Bleu Score Study

Anonymous Author(s)

ABSTRACT

Machine translation (MT) is a fast growing sub-field of computational linguistic. Until now, the most popular automatic metrics to measure the quality of MT is Bleu score. Lately, MT along with its Bleu metric has been applied to many Software Engineering(SE) tasks. In this paper, we studied Bleu score to validate its suitability for software engineering tasks. We showed that Bleu score does not reflect translation quality due to its weak relation with semantic meaning of the translated source codes. Specifically, an increase in Bleu score does not guarantee an improved in translation quality, and a good translation may have fluctuated Bleu score. [1]

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:

Anonymous Author(s). 2018. Bleu Score Study. In *Proceedings of The 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2018).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine Translation (MT) is the use of computer program to translate text or speech from one language to another. Bleu score evaluates the quality of MT by calculating the modified n-grams precision and also taking into account the length difference penalty. Traditionally, MT is only applied to natural language, but now it is also used for technical and programming language. One notable use of MT for SE tasks is Code Migration. Even with that adaptation, SE community still relies on Blue to evaluate the quality of MT. It is well known that there is a significant difference between natural language and programing language: programing language has structure, and well-defined syntax. This leads to a question as whether Blue score is suitable for SE task (Code Migration) or not. If it is, we could continue to use it. Otherwise, we need another metric that is more suitable for programing language. Hence, the answer to the question above will help researchers and developers build and evaluate MT-based Code Migration system better. Some has attempted to answer the question by stating informal arguments toward the use of Bleu for SE task []. However, up to date, there has not been any empirical evidences to formally address the problem.

[1] More abstract

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ESEC/FSE 2018, 4-9 November, 2018, Lake Buena Vista, Florida, United States

© 2018 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Bleu measures the lexical difference between machine generated code and referenced one. On the other hand, to measure the semantic similarity between them is the ultimate goal when evaluating quality of Code Migration system.

Bleu was proved to be correlated with human judgments in natural language MT systems [**?**]. However, Callison at el argued that we should not over-rely on Bleu score as an improvement in Bleu score is not sufficient nor necessary to show an in improvement in translation quality [1]. To validate the use of Bleu on SE tasks, we set up an experiment to manually judge the result of multiple MT systems and compare its to the Bleu score. Our result showed that Bleu score has weak correlation to human judgments across

2 BACKGROUND

2.1 Machine Translation and Code Migration

2.2 Metrics

Bleu (bilingual evaluation understudy) uses the modified form of n-grams precision and length difference penalty to evaluate the quality of text generated by MT compared to referenced one.

3 RESEARCH QUESTIONS AND HYPOTHESIS

3.1 RQ1

Does bleu score reflect semantic meaning of translated source code?

3.2 RQ2

If the answer to RQ1 is 'no', is Bleu correlated to Lexical representation of code?

3.3 RQ3

If the answer to RQ1 is 'no', is Bleu correlated to Syntaxtical representation of code?

3.4 Our hypothesis

Our hypothesis is that bleu score does not measure well the closeness in term of semantics between the reference and translated source code.

4 METHODOLOGY

4.1 Proof of Hypothesis

4.2 Data Collection

4.3 Settings and Metrics

5 EVALUATION

Since Bleu score is not suitable for SE task (code migration), we propose a new metric RUBY to evaluate quality of machine translation.

6 PROPOSAL

7 RELATED WORKS

8 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

REFERENCES

[1] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *In EACL*. 249–256.