Simulating CLIR Translation Resource Scarcity using High-resource Languages

Hamed Bonab, James Allan, and Ramesh Sitaraman
College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA 01003
{bonab,allan,ramesh}@cs.umass.edu

ABSTRACT

We study the impact of translation resource scarcity on the performance of cross-language information retrieval (CLIR) systems. To do that, we develop a contrastive analysis framework that uses high-resource languages to simulate low-resource languages. In the framework, we focus on parallel translation corpora and aim to better understand the factors that impact CLIR performance. We argue that both low- and high-resource corpora are needed to develop that understanding. Hence, we take the approach of starting with a true low-resource language and systematically down-sampling a high-resource language to become an artificial low-resource language—the reverse perspective of existing research. We formalize the problem as the *Resource Scarcity Simulation (RSS)* problem. We model the problem with a family of set covering problems, formulate with integer linear programming, and prove that the problem is actually **NP**-hard. To this end, we provide two greedy algorithms with polynomial complexities. We compare and analyze our approach with alternate techniques using four high-resource languages (French, Italian, German, and Finnish) down-sampled to simulate two low-resource languages (Somali and Swahili). Our experimental results suggest that language families are important for the RSS problem. We simulate Somali with German, and Swahili with Finnish, achieving 98% and 97% on the similarity percentage in terms of CLIR performance, respectively.

KEYWORDS

Low-resource languages; translation resources; language simulation

ACM Reference Format:
Hamed Bonab, James Allan, and Ramesh Sitaraman. 2019. Simulating CLIR Translation Resource Scarcity using High-resource Languages. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341981.3344236

1 INTRODUCTION

Most modern statistical approaches to cross-language information retrieval (CLIR) depend upon some form of translation table that provides a mapping from the vocabulary of the query language to that of the documents' language. Such a translation resource might be a lookup table [20] with probabilities associated with each target language word, a cross-language distributional representation (word embedding) where distance in the vector space corresponds to likelihood of translation [23, 28], or other comparable approaches [18, 24]. Each of those approaches is built from some collection of data resources, most commonly a set of parallel (or comparable [28]) texts that is used to build statistically reliable cross-language mappings of the desired type [16, 19, 25]. However, it is not always true that the more parallel text that is brought to bear, the more accurate the derived probabilities or distances are [1, 26, 30].

One issue that arises with striking regularity is what to do when the amount of parallel text (or the quantity of other translation resources) is substantially smaller than ideal. One of the collections we will discuss below has roughly two million pairs of parallel sentences, contrasting with "low resource" languages where we can have 4-16% of that. The result is a dramatic drop in the statistical reliability of the translation process and, not surprisingly, a concomitant drop in the effectiveness of downstream processes such as extraction, summarization, or (the focus of this study) retrieval. There are some studies showing the trade-off between various features of translation resources, in terms of quantity and quality, and effectiveness in the studied task [1, 10, 15, 26, 30].

We are interested, though, in a deeper understanding of *why* effectiveness drops. What is qualitatively different between the translations that result from high-resource language pairs compared to low-resource pairs? Is it vocabulary coverage in the query language or the documents' language? Is it the accuracy of the estimated translation probabilities? Is it consistent across different languages or does it depend on the languages being crossed? With a better understanding of that issue, we believe it should be possible to target resource acquisition efficiently. Given a low-resource language pair, should a researcher or system builder look for information comparable to what is on hand to improve translation probabilities? Or should it sacrifice those probabilities to provide greater coverage of the vocabulary of the target corpus or, perhaps, the vocabulary of the likely queries if those can be guessed?

To understand the impact of additional resources on CLIR quality (for example), we need corpora that are both low- and high-resource. To isolate language-specific impact, we could use massive numbers of language pairs with various amounts of data and CLIR relevance judgments, but the cost in terms of time and money is prohibitive. Instead, we take the approach of starting with a true low-resource language and systematically down-sample a high-resource language to become an artificial low-resource language—in a way that

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344236

allows us to monitor statistics of interest and compare them to the statistics of true low-resource languages. Our long-term goal is to use this approach to construct language pair corpora with understandable properties so that we can understand the impact on CLIR effectiveness of vocabulary size, vocabulary coverage, translation probabilities, morphological processing, and other items.

The goal of this study is to develop a method for simulating a low-resource language pair using a high-resource pair. We call this process *Resource Scarcity Simulation* (RSS). We know of no prior work that has explored this RSS problem. Indeed, most existing research exploits random down-sampling, an approach that is likely to preserve some statistical properties of the data but provides no insight into what is changing as resources are made scarcer, nor what other than size is comparable between the down-sampled data and a genuinely low-resource language pair. Our methods are extendable to a range of factors for optimizing the down-sampling, though we focus on vocabulary coverage and vocabulary distribution [7, 21, 26], i.e. "frequency profiling".

The rest of this paper is organized as follows. In Section 2 we provide a review of related works done previously. Section 3 provides a formal statement of the problem we address. Section 4 provides our modeling of the problem with a family of set covering problems and our proposed two greedy algorithms. Section 5 and 6 provides our experimental design and a discussion on the results. We conclude our study in Section 8 with directions on possible future works.

2 RELATED WORK

There is existing work exploring CLIR performance reliability from the perspective of available translation resources. Franz et al. [8] studied different features of translation resources possibly impacting the CLIR performance including size, domain, dialect, quality, and style. They provided evidence that the query terms' out of vocabulary rate (OOV rate) is a simple estimator of the retrieval performance and the corpus size is not always the most impacting factor [8]. Zhu and Wang [32] studied a rule-based machine translation system by decreasing the dictionary and rule base, and found that removing dictionary has a greater impact on CLIR performance. McNamee and Mayfield [17] argued that the quality of the translation resource is the most important factor, and explore various query expansion techniques with the OOV rate increased synthetically, simulating variability in the coverage of resources. Some earlier work evaluated different resources pair-wise to contrast the difference in the resource quality for the retrieval performance [9, 13, 29].

Through empirical investigations, Xu and Weischedel [30, 31] studied the impact of lexical resources on CLIR performance. In particular, they suggested a metric based on the frequency of a word in the retrieval corpus and used it for sentence selection such that with some portion of the data reasonable performance can be achieved – when compared to the original translation resource based CLIR performance. Their experimental results provided evidence for the importance of frequency distribution of terms in the retrieval corpus from the perspective of the impacting features of translation resource on the CLIR performance [30].

Talvensaari [26] studied the three major impacting factors of parallel corpora, including topical nearness, alignment quality, and size through empirical investigations. He highlighted the impact of topical nearness as the most crucial factor, and suggested that even adding noisy complementary resources to decrease the topical differences can help CLIR performance.

Some recent work on cross-lingual distributed representation construction, with different applications than CLIR, also investigated the resilience of the embedding construction methods with resource scarcity scenarios [1, 10, 15]. For example, Adams et al. [1] exploited a special down-sampling in which only the target language is scaled down and studied the impact with the aim to extend the results for a threatened language.

Most of the existing work takes a high-resource language and simulates resource scarcity by randomly down-sampling. However, no real-world resource scarcity scenario is provided in the existing investigations. Our study differs from the existing work in the following way: We investigate obstacles with low-resource languages' CLIR performance using high-resource languages in a contrastive comparison. Basically, we start with a low-resource language and aim to understand the limitations using different high-resource languages. This perspective requires that the resource scarcity situation be simulated by a high-resource language. Afterwards, using different *data augmentation* techniques, the resource scarcity scenario can be studied further. We focus on the first step and leave the latter as future work.

3 PROBLEM STATEMENT

The problem that we address is the simulation of a limited parallel corpus (*PCL*) of a low-resource language using a much richer parallel corpus (*PCH*) of a high-resource language. Our target application is the CLIR ad-hoc document ranking problem. However, our simulation problem may also be applicable to other applications in machine translation.

We define the *Resource Scarcity Simulation (RSS)* problem that takes as inputs two sets of parallel corpora, *PCH* and *PCL*, and produces a new parallel corpus PCH_d that is a "down-sampled" version of *PCH* that is "statistically similar" to *PCL*.

Figure 1 depicts our proposed solution to RSS. Let the low-resource source language of the parallel corpus *PCL* be L and the target language be E, e.g., *PCL* is a corpus of parallel sentences in the Somali language (L) and English (E). In particular, *PCL* contain m parallel sentences $\{(s_1^L, s_1^E), (s_2^L, s_2^E), \cdots, (s_m^L, s_m^E)\}$. Likewise, let the high-resource source language of the parallel corpus *PCH* be H and the target language be E, e.g., *PCH* is a corpus of parallel sentences in French (H) and English (E). *PCH* contains n parallel sentences $\{(s_1^H, s_1^E), (s_2^H, s_2^E), \cdots, (s_n^H, s_n^E)\}$, where $n \gg m$.

Note that we assume that both *PCL* and *PCH* share the target language E. In the bilingual CLIR scenario where the retrieval collection (*RC*) is a monolingual collection, and we have queries in several languages, this assumption is naturally satisfied. For example, suppose we have an English news collection as RC^E and we aim to build a system that can query RC^E with both H and L query languages.

Let $voc(.)$ and $fdist(.)$ refer to the vocabulary set and the frequency distribution for a given corpus, respectively. Note that no assumption about PCH^E and PCL^E can be made, other than being in the same language. Although ideally a parallel corpus for a

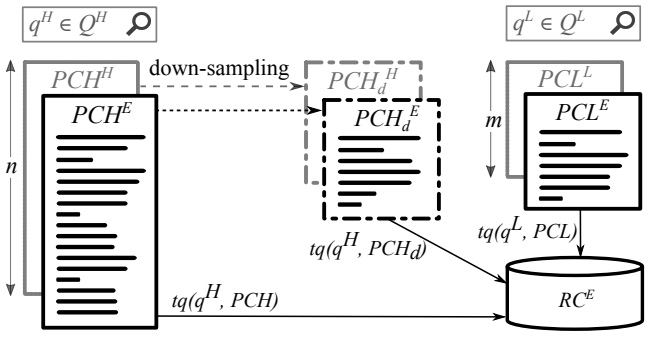

Figure 1: The overall schema of the RSS problem.

high-resource language should cover all the vocabulary in the low-resource side, i.e. $voc(PCL^E) \subset voc(PCH^E)$, in a realistic scenario we cannot make this assumption. We do assume those vocabulary sets overlap, i.e. $voc(PCL^E) \cap voc(PCH^E) \neq \emptyset$.

The output of the RSS problem is a new parallel corpus PCH_d that we construct by selecting m' sentence pairs out of n existing in PCH, i.e., $PCH_d \subset PCH$ and $m' \ll n$. We aim to do this down-sampling such that $PCH_d^E \approx PCL^E$, where PCH_d^E and PCL^E denote the sentences in the target language E in PCH_d and PCL respectively. More precisely, when we say that one set of sentences approximates ($\approx$) another, we mean that terms and their frequencies match closely, i.e., $voc(PCH_d^E) \approx voc(PCL^E)$ and $fdist(PCH_d^E) \approx fdist(PCL^E)$. It has been shown that such frequency profiling is effective for comparing corpora [21, 26]. By making $PCH_d^E \approx PCL^E$, we hypothesize that the respective sets in the source languages in H and L are also approximately similar, i.e. $PCH_d^H \approx PCL^L$. Note that while constructing PCH_d, we ensure that the sentences that we pick are distinct.

Finally, we assume that the information need is exactly same across high- and low-resource languages, i.e. $Q^L \equiv Q^H$. Let the query, $q^L \in Q^L$, be in the source language L, with constituent terms $\{q_1^L, \cdots, q_{|q^L|}^L\}$, and the document, $D^E \in RC^E$, in target language E, with constituent terms $\{d_1^E, \cdots, d_{|D^E|}^E\}$. Similar notation is defined for Q^H. As with the standard bilingual CLIR setting, the aim is to calculate score(q^L, D^E). We use statistical machine translation to construct a Translation Table (TT) for the query translation, which is built using *PCL* or *PCH*. The process of constructing TT using parallel corpora, translating search query, and the document retrieval is shown as function $tq(.\ ,\ .)$ in Figure 1.

4 MODELING RSS AS A COVERING PROBLEM

Our main insight to solving RSS is that it belongs to a family of covering problems that is well-studied in the literature [27]. While RSS itself has no prior work, a special case of RSS is the multi-set multi-cover problem that is known to be NP-complete [27]. This implies that RSS is NP-Hard and hence infeasible to solve exactly (assuming that $P \neq NP$). However, the observation that RSS is a covering problem also suggests that a greedy algorithm similar to those known for other covering problems is likely to work for RSS. Thus, the greedy algorithms presented for RSS in this section are inspired by similar algorithms known for set cover, multi-set multi-cover, and other covering problems [27].

4.1 Formulation as an Integer Linear Program

We now formulate RSS as an integer linear program (ILP) that can be viewed as a generalization of the standard ILP for the multi-set multi-cover problem. Let $U = voc(PCH^E)$ and $V = voc(PCL^E)$ represent the unique terms in each corpus. As we mentioned earlier, RSS aims to down-sample PCH^E such that its frequency profiling becomes similar to PCL^E. Each sentence $s \in PCH^E$ covers some of the terms in U. However, by down-sampling, RSS should select sentences covering V. To this end, we intend to cover the intersecting terms in $U \cap V$, and avoid terms in $U - V$.

For each sentence $s \in PCH^E$, we define a boolean variable $x_s \in \{0, 1\}$ that indicates whether or not s is chosen to be part of PCH_d^E. Let $c(s)$ be the cost of including the sentence in PCH_d^E. The objective of RSS is to minimize the total cost of all the sentences included in the down-sampled version as expressed below:

$$\min \sum_{s \in PCH^E} c(s) x_s \tag{1}$$

Note that the cost $c(s)$ can be flexibly defined in a manner that suits our application. For each term $t_k \in U$, we define a *coverage requirement* f_k that is minimum number of times the term must appear in PCH_d^E. For any term $t_k \in U \cap V$ that appears in both vocabularies, our desire would be to ensure that f_k equals the frequency $fdist(PCL^E, t_k)$ with which t_k appears in PCL^E. However, to avoid situations in which the frequency of a term t_k in PCL^E is higher than PCH^E and that would make our ILP infeasible, we formulate the desired frequency as below.

$$f_k = \min(fdist(PCH^E, t_k),\ fdist(PCL^E, t_k))$$

Let $b_{s,k}$ denote the number of times term $t_k \in U$ appears in sentence $s \in PCH^E$. We enforce the coverage requirement by adding the following constraint.

$$\sum_{s \in S} b_{s,k} x_s \geq f_k, \forall t_k \in U \cap V \tag{2}$$

We would like to avoid picking sentences that use terms $t_k \in U - V$. We capture this *term avoidance* requirement by adding the following constraint.

$$\sum_{s \in S} b_{s,k} x_s = 0, \forall t_k \in U - V. \tag{3}$$

Finally, we postulate the integrality of x_s as follows.

$$x_s \in \{0, 1\}, \forall s \in PCH^E \tag{4}$$

The RSS problem is modeled as an ILP with objective function in Eq. 1 with constraints described in Eq. 2, Eq. 3, and Eq. 4.

Theorem 1. *The RSS problem is **NP**-hard.*

Proof. We show that the multi-set multi-cover problem that is known to be NP-hard [27] is a special case of the RSS problem. In the multi-set multi-cover problem, we are given an universe U of elements and collection of multi-sets $S = \{s_1, s_2, \cdots\}$, where each multi-set s_i contains elements from U, possibly with repetitions. Each multi-set s_i also has a cost $c(s_i)$. The goal is to find a minimum cost sub-collection $S' \subseteq S$ such that frequency of an element e in

the sub-collection S' is at least a required value r_e, for all $e \in U$. Note that the multi-set multi-cover problem is itself a generalization of the classic set cover problem which does not have the frequency requirement, but is also NP-Hard.

It is now easy to see that the multi-cover multi-set problem can be reduced to RSS by simply setting each element of U in the former problem to be a term in the $voc(PCH^E)$ in the latter problem. Further each multi-set s_i of the former problem is a sentence $s \in PCH^E$ in the latter problem. The frequency requirement r_e of the former problem become the coverage requirement f_e in the RSS problem. Further, we make the $voc(PCH^E)$ equal to $voc(PCL^E)$, resulting in no term avoidance requirements. From this reduction, it is clear that any polynomial time solution to RSS will also result in a polynomial time solution for the multi-cover multi-set problem, and that is not possible unless $P = NP$. □

4.2 Greedy Algorithm

The traditional approach for designing provably good greedy algorithms for covering problems is to define a function that assigns a "price" for covering each element (term in our case). The algorithm then repeatedly picks the most cost-efficient sets (sentences in our case) that have the least price in a greedy fashion. The RSS problem differs from other known covering problems in that it incorporates a term avoidance requirement (Eq 3). However, we use an approach that is inspired by known provably-good greedy algorithms for other covering problems, although we define cost-effectiveness and price in a manner that is specific to RSS.

Algorithm 1 presents the standard greedy algorithm that repeatedly picks sentences in the decreasing order of the current price, where *cost effectiveness* can be defined by an arbitrary *price function*. We define two price functions that determine the cost-effectiveness of a sentence $s \in PCH^E$. In each iteration, the algorithm selects, from amongst the currently remaining sentences, the lowest price sentence. The *get_price()* function, line 5 of Algorithm 1, calculates the price based on our two definitions. The stopping criteria, line 2 of Algorithm 1, is determined by *satisfy()* function. We described the ideal problem constraints with the ILP formulation. However, for providing comparable numbers and simplicity in our evaluations, we select m sentences from PCH, i.e. $m' = m$.

The price $p(s)$ of a sentence s is defined to be ratio of its cost $c(s)$ and its effectiveness $e(s)$[1]. We define two different price functions by defining the effectiveness function $e(s)$ in two different ways. However, for both price functions, we use the same cost function defined below. For each sentence $s \in PCH^E$, we define a cost function, $c(s)$, based on the number of terms covering $\{U - V\}$.

$$c(s) = |voc(s) \cap (U - V)| + \alpha \tag{5}$$

α is a slack variable to control the cost of selecting a sentence. It also prohibits the cost to become zero. We use $\alpha = 1$ in our experiments as the cost of selecting a new sentence which we desire to keep it minimum. The effectiveness function, $e(.)$, is defined based on the selected sentences at the moment, C. Basically it aims to measure the novelty a new sentence brings to the terms covered by the selected sentences. We describe the two variants below.

[1]To avoid dividing by zero, we actually define price to be $p(s) = \frac{c(s)}{e(s)+\beta}$, where $\beta = 0.05$.

Algorithm 1 Greedy Solution for the RSS Problem

Input: *PCH*: High-resource PC, *PCL*: Low-resource PC
Output: C: selected sentences from *PCH*

1: $C \leftarrow \emptyset$
2: **while** *satisfy*(C, PCH, PCL) **do**
3: *cur_price* $\leftarrow \emptyset$
4: **for** $s \in PCH^E - C$ **do**
5: *cur_price* $\leftarrow$ *cur_price* $\cup$ *get_price* (s, C, PCH, PCL)
6: **end for**
7: $C \leftarrow C \cup \min($*cur_price*$)$
8: **end while**

1) Greedy Simulation (GreeSim). For distinguishing different scenarios in each iteration of the algorithm, we define three different scenarios. A term is not covered at all, the term is under-covered, and lastly the term is over-covered. The following formulation of the effectiveness deals with each scenario accordingly.

$$e(t,s) = \begin{cases} fdist(s,t), & \text{if } t \notin C \\ fdist(s,t) \times \frac{f_t - fdist(C,t)}{f_t}, & \text{if } t \in C, \text{ and } fdist(C,t) < f_t \\ 0, & \text{if } t \in C, \text{ and } fdist(C,t) \geq f_t \end{cases} \tag{6}$$

A summation over the sentence terms, $voc(s)$, in which $t \in \{U \cap V\}$ result in $e(s)$ for the GreeSim variation of the algorithm.

2) Relaxed Greedy Simulation (ReGreeSim). With the relaxed constraints, we only consider the coverage or not coverage in the currently selected sentences for the term in a given sentence.

$$e(s) = |(voc(s) \cap (U \cap V)) - voc(C)| \tag{7}$$

In particular, in this variant we relax the coverage requirement (Eq. 2) such that we only aim to cover each term t_k with a positive frequency f_k at least once, rather than at least f_k times.

Complexity Analysis. Let the price calculation for a given sentence, *get_price()* function, take constant time, $O(c)$. On each step, we select a sentence from the remaining sentences of PCH and we repeat the operation until m' sentences are selected: we have $n + (n-1) + (n-2) + \cdots + (n-m')$ operations for the greedy solution. Therefore, the complexity of the Algorithm 1 is $O(nm')$. Although we mentioned that $m' \ll n$, in the worst case scenario the complexity would be $O(n^2)$.

5 EXPERIMENTAL SETUP

In this section, we explain our CLIR system details, the data used, and the evaluation metrics of our experiments. For the high-resource languages, H, we use French, Italian, German, and Finnish as the query language[2]. For the low-resource languages, L, we use Somali and Swahili[3]. We simulate each low-resource language using any of the high-resource languages.

[2]English and German are in the Germanic language family, Italian and French are in the Romance language family, and Finnish is in the Uralic language family.
[3]Somali is in the Afro-Asiatic language family, and Swahili is in the Niger-Congo language family. Both are mostly spoken in Africa.

5.1 Retrieval System

Let the query, q, be in source language F, and the document, d, in target language E. We translate the query term-by-term into language E. We use GIZA++ [20] as the statistical machine translation toolkit. It provides a translation table with the probability of translation trained on parallel corpora. For a given query term we obtain a sorted list of T translations with the corresponding score. Formally, for the i^{th} query term, $q_i^F =< \cdots, (q_{(i,j)}^E,\ t(q_{(i,j)}^E|q_i^F)), \cdots >$, where $(1 \leq j \leq T)$ and $t(.|.)$ is the translation probability from F to E. For out-of-vocabulary (OOV) terms in queries, we set $q_i^F =< (q_{(i)}^F,\ 1.0) >$ to fall back on exact matching without translation. From this point, any monolingual ranking method (e.g., probabilistic or language modeling) can be applied [19] to calculate $score(q^F, d^E)$. We use Galago's implementation[4] of Okapi BM25 [22] with default parameters ($b = 0.75$, $K = 1.2$, and $w = 1.0$). For incorporating translation terms with the corresponding scores we exploit the Galago query language[5] – specifically, the weighted *#combine* operator is used. We use $T = 5$ in our experiments, chosen based on some preliminary experiments showing with that number of translation terms the best retrieval performance is achieved across all the languages.

5.2 Data

5.2.1 Text Pre-processing. In order to have consistent pieces of text across different resources for translation model training, queries, and test collection, we apply the following pre-processing steps. Characters are normalized by mapping diacritic characters to the corresponding unmarked characters and lower-casing. We remove non-alphabetic, non-printable, and punctuation characters from each word. The NLTK library [2] is used for tokenization and stop-word removal. No stemming is performed.

5.2.2 Query Set and Text Collection. We performed experiments on the Cross-Language Evaluation Forum (CLEF) 2000-2003 campaign [3–6] for bilingual ad-hoc retrieval tracks[6]. We aggregate all four years' track topics and query relevance judgments in order to have a higher number of queries, similar to Kraaij et al.'s experimental design [14]. The text collection for all our query languages is the Los Angeles Times (LAT94) comprising over 113k news articles.[7] We only use the *text* field of the LAT94 corpus for indexing. We apply the same text pre-processing operations on the query set and collection text. Queries are selected from $C001 - C200$ topic set for each language. For the low-resource language queries, we hired a translation organization to translate $C001 - C200$ topic set into Somali and Swahili. We share these queries with the community[8]. Queries without any relevant document are excluded, resulting in 151 queries for each language. We use only the *title* field of the queries in our experiments.

[4]https://www.lemurproject.org/galago.php
[5]https://sourceforge.net/p/lemur/wiki/Galago%20Query%20Language/
[6]http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/
[7]In the 2003 track an additional text collection was added to the evaluations, the Glasgow Herald corpus. For consistency in the evaluations, we filter out this collection with its relevance judgments.
[8]https://ciir.cs.umass.edu/download

Table 1: Statistics of the resources used for training of translation models, before down-sampling. The source language is either H or L, and the target language is English, E.

Source Lang.	#Instances	Source Voc.	English Voc.
French	1,995,528	141,338	105,182
Italian	2,267,872	275,527	201,403
German	2,547,952	577,192	263,593
Finnish	2,333,189	794,665	170,209
Swahili	306,367	126,139	65,099
Somali	98,591	131,145	42,270

5.3 Parallel Corpora (PC)

For the high-resource languages, in order to have comparable resources, in terms of size and domain of the high-resource parallel corpora, we use the Europarl sentence-aligned corpus [12]. It is extracted from the proceedings of the European Parliament, sentence-aligned with statistical methods, and includes 21 European languages. For the low-resource languages, we use a mixture of resources provided by OpenCLIR[9] program of the MATERIAL project and DARPA's LORELEI[10] project. In addition to the mentioned parallel corpora, for each of the languages pairs, we exploit the Panlex lexicon [11]—covering more than 5,700 languages. Its data acquisition strategy emphasizes high-quality lexical and broad language coverage[11]. A concatenation of Panlex with the former resources are used in our experiments. A summary of the translation resources used in our experiments for each language pair along with their vocabulary size is given in Table 1. Our low-resource languages are in the range of 4% to 16% of high-resource languages, in terms of the number of instances. However, they have reasonable vocabulary coverage.

5.4 Evaluation.

For evaluating retrieval effectiveness, we report Mean Average Precision (MAP) of the top 1000 ranked documents. We also report the Out-of-Vocabulary rate (OOV) and the similarity percentage of the down-sampled corpus, as defined below.

Out-of-Vocabulary Rate (OOV). The query OOV rate has an important role in the retrieval performance [8]. Given that for query translation in CLIR, the common practice is to use more than one translation so some may be in and others out of vocabulary, we define the following OOV rate for each query. We define OOV of a query based on its constituent terms $q_t \in q$, and average over all the queries for reporting. For q_t, we take top T translations from the translation table. Among these T ranked terms, starting with the top ranked translation term, we check whether the term exists in the retrieval collection (RC) or not. If the term exist, then the partial OOV for that term is calculated based on the rank of the term using $OOV(q_t) \leftarrow 1 - \frac{1}{rank(t(q_t))}$, the rank starts with 1 as the most probable translation. Otherwise, we set $OOV(q_t) \leftarrow 1$.

[9]https://www.nist.gov/itl/iad/mig/openclir-evaluation
[10]https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents
[11]https://dev.panlex.org/data-model/

Table 2: RSS Experimental Results with respect to Somali language. For each down-sampling method, the highest similarity percentage is marked with bold-face.

	Orig. PC		Random (5 runs)			ReGreeSim			GreeSim		
Q Lang.	MAP	OOV	MAP	OOV	Sim (%)	MAP	OOV	Sim (%)	MAP	OOV	Sim (%)
French	0.3033	0.0117	0.2748	0.1092	25.96	0.2707	0.1297	29.69	0.2646	0.1250	35.25
Italian	0.2876	0.0151	0.2683	0.0813	20.49	0.2649	0.1113	24.12	0.2612	0.1136	28.06
German	0.2851	0.0616	0.1962	0.2306	**97.05**	0.1967	0.2668	**96.51**	0.1949	0.2609	**98.47**
Finnish	0.2509	0.0730	0.1707	0.2631	60.35	0.1785	0.2697	73.87	0.1799	0.2618	76.31
Somali	0.1935	0.1673			50.96			56.05			59.52

Table 3: RSS Experimental Results with respect to Swahili language. For each down-sampling method, the highest similarity percentage is marked with bold-face.

	Orig. PC		Random (5 runs)			ReGreeSim			GreeSim		
Q Lang.	MAP	OOV	MAP	OOV	Sim (%)	MAP	OOV	Sim (%)	MAP	OOV	Sim (%)
French	0.3033	0.0117	0.2802	0.1912	24.14	0.2725	0.2542	32.18	0.2719	0.2435	32.81
Italian	0.2876	0.0151	0.2828	0.1526	6.03	0.2769	0.0626	13.38	0.2766	0.0639	13.75
German	0.2851	0.0616	0.2325	0.1332	67.87	0.2231	0.1818	80.00	0.2215	0.1807	82.06
Finnish	0.2509	0.0730	0.1947	0.1812	**70.25**	0.2090	0.1947	**96.77**	0.2086	0.1926	**97.69**
Swahili	0.2076	0.1224			42.07			55.58			56.58

Note that by this definition, a term that does not exist in the TT (i.e., cannot be the output of a translation step and we use the term itself in the translated query) but nonetheless exists in RC^E does not count toward overall OOV.

$$OOV(q) = \frac{\sum_{q_t \in q} OOV(q_t)}{|q|} \tag{8}$$

Similarity Percentage (Sim). In order to see how close the retrieval performance of the down-sampled high-resource parallel corpus, PCH_d, compared to the low-resource language's retrieval performance, we use the following measurement.

$$Sim(q) = 1 - \frac{\mid tq(q^H, PCH_d) - tq(q^L, PCL) \mid}{\mid tq(q^H, PCH) - tq(q^L, PCL) \mid} \tag{9}$$

where $tq(\ .,\ .)$ returns= the Average Precision (AP) value of the retrieval, after translating and querying the retrieval corpus. We report this as percentage in our experiments. Higher values of Sim means that the down-sampled retrieval performance is closer to the corresponding low-resource retrieval performance.

5.5 Random Baseline

To the best of our knowledge, the problem as we proposed here, is not addressed in the literature. The existing research down-samples the parallel corpus, whenever needed, by randomly taking a fraction of data. As we show through our experiments, that is not the best practice for doing the down-sampling. For the random down-sampling baseline, we randomly select $m' = m$ sentences out of PCH and repeat for five times, as described. For example, for simulating Somali, we randomly select 98k sentences out of French-English parallel corpus, repeating that five times. We report the average value of the five runs on each evaluation measurement.

6 EXPERIMENTAL RESULTS

Tables 2 and 3 present our experimental results for down-sampling four high-resource languages with respect to Somali and Swahili, respectively. We report the retrieval performance and OOV for the original resources in the first column. We also report the average similarity percentage across four high-resource languages.

For example, the first row of Table 2 presents our results for simulating of Somali using French parallel corpus. It shows that without any down-sampling, we get a MAP value of 0.3033 and OOV rate of 0.0117. Randomly taking 98k sentences out of 1.9m sentences, retrieving based on the translation table trained with the down-sampled data, and averaging over the five random runs, results in the second column for French. The results show a drop in MAP and a corresponding increase in the OOV rate. Comparing the down-sampled MAP value with the run without down-sampling, and using Eq. 9, we see that it is similar to Somali retrieval performance by a factor of $\approx$ 26%. As can be seen, with the same procedure, it is easier to simulate the characteristics of Somali with German. On average across four high-resource languages, we are able to simulate by a factor of $\approx$ 51% using multiple random down-sampling. However, with ReGreeSim and GreeSim we are able to simulate Somali language with a similarity percentage of 56% and 59%. This results suggest that it is possible to improve simulation of the resource scarcity environment with a high-resource language, when compared to random down-sampling. GreeSim is outperforming ReGreeSim in terms of simulation success in every high-resource language.

As with Table 3 relatively consistent results are achieved for simulating Swahili when compared to Somali simulation. For Swahili we have larger parallel corpus, providing evidence for Swahili's better retrieval performance, compared to Somali. We hypothesize that for the same reason, it is harder to simulate Swahili compared

to Somali with the same set of high-resource languages—compare 42% of Swahili random simulation with that of Somali as 51%. However, our greedy solutions provide consistent similarity average percentage when comparing both the scenarios—56.05% to 55.58% and 59.52% to 56.58%. Comparing ReGreeSim with GreeSim suggests that respecting the frequencies of the low-resource corpora is important in order to simulate the resource scarcity environment.

Another interesting observation is that our experiments show that it is easier to simulate Swahili using the Finnish high-resource corpus when compared to other high-resource corpora: compare that with our previous observation suggesting it is easier to simulate Somali with German parallel corpus. In addition, considering the fact that French and Italian are from the Romance language families might explain that those families are ill-suited to simulate Somali and Swahili, both in African language families. Note that this conclusion needs further investigation which we leave for future work.

In terms of MAP values, we compare German down-sampled to Somali, and Finnish down-sampled with Swahili using the two-tailed paired t-test with $p\_value < 0.05$ (i.e., 95% confidence level). We wanted to see if there is any query-wise differences between these various runs. The difference is insignificant. For example for the GreeSim simulation of German results, compared with Somali experiments, the $p\_value = 0.0649$ suggests that there is no significant difference. However, doing so with different random sub-samples shows that some of them are statistically significant. For example out of five runs of Swahili simulation with Finnish, 3 of them are statistically significant.

Comparing the OOV rates across simulations of Somali and Swahili suggests some inconsistency with the results of [8]. Particularly, with German and Finnish simulations, the OOV rate is relatively high compared to that of Somali and Swahili. For example compare the OOV rate of German simulated with GreeSim to that of Somali. It is difficult to interpret this, given that we use top-5 translations and measure OOV rate using a ranked based OOV rate definition. For this purpose, we further investigate the OOV rate reliance with a similar experiment to McNamee and Mayfield's experimental design [17] in the following section.

7 SYNTHETIC SIMULATION ANALYSIS

To study the observed OOV rate discrepancies with our experiments, we exploit the approach widely used in the literature to study the retrieval performance with resource scarcity of low-resource languages [8, 17]. For this purpose, instead of down-sampling *PCH*, one can train a translation table first, and then, by simply removing some synonym relations, synthetically increase the OOV for the queries [8]. Such an approach is relatively cheap, in terms of the computational resource requirements. However, it is a query specific solution with relatively limited possibility of generalizing the observations, in terms of understanding the limitations with the low-resource languages—since it is based on a byproduct of the parallel corpora.

As it has been suggested by McNamee and Mayfield [17] a high-resource parallel corpus is likely to have more coverage of rare terms and entities. For this reason, when dropping some of the terms from the translation table it seems that one should keep

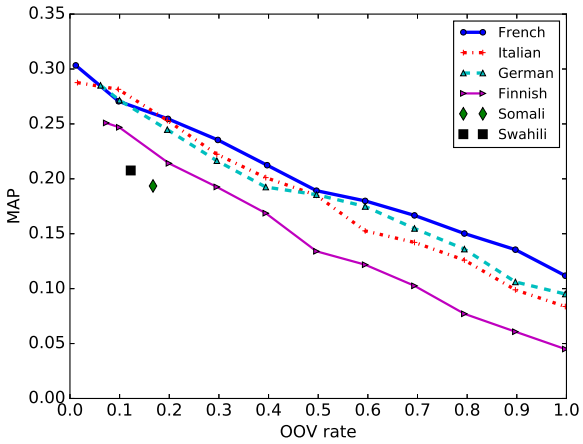

Figure 2: Sensitivity Analysis on OOV rate by synthetically dropping terms from TT.

those rare terms rather than randomly dropping. We drop from the translation table starting from the least frequent terms in the parallel corpus. Using this procedure, we increase the OOV rate of high-resource languages and measure the MAP value.

Figure 2 presents our sensitivity analysis on the OOV rate. We start with the original parallel corpus' OOV rate, as reported in Table 2, and set the synthetic OOV rate to $\{0.1, 0.2, \cdots, 1.0\}$ values. The OOV rate of 1.0 means that none of the query terms are translated using the translation table. We also plot the OOV rate of Somali and Swahili with diamond and square pointers.

In general, since the OOV rate increasing method is the same across our four high-resource languages, they all show approximately consistent reactions toward the change of OOV in terms of retrieval performance. The low-resource languages' OOV rate is also consistent with the retrieval performance; the higher OOV rate means lower retrieval performance.

One interesting observation of this experiment, which explains the inconsistency of OOV rates in Table 2 and 3, is that for having the same retrieval performance as of the low-resource language, using a synthetically increasing OOV rate of the high-resource language, it should be increased to a higher level. For example, to achieve a MAP of 0.2076 using any high-resource language, which is the MAP of Swahili queries, a higher OOV rate is required compared to that of the high-resource language.

Another interesting observation is seemingly the three clusters of behavioral pattern with OOV and MAP. These three clusters seems to be: *1)* Somali and Swahili, *2)* Finnish, and *3)* German, Italian, and French languages. We hypothesize that this might be due to the language family differences when compared to the retrieval corpus's language. However, this observation definitely needs more investigation and is left for future work. In addition, studying various dropping procedures might reveal interesting observations. Considering each of these mentioned clusters individually, an approximately linear relationship can be seen. This provides more evidence for the conclusion of McNamee and Mayfield [17]. However, our results also suggest that OOV rate is not consistent with MAP across languages.

8 CONCLUSION

We introduced a contrastive framework for studying low-resource languages, using high-resource languages. It is the reverse of the existing research in which a high-resource language is down-sampled randomly to provide artificial resource scarcity environment. Our perspective consist of a two-step procedure where a true low-resource scenario is simulated with a high-resource language, and then various translation data augmentation procedures are studied. To this end, we investigated the first step, called the *Resource Scarcity Simulation (RSS)* problem. We modeled the problem with a family of set covering problems. We formulated the RSS with ILP and proved that the problem is actually **NP**-hard. We also proposed two greedy algorithms based on our modeling of the problem.

Through our experiments, we investigated the simulation of two low-resource languages, i.e. Somali and Swahili, using four high-resource languages. We observed that the language families are important for the simulation, and particularly for CLIR retrieval performance—see Table 2 and Table 3 for evidence supporting the conclusion. For example, it is easier to simulate Somali with German compared to Italian, French, and Finnish. In addition, investigating OOV rate across our studied languages, with a synthetic increasing procedure, suggested that having the same OOV rate in two different languages may not predict their corresponding CLIR performance. However, for the same language, decreasing OOV rate may help for the CLIR performance, showing a linear relation—see Figure 2 for evidence.

As the future work, we are mainly interested to study data augmentation techniques to further investigate the resource scarcity problem with the low-resource languages. In addition, the application of the RSS problem, and the provided greedy solutions, in other domains like machine translation seems highly interesting.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Air Force Research Laboratory (AFRL) and IARPA under contract #FA8650-17-C-9118 under subcontract #14775 from Raytheon BBN Technologies Corporation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 937–947.

[2] Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proc. of the ACL Interactive Poster and Demonstration Sessions*. http://aclweb.org/anthology/P04-3031

[3] Martin Braschler. 2000. CLEF 2000 - overview of results. In *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal*. 89–101.

[4] Martin Braschler. 2001. CLEF 2001 - overview of results. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany*. 9–26.

[5] Martin Braschler. 2002. CLEF 2002 - overview of results. In *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy*. 9–27.

[6] Martin Braschler. 2003. CLEF 2003 - overview of results. In *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway*. 44–63.

[7] Leo Egghe. 2007. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology* 58, 5 (2007), 702–709.

[8] Martin Franz, J Scott McCarley, Todd Ward, and Wei-Jing Zhu. 2001. Quantifying the utility of parallel corpora. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 398–399.

[9] Alexander Fraser, Jinxi Xu, and Ralph Weischedel. 2002. TREC 2002 Cross-lingual Retrieval at BBN. In *TREC*.

[10] Ankur Gandhe, Florian Metze, and Ian Lane. 2014. Neural network language models for low resource languages. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[11] David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proc. of the Ninth International Conference on Language Resources and Evaluation, LREC'14*. 3145–3150.

[12] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[13] Wessel Kraaij. 2001. TNO at CLEF-2001: Comparing translation resources. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 78–93.

[14] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics* 29, 3 (2003), 381–419.

[15] Mikko Kurimo, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, and Tanel Alumäe. 2017. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation* 51, 4 (2017), 961–987.

[16] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR'18*. 1253–1256.

[17] Paul McNamee and James Mayfield. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 159–166.

[18] Ali Montazeralghaem, Razieh Rahimi, and James Allan. 2019. Term discrimination value for cross-language information retrieval. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19)*.

[19] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

[20] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003), 19–51.

[21] Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proc. of the workshop on Comparing corpora-Volume 9*. Association for Computational Linguistics, 1–6.

[22] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and others. 1995. Okapi at TREC-3. *NIST Special Publication* 109 (1995), 109.

[23] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902* (2017).

[24] Sheikh Muhammad Sarwar, Hamed Bonab, and James Allan. 2019. A multi-task architecture on relevance-based neural query translation. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, ACL*. Florence, Italy.

[25] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, Vol. 2. 458–463.

[26] Tuomas Talvensaari. 2008. Effects of aligned corpus quality and size in corpus-based CLIR. In *European Conference on Information Retrieval*. Springer, 114–125.

[27] Vijay V Vazirani. 2013. *Approximation algorithms*. Springer Science & Business Media.

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

[29] Jinxi Xu, Alexander M Fraser, and Ralph M Weischedel. 2001. TREC 2001 cross-lingual retrieval at BBN. In *TREC*.

[30] Jinxi Xu and Ralph Weischedel. 2003. A probabilistic approach to term translation for cross-lingual retrieval. In *Language modeling for information retrieval*. Springer, 125–140.

[31] Jinxi Xu and Ralph Weischedel. 2005. Empirical studies on the impact of lexical resources on CLIR performance. *Information processing & management* 41, 3 (2005), 475–487.

[32] Jiang Zhu and Haifeng Wang. 2006. The effect of translation quality in MT-based cross-language information retrieval. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 593–600.