Combining Contextualized and Non-contextualized Query Translations to Improve CLIR

Suraj Nair
srnair@cs.umd.edu
University of Maryland

Petra Galuscakova
petra@umiacs.umd.edu
University of Maryland

Douglas W. Oard
oard@umd.edu
University of Maryland

ABSTRACT

In cross-language information retrieval using probabilistic structured queries (PSQ), translation probabilities from statistical machine translation act as a bridge between the query and document vocabulary. These translation probabilities are typically estimated from a sentence-aligned corpus on a word to word basis without taking into account the context. Neural methods, by contrast, can learn to translate using the context around the words, and this can be used as a basis for estimating context-dependent translation probabilities. However, sparsity limits the accuracy of context-specific translation probabilities for rare words, which can be important in retrieval applications. This paper presents evidence that combining such context-dependent translation probabilities with context-independent translation probabilities learned from the same parallel corpus can yield improvements in the effectiveness of cross-language ranked retrieval.

CCS CONCEPTS

• **Information systems** → **Combination, fusion and federated search**; **Multilingual and cross-lingual retrieval**.

KEYWORDS

CLIR, machine translation

ACM Reference Format:
Suraj Nair, Petra Galuscakova, and Douglas W. Oard. 2020. Combining Contextualized and Non-contextualized Query Translations to Improve CLIR. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401270

1 INTRODUCTION

Cross-Language Information Retrieval (CLIR) is an information retrieval problem in which documents are in different language than the queries. To be able to apply the retrieval methods in such setup, queries and documents need to be first transferred into a common space, which is often done by applying translation techniques either on the queries or on the documents. Documents are typically much richer in context than queries, which can be better utilized by traditional machine translation systems. Zbib et al. [16] introduce a Neural-Network Lexical Translation Model (NNLTM) which uses the contextual information in the documents to produce translations that are matched with the query terms. They show that their model which utilizes several probable translations jointly with their probabilities outperforms one-best document translations of the documents either created by an automatic translation system or by a human translator. In this work, we first focus on further improvement of the Zbib et al. models. Next, we investigate joint approaches which combine NNLTM and Probabilistic Structured Queries (PSQ) [3]. PSQ only uses information about the query words without any context, but similarly to NNLTM, it also provides several translation alternatives together with their probabilities. As both these models provide multiple translation possibilities which might also include numerous noisy translations, combining them may help to mitigate the noise.

There are two main contributions in this paper: 1) we further improve the model introduced by Zbib et al. and test it on new evaluation collections, and 2) we further outperform that model alone when we combine the documents retrieved by this system with the documents returned by PSQ using post-retrieval and in-retrieval evidence combination methods. We show these further improvements from combination on three test collections of moderate size, two of which were not available to Zbib et al.

2 METHODS

This section describes details of the used translation approaches and applied combination techniques.

2.1 Contextual translation

As compared to ambiguous short queries, documents have more context available, and that context is typically leveraged by neural machine translation (NMT) systems to generate contextualized translations of document terms. Typically, in a sequence-to-sequence model, the target words are generated in a sequential manner using the source context and the previous target words. Devlin et al. [4] uses a similar approach to produce contextualized word translation probabilities conditioned on source and target word context. Later Zbib et al. extended the work to produce word translations conditioned only on the source word context, which were used to perform CLIR. In this work we use their NNLTM which trains on the alignment output from word aligners to estimate contextual word translation probabilities. Specifically, for an aligned word pair $f_i \leftrightarrow e_j$, it uses a contextual window of k terms around the source word f_i to predict the target word e_j. NNLTM consists of an embedding layer that maps the $2k + 1$ source words to separate embeddings which are concatenated and fed to a single feedforward layer. The final layer produces a softmax distribution

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401270

	SW		SO		LT	
	Val	Eval	Val	Eval	Val	Eval
# query	300	1000	300	1000	300	1000
# doc	996	10435	1041	10717	1047	10203
# rel/query	1.89	15.66	2.42	13.46	2.58	12.96
Avg. Doc. Len.	384.28	407.90	315.82	370.70	373.49	403.15

Table 1: CLIR test collection statistics.

	SW	SO	LT
#sentences	257k	88k	732k
#source words	1,706k	1,678k	11,864k
#target words	1,795k	1,665k	16,149k

Table 2: Parallel corpus statistics

of *contextual* probabilities $P(e_j|f_{i-k}..f_i..f_{i+k})$ over the target vocabulary and the model is optimized using cross-entropy loss using one-hot representation for the target word e_j

2.2 Non-contextual translation

We use the PSQ approach [3] to estimate term counts for a query q in a document d using the translation probabilities of a query word given the document terms. These translation probabilities are generated using a statistical machine translation (SMT) model trained on parallel text. We call these probabilities *non-contextual* since they are learnt without taking into account either the source or the target context.

2.3 Evidence Combination

Our goal in this work is to find if the performance of the context-dependent system can be further improved by a combination with a context-independent translation systems. To answer that, we specifically explore two types of evidence combination: *post-retrieval* and *in-retrieval* system combination.

The post-retrieval system combination is a method for combining the sets of documents acquired by different systems. In this work, we make use of *CombMNZ* [5], a widely used data fusion method which utilizes the scores of the documents returned by these systems.

CombMNZ uses the sum of document scores produced by different retrieval systems and multiplies it by a parameter n_d denoting the number of systems that marks the document d as relevant to the query:

$$sc^{CombMNZ}(d) = n_d * \sum_{i=1}^{n} sc^{s_i}(d) \quad (1)$$

Thanks to this, CombMNZ promotes the documents which are returned by multiple systems and can be thus expected to be especially helpful in our setup.

In contrast to post-retrieval combination which combines information about documents, in-retrieval system combination refers to a combination in which word translation probabilities are directly combined at query time. Application of this type of combination is straightworward in our setup where we have two translation approaches, each providing n-best translations with assigned probabilities. A variant of CombMNZ is used in this case:

$$sc^{CombMNZ}(w) = n_w * \sum_{i=1}^{n} sc^{s_i}(w) \quad (2)$$

for the combination of context-dependent and context-independent translation probabilities. $sc^{s_i}(w)$ is the probability of the translation of the word w by the i^{th} system, and n_w is a count of translation evidence sources that include the word. This helps in augmenting the computed weight for a translation that occurs in multiple sources. The combined weights are then used in the retrieval model explained later in section 3.4.

3 EXPERIMENT SETUP

This section describes details of the test evaluation setup, training data and the CLIR models used.

3.1 Test Collection

To evaluate our methods, we use CLIR test collections for three low-resource languages, Swahili (SW), Somali (SO) and Lithuanian (LT). These collections were created as a part of the IARPA MATERIAL program[1]. The queries are in English and the documents are either in Swahili, Somali, or Lithuanian. Though the MATERIAL collection also contains audio recordings, we only use text documents for the experiments in this paper. See Table 1 for collection statistics.

The MATERIAL queries are divided into three main types: simple, conceptual and hybrid. The description of the queries can be found in [15]. The Validation (Val) set is the union of the IARPA DEV, ANALYSIS1 and ANALYSIS2 collections and the queries from IARPA query set Q1 are applied to this set. The Evaluation (Eval) documents are the union of the IARPA EVAL1, EVAL2 and EVAL3 collections and the queries from query sets Q2 and Q3 are applied to these documents.

3.2 Training setup for PSQ and NNLTM

Our approach for obtaining the parallel texts and training the neural model closely follows the approach by Zbib et al. Both the NNLTM and the PSQ models are trained on the same parallel data for a given language pair. For MATERIAL languages, we use the bitext available in the IARPA BUILD pack which contains roughly 25-44k parallel sentences. In addition, we collect available parallel texts from OPUS[2] for Swahili and Somali. For Lithuanian, we use sentences from Europarl.[3] We also mine dictionaries from Panlex[4] and Wiktionary[5] and append them to the parallel corpora. The parallel data is tokenized using Moses [7] toolkit, lowercased and preprocessed to strip any punctuation, digits, and accents from the characters. The preprocessed data is then used to train GIZA++ [10] and the Berkeley Aligner [6]. The alignment outputs from the two aligners are concatenated to estimate a unidirectional lexical translation probabilities (i.e., the probability of an English query word given a non-English document word). Table 2 lists the size of the

[1] https://www.iarpa.gov/index.php/research-programs/material/material-baa
[2] http://opus.nlpl.eu/
[3] https://www.statmt.org/europarl/
[4] https://panlex.org/snapshot/
[5] https://dumps.wikimedia.org/

Model	MAP
Word-level NNLTM (Zbib et al.)	0.263
Word-level NNLTM replicate	0.246
+*top50*	0.252
+*min_tf* +*stopwords*	0.259
+*label smoothing*	0.266

Table 3: Results of different NNLTM model versions on Swahili Eval dataset used in Zbib et al.

parallel corpus used to generate the word alignments. Two translation tables are generated; unstemmed foreign word to unstemmed English word and unstemmed foreign word to stemmed English word.

NNLTM is first trained on unstemmed word alignments using the same hyperparameters as the ones used in the Zbib et al. paper. Based on their experiments, the source context window size is set to 1. The model is trained for 20 epochs with a batch size of 512 using Adam optimizer and a learning rate of 0.001. The dropout probability is set to 0.8 and the source vocabulary size is restricted to 30,000 most frequent tokens. At the inference time, for a given document term and the context surrounding it, we store the top-10 NNLTM's output words and their contextualized probabilities.

3.3 Improvements to NNLTM

We experiment with changing several hyperparameters associated with the NNLTM that might affect the retrieval performance. The number of contextualized translations stored is increased from 10 to 50 (*top 50*). In addition to that, we remove samples from training that have either English *stopwords* as the target or occur less than 5 times (*min_tf*). We also employ a regularization technique called *label smoothing* [11] which avoids the model to get too overconfident in its predictions. This technique involves smoothing the one-hot target labels with a uniform distribution over the target vocabulary size. These smoothed target labels are then used to train the model. We use 0.1 as the value for label smoothing parameter. The effects of these changes are further analyzed in Section 4.1

3.4 Retrieval model

For testing our combination approaches, we use a modified version of PSQ-based HMM model [9] to perform retrieval. The foreign state of the HMM model is replaced with the Probabilistic Term Occurrence model (PTO) [16], as shown in Equation 3. PTO models the relevance assumption for simple queries in the MATERIAL collections, which are far more common than conceptual queries in those collections. To do this requires finding relevant documents that contain the translation of query terms at least once in the document of interest. Assuming a query term q consists of N terms $t_1 ....t_N$, the document relevance probability is modeled as

$$p(q|doc) = \prod_{n=1}^{N} \left[\alpha P(t_n|\theta_e) + (1-\alpha) \left(1 - \prod_{f \in doc} (1 - p(t_n|f)) \right) \right] \tag{3}$$

where θ_e represents the HMM state which produces English words. This is a back-off unigram language model which is estimated from a large English corpus, Google's One Billion Word collection [1]. We use $\alpha = 0.1$. The query term counts t_n in document d are generated from the foreign words f which are mapped to t_n using the translation probabilities $P(t_n|f)$. Further, PTO models these counts between 0 (no translation occurred) to 1 (multiple translation occurs). These counts can be computed by either using the context-dependent or independent translation probabilities. In an *in-retrieval* combined system, these counts generated from context-dependent and independent translation probabilities are combined together using the CombMNZ fusion as explained in Section 2.3. Alternatively, these individual systems can be combined together after the retrieval is performed.

4 RESULTS

We first describe the comparison of our basic and enhanced model with the results achieved by Zbib et al. and then combine the enhanced NNLTM model with the PSQ approach using two described combination methods.

4.1 NNLTM model enhancement

Results from all of our changes in the original NNLTM model are summarized in the Table 3. For ease of comparison, we ran these experiments (but not the system combination experiments below) on the same setup as the one used in Zbib et al. (the IARPA Eval set, with query sets Q1 and Q3) for Swahili.[6] The Zbib et al. model is compared with our replicated model which uses the same configuration as described in the paper. However, the models differ in the training data described in Section 3.2, which, we believe, causes the differences between the original and replicated model. Specifically, we do not have access to the LORELEI [2] Swahili data used to train the original model. However, the replicated model with the described features (see Section 3.3) outperforms the original model, and thus we believe that the model with the same training data would outperform their original model as well.

4.2 System combination

The enhanced NNLTM model was further combined with the PSQ non-contextual translations using both post-retrieval and in-retrieval combination methods. The enhanced model achieving a Mean Average Precision (MAP) of 0.266 in the Table 3 on the Q1+Q3 query sets is achieving a MAP of 0.272 on our Eval set with Q2+Q3 (i.e., the same documents, but some different queries). The results for individual systems and their combinations are shown in Table 4.

As the stemmed PSQ system in most cases outperforms its unstemmed version, stemmed PSQ system is used in our experiments. This stemmed PSQ system is outperforming the NNLTM model

[6] We only use Swahili results as it is the only Eval set used in [16].

Model	Comb	EN→SW		EN→SO		EN→LT	
		Val	Eval	Val	Eval	Val	Eval
NNLTM	-	0.362	0.272	0.277	0.165	0.442	0.304
PSQ	-	0.368	0.252	0.267	0.147	0.534	0.359
PSQ+NNLTM	Post	0.373	**0.282***	**0.298**	**0.172***	0.540	0.381*
	In	**0.375**	0.275*	0.294	0.169*	**0.561***	**0.393***

Table 4: Mean Average Precision of the NNLTM and PSQ models on the MATERIAL Val and Eval collections. Bold indicates best results per column, * indicates statistically significant improvement over both single systems in the combination. Two-tailed Wilcoxon signed rank test with p<0.01 is applied.

on Lithuanian, which is the language with the most available resources among the tested languages. NNLTM is outperforming PSQ on Swahili and Somali, except on the very small Swahili Val set. The post-retrieval combination on each Eval collection significantly outperforms the individual systems which are being combined, with the largest differences achieved on the Lithuanian. In the case of the Lithuanian and Swahili Val sets, in-retrieval combination further improves these results. The strong performance of both combination methods confirms our assumption that the noise which can emerge as the systems work with n-best possible translations can be effectively surpassed by combining multiple such systems.

5 RELATED WORK

Zhou et al. [19] provides an excellent survey on the translation techniques mainly used to perform cross language retrieval. The preferred approach is to translate queries to the document language and perform monolingual retrieval. Early works [8] though point to the evidence that it is often better to build hybrid system that involve both query and document translation. Recent works [14, 17, 18] involve neural approaches to creating document representations used to perform retrieval. Türe et al. [13] combines the query translations from two complementary systems, PSQ and hierarchical phrase-based translation system. The choice of combination is based on a linear interpolation of probabilities. Using supervised learning methods to learn optimal set of combination weights for each query separately has also proven to be effective [12]. In contrast to these works, we focus on combining translation learned from parallel text with a contextualized neural system using well-known data fusion method in a low-resource language setting.

6 CONCLUSION AND FUTURE WORK

This paper proposes the combination of two complementary sources of evidence: contextualized word translations produced by a neural lexical translation model and non-contextualized translation generated by a statistical machine translation system. We demonstrate the effectiveness of our approach through two combination techniques: in-retrieval and post-retrieval, both of which produce statistically significant improvements over the individual systems. Using our approach, MAP gains of 4%, 4% and 9% are observed in the Eval collections for three low-resource language collections, Swahili, Somali and Lithuanian respectively. We also show that the NNLTM can be tuned further to improve CLIR performance through regularization technique and careful hyperparameter selection.

In the future, we plan to test both the NNLTM and the combination approach on CLEF bilingual ad-hoc retrieval collections. We also plan to perform broad range of tuning experiments for the NNLTM hyperparameters taking into account the OOV rate and document terms coverage.

ACKNOWLEDGMENTS

This research has been supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies of ODNI, IARPA, or the U.S. Government.

REFERENCES

[1] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, et al. 2013. One billion word benchmark for measuring progress in statistical language modeling. (2013). arXiv:1312.3005
[2] C. Christianson, J. Duncan, and B. Onyshkevych. 2018. Overview of the DARPA LORELEI Program. *Machine Translation* (2018).
[3] K. Darwish and D. W. Oard. 2003. Probabilistic structured query methods. In *SIGIR*.
[4] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, et al. 2014. Fast and robust neural network joint models for statistical machine translation. In *ACL*.
[5] E. A. Fox and J. A. Shaw. 1994. Combination of multiple searches. In *TREC*.
[6] A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. 2009. Better word alignments with supervised ITG models. In *ACL&AFNLP*.
[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.
[8] J. S. McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval?. In *ACL*.
[9] D. R. Miller, T. Leek, and R. M. Schwartz. 1999. BBN at TREC7: Using hidden Markov models for information retrieval. (1999).
[10] F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* (2003).
[11] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. (2017). arXiv:1701.06548
[12] F. Türe and E. Boschee. 2014. Learning to translate: a query-specific combination approach for cross-lingual information retrieval. In *EMNLP*.
[13] F. Türe, J. Lin, and D. W. Oard. 2012. Combining statistical translation techniques for cross-language information retrieval. In *COLING*.
[14] M. Yarmohammadi, X. Ma, S. Hisamoto, M. Rahman, Y. Wang, et al. 2019. Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings. In *EACL*.
[15] I. Zavorin, A. Bills, C. Corey, M. Morrison, A. Tong, et al. 2020. Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages. In *LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech.*
[16] R. Zbib, L. Zhao, D. Karakos, W. Hartmann, J. DeYoung, et al. 2019. Neural-network lexical translation for cross-lingual IR from text and speech. In *SIGIR*.
[17] R. Zhang, C. Westerfield, S. Shim, G. Bingham, A. Fabbri, et al. 2019. Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations. In *ACL*.
[18] L. Zhao, R. Zbib, Z. Jiang, D. Karakos, and Z. Huang. 2019. Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval. In *DeepLo*.
[19] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman. 2012. Translation techniques in cross-language information retrieval. *CSUR* (2012).