Bayesian Optimisation for Machine Translation

Yishu Miao[1] **Ziyu Wang**[1] **Phil Blunsom**[1,2]
[1]Department of Computer Science, University of Oxford
[2]DeepMind Technologies
{yishu.miao,ziyu.wang,phil.blunsom}@cs.ox.ac.uk

Abstract

This paper presents novel Bayesian optimisation algorithms for minimum error rate training of statistical machine translation systems. We explore two classes of algorithms for efficiently exploring the translation space, with the first based on N-best lists and the second based on a hypergraph representation that compactly represents an exponential number of translation options. Our algorithms exhibit faster convergence and are capable of obtaining lower error rates than the existing translation model specific approaches, all within a generic Bayesian optimisation framework. Further more, we also introduce a random embedding algorithm to scale our approach to sparse high dimensional feature sets.

1 Introduction

State of the art statistical machine translation (SMT) models traditionally consist of a small number ($<$20) of sub-models whose scores are linearly combined to choose the best translation candidate. The weights of this linear combination are usually trained to maximise some automatic translation metric (e.g. BLEU) [1] using Minimum Error Rate Training (MERT) [2, 3] or a variant of the Margin Infused Relaxed Algorithm (MIRA) [4, 5]. These algorithms are heavily adapted to exploit the properties of the translation search space. In this paper we introduce generic, effective, and efficient Bayesian optimisation (BO) algorithms [6, 7] for training the weights of SMT systems for arbitrary metrics that outperform both MERT and MIRA. To our knowledge this is the first application of BO in natural language processing (NLP) and our results show that their may be significant scope for using BO to tune hyperparameter in a range of NLP models.

The linear model popular for SMT systems [2] is parametrised in terms of a source sentence $\boldsymbol{f}$, target translation $\boldsymbol{e}$, feature weights w_k and corresponding feature functions $H_k(\boldsymbol{e}, \boldsymbol{f})$ (including a language model, conditional translation probabilities, etc.). The best translation is selected by,

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \{\sum_{k=1}^{K} w_k H_k(\boldsymbol{e}, \boldsymbol{f})\}. \tag{1}$$

Since the translation metrics (e.g. BLEU score) can only be evaluated between the selected translations and reference translations (i.e. the standard manual translations from the parallel training data), meanwhile decoding new translations following Equation 1 is very time consuming, we cannot tune the linear weights directly as in ordinary classification tasks. The most common approach is an iterative algorithm MERT [3] which employs N-best lists (the best N translations decoded with a weight set from a previous iteration) as candidate translations $\boldsymbol{C}$. In this way, the loss function is constructed as $E(\bar{\boldsymbol{E}}, \hat{\boldsymbol{E}}) = \sum_{s=1}^{S} E(\bar{\boldsymbol{e}}_s, \hat{\boldsymbol{e}}_s)$, where $\bar{\boldsymbol{e}}$ is the reference sentence, $\hat{\boldsymbol{e}}$ is selected from N-best lists by $\hat{\boldsymbol{e}}_s = \arg\max_{\boldsymbol{e} \in \boldsymbol{C}} \left\{ \sum_{k=1}^{K} w_k H_k(\boldsymbol{e}, \boldsymbol{f}_s) \right\}$ and S represents the volume of sentences. By exploiting the fact that the error surface is piece-wise linear, MERT iteratively applies line search to find the optimal parameters along the randomly chosen directions via Equation 2, generating new

N-best lists until convergence (no change happened in the new N-best lists),

$$\hat{\boldsymbol{w}} \quad = \quad \arg\min_{\boldsymbol{w}} \left\{ \sum_{s=1}^{S} E\left(\bar{\boldsymbol{e}}_s, \arg\max_{\boldsymbol{e} \in \boldsymbol{C}} \left\{ \sum_{k=1}^{K} w_k H_k(\boldsymbol{e}, \boldsymbol{f}_s) \right\} \right) \right\}. \tag{2}$$

Hypergraph, or lattice, MERT [8, 9] aims to tackle problems caused by the lack of diversity in N-best lists. A hypergraph [10] efficiently encodes the exponential translation space explored by the beam-search translation decoder. The line search can then be carried out on the edges of the hypergraph, instead of the translations in the N-best lists. And dynamic programming is used to find the upper envelope of the hypergraph corresponding to the maximum scoring translation. Prior work [8, 9] showed that hypergraph MERT is superior to the original N-best algorithm both in speed of convergence and stability. MIRA is an online large-margin learning algorithm that applies a different strategy to MERT. It enforces a margin between high and low loss translations and enables stochastic gradient descent to be used to update parameters. A disadvantage of this approach is that it requires the global BLEU score, which is a non-linear function of local translation candidate statistics, to be approximated by a linear combination of sentence level BLEU scores.

In this paper, however, our BO algorithms treat the loss function as a black-box function so that we could directly query the function value without the cumbersome and inefficient work of constructing an error surface for random directions. Instead of applying BO to the whole SMT pipeline, which would require expensive decoding of new translations with every parameter set sampled, our BO algorithms only decode new translations after obtaining the best parameters on fixed N-best lists or hypergraphs. Hence our algorithms iteratively run Gaussian processes on the sub-models and only a few decoding iterations are required to reach convergence. The experiments in Section 3 illustrate the superiority of our algorithms both in translation quality and speed of convergence.

2 Bayesian Optimisation Tuning Algorithms

Algorithm 1 describes our hypergraph algorithm (HG-BO). The N-best algorithm (NBL-BO) is similar to HG-BO and can be derived from Algorithm 1 by replacing the hypergraphs with N-best lists. In HG-BO, both $\boldsymbol{w}_i$ and $\boldsymbol{x}_j$ represent the weights of the linear model. The weights $\boldsymbol{w}_i$ are used to produce the hypergraphs $\boldsymbol{H}_i$, while $\boldsymbol{x}_j$ are the weights sampled from the GP to compute the BLEU score (i.e. objective function value) for a fixed set $\boldsymbol{H}_i$. Since $\boldsymbol{H}_i$ remains unchanged during an iteration of Bayesian optimisation, the BLEU score calculated for the fixed hypergraphs approximates the true BLEU score that would be achieved if the translation system were run with $\boldsymbol{x}_j$. This introduces some noise owing to the variance between $\boldsymbol{w}_i$ and $\boldsymbol{x}_j$.

As depicted in Fig. 1, a key aspect of Algorithm 1 is that we place a bound (blue area) around $\boldsymbol{w}_i$ and only consider samples inside this region. The sample with the highest BLEU score will then be

Algorithm 1: Hypergraph BO

Input : Initial weights $\boldsymbol{w}_0$, source sentences $\boldsymbol{F}$, reference sentences $\bar{\boldsymbol{E}}$.
Output: Final weights $\boldsymbol{w}_f$
for $i = 0; i < maxIter; i = i + 1$ **do**
  Decode hypergraphs $\boldsymbol{H}_i$ using $\boldsymbol{w}_i$;
  Generate search bound $\boldsymbol{B}_i$= $\{w_i^k \in \boldsymbol{w}_i | w_i^k - b \leqslant x_i^k \leqslant w_i^k + b\}$;
  Initialise candidate point set $\mathcal{X}$ in bounded search area $\boldsymbol{B}_i$ of the Gaussian process;
  for $j = 0; j < maxBOIter; j = j + 1$ **do**
    $\boldsymbol{x}_j = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{EI}(\boldsymbol{x})$;
    Reweight hypergraphs $\boldsymbol{H}_i$ by $\boldsymbol{x}_j$;
    Generate translation set $\hat{\boldsymbol{E}}_j$ by Viterbi algorithm;
    $y_j = \mathrm{BLEU}(\bar{\boldsymbol{E}}, \hat{\boldsymbol{E}}_j)$;
    Update GP with $(\boldsymbol{x}_j, y_j)$;
  $\boldsymbol{w}_{i+1} = \boldsymbol{x}_{best}$;
Return $\boldsymbol{w}_i$

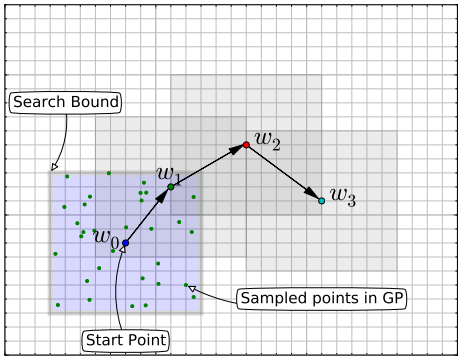
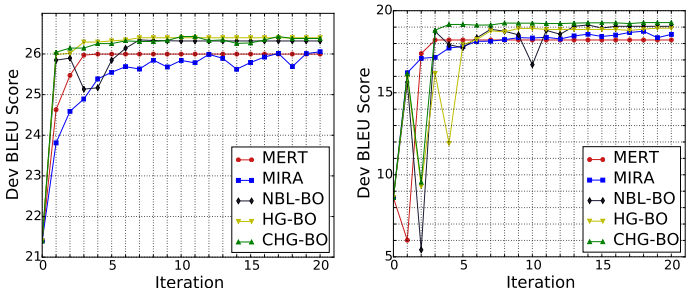


Figure 1: Bounded search in 2 dimensions.

Language	French-English (fr-en)			Spanish-English (es-en)			German-English (de-en)			Czech-English (cs-en)		
Dataset	Dev (variance)	Test-1	Test-2	Dev (variance)	Test-1	Test-2	Dev (variance)	Test-1	Test-2	Dev (variance)	Test-1	Test-2
MERT	26.1 (2×10^{-2})	26.8	26.5	29.5 ($\mathbf{1\times10^{-4}}$)	28.2	30.2	21.1 (6×10^{-1})	18.9	20.2	17.7 (5×10^{-1})	17.8	16.9
MIRA	26.0 (1×10^{-3})	26.8	26.5	29.2 (1×10^{-3})	**28.5**	**30.7**	20.9 (1×10^{-2})	18.9	20.3	18.4 (4×10^{-3})	18.7	17.7
NBL-BO	**26.4** (6×10^{-5})	26.7	26.5	29.7 (1×10^{-2})	28.1	30.4	22.0 (2×10^{-3})	**19.8**	**21.0**	18.8 ($\mathbf{2\times10^{-3}}$)	18.8	17.3
HG-BO	**26.4** ($\mathbf{3\times10^{-5}}$)	26.8	26.7	**29.9** ($\mathbf{1\times10^{-4}}$)	28.0	30.1	**22.2** ($\mathbf{1\times10^{-5}}$)	**19.8**	20.9	19.1 (1×10^{-2})	19.1	17.7
CHG-BO	**26.4** (3×10^{-3})	**26.9**	**26.8**	**29.9** (2×10^{-3})	28.3	30.4	22.1 (3×10^{-2})	19.7	20.9	**19.2** ($\mathbf{2\times10^{-3}}$)	**19.3**	**17.8**

Table 1: Translation Performance (BLEU) score



(a) *fr-en* (b) *cs-en*

Figure 2: Convergence of different models

Model	Time(h)	Iteration
MERT	4	5
MIRA	4	20
NBL-BO	1.5	5
HG-BO	1.5	5
CHG-BO	2	5

Table 2: Time consumption

used to decode new hypergraphs for the next iteration of BO. Intuitively, to speed up convergence, we would like the search space of BO to be as large as possible. When the search space is too large, however, a sampled $\boldsymbol{x}_j$ could be so far from $\boldsymbol{w}_i$ that the generated translations would become unreliable thus leading to noisy BLEU measurements. HG-BO is preferable to NBL-BO as it weighs the translations directly in the hypergraphs, which encode an exponentially larger space of translations than the N-best lists, and thus noise is diminished. To further expand the translation space searched at each iteration, we present a variant cumulative hypergraph BO algorithm (CHG-BO) which combines hypergraphs from one previous and current iterations in order to trade stability and speed of convergence with memory usage.

Similar to MERT, our BO algorithms become less reliable when the number of features in the linear model exceeds 30. Hence, we introduce a variant of random embedding Bayesian optimisation (REMBO) [11] into our hypergraph algorithm (HG-REMBO) to tackle the large scale training problem. The original REMBO generates a random matrix $\boldsymbol{A} \in R^{h\times l}$ to map the sample $\boldsymbol{x} \in R^h$ from high dimensional space to a point $\boldsymbol{z} \in R^l$ in low dimensional space. The objective function to be optimised then becomes $g(\boldsymbol{z}) = f(\boldsymbol{A}\boldsymbol{z})$. Instead of $\mathbf{A}$, we used a regularised random matrix $\bar{\boldsymbol{A}}$ where $\bar{A}_{mn} = \frac{A_{mn}}{\|\boldsymbol{A}_m\|_1}$ and transform the objective function to $g(\boldsymbol{z}) = f(\bar{\boldsymbol{A}}\boldsymbol{z} + \boldsymbol{w})$, where $\boldsymbol{w}$ are the weights producing the hypergraphs. $\boldsymbol{w}$ would remain constant during Bayesian optimisation. In this way, BO can be carried out in the low dimensional space and the regularisation of $\mathbf{A}$ ensures that each update of the weights remains in a bounded domain.

3 Experiments

We implemented our models using *spearmint* [7][1] and the *cdec* SMT decoder [12][2]. The datasets are from WMT14 shared task,[3] all tokenized and lowercased. We employ ARD Matern 5/2 kernel and EI acquisition function. The *cdec* implementations of hypergraph MERT [9] and MIRA[13] are used as benchmarks.

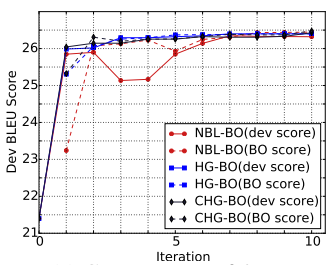
The experiment[4] results in Table 1, averaged over 3 runs, show that our BO algorithms always achieve a higher training objective score than MERT and MIRA, and in most cases a higher test BLEU score. Fig.2 illustrates the convergence w.r.t. the development BLEU score and Fig. 2b

[1] https://github.com/JasperSnoek/spearmint

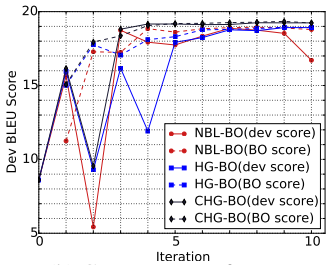
[2] http://www.cdec-decoder.org/

[3] http://www.statmt.org/wmt14/translation-task.html

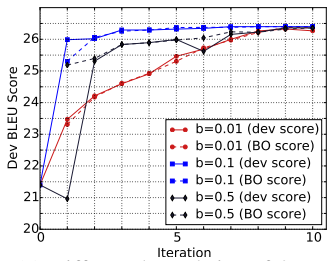
[4] The 4-gram language model is trained on *europarl*, *news-crawl* and *news-commentary* sections, translation grammar is extracted from *news-commentary*, while *news-test* 2010 is used for BO, *news-test* 2011 and 2012 are used for testing. We use 18 default *cdec* features and the same initial weights on one machine with 10 processors and trained for 20 iterations. The BO bound size is 0.1 and the number of BO iterations is 100.



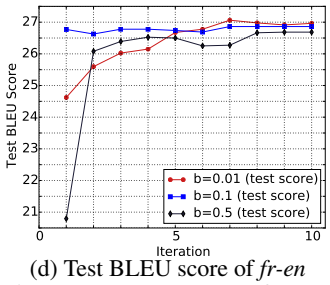
(a) Convergence of *fr-en*



(b) Convergence of *cs-en*



(c) Different bound size of *fr-en*



(d) Test BLEU score of *fr-en*

Figure 3: Convergence of BO algorithms

Model	French (fr-en)			Spanish-English (es-en)		
Dataset	Dev (variance)	Test-1	Test-2	Dev (variance)	Test-1	Test-2
MIRA	**26.8** ($\mathbf{3\times10^{-3}}$)	**27.0**	**30.6**	**30.0** (4×10^{-3})	**28.3**	26.5
HG-REMBO	26.5 (2×10^{-2})	26.9	**30.6**	29.8 ($\mathbf{1\times10^{-5}}$)	**28.3**	**26.8**

Table 3: Adding four sparse feature functions, 1) rule identifier, 2) rule shape, 3) bigrams on source side and 4) target side of rules, increases the number of features from 18 to 56,396 in *es-en* and 207,952 in *fr-en*. More details about the rules can be found in [14].

shows a particular case where the imperfect starting weights cause a violent fluctuation initially. CHG-BO quickly reaches the plateau in 5 iterations but NBL-BO dips again at the 10th iteration.

Table 2 illustrates the efficiency of the BO algorithms. They consistently obtain a good weight set within 5 iterations, but the best one is always achieved after 7 iterations. This suggests setting the maximum number of iterations to 10 in order to ensure a good result. Our BO tuning algorithms only take advantage of multiple processors for decoding, thus there still exists some space to further improve their efficiency.

Fig. 3a and 3b indicate the comparison of development score and BO score[5] at each iteration in *fr-en* and *cs-en*, which again demonstrates the advantage of CHG-BO on stability over NBL-BO and HG-BO. Fig. 3c and 3d compare the models with different bound size: $b = 0.01$ is able to achieve a development and test BLEU score as good as $b = 0.1$ with more iterations, but $b = 0.5$ performs worse on the test dataset. Thus too large search bound may introduce too much noise which in turn affects the translation performance.

Table 3 shows the experiments on a large number of sparse features. We modify HG-REMBO into a two step coordinate ascent processes in order to stabilise the updates of the core default feature weights. First, we optimise the default 18 features, then we fix them and generate a regularised random matrix to update the large scale sparse features in the low dimensional space. Table 3 demonstrates that HG-REMBO is able to carry out large scale discriminative training, performing almost on par with MIRA. Although HG-REMBO loses its advantage on speed of convergence as it requires multiple runs to generate a good transformation matrix, these results illustrate the potential of applying REMBO on statistical machine translation systems.

4 Conclusion

We introduce novel Bayesian optimisation (BO) algorithms for machine translation. Our algorithms exhibit faster convergence and achieve higher training objectives and better translation quality than existing translation model specific approaches. We further demonstrate that by incorporating the method of random embeddings it is viable to employ Bayesian optimisation to carry out large sale training with a high number of sparse features. This initial investigation also suggests that BO has great potential for general natural language processing tasks.

5 Acknowledgements

This work was supported by a Xerox Foundation Award and EPSRC grant number EP/K036580/1.

[5]BO score is the best BLEU score achieved by Gaussian processes on fixed N-best lists or hypergraphs.

References

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[2] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 295–302, Association for Computational Linguistics, 2002.

[3] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 160–167, Association for Computational Linguistics, 2003.

[4] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *The Journal of Machine Learning Research*, vol. 3, pp. 951–991, 2003.

[5] D. Chiang, "Hope and fear for discriminative training of statistical translation models," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1159–1187, 2012.

[6] E. Brochu, V. M. Cora, and N. De Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *arXiv preprint arXiv:1012.2599*, 2010.

[7] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

[8] W. Macherey, F. J. Och, I. Thayer, and J. Uszkoreit, "Lattice-based minimum error rate training for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 725–734, Association for Computational Linguistics, 2008.

[9] S. Kumar, W. Macherey, C. Dyer, and F. Och, "Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 163–171, Association for Computational Linguistics, 2009.

[10] L. Huang, "Advanced dynamic programming in semiring and hypergraph frameworks," *COLING, Manchester, UK*, 2008.

[11] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. De Freitas, "Bayesian optimization in high dimensions via random embeddings," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1778–1784, AAAI Press, 2013.

[12] C. Dyer, J. Weese, H. Setiawan, A. Lopez, F. Ture, V. Eidelman, J. Ganitkevitch, P. Blunsom, and P. Resnik, "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models," in *Proceedings of the ACL 2010 System Demonstrations*, pp. 7–12, Association for Computational Linguistics, 2010.

[13] D. Chiang, K. Knight, and W. Wang, "11,001 new features for statistical machine translation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 218–226, Association for Computational Linguistics, 2009.

[14] P. Simianer, S. Riezler, and C. Dyer, "Joint feature selection in distributed stochastic learning for large-scale discriminative training in smt," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 11–21, Association for Computational Linguistics, 2012.

[15] A. Krause and C. S. Ong, "Contextual gaussian process bandit optimization," in *Advances in Neural Information Processing Systems*, pp. 2447–2455, 2011.