Improving Statistical Machine Translation Performance by Oracle-BLEU Model Re-estimation

Dakwale, P.; Monz, C.

DOI
10.18653/v1/P16-2007

Publication date
2016

Document Version
Final published version

Published in
The 54th Annual Meeting of the Association for Computational Linguistics : ACL 2016

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Aug 2024
Improving Statistical Machine Translation Performance
by Oracle-BLEU Model Re-estimation

Praveen Dakwale
Informatics Institute
University of Amsterdam
p.dakwale@uva.nl

Christof Monz
Informatics Institute
University of Amsterdam
c.monz@uva.nl

Abstract

We present a novel technique for training translation models for statistical machine translation by aligning source sentences to their oracle-BLEU translations. In contrast to previous approaches which are constrained to phrase training, our method also allows the re-estimation of reordering models along with the translation model. Experiments show an improvement of up to 0.8 BLEU for our approach over a competitive Arabic-English baseline trained directly on the word-aligned bitext using heuristic extraction. As an additional benefit, the phrase table size is reduced dramatically to only 3% of the original size.

1 Introduction

In phrase-based SMT, the phrase pairs in the translation model are traditionally trained by applying a heuristic extraction method (Och and Ney, 2000) which extracts phrase pairs based on consistency of word alignments from a word-aligned bilingual training data. The probabilities of the translation model are then calculated based on the relative frequencies of the extracted phrase pairs.

A notable shortcoming of this approach is that the translation model probabilities thus calculated from the training bitext can be unintuitive and unreliable (Marcu and Wong, 2002; Foster et al., 2006) as they reflect only the distribution over the phrase pairs observed in the training data.

However, from an SMT perspective it is important that the models reflect probability distributions which are preferred by the decoding process, i.e., phrase translations which are likely to be used frequently to achieve better translations should get higher scores and phrases which are less likely to be used should get low scores. In addition, the heuristic extraction algorithm generates all possible, consistent phrases including overlapping phrases. This means that translation probabilities are distributed over a very large number of phrase translation candidates most of which never lead to the best possible translation of a sentence.

In this paper, we propose a novel solution which is to re-estimate the models from the best BLEU translation of each source sentence in the bitext. An important contribution of our approach is that unlike previous approaches such as forced alignment (Wuebker et al., 2010), reordering and language models can also be re-estimated.

2 Related Work

The forced alignment technique of Wuebker et al. (2010) forms the main motivation for our work. In forced alignment, given a sentence pair \((F, E)\), a decoder determines the best phrase segmentation and alignment which will result in a translation of \(F\) into \(E\). The best segmentation is defined as the one which maximizes the probability of translating the source sentence into the given target sentence. At the end, the phrase table is re-estimated using the phrase pair segmentations obtained from forced decoding. Thus forced alignment is a re-estimation technique where translation probabilities are calculated based on their frequency in best-scoring hypotheses instead of the frequencies of all possible phrase pairs in the bitext. However, one limitation of forced alignment is that only the phrase translation model can be re-estimated since it is restricted to align the source sentence to the given target reference, thus fixing the choice of reordering decisions.

A similar line of work is proposed by Lambert et al. (2011) and Schwenk et al. (2011) who use a self-enhancing strategy to utilize additional mono-
lingual source language data by aligning it to its target language translation obtained by using an SMT system to rank sentence translation probabilities. However, the main focus of their work is translation model adaptation by augmenting the bitext with additional training data and not the re-estimation of the translation models trained on the parallel data.

In this work, we propose that aligning source sentences to their oracle BLEU translations provides a more realistic estimate of the models from the decoding perspective instead of aligning them to high-quality human translations as in forced decoding.

Another relevant line of research relates tuning (weight optimisation), where our work lies between forced decoding (Wuebker et al., 2010) and the bold updating approach of (Liang et al., 2006). However, our approach specifically proposes a novel method for training models using oracle BLEU translations.

3 Model Re-estimation

The idea of our approach is to re-estimate the models with n-best oracle-BLEU translations and sentence alignments resulting from decoding the source sentence. Given a source and its reference translation, the oracle-BLEU translation is defined as the translation output with highest BLEU score. Oracle BLEU translations have been previously used for different analytical purposes in SMT (Srivastava et al., 2011; Dreyer et al., 2007; Wiesniewski et al., 2010).

Figure 1 shows example of word alignment obtained from EM training, segmentations and alignment obtained from forced decoding and oracle-BLEU re-estimation.

3.1 Oracle BLEU

Ideally, one would like to re-estimate translation models directly from the n-best BLEU translations. However, there are two problems in calculating BLEU for individual sentence: First, as discussed in (Chiang et al., 2008), BLEU is not designed to be used for sentences in isolation where it can exhibit rather volatile behavior. Hence, following their work and (Watanabe et al., 2007), we calculate BLEU for a sentence in the context of an exponentially-weighted moving average of previous translations. We briefly discuss the computation from (Chiang et al., 2008) as follows: Given a source sentence f, and its reference translation r, for an n-best translation e∗, let c(e) be defined as the vector of target length |e|, source length |f|, reference length |r|, and the number of n-gram matches between e and r, then two pseudo document parameters O and O_f are defined as:

\[ O \leftarrow 0.9 \cdot \left( O + c(e^*) \right), \quad O_f \leftarrow 0.9 \cdot \left( O_f + |f| \right) \] (1)

O is an exponentially-weighted moving average of the vectors from previous sentences and O_f is the correction of source length with respect to the previous sentences. Then the BLEU score for a sentence pairs (f,r) and translation e∗ is defined as:

\[ B(e; f, r) = (O_f + |f|) \cdot \text{BLEU}(O + c(e^*; r)) \] (2)

The second problem as discussed in Chiang et al. (2008) is that due to noise in the training data, a high-BLEU translation may contain certain rules which are unlikely to be used by the model. Hence
following them, we use a weighted combination of
BLEU and model score to select the n-best list:

\[ e^* = \arg\max_e (B(e) - \mu \cdot (B(e) - h(e)d)) \]  

(3)

where B(e) and h(e) are the BLEU and model
scores of the candidate translation and \( w \) is the
optimised weights for the models, \( \mu \) controls the
preference between BLEU and model scores to
determine oracle translations. We set \( \mu=0.5 \) to
balance between BLEU scores almost as high as
the max-BLEU translations, while staying close
to translations preferred by the model. We also
conducted a set of experiments with \( \mu=0 \) (pure or
absolute BLEU) in order to verify the necessity
for the optimal combination. The lower scores
for this setting as compared to the baseline veri-
fied that using only the best BLEU translation in-
deed degrades the performance of the re-estimated
models. This finding for the optimal value of
\( \mu \) has also been established in (Chiang et al., 2008)
through a series of experiments.

3.2 Training

For obtaining the oracle-BLEU translations, we
first train the translation models from the bitext
using the standard pipeline of word alignment
and heuristic extraction. Along with the phrase
translation and language models, we also train
a bilingual language model (BiLM) (Niehues et
al., 2011; Garmash and Monz, 2014), as well as
directed graph (Tillman, 2004) and hierarchical re-
ordering models (Galley and Manning, 2008). We
use a BiLM specifically as an instance of a re-
ordering model in order to determine the effect of
re-estimating re-ordering decisions from oracle-
BLEU translations.

We use the decoder trained on these models to
translate the training bitext. Along with the 1-
best translation (based on model scores), we also
store search graphs or lattices generated during
the translations process. Using the target sen-
tences, we convert the translation lattice to an
isomorphic oracle-BLEU lattice which has the
same set of nodes but the edges represent BLEU
score differences corresponding to each transition.
Finally, we extract n-best candidate translations
from the graphs ranked on BLEU score as de-
fined in Equation (3). Using the word alignments
from the initial phrase table, we extract the align-
ments between each source sentence and each of
their n-best oracle-BLEU translations. Finally, we
re-train the phrase translations, re-ordering and
BiLM on these translations and alignments.

3.3 Avoiding over-fitting

Re-estimation of the translation models from the
n-best translation of the bitext could re-enforce
the probabilities of the low frequency phrase pairs
in the re-estimated models leading to over-fitting.
Within forced decoding, Wuebker et al. (2010) ad-
dress this problem by using a leave-one-out ap-
proach where they modify the phrase translation
probabilities for each sentence pair by remov-
ing the counts of all phrases that were extracted
from that particular sentence. However, in our ap-
proach, we do not impose a constraint to produce
the exact translation, instead we use the highest
BLEU translations which may be very different
from the references. Thus it is not strictly nec-
ecessary to apply leave-one-out in our approach as
a solution to over-fitting. Instead, we handle the
problem by simply removing all the phrase pairs
below a threshold count which in our case is 2,

\[ \phi_{init} = \phi_{baseline} - \phi_{C(e,f)} < 2 \]  

(4)

therefore removing phrase pairs with high proba-
bility but low frequency.

4 Experimental set up

Our experiments are carried out for an Arabic-
English parallel corpus of approximately 1 million
sentence pairs. We establish a baseline system by
training models on this bitext and then compare
this to a forced decoding implementation and to
oracle-BLEU re-estimation using the same bitext.

4.1 Baseline and forced decoding

The initial training corpus we use is a collection
of parallel sentences taken from OpenMT data
sources released by the LDC.

Phrase table, distortion models and the lexical
BiLM are trained with initial alignments obtained
using GIZA++ (Och and Ney, 2003). The En-
lish 5-gram target language model is trained with
Kneser-Ney smoothing on news data of nearly
1.6B tokens. We use an in-house phrase-based
SMT system similar to Moses. For all settings
in this paper, weights were optimized on NIST’s
MT04 data set using pairwise ranked optimization
(Hopkins and May, 2011).

For forced alignment we use the existing imple-
mentation within the Moses SMT toolkit (Koehn
Table 1: Performance of our oracle-BLEU re-estimation with varying size \( n \) of \( n \)-best lists for the MT09 test set. ▲/▼ indicates a statistically significant gain/drop at \( p < 0.01 \) and \( \triangle/\triangledown \) at \( p < 0.05 \). Values in brackets show gains over the baseline.

<table>
<thead>
<tr>
<th>n</th>
<th>50.1</th>
<th>50.1</th>
<th>50.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>n=10</td>
<td>50.5</td>
<td>50.5</td>
<td>50.5</td>
</tr>
<tr>
<td>n=100</td>
<td>49.6</td>
<td>49.6</td>
<td>49.6</td>
</tr>
</tbody>
</table>

Table 2 provides a comparison between BLEU improvements achieved by forced decoding \( (n = 100 \) best) and our oracle-BLEU re-estimation approach \( (n = 1 \) best) over the baseline for different models. One can see in Table 2 that while phrase table re-estimation drops substantially for forced decoding for all test sets (up to -1.4 for MT09), oracle-BLEU phrase table re-estimation shows either slight improvements or negligible drops compared to the baseline. For the linear interpolation of the re-estimated phrase table with the baseline, forced decoding shows only a slight improvement for MT06, MT08 and MT09 and still suffers from a substantial drop for MT05. On the other hand, oracle-BLEU re-estimation shows consistent improvements for all test sets with a maximum gain of up to +0.7 for MT06. It is important to note here that although linear interpolation extinguishes the advantage of a smaller phrase table size obtained by re-estimation, the improvement achieved by interpolation for oracle-BLEU re-estimation are significantly higher as compared to forced decoding.

An important novelty of oracle-BLEU re-estimation is that it also allows for re-training of other models alongside the phrase table. Here we provide the results for the re-estimation of a BiLM. For all test sets, BiLM re-estimation provides additional improvements over simple phrase table interpolation, demonstrating that re-estimation of re-ordering models can further improve translation performance. The last row of Table 2 shows that the re-estimated BiLM on its own adds BLEU improvement of up to +0.5 (for MT09). The highest BLEU improvement of +0.8 is achieved by using a re-estimated BiLM and an interpolated phrase table. Note that re-estimation of BiLM or re-ordering models is not possible for forced decoding due to the constraint of having to match the exact reference. For an additional anal-
In this paper, we proposed a novel technique for improving the reliability of SMT models by model re-estimation from oracle-BLEU translations of the source sentences in the bitext. Our experimental results show BLEU score improvements of up to +0.8 points for oracle-BLEU re-estimation over a strong baseline along with a substantially reduced size of the re-estimated phrase table (3.3% of the baseline). An important novelty of our approach is that it also allows for the re-estimation of re-ordering models which can yield further improvements in SMT performance as demonstrated by the re-estimation of a BiLM.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218. We thank Arianna Bisazza and the anonymous reviewers for their comments.
References


