phrase-based machine translation quality: a case study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas: 257–267.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. 2016. Findings of the 2016 Conference on Machine Translation. ACL 2016 First Conference on Machine Translation (WMT16). The Association for Computational Linguistics: 131–198.

Guillou, L. and Hardmeier, C. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016): 636-643.

Isabelle, P., Cherry, C., and Foster, G. 2017. A Challenge Set Approach to Evaluating Machine Translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages: 2486–2496

Koh, S. R., Maeng, J. N., Lee, J. Y., Chae, Y. S., and Choi, K. S. 2001. A Test Suite for Evaluation of English-to-Korean Machine Translation Systems. MT Summit:

Lommel, A., Burchardt, A., Popovic, M., Harris, K., Avramidis, E., and Uszkoreit, H. 2014. Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data. Proceedings of the 17th Annual Conference of the European Association for Machine Translation: 165-172.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA: 311–318.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao T. 2008. Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points. Proceedings of the 22nd International Conference on Computational Linguistics: 1121-1128

Linguistic Test Points	**NMT1**	**NMT2**
Article	60.00%	60.00%
Noun	82.50%	82.50%
Pronoun	53.33%	75.00%
Adjective	50.00%	75.00%
Adverb	57.50%	67.50%
Preposition	63.33%	68.33%
Verb	64.29%	74.29%
Relative pronoun	80.00%	80.00%
Conjunction	92.50%	80.00%
Symbol	55.00%	45.00%
Number	28.33%	95.00%
Infinitive phrase	57.50%	75.00%
Participial construction	36.67%	55.00%
Gerund	45.00%	70.00%
Idiomatic expression	50.00%	85.00%
Numeric expression	25.00%	85.00%
Type	60.00%	77.50%
Negation	90.00%	95.00%
Special	45.00%	80.00%
Comparative	50.00%	75.00%
Subjunctive	40.00%	85.00%
Passive	65.00%	60.00%
Mood	55.00%	60.00%
Parenthesis	55.00%	55.00%
Ellipsis	65.00%	75.00%
Coordination	70.00%	90.00%
Tense	65.00%	78.33%
Part-of-speech ambiguity	62.50%	95.00%
Structural ambiguity	42.50%	70.00%
Collocation	57.50%	80.00%
Polysemy	82.50%	92.50%
Natural translation	47.50%	62.50%

Table 4: Detailed Automatic Evaluation by Linguistic Test Points in English-to-Korean NMT Systems

Linguistic Test Points	NMT1	NMT2
Article	10.00%	20.0%
Noun	40.00%	35.0%
Pronoun	33.33%	43.3%
Adjective	20.00%	36.7%
Adverb	35.00%	35.0%
Preposition	20.00%	33.3%
Verb	48.57%	51.4%
Relative pronoun	20.00%	20.0%
Conjunction	50.00%	40.0%
Symbol	40.00%	40.0%
Number	33.33%	53.3%
Infinitive phrase	30.00%	45.0%
Participial construction	6.67%	20.0%
Gerund	40.00%	40.0%
Idiomatic expression	20.00%	30.0%
Numeric expression	40.00%	30.0%
Type	60.00%	65.0%
Negation	30.00%	50.0%
Special	60.00%	50.0%
Comparative	50.00%	60.0%
Subjunctive	20.00%	10.0%
Passive	20.00%	30.0%
Mood	40.00%	80.0%
Parenthesis	0.00%	10.0%
Ellipsis	30.00%	40.0%
Coordination	20.00%	30.0%
Tense	43.33%	53.3%
Part-of-speech ambiguity	40.00%	50.0%
Structural ambiguity	10.00%	15.0%
Collocation	45.00%	45.0%
Polysemy	60.00%	70.0%
Natural translation	10.00%	20.0%

Table 7: Detailed Automatic Evaluation by Linguistic Test Points in Korean-to-English NMT Systems