Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 269–281

Climbing Mount BLEU: The Strange World of Reachable High-BLEU Translations

Aaron SMITH[1,2], Christian HARDMEIER[1], Jörg TIEDEMANN[3]

[1] Uppsala University
[2] Convertus AB, Uppsala, Sweden
[3] University of Helsinki

aaron.smith@convertus.se, christian.hardmeier@lingfil.uu.se, jorg.tiedemann@helsinki.fi

Abstract. We present a method for finding oracle BLEU translations in phrase-based statistical machine translation using exact document-level scores. Experiments are presented where the BLEU score of a candidate translation is directly optimised in order to examine the properties of reachable translations with very high BLEU scores. This is achieved by running the document-level decoder Docent in BLEU-decoding mode, where proposed changes to the translation of a document are only accepted if they increase BLEU. The results confirm that the reference translation cannot in most cases be reached by the decoder, which is limited by the set of phrases in the phrase table, and demonstrate that high-BLEU translations are often of poor quality.

Keywords: Statistical machine translation, oracle decoding, BLEU, Docent

1 Introduction

This paper presents a method for finding oracle translations in phrase-based (PB) statistical machine translation (SMT) using exact document-level BLEU scores. The method, which we call BLEU decoding, is implemented in the document-level machine translation decoder Docent. BLEU decoding is a stochastic hill climbing algorithm: changes are proposed by the decoder to an initial translation and only accepted if they increase BLEU.

Analysing the translations obtained in this way we corroborate previous research on the problem of reference reachability: perfect BLEU scores, corresponding to the decoder finding the reference translation exactly, are rarely possible; meanwhile we add to the extensive literature on problems and biases with the BLEU metric itself, showing for the first time clear examples of sentences from documents with high BLEU scores with obvious poor translation quality.

The paper is structured in the following manner: Section 2 describes the BLEU metric, Section 3 presents the Docent decoder and BLEU decoding, Section 4 details

experiments carried out with BLEU decoding and presents their results, while Section 5 comprises a discussion.

2 BLEU

The BLEU score, introduced by Papineni et al. (2002), is a metric for evaluating the quality of a candidate translation by comparing it to one or more reference translations. For $1 \leq n \leq N$, where normally $N = 4$, each n-gram in each candidate sentence is checked against all of the references in order to calculate precision. To count towards precision, the candidate n-gram need only appear in one of the references; this helps to account for possible variations in style and word choice. However, the same n-gram appearing more than once in the candidate is only counted multiple times if it also appears multiple times in a single reference. BLEU is then based on the geometric average of these so-called modified n-gram precisions p_n.

As multiple references are employed in calculating BLEU, it is difficult to take recall into account, which could lead to short sentences scoring unfairly highly. To prevent this from occurring, a brevity penalty is introduced, lowering the BLEU score for cases where the length of the candidate translation c is less than the length of the reference translation r. The equation for BLEU is as follows:

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

Obvious problems with BLEU are that it gives all words equal weighting and harshly punishes synonyms and elaborations, as well as words such as ‘thus’ or ‘however’ spliced occasionally into a text (see Callison-Burch et al. (2006) for a full discussion of these shortcomings). Chiang et al. (2008) meanwhile describe several situations where they are able to obtain highly dubious improvements in BLEU. They point out, for example, that if translating multiple genres at the same time, one can generate longer sentences within a specific genre where the translation quality is known to be higher, and shorter sentences in other more difficult genres. This will generate higher overall BLEU scores due to the fact that the brevity penalty works on whole documents rather than sentence-by-sentence, but the final translation quality would clearly have been higher if combined systems had been used, each optimised for a particular genre.

Despite these and other issues, however, BLEU has been shown to correlate extremely well with human judgement of translation quality in many cases (Agarwal and Lavie, 2008; Farrús et al., 2012). There have been a lot of recent efforts to develop more sophisticated metrics that counteract some of BLEU’s weaknesses (Macháček and Bojar, 2013), but for the time being it remains ubiquitous in SMT. For this reason, the computation of oracle BLEU hypotheses is an active field (Wisniewski et al., 2010; Sokolov et al., 2012). Oracle BLEU hypotheses are those in the search space of a PBSMT decoder with the highest BLEU scores. Ultimately we want our translation systems to find these hypotheses on unseen data; calculating them when a reference is available can help identify deficiencies in current systems and facilitate the development of new techniques. BLEU oracles are also useful during feature-weight tuning,

though it has been pointed out that relying too heavily on BLEU here can lead to poor results (Liang et al., 2006; Chiang, 2012).

3 Docent

Docent is a decoder for phrase-based SMT (Hardmeier et al., 2013). In Docent's search algorithm, feature models have access to a complete translation of a whole document at all stages of the search process. The algorithm is a stochastic variant of standard hill climbing: at each step, the decoder generates a successor of 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 translation only if it has a better score than the previous translation. 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 original motivation behind Docent was to facilitate the development of models with cross-sentence dependencies. A classic problem is that of pronominal anaphora resolution: identifying the antecedents of pronouns in order, for example, to correctly translate from English into languages that have grammatical gender for inanimate nouns. This type of problem is very difficult to solve in standard SMT decoders, which have hard-wired assumptions of sentence independence.

The standard tool-kit of sentence-level models, such as the phrase table, n-gram language models and distortion cost are implemented in Docent, along with document-level models including a length parity model, a semantic language model and several readability models. The initial translation can be created either by generating a random segmentation and taking random translations from the phrase table in monotonic order, or by a run from Moses.

Docent is not designed to perform better than Moses when only sentence-level features are used; its advantage lies in the ability to use features that disable recombination. Information about Docent's performance can be found in Hardmeier et al. (2012).

3.1 BLEU decoding

BLEU decoding is the name we have given to a particular mode of decoding in Docent whereby proposed changes to the translation are only accepted by the decoder if they result in an increase in the BLEU score. A new feature model, `BleuModel`, was implemented in Docent. Before decoding begins, `BleuModel` processes and stores the lengths of the reference translations, as well as the lengths of individual sentences within those translations and n-gram counts for $1 \leq n \leq 4$. Once an initial candidate translation for each document has been created, `BleuModel` calculates the BLEU score. The clipped counts for each sentence, required to calculate BLEU, are recorded along with the length of the candidate translation. In this way the counts for a particular sentence need only be updated when Docent proposes a change to that sentence; this makes `BleuModel` a particularly efficient feature model.

In the following section experiments are carried out in *pure* BLEU-decoding mode in Docent, that is to say the weights of all standard feature functions are set to zero, and

only changes to the translation that increase BLEU are accepted. The aim is to examine the properties of translations with very high BLEU scores that are reachable by the decoder.

4 Experiments

A German-English Moses translation model was trained on just over 1.5 million sentences from Europarl v7. The test data was a set of 3052 sentences from the newstest2013 data, divided into 52 separate documents. Two types of experiments were carried out, firstly with the candidate translation initialised by running Moses (with a 5-gram language model trained with KenLM on 2.2 million Europarl sentences and feature weights tuned using MERT on a development set of 2525 sentences from the newstest2009 data), and secondly by random initialisation (i.e. random segmentation and random phrase translation). Docent was then run in BLEU-decoding mode: only changes to the translation that increased BLEU were accepted. Model and BLEU scores were monitored at exponentially increasing intervals, after iterations $2^8, 2^9, ..., 2^{25}$. The motivation for this sampling is that many more proposed changes to the translation are accepted in the beginning: as decoding progresses and the translation improves, there are simply more iterations between each interesting event.

4.1 Moses-based initial translation

Fig. 1 shows how BLEU scores evolve across the 52 test documents during decoding from initial translations produced by Moses. The initial BLEU scores after Moses de-

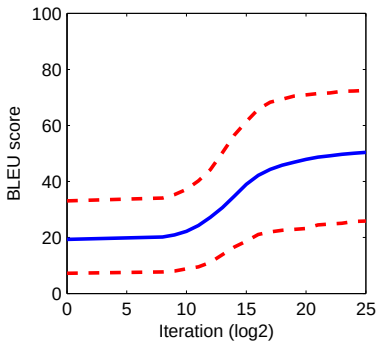

Fig. 1. Progression of the minimum, mean, and maximum BLEU scores across 52 test documents during BLEU decoding from an initial configuration based on a Moses run.

coding ranged from 7.2 to 33.1, with a mean of 19.3; after subsequently running Docent in BLEU-decoding mode, the mean had increased to 50.4, with a range from 25.9 to 72.5. A substantial and consistent increase in BLEU, as expected, is thus observed.

Given the huge increase in mean BLEU score from 19.3 to 50.4, conventional wisdom would say that the quality of the translations after BLEU decoding should be much higher. However, looking at our BLEU-decoded documents it quickly became clear that this was not the case: many of the translations appeared to have deteriorated in quality. To confirm this, we evaluated the first 100 sentences from the test data, randomising the order in which the two competing translations were presented so that it was not possible to know which translation was which, and judged which of the two was of better translation quality. We found that the Moses translation was judged to be superior in 59 cases, the BLEU-decoded translation in 23, and in 18 cases the two translations were judged to be of equal quality.

This is a striking result that deserves restating: despite an increase in mean BLEU score from 19.3 to 50.4, the translations are worse in 59 out of 100 sentences studied. Moreover, it is fair to say that sentences that got worse often got a lot worse, whereas sentences that improved generally did so only marginally. Although we have only studied 100 sentences systematically, it is clear to us that this pattern holds over the whole test set, and even in other experiments with different data sets and language pairs. Let us take a look at some demonstrative examples to understand how this can happen:

(Example 1)

SRC: *in diesem sinne untergraben diese maßnahmen teilweise das demokratische system der usa .*

REF: *in this sense , the measures will partially undermine the american democratic system .*

MOS: ***in this sense ,*** *undermine these measures in the* ***democratic system*** *of the united states .*

BLEU: *the* ***democratic system*** ‖ ***in this sense , the measures*** ‖ ***partially undermine the american*** *.*

The fragments in bold show n-grams for $n \geq 2$ where the Moses and BLEU translations match the reference. The pipe symbol ‖ is used to separate contiguous non-overlapping n-gram matches. We see here by comparing to the reference (REF) that the Moses translation (MOS) is quite poor, with *these measures* appearing as the object, rather than subject, of the verb *undermines*. With some effort, however, the true sense of the phrase can be understood from this translation. This is not the case, however, with the BLEU-optimised translation, which is completely unintelligible. The problem is that BLEU decoding has worked hard to increase the number of n-gram matches, leading to the phrase *partially undermine the american*, which unbeknown to BLEU needs to be followed by *democratic system* to retain the meaning of the original sentence. The Moses translation meanwhile includes *the democratic system of the united states*, a perfectly acceptable equivalent to *the american democratic system*, but one that BLEU decoding does not like.

BLEU decoding produces an even more nonsensical translation in the following example:

(Example 2)

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*** *.*

BLEU: ***the important thing is to have a*** *doctor performed but , with* ***to take it .*** *talking* ***to determine whether or not to*** *s*

Again we see that while the original Moses translation, although far from perfect, has some merit, the BLEU-decoded version is junk. 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 eight such matches. The BLEU translation also has a higher unigram precision; indeed, for all $1 \leq n \leq 4$, the number of matching n-grams is much higher in the BLEU translation than the Moses translation.

In a third example BLEU decoding does in fact produce an intelligible translation:

(Example 3)

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 .*

BLEU: ***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, which lacks a direct translation of *krebsarten* (*cancers* or *types of cancer*). After BLEU decoding the sentence has been transformed: it now matches the whole of the reference, but with the word *cancers* added after the full-stop. It is straightforward to see why the BLEU translation leads to a higher BLEU score: the extra couple of tokens at the end of the matching fragment increase the precision for all n-grams. It is in many ways the reference itself here which is the problem: BLEU decoding has been tricked into trying to mimic a less-literal reference translation rather than stick with a perfectly valid translation from the standard log-linear model. Despite being intelligible and matching the reference, it is highly doubtable that there is any benefit to a system finding this translation over the Moses translation.

4.2 Model scores during BLEU decoding

In the standard setting for statistical machine translation, we decode to maximise the combined scores of a set of features, then use BLEU as an independent evaluation metric. In pure BLEU-decoding mode we are able to turn the tables somewhat, and look at what happens to the model score as decoding proceeds. Of course, BLEU has been shown to correlate better with translation quality than model score, but we would still

expect the two to correspond to some extent: this is why we normally build our systems around this set of features. With this in mind, Fig. 2 shows how the model score, for a standard set of features with MERT-tuned weights, varies as BLEU increases.

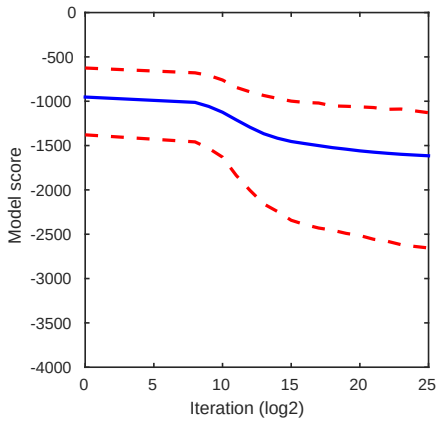

Fig. 2. Progression of the minimum, mean, and maximum model scores across 52 test documents during BLEU decoding from an initial configuration based on a Moses run.

We observe that the model scores decrease as decoding progresses and BLEU increases; Docent in BLEU-decoding mode is able to find translations with high BLEU scores that score poorly on the traditional set of PBSMT features. The Moses-based initialisation procedure works of course to maximise the model score, so it would be unrealistic to expect it to increase much more during BLEU decoding, unless we had reason to believe that there was significant search error in the Moses decoding process. The fact that the model score drops in this way however adds weight to the point made earlier by the example sentences, that we have high BLEU scores but many poor quality sentences. These results suggest that by letting BLEU run wild, we move far away from the part of the search space containing good translations.

4.3 Random initial translation

Fig. 3 shows how the BLEU scores evolve among the 52 test documents during decoding from an random initial translation. We again observe a large increase in BLEU scores; on this occasion the mean BLEU score at the beginning of the decoding process was 3.6 (with range 0.0 to 6.6); after running Docent in BLEU-decoding mode it had increased to 50.2 (with range 24.9 to 71.5). The figure for the mean at the end of decoding is very similar to that of 50.4 obtained when decoding from Moses-based initial translations, suggesting that the initial translation does not have a great effect on the final result.

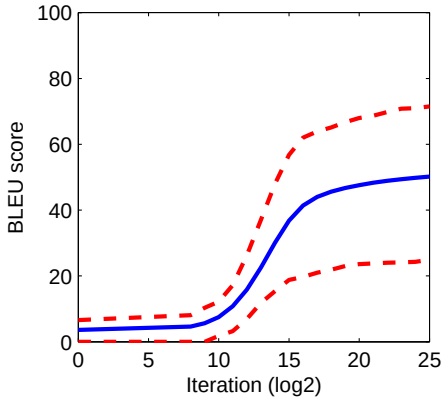

Fig. 3. Progression of the minimum, mean, and maximum BLEU scores across 52 test documents during BLEU decoding from a random initial configuration.

We can now go back to an example sentence from the previous experiment and add two new translations: the random initial translation (RAND) and the revised version of this after BLEU decoding (BLEU2):

(Example 2)

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*** *.*

BLEU: ***the important thing is to have a*** *doctor performed but , with* ***to take it .*** *talking* ***to determine whether or not to*** *s*

RAND: *most important of all has it , which from his own medical with talking about with a view to set ,* ***whether or not*** *report implement to be* ***or not*** *||* ***it .***

BLEU2: *talking* ***but the important thing is to*** *its* ***to have a*** *||* ***doctor to determine whether or not to take it .*** *or report implement to*

While BLEU and BLEU2 are not identical, they are strikingly similar in that they share many phrases and contiguous sets of words, as well as the property that they make very little sense. This suggests that the type of translation in which BLEU decoding results is independent of the initial translation; the initial translations – MOS and RAND – of BLEU and BLEU2 are clearly very different from each other. It is also interesting to compare BLEU2 with its antecedent RAND. While neither of these translations can be said to convey much of the sense of the original German sentence, it could perhaps be argued that BLEU2 is slightly more sensical than RAND. Perhaps the chunks that

match the reference do actually help to bring through some trace of meaning. One way to compare the random initial translation to the BLEU-decoded version is to look again at the model scores.

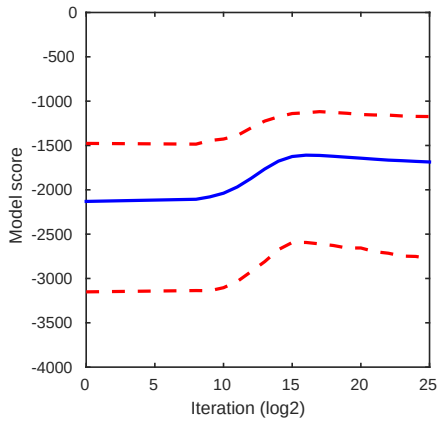

Fig. 4. Progression of the minimum, mean, and maximum model scores across 52 test documents during BLEU decoding from a random initial configuration.

Fig. 4 shows a slight increase in model scores at the beginning of decoding, followed by a gradual decline, but with final values still above the initial translation. We can therefore draw the conclusion that BLEU decoding from a random initial translation does result in translations that are slightly better, in some meaningful sense, than the initial translation. It is however clear by looking at the example sentences that the improvement in quality is nowhere near that which would be normally be expected given the jump in mean BLEU score from 3.6 to 50.2.

4.4 BLEU decoding towards reachable translations

We saw in the previous sections that the mean BLEU score after 2^{25} iterations of BLEU decoding was 50.4 when the initial translation came from Moses, and 50.2 when the initial translation was randomly chosen. While these are undoubtedly high BLEU values, they are still a long way from 100, which would represent the decoder finding the reference translation exactly. It is natural to wonder why this is the case; what is stopping the BLEU score getting much higher. BLEU decoding bears some resemblance to the technique of forced decoding, where the training data is decoded in such a way that guarantees the reference be found, in order to re-calculate phrase translation probabilities. Wuebker et al. (2010) reported being able to match the reference 95% percent of the time, while Foster and Kuhn (2012) report slightly lower performance. Note however that in these cases it is the same training data used for the original phrase extraction that is force-decoded, unlike in our case where BLEU decoding is carried out on a separate test/development data.

There are two obvious candidates to explain the failure of BLEU decoding to find the reference exactly. One is the availability of the right phrases in the phrase table. Reference reachability has long been known to be a problem in PBSMT (Liang et al., 2006). This is also the problem in forced decoding, where despite the fact that the phrases are extracted from the same data being decoded, it is not always possible to force-decode every sentence (Foster and Kuhn, 2012). Another possibility might be that the decoder's hill-climbing algorithm tends to get stuck in local maxima. The fact that the initial configuration apparently plays no role speaks against this hypothesis, but not definitively. Another test that can be carried out is to give the decoder a pseudo-reference translation, that is not really a true reference at all, but simply another random translation generated by Docent. As Docent generates this translation from phrases in the phrase table, it is guaranteed that the reference is theoretically reachable by the decoder.

The experimental set-up was similar to that described in Section 4, the only difference being the switch from the genuine reference translation to a simulated reference generated randomly by Docent. The 52 test documents were decoded from random initial translations.

The average BLEU score before decoding was 6.8, with range from 4.2 to 12.9; after 2^{25} iterations it was 98.4, with range 96.8 to 99.6 (Fig. 5). The contrast between Fig. 5

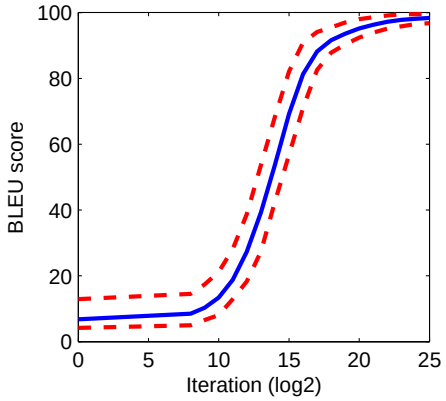

Fig. 5. Progression of the minimum, mean, and maximum BLEU scores across 52 test documents during BLEU decoding towards a reachable configuration.

and Fig. 3 is striking. These results confirm that, when the necessary phrases are in the phrase table, the decoder in BLEU-decoding mode is able to get extremely close to the reference translation. Note that the monotonic way in which random initial translations are generated in Docent makes it somewhat easier for the decoder to find the reference translations than in a real-world case where extensive reordering is necessary. We also tested however decoding from a random initial translation towards a Moses translation

as the pseudo-reference, where more reordering is likely to be necessary, and again found BLEU scores in the high 90s.

This suggests that the lower BLEU scores for decoding towards genuine reference translation are as high, or almost as high, as they can possibly get, on this data given the phrases in the phrase table. BLEU scores near to 100 are simply impossible on this data set given a genuine reference translation and the phrases available to the decoder. The poor quality translations demonstrated by the examples in the previous section are therefore probably as good as it gets in terms of BLEU score: there are no translations that the decoder can conceivably reach with significantly higher BLEU scores.

5 Discussion

This paper presented BLEU decoding, a method for finding oracle BLEU translations using exact document-level scores for a phrase-based SMT decoder. Previous attempts to find oracle translations, for use in feature-weight tuning for example, have relied on sentence-level approximations to BLEU. As other authors have shown, however, optimising BLEU at the sentence-level and document-level are not always equivalent (Chiang et al., 2008).

By performing experiments with BLEU decoding, we explored the view from the top of Mount BLEU, examining in detail high-BLEU regions of the search space. While it might be assumed that translations in this region would be of high quality, results presented here show this not to be the case if the reference translation is not reachable by the decoder. Despite an increase in mean BLEU score from 19.3 to 50.4 across 52 documents from newstest2013 translated in BLEU-decoding mode from an initial translation generated by Moses, there was a clear drop in translation quality (59 out of 100 sentences were judged to be worse, and only 23 judged better). We observed long n-gram matches interleaved with strings of nonsense, leaving many sentences unintelligible. This makes sense given how BLEU works, favouring long n-gram matches and saying nothing about parts that do not match. An even larger increase in mean BLEU score, from 3.6 to 50.2, was observed when decoding from a random initial translation, but results were similar in terms of translation quality.

What do these results say about BLEU as an evaluation metric? Initial impressions might suggest that the evidence presented here is damning for BLEU: it has been clearly shown that it can be 'cheated': very bad translations can get high BLEU scores. This is not the first time problems with BLEU have been highlighted (Callison-Burch et al., 2006; Chiang et al., 2008), and research into better metrics is a very active field (Macháček and Bojar, 2013). However, it must be remembered that the experiments presented here used BLEU in a very different fashion from that for which it was designed. Papineni et al. (2002) demonstrated clearly that when translation quality is manipulated as the independent variable in experiments, there is a strong correlation with BLEU as the dependent variable. This does not imply, and indeed the opposite has been shown in this paper, that manipulating BLEU as an independent variable will necessarily result in high quality translations.

Another way of saying this is as follows: if translations are produced independently of BLEU, then BLEU is often a good metric to distinguish their quality; however this

does not imply that actively looking for translations with high BLEU score will result in high quality. There is clearly a high-BLEU area of the search space with low quality translations. This problem has previously been encountered by researchers working on feature-weight tuning (Liang et al., 2006; Chiang, 2012). Searching for weights that produce high BLEU scores on development data is a central part of many standard tuning algorithms such as MERT (Och, 2003), PRO (Hopkins and May, 2011) and MIRA (Watanabe et al., 2007). In reality feature models are selected in such a way that ending up in this strange region of the search space is unlikely, but if we blindly optimise feature weights using BLEU, we could run the risk of moving dangerously close.

Despite these problems, BLEU decoding may still have the potential to be applied during tuning to improve translation quality. We have seen that in its pure form, BLEU decoding leads us far from the area of the search space containing good translations, and optimising our models towards finding these regions is unlikely to be a good idea. However, by combining BLEU decoding with other regular SMT features, we may be able to keep the decoder in higher-quality areas of the search space, using the BLEU feature model to find the best translations within this constrained region. The principle behind BLEU decoding can also be implemented for other translation metrics, either to include as additional features in the tuning process, or in order to stress test the metric itself.

References

A. Agarwal and A. Lavie. 2008. METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

D. Chiang, S. DeNeefe, Y. S. Chan, and H. T. Ng. 2008. Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619.

D. Chiang. 2012. Hope and Fear for Discriminative Training of Statistical Translation Models. *Journal of Machine Learning Research*, 13:1159–1187.

M. Farrús, M. R. Costa-jussà, and M. Popović. 2012. Study and Correlation Analysis of Linguistic, Perceptual and Automatic Machine Translation Evaluations. *Journal of the American Society for Information Science and Technology*, 63(1):174–184.

G. Foster and R. Kuhn. 2012. Forced Decoding for Phrase Extraction. Technical Report, Université de Montreal.

C. Hardmeier, J. Nivre, and J. Tiedemann. 2012. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190.

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.

M. Hopkins and J. May. 2011. Tuning as Ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362.

P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics.*

M. Macháček and O. Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51.

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*, pages 160–167.

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.

A. Sokolov, G. Wisniewski, and F. Yvon. 2012. Computing Lattice BLEU Oracle Scores for Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 120–129.

T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. 2007. Online Large-Margin Training for Statistical Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 764–773.

G. Wisniewski, A. Allauzen, and F. Yvon. 2010. Assessing Phrase-Based Translation Models with Oracle Decoding. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 933–943.

J. Wuebker, A. Mauser, and H. Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In *Proceeding of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 475–484.

Received May8, 2016 , accepted May 15, 2016