The Challenges of Optimizing Machine Translation for Low Resource Cross-Language Information Retrieval

Constantine Lignos[1], Daniel Cohen[2], Yen-Chieh Lien[2], Pratik Mehta[1], W. Bruce Croft[2], Scott Miller[1]
[1]Information Sciences Institute, University of Southern California
[2]Center for Intelligent Information Retrieval, University of Massachusetts Amherst
lignos@brandeis.edu
{dcohen,ylien,psmehta,croft}@cs.umass.edu,
smiller@isi.edu

Abstract

When performing cross-language information retrieval (CLIR) for lower-resourced languages, a common approach is to retrieve over the output of machine translation (MT). However, there is no established guidance on how to optimize the resulting MT-IR system. In this paper, we examine the relationship between the performance of MT systems and both neural and term frequency-based IR models to identify how CLIR performance can be best predicted from MT quality. We explore performance at varying amounts of MT training data, byte pair encoding (BPE) merge operations, and across two IR collections and retrieval models. We find that the choice of IR collection can substantially affect the predictive power of MT tuning decisions and evaluation, potentially introducing dissociations between MT-only and overall CLIR performance.

1 Introduction

For cross-language information retrieval scenarios involving queries in a higher-resourced language and documents in a lower-resourced language, direct training of a cross-language IR system (Litschko et al., 2018; Sasaki et al., 2018; Vulić and Moens, 2015) is typically infeasible due to insufficient data in the document language.

A practical solution is to use machine translation to translate documents into the language of the queries, enabling the use of a traditional monolingual IR system trained on the higher-resourced language. Prior work (Pecina et al., 2014; Nikoulina et al., 2012) demonstrates the effectiveness of using an MT system for query translation within a CLIR framework, but increases in MT performance measured by BLEU (Papineni et al., 2002) do not necessarily correlate with IR performance. Taking a document translation approach, Hieber and Riezler (2015) show that by enforcing a bag of words constraint on the MT decoding process, a translated document can provide richer input to an MT-IR system. However, these prior works do not fully investigate the effectiveness of MT metrics in predicting performance of a downstream IR model.

In this paper, we explore the consequences of performing information retrieval over machine translation and address three questions regarding the performance of the resulting system. First, when making MT tuning decisions such as selecting the number of BPE merge operations, will a value that improves MT performance generally increase IR performance? Second, when evaluating MT performance, can standard metrics be adapted to better correlate with downstream IR performance? Finally, what is the general relationship between the performance of MT and a downstream IR task, and how is it affected by the choice of retrieval model?

2 Experiment Design

To address the above questions, we trained an MT system in a large number of configurations, used these models to produce translated English collections of varying quality, and then compared the performance of retrieval over translated collections to retrieval over the matching source English documents. While an MT-IR system of this type is most appropriately used on lower-resourced languages, the resources needed to perform such a study using publicly available data thus far only exist in higher-resourced languages. To simulate a lower-resourced setting, we used a small portion of the MT training data available from higher-resourced languages, ablated it into smaller subsets, and did not rely on any other language resources. By training translation models using varying amounts of training data, we are able to explore performance across a wide range of resource levels and examine

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3497–3502,
Hong Kong, China, November 3–7, 2019.

the reliability of performance patterns. The software and data required to replicate the experiments reported here are available from https://github.com/ConstantineLignos/mt-clir-emnlp-2019.

2.1 Machine Translation

We trained German-English and Czech-English translation models using the News Commentary (NC) dataset version 13 (Tiedemann, 2012) as provided in the WMT18 training data. This dataset was chosen to ensure no overlap between MT training data and the Wikipedia and Europarl IR test sets. Up to 200k sentences were selected for training, 5k for validation, and 5k for testing. The training data was ablated from 200k to 25k sentences in increments of 25k to create 8 training sets of decreasing sizes, simulating a wide range of low-resource conditions. The MT training scenarios varied by language, training data size, and BPE configuration. For each training set, a joint (source and target languages) BPE (Sennrich et al., 2016) was learned using two different numbers of merge operations (hereafter *BPE size*), 16k and 32k.

We selected fairseq (Gehring et al., 2017) as the machine translation system for our experiments. The MT system was trained using the data preprocessing approach and parameters given in the documentation[1]. Moses was used for tokenization of all data. We used the `fconv_iwslt_de_en` fully convolutional architecture with a single set of hyperparameters that provided for robust convergence on all configurations: learning rate 0.25, gradient clip threshold (parameter `clip-norm`) 0.1, dropout 0.2, and max-tokens 4000.[2] The validation set was provided during training to identify stopping conditions. The epoch that yielded the highest validation set performance was used to decode the test data. Test data was decoded with a beam size of 5. BLEU scores were computed using `sacreBLEU` (Post, 2018), with the configuration *BLEU+case.lc+numrefs.1+smooth.none+tok.13a+version.1.2.12*.

[1] https://fairseq.readthedocs.io/en/latest/getting_started.html#training-a-new-model

[2] We also experimented with a transformer-based architecture. While it did produce BLEU scores several points higher for the highest data conditions, it did not reliably converge with smaller amounts of training data, typically producing single-digit BLEU scores. While further tuning would have likely addressed this issue, we opted to report results using the fully convolutional architecture and constant hyperparameters.

2.2 Information Retrieval

We selected two multilingual collections for retrieval experiments, Europarl and Wikipedia. We created a collection of the 2,330 non-empty Europarl V7 (Koehn, 2005) transcripts that were present in English, German, and Czech.[3] As there are no queries and relevance judgments specific to Europarl, we used GOV2 TREC topics 701–850 as domain-relevant English queries and treated the top 100 documents retrieved by BM25 for each query as the relevant documents for evaluation.

Wikipedia was selected to provide a second collection for comparison, specifically one large enough to support training a neural retrieval model. To create relevant query-document pairs for training and evaluation, we used titles as queries and treated each document as the sole relevant document for its own title (Sasaki et al., 2018). For evaluation, we selected 5k articles with content in all three languages. We constrained selection such that the shortest article across the three languages had at least 500 words, and the longest article was not more than three times as long as the shortest article. While the information content for the "same" article will not be identical across languages, this constraint helps limit cross-language information differences caused by article length.

Models. As *tf.idf* is still a competitive performer compared to neural models for ad-hoc retrieval (Guo et al., 2019), we evaluated both models to provide insight into how each responds to the noise introduced by MT. We used Okapi-BM25 (Jones et al., 2000) to represent a term-based approach, and the Duet architecture introduced by Mitra et al. (2017) to represent a neural model. The Duet architecture was chosen because it transforms the document's text into the most common character n-grams within a corpus, making it robust to subword translation errors that are likely to be caused by using BPE in a low resource setting. Duet has been shown to perform comparably to more recent models both for ad-hoc and passage length retrieval (Mitra and Craswell, 2019).

The Duet model was trained on English

[3] Europarl transcripts were considered empty and excluded from the collection if for any language they were absent or appeared to have no content, such as merely noting that a session did or did not occur without containing any session content. The method for identifying trivial transcripts was that any lines consisting of only whitespace, XML, or a parenthetical were removed. If fewer than three lines remained, the transcript was considered empty and excluded.

Wikipedia using the hyperparameters and loss function suggested by Mitra et al. (2017) rather than weakly supervised BM25 scores (Dehghani et al., 2017), which would produce a smoothed *tf.idf* neural BM25 model that behaves differently than a standard neural IR model. We also experimented with a word-based representation—using GLoVE (Pennington et al., 2014) embeddings—as input to Duet instead of using the most common character n-grams. It performed worse (highest RBO .091), likely due to being more sensitive to unusual subword unit combinations generated by the MT system.

Evaluation. The goal of our IR evaluation is to compare the ranking produced by retrieval over translated documents against that produced by retrieval over the English documents. We used the extrapolated rank-biased overlap (RBO, Webber et al., 2010) as our primary metric, defined as

$$RBO_{EXT}(S,T,p,k) = \frac{X_k}{k}\cdot p_k + \frac{1-p}{p}\sum_{d=1}^{k}\frac{X_d}{d}\cdot p_d$$

where k is the current position in a rank list, S, T are the two rank lists being compared, and X_i is the size of the intersection of the two lists at depth i. This results in an RBO value of 1 being the upper bound when the rank list S is the same as T. RBO is based on a probabilistic user model, facilitates indefinite rankings, and allows for incomplete sets to be compared, unlike Kendall's tau and Spearman's rho. We used the suggested parameter $p = 0.98$, where the top 50 documents of each rank list contain 86% of the weight. We also computed mean average precision (MAP) as a standard IR metric for comparison.

The BM25 model was evaluated against both the Europarl and Wikipedia collections. However, to avoid the performance degradation caused by cross-collection evaluation (Cohen et al., 2018), we only evaluate the Wikipedia-trained neural model on the Wikipedia evaluation collection. Due to the large number of conditions we evaluate for both MT and IR performance, results are shown in figures, while full tables are provided in Appendix A.1.

3 Results

Unsurprisingly, using larger amounts of MT training data leads to higher MT and end-to-end performance; we focus on the subtleties of performance and not this generality. Figure 1 shows lowercase

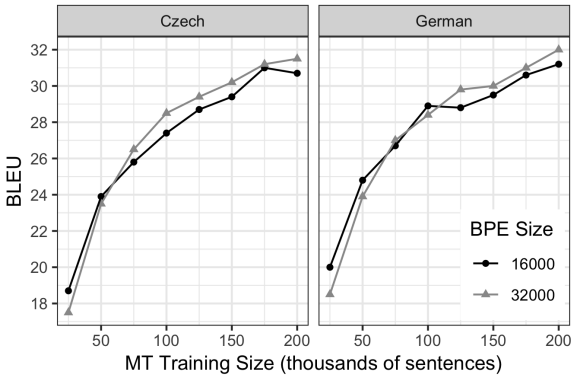

Figure 1: MT test BLEU for all training configurations.

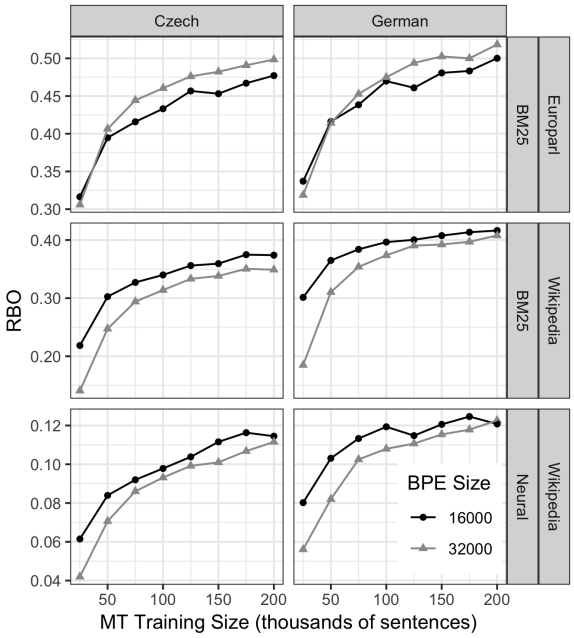

Figure 2: RBO for retrieval models on Europarl and Wikipedia across all MT training configurations.

BLEU on the MT test set for all training conditions; BLEU scores ranged 17.5–31.5 for Czech-English and 18.5–32.0 for German-English. Figure 2 shows RBO for retrieval over the output of each MT training configuration. While the relationship between MT training size and RBO is similar across languages, collections, and models, the performance of neural retrieval over Wikipedia is far lower than that of BM25, which we discuss below.

Collection/Model	Match	Mismatch
Europarl BM25	14	2
Wiki BM25	5	11
Wiki Neural	6	10

Table 1: Counts of MT/IR performance tuning matches and mismatches when tuning BPE size.

3.1 Impact of BPE

We now address our first question: when making MT tuning decisions such as selecting the number of BPE merge operations, will a value that improves MT performance generally increase IR performance? We explore tuning BPE because it is commonly used in modern neural MT systems, and it is likely to have substantial end-to-end impact as it controls the subword input and output vocabulary. To address this question, we analyzed how often the directionality of the change in BLEU matched that of RBO when comparing across BPE size conditions. We categorized each configuration as a *match* if changing BPE size from 16k to 32k had an effect in the same direction for both MT (BLEU) and IR (RBO) performance, and categorized it as a *mismatch* otherwise. As shown in Table 1, the collection determined the probability of matches; for Europarl, BPE tuning decisions that increase BLEU reliably increase RBO, but for Wikipedia, more often than not tuning BPE to increase BLEU will *decrease* RBO.

Examining the MT and IR performance separately makes the cause of these tuning mismatches more clear. For MT, Figure 1 shows that 16k BPE performed slightly better at the lowest training data points, and 32k performed better at the highest. For IR, the effect of BPE varies across collections. On Europarl, the effect of BPE code size on IR performance is similar to that of MT. Increasing BPE size from 16k to 32k generally leads to better performance, with a mean MAP increase of .010 across all sizes. On Wikipedia, the 32k BPE size almost always performs worse, with mean MAP decreases of .056 (BM25) and .010 (neural). The difference between the MAP scores across the BPE configurations was statistically significant as determined by a Wilcoxon signed-rank test, a non-parametric paired sample test: Europarl BM25 ($p = .023$); Wikipedia BM25 ($p < .001$) and neural ($p < .001$).

We conclude that the nature of the queries, documents, and relevance characteristics of the downstream IR task plays a critical role when selecting BPE size or making other MT tuning decisions. Depending on the collection, attempts to improve MT may end up hurting end-to-end performance.

3.2 Evaluating MT for End-to-end Performance

We now turn to our second question: when evaluating MT, can standard metrics be adapted to better correlate with downstream IR performance? To explore this question, we computed correlations using Kendall's tau between RBO and the following: BLEU, BLEU's component 1–4-gram precisions, and variations of BLEU that remove punctuation, stem words using the Porter stemmer, or do both (as is common in term-based retrieval). Tau was computed separately within each combination of language, dataset, and retrieval model, measuring the correlation with RBO across changes in MT performance due to language, training data size, and BPE size.

Coll./Model	BLEU	P1	P2	P3	P4
Europarl BM25	.873	**.881**	.867	.864	.862
Wiki BM25	**.722**	.689	.708	.714	.711
Wiki Neural	**.789**	.756	.775	.781	.778

Table 2: Mean Kendall's Tau correlation with RBO for lowercased BLEU and each n-gram precision (P1=1-gram, etc.), by collection/model.

Coll./Model	BLEU	No Punc.	Stem	Both
Europarl BM25	.873	.870	**.875**	.873
Wiki BM25	.722	**.728**	.717	.722
Wiki Neural	**.789**	.787	.783	.781

Table 3: Mean Kendall's Tau correlation with RBO for lowercased BLEU and variations, by collection/model.

Table 2 shows that the downstream IR task affects which MT measures correlate best with RBO. For Europarl, the best correlation with RBO was achieved by unigram precision, with correlation decreasing as larger n-grams are evaluated. For Wikipedia, BLEU provided the best correlation with RBO, and larger n-gram precisions correlated better with RBO than unigram precision. Table 3 gives correlations for the other variations of BLEU we explored; none led to a reliable improvement over standard lowercased BLEU, but results differed across collections/models.

We conclude that none of the n-gram precision components of BLEU or variations on it provide consistently better correlations with IR performance. However, given a specific collection and model, it is likely one of the alternatives we explored here or other metrics (e.g. ROUGE) can be slightly more predictive than standard BLEU.

3.3 IR Model Sensitivity

Finally, we explore our third question: what is the general relationship between the performance of MT and a downstream IR task, and how is it affected by the choice of retrieval model? Figure 3

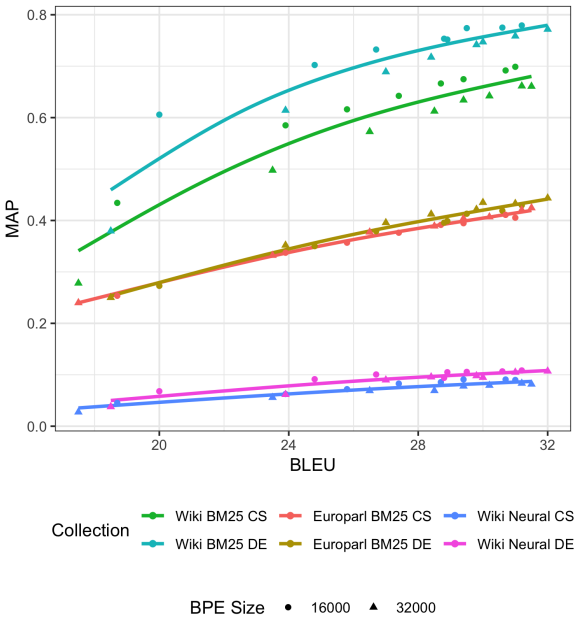

Figure 3: MAP for retrieval models across all collections and MT training configurations. GAM fit lines are provided for each combination of collection, retrieval model, and language (CS, Czech; DE, German).

shows MAP and BLEU across all tested configurations. The relationship between them is approximately linear above the lowest data points. Using linear regression to predict MAP from BLEU while accounting for variance due to language and BPE size, a one-point increase in BLEU was estimated to yield 1.33 points of MAP for Europarl BM25, 2.39 for Wikipedia BM25, and 0.42 for Wikipedia neural.

The neural Wikipedia model (highest RBO 0.12; highest MAP 0.11) significantly trailed BM25 (0.42; 0.78). Low RBO values show that the neural model degraded much more than BM25 when presented with MT as input. Given the regression described above, we can explore the BLEU score needed to match the MAP value of 0.20 that the same model attained retrieving over the English articles. Extrapolating generously, that would be achieved with a BLEU of approximately 53, a value unattainable in a low-resource scenario. As neural IR models are sensitive to shifts in the input distribution (Cohen et al., 2018), the most likely path for effective neural retrieval in a low-resource MT-IR setting is to train retrieval in a way that adapts to the idiosyncrasies of MT output, unlike the model we evaluated.

4 Conclusions

We conclude that there is no substitute for end-to-end, task-specific tuning when attempting to improve MT-IR system performance by increasing MT quality. As demonstrated by the effect of BPE tuning, changes made to the MT model can have opposite effects on performance when different IR collections are considered. The performance degradation of the evaluated neural IR model when retrieving over MT suggests that it is unlikely that off-the-shelf neural IR will be able to function adequately in an MT-IR system. Adapting neural models for use in an MT-IR setting and addressing the properties of MT output most harmful to their performance are promising avenues for future work to enable low-resource CLIR.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Daniel Cohen, Bhaskar Mitra, Katja Hofmann, and W. Bruce Croft. 2018. Cross domain regularization for neural ranking models using adversarial learning. In *The 41st International ACM SIGIR Conference on Research #38; Development in Information Retrieval*, SIGIR '18, pages 1025–1028.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 65–74.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of ICML*.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *Information Processing & Management*.

Felix Hieber and Stefan Riezler. 2015. Bag-of-words forced decoding for cross-lingual information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1182.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing & Management*, 36(6):779–808.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86.

Robert Litschko, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 1253–1256.

Bhaskar Mitra and Nick Craswell. 2019. An updated duet model for passage re-ranking. *arXiv preprint arXiv:1903.07666*.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1291–1299.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 109–119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth JF Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, et al. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165–185.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 458–463.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Workshop abstracts: Eighth International Conference on Language Resources and Evaluation*.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 363–372.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38.