BLEU Is Not the Colour: How Optimising BLEU Reduces Translation Quality

Aaron Smith, Christian Hardmeier, Jörg Tiedemann

Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
aaronsmith540@gmail.com

Abstract

In this paper we present experiments in statistical machine translation (SMT) where the BLEU score is optimised with the aim of providing high-quality translations to use as data points in parameter tuning. This is achieved by running the document-level SMT decoder Docent in BLEU-decoding mode, where proposed changes to the translation of a document are checked against a reference, and only accepted if they increase BLEU. The expected increase in translation quality is however not forthcoming. Our results suggest that BLEU is not always a reliable metric; conversely, directly optimising for BLEU can seriously damage translation quality.

1. Introduction

This paper presents and discusses results from experiments with the document-level SMT decoder Docent in BLEU-decoding mode. In this mode, the weights of all standard feature functions are set to zero, and only changes to the translation that increase the BLEU score, calculated by comparison with a reference translation, are accepted. The initial aim in running these experiments was to find good translations to use as data points in parameter tuning for Docent. Our results, however, show that by optimising for BLEU, translation quality actually goes down. This throws into question the paradigm of optimising weights against BLEU, and moreover of using BLEU as an objective measure in the assessment of translation quality.

1.1 BLEU

The BLEU metric (Papineni et al., 2002) is widely used to evaluate the quality of statistical machine translation. It works by calculating the geometric mean of the precision p_n of n-grams, where normally $1 \leq n \leq 4$, by comparing the proposed translation to one or more reference translations. To ensure that short sentences with high precision cannot cheat the system, a brevity penalty is introduced, depending on the lengths of the reference translation r and the candidate translation c:

$$\text{BLEU} = \min\left(\exp\left(1 - r/c\right), 1\right) \cdot \exp\left(\sum_{n=1}^{N} \frac{\log p_n}{N}\right) \quad (1)$$

An obvious flaw is that BLEU gives equal weighting to all words: the incorrect translation of a pronoun, for example, is penalised exactly the same as a noun that appears in slightly the wrong form, although the effect on understanding may be much greater in one case than the other. BLEU also harshly punishes synonyms and elaborations, as well as words such as 'thus' or 'however' spliced occasionally into a text. Despite these and other issues, BLEU has been shown to correlate extremely well with human judgement of translation quality in certain cases (Papineni et al., 2002). There have been a lot of recent efforts to develop more sophisticated metrics that counteract some of BLEU's weaknesses (Macháček and Bojar, 2014), but for the time being it remains ubiquitous in SMT.

1.2 Docent

Our document-level SMT decoder Docent (Hardmeier et al., 2013) implements a search procedure based on local search. At any stage of the search process, the document state consists of a complete translation, making it easy for feature models to access the complete document with its current translation at any point in time. The search algorithm is a stochastic variant of standard hill climbing. In the experiments presented here, we use the hill climbing decoder directly to optimise the BLEU score of the output measured against a reference translation. At each step, the decoder generates a successor of the current search state (i.e. the current translation) by randomly applying one of a set of state-changing operations at a random location in the document, and accepts the new state only if it has a better score than the previous state. Implemented operations include changing the translation of a phrase, changing the word order by swapping the positions of two phrases or moving a sequence of phrases, and resegmenting phrases. The initial translation can be either generated randomly or based on a run from Moses.

2. Experiments

We trained a German-English Moses translation model on just over 1.5 million sentences from Europarl v7, downloaded from www.statmt.org/wmt13. A 5-gram English language model (LM) was trained on just over 2.2 million sentences from the same source, while feature weights were tuned on a set of 2525 sentences from the newstest2009 data using the Mert algorithm (Och, 2003). The test data was a set of 3052 sentences from the newstest2013 data. This data set contains document markup, enabling us to feed 52 separate documents (an average of 59 sentences per document) to the Docent decoder.

For each document, the initial translation consisted of the output from running Moses. Docent was then run in BLEU-decoding mode: only changes to the translation that

increased BLEU were accepted. We allowed the decoder to run for 1,000,000 iterations for each document.

The mean BLEU score after Moses decoding, before running Docent, was 19.4; after running Docent in BLEU-decoding mode, it had increased to 47.8. We were able to track the feature scores for standard SMT features as the BLEU scores increased. Fig. 1, for example, shows the results for the LM; each line represents a single document. When we start the decoder with a translation obtained from

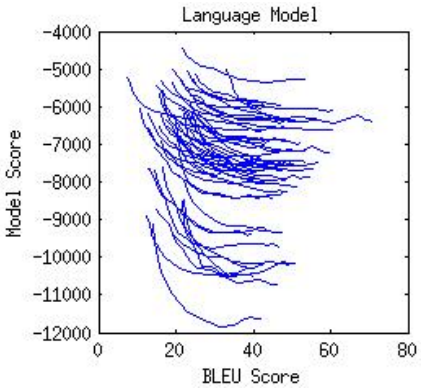

Figure 1: *Language model scores v. BLEU scores for 52 documents run first through the Moses decoder and then Docent in BLEU-decoding mode. See text for details.*

running Moses, the LM scores decrease as BLEU goes up. This shows that Docent in BLEU-decoding mode is able to find translations with high BLEU scores that score poorly on the LM feature. A natural assumption here might be that the problem lies in the LM: perhaps we are finding high quality translations with lower LM scores than those produced by Moses. But this turns out not to be the case. Looking at the Docent output it becomes clear that high BLEU scores have been achieved *in spite of* worse translation quality. Let us take a demonstrative example, considering the source (SRC), reference (REF), Moses translation (MOS), and BLEU-optimised translation (BLU):

SRC: *am wichtigsten ist es aber , mit seinem arzt zu sprechen , um zu bestimmen , ob er durchgeführt werden sollte oder nicht .*

REF: *but the important thing is to have a discussion with your doctor to determine whether or not to take it .*

MOS: *the most* ***important thing is*** *, however , with his* ***doctor to*** *speak , in order* ***to determine whether*** *it should be carried out* ***or not*** *.*

BLU: *with* ***but the important thing is to*** *its a* ***doctor to*** *mention* ***to determine whether or not to*** *be implemented* ***it*** ***.*** *to*

We see here that Moses translation, while not perfect, carries at least most of the meaning across from the original sentence. The BLEU-optimised translation, meanwhile, is unintelligible. The fragments in bold show n-grams for $n \geq 2$ where the Moses and BLEU translations match the reference. It is telling that there are no 4-gram matches at all in the Moses translation, while the long matching fragments in the BLEU translation ensure that there are as many as six such matches. The BLEU translation also has a higher unigram precision; indeed, for $1 \leq n \leq 4$, the number of matching n-grams is much higher in the BLEU translation than the Moses translation.

Another example exposes further the internal workings when we optimise solely towards BLEU:

SRC: *es ist auch ein risikofaktor für mehrere andere krebsarten .*

REF: *it is also a risk factor for a number of others .*

MOS: *there* ***is also a risk factor for a number of*** *other types of cancer .*

BLU: ***it is also a risk factor for a number of others .*** *cancers*

In this example the Moses translation is actually very good; a more literal translation of the source sentence than the reference. After BLEU-decoding, however, the sentence has been transformed: it now matches the whole of the reference identically, but with the word *cancers* added after the full-stop. It is straightforward to see why the BLEU translation has a higher BLEU score: the extra couple of tokens at the end of the matching fragment increase the precision for all n-grams. These examples are typical of BLEU-decoding: sometimes, but not always, fragments of sense; but mostly nonsense.

3. Discussion and Future Work

The results presented here suggest that by letting BLEU run wild, we move far away from the part of the search space containing good translations. It is perhaps not surprising that the LM scores decrease: Moses of course works directly to optimise these. Indeed, we see the same trend for several other features, including the standard translation models. The surprise was rather that the translation quality takes such a hit: while the average BLEU score jumps massively from 19.4 to 47.8, translation quality clearly goes down. These results have been confirmed on other data sets with different text types and language pairs.

It is worth reiterating that these experiments required reference translations to calculate the BLEU score for proposed changes to the translation. It is not yet clear exactly what the general implications are for SMT. Is it possible to tune a system that gets high BLEU scores with low translation quality using SMT features that do not rely on the presence of a reference? And what are the implications for tuning at the document level? One further experiment that would be interesting to see is what kind of results are produced when we *combine* BLEU decoding with the standard features. Will the decoder then be able to produce translations with high BLEU scores *and* high quality? We know that the reference translation itself fulfills both these criteria: its BLEU score is of course 1. The reference translation will not in most cases be reachable by the decoder; one might assume however that BLEU-decoding would bring us 'near' to the reference translation, in some linguistically meaningful sense of the word *near*. The results presented here show that this is not true; a BLEU score more similar to that of the reference (i.e. a higher BLEU score) does not necessarily imply a translation closer to the reference in the sense that a human would judge the translation to be of better quality. It remains to be seen if we need to use auxiliary scoring systems, such as those proposed in Macháček and Bojar (2014), to counteract the weaknesses of BLEU and guarantee high quality translations.

References

C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria.

M. Macháček and O. Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland USA.

F. J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg (Pennsylvania, USA).

K. Papineni, S.Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceeding of the 40th Annual Meeting of the Association for Computationial Linguistics*, pages 311–318, Philadelphia (Pennsylvania, USA).