iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems

Nitin Madnani
Text, Language and Computation
Educational Testing Service
Princeton NJ 08542
nmadnani@ets.org

***Abstract*—Machine Translation (MT) systems are evaluated and debugged using the BLEU automated metric. However, the current community implementation of BLEU is not ideal for MT system developers and researchers since it only produces textual information. I present a novel tool called iBLEU that organizes BLEU scoring information in a visual and easy-to-understand manner, making it easier for MT system developers & researchers to quickly locate documents and sentences on which their system performs poorly. It also allows comparing translations from two different MT systems. Furthermore, one can also choose to compare to the publicly available MT systems, e.g., Google Translate and Bing Translator, with a single click. It can run on all major platforms and requires no setup whatsoever.**

I. Introduction

Machine translation (MT) is the process of automatic translation from one natural language into another using computers. The current trend in MT is towards statistical machine translation (SMT) systems that apply a learning algorithm to a large body of previously translated text, known as a *parallel corpus* or a *bitext*. It is assumed that the learner can generalize from these already translated examples and *learn* how to translate unseen sentences. SMT systems were pioneered two decades ago by IBM researchers and used statistical models that translated a single source language word at a time [1]. Today, SMT systems have advanced the state-of-the art significantly and employ more complex models, e.g., *phrase-based* models that can translate entire phrases—contiguous sequences of words—together rather than each individual word [2], [3] and *syntax-based models* that use synchronous context free grammars to model the translation as a parsing process [4], [5].

The most significant reason for the rapid advances made in SMT has been the development and use of automated metrics for evaluation of translation quality. The goal of any such metric is to assess whether the translation *hypothesis* produced by a system is semantically equivalent to the sentence that was translated. However, the cross-lingual nature of this goal makes it quite challenging. Therefore, most MT metrics try to measure whether the hypothesis is semantically equivalent to a human-authored *reference* translation instead. Using such metrics can quickly provide an assessment of system performance both for SMT system development (internal to a research group) and for comparison of multiple SMT systems on a shared translation task [6]. The most common MT metric currently used for these purposes is the BLEU metric which we describe next.

II. The BLEU Metric

The BLEU metric measures the n-gram (n=1 to 4) precisions of the hypothesis against the reference. For example, if we consider the sentence *the cat sat on the mat* to be the hypothesis and *the cat stood on the mat* to be the reference: the unigram precision will be 5/6 (*the*, *cat*, *on*, *the*, *mat*), the bigram precision 3/5 (*the cat*, *on the*, *the mat*), the trigram precision 1/4 (*on the mat*) and the 4-gram precision 0. BLEU also incorporates other ideas such as (a) counting any word in the hypothesis no more than it occurs in the reference (to avoid unfair credit to nonsensical hypotheses like *the the the the the*) (b) a brevity penalty to prevent extremely short hypotheses from getting high precision scores and (c) smoothing precisions to ensure that no sentence gets a zero score. The final BLEU score is the product of the geometric mean of the *cumulative* n-gram precisions and the brevity penalty. Scores can be computed at the sentence level, at the document level, and at the system level across all documents.

Measuring n-gram precisions might appear to be a poor proxy for determining semantic equivalence. However, with *multiple* reference translations, BLEU has been shown to have reasonable correlations with human judgments of semantic equivalence. More details can be found in [7]. While other, more informative, MT metrics have been proposed, BLEU remains the most popular metric in the community.

III. iBLEU

A. Motivation

When debugging MT systems, one would like to:

(a) Examine document and sentence level BLEU scores, in addition to the system level score, so as to locate documents & sentences that the system performs poorly on. Ideally, these scores should be presented visually. Furthermore, once a specific document or a sentence has been located, it should be easy to

examine the respective hypothesis and reference to determine the exact cause of poor performance.

(b) Compare two different MT systems (or two different versions of the same system) to find documents & sentences with the highest differences.

(c) Compare the translation produced by a system (or two systems) for any of the sentences in the data set to a high-performing publicly available SMT system, e.g., Google Translate or Bing Translator.

However, the most common implementation of BLEU is a command line script which does not provide these facilities.[1]. The best one can do with it is to output system and document level scores to a text file along with the system score. However, the non-trivial task of using these scores to construct the ideal debugging environment then falls on the MT system developer or researcher.

iBLEU was designed from the ground up for the purpose of visualizing and debugging SMT system output. The two primary considerations were that (a) to present the BLEU score information in an intuitively visual manner and (b) to be able to download and run it on any major platform without any setup. To this end, iBLEU is implemented using the latest web technologies (HTML5, CSS3 and JavaScript 1.3) and can run on any compliant browser, particularly the free and open-source Mozilla Firefox browser (v4.0 or higher). iBLEU can run entirely in offline mode and an internet connection is only necessary when using Google or Bing for comparison. iBLEU uses the same input format as the NIST MTEval script so no additional work is necessary to pre-process the input files.

B. Video Demo & Code

A video demo highlighting all functions of iBLEU is available at `http://bit.ly/ibleu-demo`. The core of iBLEU has been written in JavaScript 1.3 and is about 3-4 times faster than the NIST MTEval script (written in perl) on the same data sets. The entire code for iBLEU has been released under the MIT license.[2] Of particular interest might be `bleu.js`, an implementation of the BLEU metric in JavaScript that could prove useful in other web-based MT projects. The website also contains a detailed FAQ that explains all of the design choices and possible future work on improving iBLEU.

IV. Related Work

I am not aware of any related work that deals directly with visualizing document and system level BLEU scores in an intuitive manner. However, there has been work on visualizing data structures internal to specific types of SMT systems [8], [9]. There has also been work on interactive machine translation, i.e., allowing human translators to collaboratively create new translations starting with SMT system translations [10].

V. Summary

I presented a novel tool called iBLEU that provides BLEU metric scoring information in a visual and easy-to-understand manner, making it significantly easier for SMT system developers and researchers to quickly locate documents and sentences on which the system is underperforming. iBLEU also allows comparing translations from two different MT systems. It also allows the SMT system(s) under examination to be compared to the publicly available SMT systems, such as Google Translate and Bing Translator, with just a single click. It is significantly faster than the implementation that is currently used by the community and provides much more information in a more organized fashion. It requires absolutely no setup and can run on every major platform.

[1] MTEval script v13a, `www.itl.nist.gov/iad/mig/tools/`

[2] http://ibleu.googlecode.com

References

[1] A. L. Berger, P. F. Brown, S. D. Pietra, V. J. D. Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, and L. Ures, "The Candide System for Machine Translation," in *Proceedings of HLT*, 1994.

[2] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proceedings of HLT-NAACL*, 2003.

[3] D. Marcu and W. Wong, "A Phrase-Based, Joint Probability Model for Statistical Machine Translation," in *Proceedings of EMNLP*, 2002.

[4] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[5] A. Zollmann and A. Venugopal, "Syntax Augmented Machine Translation via Chart Parsing," in *Proceedings of the Workshop on Statistical Machine Translation*, 2006, pp. 138–141.

[6] NIST, "NIST Open Machine Translation (MT) Evaluation," Information Access Division, 2008, http://www.nist.gov/speech/tests/mt/.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceeding of ACL*, 2002.

[8] S. DeNeefe, K. Knight, and H. H. Chan, "Interactively Exploring a Machine Translation Model," in *Proceedings of ACL (Demos)*, 2005, pp. 97–100.

[9] J. Weese and C. Callison-Burch, "Visualizing Data Structures in Parsing-based Machine Translation," *Prague Bulletin of Mathematical Linguistics*, vol. 93, pp. 127–136, 2010.

[10] J. Albrecht, R. Hwa, and G. E. Marai, "Correcting Automatic Translations through Collaborations between MT and Monolingual Target-language Users," in *Proceedings of EACL*, 2009, pp. 60–68.