CROSS LANGUAGE INFORMATION RETRIEVAL FOR LANGUAGES WITH SCARCE RESOURCES

Christian E. Loza

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2009

APPROVED:

Rada Mihalcea, Major Professor
Paul Tarau, Committee Member
Miguel Ruiz, Committee Member
Armin Mikler, Graduate Advisor
Krishna Kavi, Chair of the Department of Computer Science and Engineering
Michael Monticino, Interim Dean of the Robert B. Toulouse School of Graduate Studies

Our generation has experienced one of the most dramatic changes in how society communicates. Today, we have online information on almost any imaginable topic. However, most of this information is available in only a few dozen languages. In this thesis, I explore the use of parallel texts to enable cross-language information retrieval (CLIR) for languages with scarce resources. To build the parallel text I use the Bible. I evaluate different variables and their impact on the resulting CLIR system, specifically: (1) the CLIR results when using different amounts of parallel text; (2) the role of paraphrasing on the quality of the CLIR output; (3) the impact on accuracy when translating the query versus translating the collection of documents; and finally (4) how the results are affected by the use of different dialects. The results show that all these variables have a direct impact on the quality of the CLIR system.
Copyright 2009

By

Christian E. Loza
## CONTENTS

CHAPTER 1. INTRODUCTION

1.1. Problem Statement 1

1.2. Contribution 2

CHAPTER 2. A BRIEF REVIEW OF INFORMATION RETRIEVAL, MACHINE TRANSLATION AND NLP FOR SCARCE RESOURCE LANGUAGES 4

2.1. Information Retrieval 4

2.1.1. Definition of Information Retrieval 4

2.1.2. Formal Definition and Model 5

2.1.3. Boolean Model 5

2.1.4. Vector Space Model 6

2.1.5. Probabilistic Model 8

2.1.6. Other Models 9

2.2. Machine Translation 9

2.2.1. Development of Machine Translation 9

2.2.2. Approaches to MT 11

2.2.3. Software Resources for Machine Translation 11

2.3. Cross-language Information Retrieval 15

2.3.1. The Importance of Cross Language Information Retrieval 15

2.3.2. Main Approaches to CLIR 16

2.3.3. Challenges Associated with CLIR for Languages with Scarce Resources 16

2.4. Related Work and Current Research 18

2.4.1. Computational Linguistics for Resource-Scarce Languages 18

2.4.2. Focusing on a Particular Language 18

2.4.3. NLP for Languages with Scarce Resources 19

2.4.4. CLIR for Languages with Scarce Resources 19
CHAPTER 1
INTRODUCTION

1.1. Problem Statement

One of the most important changes in our society and the way we live in the last decades is the availability of information, as a result of the development of the World Wide Web and the Internet.

English being the *lingua franca* of the modern era, and United States the place where the Internet was born, most of the current content in the World Wide Web is written in this language.

Along with English, major European languages, like Spanish, French and German, and others, like Chinese, Russian and Japanese are examples of the few dozen languages with significant amount of content available on the Web, in direct relationship with the number of speakers of those languages with regular access to the Web.

Figure 1.1, based on information reported in [8], shows that most of the languages in the world have relatively few speakers.

![Figure 1.1. Distribution of number of speakers.](image)
As a result of the availability of content, out of all the languages spoken in the world today, only a few dozen languages have a significant number of computational resources developed for them. We can say that most of the current spoken languages have little or no computational resources.

I will call the first set of languages “languages with computational resources,” as opposed to the rest of the languages spoken in the world, which I call “scarce resource languages.”

We want to study methods to develop standard natural language processing (NLP) computational resources for these languages, and see how different approaches producing them affect other tasks. In particular, in this thesis I concentrate on the task of information retrieval.

I chose Quechua for our study, because it has fourteen million of speakers, but most of them have little or no access to the Web, making it a scarce resource language.

As shown in earlier works as [1] and [5], even gathering the initial resources is a difficult task when creating computational resources for a language that does not have them.

The quantity of documents for any language is generally proportional to the number of its speakers. Most of the scarce resource languages have relatively few speakers and few written documents available, which increases the difficulty of obtaining them.

Other problems to develop initial resources are the availability of one unique written form of the language, the availability of an alphabet, the presence of dialects, and how to use different character sets with current software tools.

All these problems are common to most scarce resource languages. Many researchers have proposed different approaches to them, according to the language that was the focus of the study.

1.2. Contribution

In this thesis, we analyzed specific aspects of the construction and use of parallel corpora for information retrieval.
More specifically, we want to discuss the following:

(i) How much does the amount of parallel data affect the precision of information retrieval?

(ii) Is it possible to increase the precision of the information retrieval results using paraphrasing?

(iii) Is there a significant difference if we translate the query or the collection of documents? How is this affected by the other factors?

(iv) Is it possible to use information retrieval across different dialects of the same language? How much is the precision affected by this?
2.1. Information Retrieval

Information retrieval (IR) is defined as the task to retrieve certain documents from a collection that satisfy a query or an information need [2].

In the context of computer science, IR is a branch of natural language processing (NLP) that studies the automatic search of information and all the sub tasks associated with it.

Previous remarkable events in the development of IR include the development of the mechanical tabulator based on punch cards in 1890 by Herman Hollerith, to rapidly tabulate statistics, Vannevar Bush’s ‘As We May Think’, that appeared in *Atlantic Monthly* in 1945, and Hans Peter Luhn’s (researcher engineer at IBM) work in 1947, on mechanized punch cards for searching chemical compounds. More recently, in 1992 the first Text Retrieval Conference (TREC) took place.

One the most important development of IR to date has been the development of massive search engines the last decade. A decade ago libraries were still the main place to search for information, but today we can find information on almost any topic using the Internet.

2.1.1. Definition of Information Retrieval

IR studies the representation, storage, organization of, and access to the information items [2].

A *data retrieval* system receives a query from a user, and returns an information item or items, sorted, with a degree of relevance to the query.

An *information retrieval system* tries to retrieve the information that will answer the particular query from the collection.
The availability of resources and information on the Web has changed how we search for information. The Internet not only has created a vast amount of people looking for information, it also has created a greater number of people writing contents in all subjects.

2.1.2. Formal Definition and Model

IR can be defined using the following model [2]:

\[
[D, Q, F, R(q_i, d_j)]
\]

where:

(i) \(D\) is a set of logical representations for the documents

(ii) \(Q\) is a set composed of logical representations of the user information needs, called queries.

(iii) \(F\) is a framework for modeling the documents, queries and their relationships.

(iv) \(R(q_i, d_j)\) is a ranking function

There are three main models in IR: The Boolean model, the vector model, and the probabilistic.

2.1.3. Boolean Model

The Boolean model is based on Boolean algebra, the queries are specified in boolean terms, using Boolean operators as AND and OR. This gives this model simplicity and clarity, and is easy to formalize and implement.

Unfortunately, these same reasons also present disadvantages. Because the model is binary, it prevents any notion of scale, which has the effect to decrease the performance of the information system in terms of quality.
The representation of a document consists of a set of terms weights assumed to be binary. A query term is composed of index terms linked by the boolean operators, like AND, OR, NOT.

The Boolean model can be defined as following[2]:

1. \( D \) The terms can either be present (True) or absent (False).
2. \( Q \) The queries are boolean expressions, linking terms with operations.
3. \( F \) The framework is the boolean algebra, and it’s operators AND, OR and NOT.
4. The similarity function is defined in the context of the Boolean Algebra as:

\[
sim(d_j, q) = \begin{cases} 
    \text{True} & \text{if } \exists q_{cc} \mid (q_{cc} \in d_{dnf} \land (\forall_{k}, g_i(d_j)) = g_i(q_{cc})) \\
    \text{False} & \text{otherwise}
\end{cases}
\]

As shown above, the relevance criteria is a Boolean value, true (relevant) or false (not relevant). Other disadvantage of this model is that the ranking function doesn’t provide information regarding how relevant is a document, so is not possible to compare how relevant it is relative to others. This makes this model less flexible than the other two models.

As a result of this disadvantages, but taking into account all the advantages, the Boolean model is generally used in conjunction with the other models, or has been modified to show a degree of relevance, as in the fuzzy Boolean model.

2.1.4. Vector Space Model

The vector space model presents a framework where a document is represented as a vector in a \( N - \text{dimensional} \) space [18]. Each dimension of the vector is defined as \( w_i \), and is a non-negative non binary number. The space has \( N \) dimensions corresponding to the unique terms present in the collection. The query vector is also defined as a vector in the \( N - \text{dimensional} \) space.

1. \( D \) The \( N \) terms are dimensions, the documents are vectors in this space.
2. \( Q \) The queries \( q_j \) are also vectors in this \( N - \text{dimensional} \) space.
(iii) F The framework is defined in the context of the vector analysis and its operations. One of the most commonly used metrics of distance between vectors is the cosine similarity, as defined below.

\[ \text{sim}(d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} \]  

(3)

The main advantage of using this framework is that we can define the similarity between queries and documents as a function of similarity between two vectors. There are many functions that can give us a measure of distance between two vectors, been the cosine similarity one of the most used for IR purposes.

2.1.4.1. Weight Metrics

There are many methods to calculate the weights \( w_i \) for the \( N \) dimensions of each document, as shown in [18], being an active research topic. For our experiments, we used the term frequency - inverse document frequency, \( tf - idf \).

We define the term frequency (tf) \( tf_{i,j} \) as:

\[ tf_{i,j} = \frac{f_{i,j}}{\max_k f_{k,j}} \]  

(4)

where \( f_{i,j} \) is the frequency of the term \( i \) in the document \( j \), normalized dividing by the maximum frequency observed of a term \( k \) in the document \( d_j \).

The inverse document frequency (idf) is defined as:

\[ idf_i = \log \frac{N}{f_i} \]  

(5)
The *tf–idf* weighting scheme is defined as:

\begin{equation}
   w_{i,j} = tf_{i,j} \times idf_i = \frac{f_{i,j}}{\max_k f_{k,j}} \times \log \frac{N}{f_i}
\end{equation}

2.1.5. Probabilistic Model

Defined in the context (framework) of probabilities, the Probabilistic Model tries to estimate the probability that specific document is relevant to a given query.

The model assumes there is a subset $R$ of documents which are the ideal answer for the query $q$. Given the query $q$, the model assigns measure of similarity to the query.

Formally, we have the following.

(i) $D$ The terms have probability weights, $w \in \{0, 1\}$.

(ii) $Q$ The queries are subsets of index terms.

(iii) $F$ The framework is defined in the context of the probability theory. as:

Let $R$ be the set of documents know to be relevant.

Let $\overline{R}$ be the complement of $R$.

$P(R | d_j)$ is the probability of a document to be relevant.

\begin{equation}
   \text{sim}(d_j, q) = \frac{P(d_j \mid R) \times P(R)}{P(d_j \mid \overline{R}) \times P(\overline{R})}
\end{equation}

One of the advantages of this model is that the documents are ranked according to their probability to be relevant. One of the many disadvantages of this model is that it requires an initial classification of documents in relevant and not relevant (which is the original purpose), and the assumption that the terms are independent of each other.
2.1.6. Other Models

Most of current approaches in the field of IR build on top of one or more of the previous models. We can mention, among other models, the fuzzy set model and the extended Boolean model.

2.2. Machine Translation

Machine translation (MT) is a field in NLP that investigates ways to perform automatic translation between two languages.

In the context of this task, the original language is called the source (S) language, and the desired result language is called the target (T) language.

The naive approach is to replace words from the source language with words from the target language, based on a dictionary.

During the past two decades, research has been active, and many new methods have been proposed.

2.2.1. Development of Machine Translation

The history of machine translation can be traced to Rene Descartes and Gottfried Leibniz. In 1629, Descartes proposed a universal language, which could have a symbol for equivalent ideas in different languages. In 1667, in Preface to the General Science Gottfried Leibniz proposed that symbols are important for human understanding, and made important contributions in symbolic reasoning and symbolic logic.

In the past century, in 1933 the Russian Petr Petrovich Troyanskij, patented a device for translation, storing multilingual dictionaries and continuing his work for 15 years. He used Esperanto to deal with grammatical roles between languages.

During the Second World War, scientists worked with the task to break codes and perform cryptography. Mathematicians and early computer scientists tough that the task of MT was
very similar in essence to the task of breaking encrypted codes. The use of terms as 'decoding' a text came during this period of time.

In 1952, Yehoshua Bar-Hillel, MIT’s first full-time MT researcher, organized the first MT research conference. In 1954, at Georgetown University the first public demonstration of a MT system is held, translating 49 sentences Russian to English. In 1964, the Academy of Science creates the Automatic Language Processing Advisory Committee, ALPAC, to evaluate the feasibility of MT.

One of the best known events in the history of MT is the publication of ALPAC’s report in 1966, having an undeniable negative effect on the MT research community. The report concluded that after many years of research, the task of MT hasn’t produced useful results, and qualified it as hopeless. The result was a substantial funding cut for MT research in the following years, therefore, in the immediate following decades, research on MT was somehow slow.

In 1967, L. E. Baum, T. Petrie, G. Soules and N. Weiss introduced the expectation maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, which, after the Shannon Lecture by Welch, became the Baum-Welch algorithm to find the unknown parameters of a hidden Markov model (HMM). The Baum-Welch algorithm is a generalized expectation-maximization (GEM) algorithm.

Although many consider that stochastic approaches for MT began in the 1980s, short contributions during the 1950s already pointed in this direction. Most systems were based on mainframe.

At the end of the 1980s, a group in IBM lead by P. Brown developed statistical models in [4], which lead to the widely used IBM models for Statistical Machine Translation (SMT).

During the 1990s is the development of machine translation systems for personal computers. At the end of the 1990s, many websites started offering MT services.
In the 2000s machines became more powerful and more data was available, making many approaches for SMT feasible.

2.2.2. Approaches to MT

The main paradigms for MT can be classified in three groups: Rule based MT, Corpus Based MT, and hybrid approaches.

2.2.2.1. Rule Based Machine Translation

RBMT is a paradigm that generally involves many steps or processes which analyze the morphology, syntactic and semantics of an input text, and use some mechanism based on rules to transform, using an internal structure or some interlingua the source language to the target language.

2.2.2.2. Corpus Based Machine Translation

The corpus based machine method involves the use of bilingual corpus to extract statistical parameters and applying them to models.

- Statistical Machine Translation (SMT)
  Using a parallel corpus, parameters are determined using the parallel corpus, and they are applied to their models. These methods were reintroduced by [4].
- Example based Machine Translation (EBMT)
  This approach is comparable to the case-based learning, with examples extracted from a parallel corpus.

2.2.3. Software Resources for Machine Translation

2.2.3.1. GIZA++

As part of the Egypt project, GIZA is used for analyzing bilingual corpora, which includes the implementation of the algorithms described in [4].
This tool was developed as part of the SMT system EGYPT. In addition to GIZA, GIZA++ was developed as an extension that includes Models 4 and 5, by Franz Josef Och [15].

It implements the Baum-Welch Forward-Backwards algorithm, with empty words and dependency on word classes.

2.2.3.2. MKCLS

MKCLS implements language modelling, and is used to train word classes, using a maximum-likelihood criterion [14]. The task of this model is to estimate the probability of a sequence of words, \( w_1^N = w_1, w_2, \ldots, w_N \). To approximate this, we can use \( Pr(w_1^N) = \prod_{i=1}^{N} p(w_i|w_{i-1}) \).

When we want to use this to estimate bigram probabilities (\( N = 2 \)), we will have problems with data sparseness, since most bigrams will not be seen. To solve this problem, we can create a function \( C \) that maps words \( w \) to classes \( C \), in this way:

\[
(8) \quad p(w_1^N|C) := \prod_{i=1}^{N} p(C(w_1)|C(w_{i-1})) \cdot p(w_i|C(w_1))
\]

This is obtained with a maximum likelihood approach:

\[
(9) \quad \hat{C} = \arg \max_{C} p(w_1^N|C)
\]

As described in [14], we can arrive to the following optimization, following [16].

\[
(10) \quad LP_1(C, n) = - \sum_{C,C'} h(n(C|C')) + 2 \sum_{C} h(n(C))
\]

\[
(11) \quad \hat{C} = \arg \max_{C} LP_1(C, n)
\]
MKCLS uses alignment templates extracted from the parallel corpus, which contain an alignment between the sentences.

The implementation of the template model was an early development of the phrase translation, which is used in current machine translation systems.

2.2.3.3. Moses

Moses, as GIZA++, is an implementation of a statistical machine translation system. It’s a factored phrase-based beam-search decoder. The phrase-based statistical translation model [10] is different since it tries to create sequences of words instead of aligning words.

Phrase-based models map text phrases without any linguistic information.

We model this problem as a noisy channel problem.

\[
p(e|f) = \frac{p(f|e)}{p(f)}
\]

Using the Bayes Rule we can rewrite this expression to the following.

\[
\arg\max_e p(e|f) = \arg\max_e \frac{p(f|e)}{p(e)}
\]

Each sentence in this case is transformed into a sequence of phrases, with the following distribution.

\[
\varphi(f_i|e_i)
\]

Using a parallel text line aligned, we produce an alignment between the words. To do this step, we use GIZA++.
For the example, we used the book of Revelation, chapter 22, versicle 21. In figure 2.2 and figure 2.2, we can see the translation alignment for English to Spanish and Spanish to English of this verse, which is the last verse of the Bible, using the New King James Version and the Dios Habla Hoy translations respectively.

- e: the grace of our lord jesus christ be with you all
- f: que el señor jesus derrame su gracia sobre todos

![Figure 2.1. Translations English to Spanish, and Spanish to English.](image)

![Figure 2.2. The intersection.](image)

Using the word alignment, several approaches try to approximate the best way to learn phrase translations.
2.3. Cross-language Information Retrieval

Cross-language information retrieval (CLIR) is the task of performing IR when the query and the collection are in two different languages.

One of the main motivations for the development of CLIR is the availability of information in English, and the need to make this resource available to other languages.

2.3.1. The Importance of Cross Language Information Retrieval

CLIR is important because most of the contents available online are written mainly in a couple dozen of languages, and they are not available for thousands of other languages, for which little or no content is available.

For example, a query in English in a search engine for ‘cold sickness’ would return 2,130,000 results, while the same search in Spanish, 'resfrio', will return only 106,000 documents. A search for this same string in Quechua for 'Qhonasuro' will show no documents available in the index.

A query in English for the word 'answer' will return 511 million relevant document. The query for the Spanish word 'respuesta' returns 76 million results. The same query for the word 'kutichiy' returns 638 documents.

A query in English for the words 'blind wiki' will return 2.4 million results, with 30 different senses in Wikipedia. The same query in Spanish, for 'ciego' will return 71 thousand relevant documents, with 10 different senses. The query for the word 'awsa' in Quechua will return two thousand results, without any entry in a Wiki page.

These examples show that although there is a lot of useful information on the Internet, it may not be accessible in most languages.

Unesco has estimated that today, around six thousand languages are disappearing, or at risk to disappear, out of close to seven thousand live languages spoken around the world [21].

A language not only constitutes a way of human communication, it also constitutes a repository of cultural knowledge, a source of wisdom and a particular way to interpret and to
see the world. In this context, this work is an effort to provide tools for native speakers to use their languages to access the vast amount of information available for other languages. This also constitutes a part of the effort to encourage the use of modern tools in different languages, other than the few major languages.

Creating modern tools like an IR system in different languages constitutes another way to preserve the languages. This can promote the use of one language, given that there are documents accessible to the people that speak it.

2.3.2. Main Approaches to CLIR

Depending on what we choose to translate, there are two main approaches for CLIR.

(i) Query translation: The main advantage of this method is that it is very efficient for short queries. The main disadvantage of this approach is that it is difficult to disambiguate query terms, and it is difficult to include relevance feedback.

(ii) Document translation: This approach involves translating the collection into the query language. This step can be done once, and only applied to new documents of the collection. This approach is difficult if the collection is very large, and the number of target languages increase, since each document will need to be translated into all target languages.

2.3.3. Challenges Associated with CLIR for Languages withScarce Resources

2.3.3.1. The Alphabet

To perform MT and CLIR and MT with any pair of languages, the first factor to consider is the written form of that language. A native written system may or may not be present, and in many cases, transliteration is needed to map words from one written system to the other. While some languages have already transliteration rules, the complexity of this task is important, but is beyond the present experiment.
In the absence of a native alphabet, many languages have adopted a borrowed alphabet from a different language. In the case of Quechua, it adopted the written alphabet of Spanish (which is based on the Latin alphabet). As a result, some phonetics have been represented differently according to different linguists:

- **Five-vowel system** This is the old approach, based on a Spanish orthography. Supported almost exclusively by the “Academia Mayor de la Lengua Quechua”.
- **Three-vowel system** This approach uses only the three vowels a, i and u. Most modern linguists support this system.

2.3.3.2. The Encoding

Another particular task on the development of resources for a scarce resource language at the implementation level will involve definitely the task to use different encodings in the application.

This sub task can be very challenging, since this issue doesn’t involve only problems with software and programs developed by other researchers, but also some times includes deciding many aspects that are specific to the language being studied.

For example, while it is easier to agree on the alphabetic order of characters in English and languages which depend on the Latin alphabet, this task can have different difficulties when the language is not English. Quechua uses the Spanish alphabet, but is very common to include the symbol ‘‘’. While there are rules on how to classify this symbol in Spanish, they may or may not make sense for Quechua, taking into account the phonetics of the symbol.

In the specific case of Quechua, many systems for phonetic transcription where present, and there was still some ongoing debate about this topic among linguists and native speakers.
2.4. Related Work and Current Research

2.4.1. Computational Linguistics for Resource-Scarce Languages

As we mentioned earlier, one of the first problems researchers face when building new NLP resources for a scarce resource language is the availability of written resources, native speakers and general knowledge for a language.

There are two possible approaches to perform this task.

(i) To work on a particular language, creating resources particular for it, working with experts to develop methods most of the times only applicable to the particular language.

(ii) To create methods that would be applicable with little effort to any language.

2.4.2. Focusing on a Particular Language

Specific methods have been researched and developed that were focused on a particular scarce resource language. [20] describes the process of creating resources for two languages, Cebuano and Hindi. Cebuano is a native language of the Philippines.

Most of the experiments done for languages with scarce resources work on the same initial grounds, which are to create as much resources as possible to start using traditional methods and models.

For example, [12] describe a shared task on word alignment, were seven teams were provided training and evaluation data, and submitted word alignments between English and Romanian, and English and French.

Other projects have focused as well the creation of resources for Quechua. For example, [5] described the efforts to create NLP resources for Mapudungun, an indigenous language spoken in Chile and Argentina, and Quechua. In this experiment, EBMT and RBMT systems were developed for both languages. As the other experiments mentioned, one of the main problems is focused on building the initial resources.
2.4.3. NLP for Languages with Scarce Resources

The construction of NLP resources for languages with scarce resources has been studied lately. For example, [6] proposes an unsupervised method for POS acquisition.


2.4.4. CLIR for Languages with Scarce Resources

One of the common tasks of CLIR is the experimentation with languages with few or no resources.

Nevertheless, it is necessary to note that this experiments have targeted most of the times languages of interest with many speakers. While TREC-2001 and TREC-2002 have focused on Arabic, other experiments have been also provided for languages with few or no resources at all. For example, DARPA organized in 2003 a competition where teams had to build from scratch a MT system for a suprise language, Indi in the context of the Translingual Information Detection, Extraction and Summarization (TIDES). In this experiments, researchers needed to port and created resources for a new language in a relative short amount of time, in this case, one month.

2.4.5. Cross Language Information Retrieval with Dictionary Based Methods

Earlier experiments worked with dictionary-based methods for CLIR. One of this works was presented by Lisa Ballesteros [3].

In this work, they found that Machine Readable Dictionaries would drop about 50% the effectiveness of retrieval due to ambiguity.

They reported that local feedback prior to translation improves precision. They also showed that this affects more longer queries, attributing the effect due to the reduction of irrelevant translations.
The combination of both methods leaded to better results both in short queries and long queries.

2.4.6. Cross Language Information Retrieval Experiments

Ten groups participated on the TREC-2001, which had to retrieve Arabic language documents based on 25 queries in English.

Since that was the first year that many resources in English were available, new approaches were tried[7].

Some results on this experiment were that query length did not improve the retrieval effectiveness.

In the TREC-2002, nine teams participated in the CLIR track [13].

The monolingual runs were mentioned as a comparative baseline to which cross-language results can be compared.

As it happened a year before, the teams observed substantial variability in the effectiveness on a topic to topic basis.

2.4.7. Word Alignment for Languages with Scarce Resources

More recently, the ACL organized a task[9] to align words for languages with scarce resources. The pairs were English-Inuktitut, English-Hindi and Romanian-English.

One of them most interesting outcomes of this work, relative to this study, is in the results for English-Hindi. The use of additional resources resulted in absolute improvements of 20%, which was not the case for languages with larger training corpora.
CHAPTER 3

EXPERIMENTAL FRAMEWORK

3.1. Metrics

For our experiments, we used precision and recall to evaluate the effectiveness of the results of our experiments.

In CLIR, the quality of the translation affects directly the results obtained using IR, but it’s different from the quality measured only by MT methods.

One of the main reasons for this is that most IR methods use bag of words approach, as opposed to MT, where the order of the words is a factor.

We used the normal metrics, precision and recall to evaluate the results returned by the systems.

To be able to analyze better the results, we also calculated the precision at other points (different from the 11 normal points), taking into account a progressive inclusion of results.

\[
P = \frac{(\text{Relevant documents})}{(\text{Documents returned})} \quad R = \frac{(\text{Relevant documents})}{(\text{Total relevant documents})}
\]

In the formula, in order to calculate more points, we made the total number of relevant documents vary accordingly.

3.2. Data Sources

To evaluate how different quantities of parallel text affects the translation quality, we used the following data sources:

3.2.1. Time Collection

The Times magazine collection is a set of 423 documents, associated with 83 queries, for which the relevance was manually assigned.
This collection contained originally 425 documents, with two of them repeated, document 504 and document 505 according to the original numeration. Both documents were not used for this reason.

There is a total of 324 relevant documents associated with the 83 queries.

3.2.1.1. Preprocessing

In order to use this collection, we needed to preprocess the files. The collection’s documents are distributed in a single file, which has to be parsed and transformed to be usable.

The files were numbered in a unspecified order. For this experiment, they were rearranged and indexed. The mapping is detailed in Appendix ‘A’.

3.2.2. Parallel Text

Many sets of parallel text were used on this experiment. The main source of parallel text used was the Bible.

3.2.2.1. The Bible as a Source of Parallel Text

Many authors have described advantages and disadvantages of using the Bible as a source of parallel text, for example [17].

The Bible is a source of high quality translation that is available in several languages. According to the United Bible Society [19], there are 438 languages with translations of the Bible, and 2,454 languages with at least a portion of the Bible translated, making it the parallel corpus that is available in more languages.

Table 3.1 contains all the versions that were used in this experiment.

All the Bibles required a preprocessing to be usable for this experiment.

This process involved many manual steps. The final format of the Bibles, for post processing was defined as the following:

Every new line involving a new versicle, was denoted with a line starting with the following: [a, b, c, d, e] Text of the versicle.
Table 3.1. List of English translations of the Bible used for this experiment.

<table>
<thead>
<tr>
<th>ID</th>
<th>Language</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>008</td>
<td>English</td>
<td>American Standard Version</td>
</tr>
<tr>
<td>009</td>
<td>English</td>
<td>King James Version</td>
</tr>
<tr>
<td>015</td>
<td>English</td>
<td>Youngs Literal Translation</td>
</tr>
<tr>
<td>016</td>
<td>English</td>
<td>Darby Translation</td>
</tr>
<tr>
<td>031</td>
<td>English</td>
<td>New International Version</td>
</tr>
<tr>
<td>045</td>
<td>English</td>
<td>Amplified Bible</td>
</tr>
<tr>
<td>046</td>
<td>English</td>
<td>Contemporary English Version</td>
</tr>
<tr>
<td>047</td>
<td>English</td>
<td>English Standard Version</td>
</tr>
<tr>
<td>048</td>
<td>English</td>
<td>21st Century King James Version</td>
</tr>
<tr>
<td>049</td>
<td>English</td>
<td>New American Standard Bible</td>
</tr>
<tr>
<td>050</td>
<td>English</td>
<td>New King James Version</td>
</tr>
<tr>
<td>051</td>
<td>English</td>
<td>New Living Translation</td>
</tr>
<tr>
<td>063</td>
<td>English</td>
<td>Douay Rheims 1899 American Edition</td>
</tr>
<tr>
<td>064</td>
<td>English</td>
<td>New International Version U K</td>
</tr>
<tr>
<td>072</td>
<td>English</td>
<td>Todays New International Version</td>
</tr>
<tr>
<td>074</td>
<td>English</td>
<td>New Life Version</td>
</tr>
<tr>
<td>076</td>
<td>English</td>
<td>New International Readers Version</td>
</tr>
<tr>
<td>077</td>
<td>English</td>
<td>Holman Christian Standard Bible</td>
</tr>
<tr>
<td>078</td>
<td>English</td>
<td>New Century Version</td>
</tr>
<tr>
<td>998</td>
<td>Quechua Bolivia</td>
<td>Quechua Catholic Bolivia</td>
</tr>
<tr>
<td>999</td>
<td>Quechua Cuzco</td>
<td>Quechua Catholic Cuzco</td>
</tr>
</tbody>
</table>

Where:

a) Bible translation ID  
b) Book number  
c) Chapter number  
d) Versicle number  
e) Type of annotation

The different types of annotation are:

4) Title  
5) Sub title  
6-10) Warning of duplicate versicles

We introduced lines to be skipped in the document, which contained annotation of the Bible, as many chapters have a name, and also, there are many titles for some stories. This titles were commented, and although they are present in the Bible file, they have a different notation, to be skipped by the post processing.

The versions of the Bibles contained many particularities, which are not relevant in this context. Nevertheless, some of them had to be accounted in order to process the parallel files.

3.2.2.2. Catholic and Protestant Versions

We used both Catholic and Protestant versions of the Bible. The difference between them, in this context, is the number of books included, and the length of some of the books.
We include a listing of the books of the Bible used in both versions.

3.2.2.3. Joined Versicles

Some of the Bibles contained versicles that were, for translation purposes, joined together into one. Some versions specifically had more joined versicles than others. This seemed to be arbitrary, since on inspection, the joined versicles didn’t match from version to version.

In order to create parallel files, we joined together the versicles when creating the parallel files for the parallel Bible.

In some cases, new joined versicles corresponded to versicles that were not joined in the other version. We grouped all the versicles that needed to be grouped so there was a parallel number of corresponding versicles in both sides.

This process is dynamic, since different versions of the Bible are grouped differently, in a case by case basis.
Table 3.2. Comparison of content between Catholic and Protestant Bibles.

<table>
<thead>
<tr>
<th>Book number</th>
<th>Book name</th>
<th>Versicles Catholic Bible</th>
<th>Versicles Protestant Bible</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Genesis</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>Exodus</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Leviticus</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>Numbers</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>Deuteronomy</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>Joshua</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Judges</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>Ruth</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1 Samuel</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>2 Samuel</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>11</td>
<td>1 Kings</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>2 Kings</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>13</td>
<td>1 Chronicles</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>2 Chronicles</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>15</td>
<td>Ezra</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>Nehemiah</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>Tobit</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>Judith</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>Esther</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>1 Maccabees</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>2 Maccabees</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>Job</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>23</td>
<td>Psalm</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>24</td>
<td>Proverbs</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>25</td>
<td>Ecclesiastes</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>26</td>
<td>Song</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>27</td>
<td>Wisdom</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>Sirach</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>Isaiah</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>30</td>
<td>Jeremiah</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>31</td>
<td>Lamentations</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>32</td>
<td>Baruch</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>Ezekiel</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>34</td>
<td>Daniel</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>35</td>
<td>Hosea</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>36</td>
<td>Joel</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>37</td>
<td>Amos</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>38</td>
<td>Obadiah</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>Jonah</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>Micah</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>41</td>
<td>Nahum</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>42</td>
<td>Habakkuk</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>43</td>
<td>Zephaniah</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>44</td>
<td>Haggai</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>Zechariah</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>46</td>
<td>Malachi</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>Matthew</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>48</td>
<td>Mark</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>49</td>
<td>Luke</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>50</td>
<td>John</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>51</td>
<td>Acts</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>52</td>
<td>Romans</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>53</td>
<td>1 Corinthians</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>54</td>
<td>2 Corinthians</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>55</td>
<td>Galatians</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>56</td>
<td>Ephesians</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>57</td>
<td>Philippians</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>58</td>
<td>Colossians</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>59</td>
<td>1 Thessalonians</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>60</td>
<td>2 Thessalonians</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Book number</td>
<td>Book name</td>
<td>Versicles Catholic Bible</td>
<td>Versicles Protestant Bible</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>61</td>
<td>1 Timothy</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>62</td>
<td>2 Timothy</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>63</td>
<td>Titus</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>64</td>
<td>Philemon</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>65</td>
<td>Hebrews</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>66</td>
<td>James</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>67</td>
<td>1 Peter</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>68</td>
<td>2 Peter</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>69</td>
<td>1 John</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>70</td>
<td>2 John</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>71</td>
<td>3 John</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>72</td>
<td>Jude</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>Revelation</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>
CHAPTER 4

EXPERIMENTAL RESULTS

4.1. Objectives

The objective of our experiments is to measure the impact of different variables on the quality of the results obtained doing CLIR, in the context and limitations of scarce resources languages.

Namely, in the first experiment we want to analyze the effect of the amount of parallel data on precision and recall. We contrast the quality of IR results on different corpus sizes. In the second experiment, we analyze if paraphrasing improves the results. In this experiment, we contrast the use of more than one version of the English Bible, and we align this to the Quechua Bibles. In the third experiment, we compare the results of translating the collection against translating the query. In the forth experiment, we analyze the possibility to use MT on similar dialects, and how this would affect the IR results. In the fifth and last experiment, we want to analyze if removing punctuation will affect in any way the results of the experiment. We remove punctuation in the parallel corpus and we contrast the results in different situations for experiment.

To see the upper bound of the experiments, we performed the IR experiment with the collection and query in English. Figure 4.1 shows the results using the vector model space doing monolingual IR on the collection, which is the upper bound of our CLIR experiment. We contrast this results with the baseline, which is to pick any document from the list.

As the size of the corpus grows, the precision and recall improved, and as a direct result, the results of the IR task improved. In this experiment, we want to analyze the impact of the increment of data available for the parallel corpus, and how much quantity will make an impact on the results.
For all the experiments, when we refer to the set of all Bibles, we refer to the aggregation of the following English Bibles:

(i) American Standard Version
(ii) King James Version
(iii) Youngs Literal Translation
(iv) Darby Translation
(v) New International Version
(vi) Amplified Bible
(vii) Contemporary English Version
(viii) English Standard Version
(ix) 21st Century King James Version
(x) New American Standard Bible
(xi) New King James Version
(xii) New Living Translation
(xiii) Douay Rheims 1899 American Edition
(xiv) New International Version U K
(xv) Today’s New International Version
(xvi) New Life Version
(xvii) New International Readers Version
(xviii) Holman Christian Standard Bible
(xix) New Century Version

The entire corpus has a total of 574,296 lines.
When we refer to the set of four Bibles, we refer to the aggregation of the following English Bibles:

(i) American Standard Version
(ii) Contemporary English Version
(iii) English Standard Version
(iv) New International Readers Version

The corpus of four Bibles has a total of 118,762 lines.
The set of two Bibles is the aggregation of the following English Bibles:

(i) American Standard Version
(ii) English Standard Version

The corpus of two Bibles has a total of 58,711 lines.
When we refer to the set of one Bible, we refer to the following Bible:

(i) English Standard Version

This corpus has a total of 27,620 lines.
For the experiments involving only sections of 5,000 lines, 10,000 lines, 15,000 lines, 20,000 lines and 25,000 lines, we used in all cases the English Standard Version.
4.2. Experiments

4.2.1. Experiment 1

<table>
<thead>
<tr>
<th>Analyze the effect of the size of the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation of the collection</td>
</tr>
<tr>
<td>Punctuation was not removed</td>
</tr>
</tbody>
</table>

We used the English Standard Version and the Quechua Cuzco Version of the Bible in order to construct the parallel files.

We aligned parallel versions of the two Bibles, and prepared the parallel files. In this case, we translated the queries from Quechua to English, using the different translators created based on different quantities of lines from the corpus, in sections of 5,000 lines. We evaluated precision and recall of the results obtained using the vector model.

![Figure 4.2: Increments of 5,000 lines for the parallel corpus.](image)

One thing that we observed is that there is a very clear difference between the results using the complete Bible, 27,610 lines, and the results obtained using only 25,000 parallel lines. This difference is less evident between the other sets, even when the difference of
parallel data is several times greater. One possible reason for this difference is that the New Testament uses words that are more likely to appear in a current text.

In general terms, more data positively affects the quality of IR results.

In table 4.1 we calculated the F-measure at different recall points. Then, we normalized this values with the upper bound, to see the impact of the increase of the size of the corpus.

<table>
<thead>
<tr>
<th>Recall point</th>
<th>5k</th>
<th>10k</th>
<th>15k</th>
<th>20k</th>
<th>25k</th>
<th>27.6k</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.9731</td>
<td>0.9590</td>
<td>0.9590</td>
<td>0.9567</td>
<td>0.9567</td>
<td>0.9975</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9203</td>
<td>0.9203</td>
<td>0.9153</td>
<td>0.9170</td>
<td>0.9119</td>
<td>0.9522</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8568</td>
<td>0.8142</td>
<td>0.8604</td>
<td>0.8554</td>
<td>0.8279</td>
<td>0.8889</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7995</td>
<td>0.7889</td>
<td>0.7820</td>
<td>0.7925</td>
<td>0.7943</td>
<td>0.8497</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7644</td>
<td>0.7133</td>
<td>0.6791</td>
<td>0.6910</td>
<td>0.7203</td>
<td>0.8224</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6030</td>
<td>0.5315</td>
<td>0.5232</td>
<td>0.5264</td>
<td>0.5469</td>
<td>0.6307</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4013</td>
<td>0.3217</td>
<td>0.3346</td>
<td>0.3217</td>
<td>0.3526</td>
<td>0.3685</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3534</td>
<td>0.3003</td>
<td>0.3112</td>
<td>0.2876</td>
<td>0.3085</td>
<td>0.3157</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3878</td>
<td>0.3650</td>
<td>0.3650</td>
<td>0.3650</td>
<td>0.3739</td>
<td>0.4105</td>
</tr>
<tr>
<td>µ</td>
<td>0.6733</td>
<td>0.6349</td>
<td>0.6366</td>
<td>0.6338</td>
<td>0.6437</td>
<td>0.6929</td>
</tr>
</tbody>
</table>

Figure 4.3. Contrasting 5000 lines with 27600 lines (complete Bible).
4.2.2. Experiment 2

Analyze the effect of paraphrasing.
Quechua Version used: Cuzco Version.
Translation of the collection.

In the second experiment we want to evaluate if paraphrasing improved precision and recall. Paraphrasing increases the size of the parallel corpus, but is aligned in all cases with the corresponding Quechua Bible, of which we only had one version.

For this reason, in figure 4.4 we compared the results for one Bible, the combination of two Bibles, and the combination of four.

![Figure 4.4. Query translation.](#)

We have to note that the increase of parallel data came from paraphrasing. We used different versions of the Bible in English, and we aligned all of them to the Quechua Bibles. We used only one of the Quechua Bibles at a time.

The results show an increase in the precision when using different versions of the Bible in the same language.

In figure 4.4 we can observe that the best performing corpus is the set of four Bibles.
In figure 4.5 we repeated the same experiment removing the punctuation, and again the best results belonged to the set of four Bibles.

We also compared both sets of results, and the best results came from the set of four Bibles, removing punctuation, as shown in figure 4.6.
Table 4.2. Normalized F-measure at different recall points.

<table>
<thead>
<tr>
<th>Recall point</th>
<th>1 bible</th>
<th>2 bibles</th>
<th>4 bibles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.9567</td>
<td>0.9567</td>
<td>0.9567</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9102</td>
<td>0.9170</td>
<td>0.9186</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8493</td>
<td>0.8198</td>
<td>0.9011</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7583</td>
<td>0.7601</td>
<td>0.8219</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7099</td>
<td>0.6791</td>
<td>0.7852</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5770</td>
<td>0.5342</td>
<td>0.6121</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3969</td>
<td>0.3994</td>
<td>0.4740</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3507</td>
<td>0.3765</td>
<td>0.4546</td>
</tr>
<tr>
<td>0.9</td>
<td>0.4268</td>
<td>0.4080</td>
<td>0.4544</td>
</tr>
<tr>
<td>μ</td>
<td>0.6595</td>
<td>0.6501</td>
<td>0.7087</td>
</tr>
</tbody>
</table>

In table 4.2 we can clearly see that although the corpus of 2 bibles performed nearly 1 % worst than the corpus of 1 bible, the best performance was for the corpus of 4 bibles, almost 5 % better than the others.
4.2.3. Experiment 3

Analyze the effect of translating collection vs query.

Quechua Version used: Cuzco Version.

One of the research questions we wanted to answer was if it was better to translate the query or the collection, or if both approaches would give the same result. In this experiment, we compare results of the experiments translating the query, and contrasting that result with the translation of the collection.

This experiment was constructed using the corpus of 5,000 lines, one Bible, and finally four Bibles, using the Cuzco dialect of the Quechua Bible without removing punctuation.

Figure 4.7 shows that translating the collection with 5,000 lines gave better results.

Figure 4.7. Query translation and collection translation (5k lines).

Figure 4.8 shows the same experiment using the corpus of one Bible, and we can see clearly that translating the collection yields better results too.

Lastly, we repeated the experiment with only 5,000 lines of parallel text. Figure 4.9 shows also that the translation the collection performed better than the query translation also in this case.
4.2.4. Experiment 4

Analyze the effect of using a different dialect

Quechua Version used: Cuzco Version and Bolivia
In this experiment, we want to see how much impact would make to use a different dialect for IR purposes. We want to see how much impact would have to perform the experiment using a different dialect in the query and the collection.

The Time collection was translated by human translators that spoke Quechua Cuzco. In this experiment, we contrast the impact of creating translators using the same dialect, Quechua Cuzco, with translators using a different dialect, with is Quechua Bolivia.

In figure 4.12 we translated the collection for the first set of experiments. In the next experiment, we used only one Bible, as shown in figure 4.11. In the following experiment, we used only five thousand lines of parallel text.
Figure 4.10. Using different dialects, translating collection for 5k lines corpus.

Figure 4.11. Using different dialects, translating collection for one bible corpus.

Only in the last experiment we can see a different behavior, where both dialects perform almost equally.

In the next set of experiments, we used query translation.
In figures 4.13, 4.14 and 4.15 the results don’t show a significant difference between both. We believe that this result is caused because keywords will be similar on both languages, and the difference between both will be less relevant if we translate a smaller amount of text, which is the case in query translation.
Figure 4.14. Using different dialects, translating query for one bible corpus.

Figure 4.15. Using different dialects, translating query for four bibles corpus

Table 4.6. F-measure average using different dialects on different sizes of corpus, with collection translation.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Cuzco dialect</th>
<th>Bolivian dialect</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k</td>
<td>0.1804</td>
<td>0.1810</td>
<td>0.2990</td>
</tr>
<tr>
<td>one bible</td>
<td>0.2036</td>
<td>0.1750</td>
<td>16.3466</td>
</tr>
<tr>
<td>four bibles</td>
<td>0.2009</td>
<td>0.1809</td>
<td>11.0547</td>
</tr>
</tbody>
</table>
Table 4.7. F-measure average using different dialects on different sizes of corpus, with query translation.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Cuzco dialect</th>
<th>Bolivian dialect</th>
<th>( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k</td>
<td>0.1783</td>
<td>0.1741</td>
<td>2.4258</td>
</tr>
<tr>
<td>one bible</td>
<td>0.1816</td>
<td>0.1750</td>
<td>3.7862</td>
</tr>
<tr>
<td>four bibles</td>
<td>0.1729</td>
<td>0.1809</td>
<td>4.4234</td>
</tr>
</tbody>
</table>

Table 4.6 shows clearly that there is an improvement when using the correct dialect when we do collection translation. Table 4.7 shows that there is no noticeable difference if we use different dialects when the query is translated.
4.2.5. Experiment 5

Analyze the effect of removing punctuation
Quechua Version used: Cuzco Version

During the previous experiments, we noticed that many words were aligned with punctuation. In this experiment, we wanted to investigate if removing punctuation would affect in any way the quality of the translation. For the experiment, we removed all punctuation signs.

In figures 4.16 and 4.17 we translated the collection. We can see that the results are almost similar, and there is no measurable difference between both.

![Graph showing precision and recall for Corpus with symbols and without symbols (four bibles)](image)

Figure 4.16. Removing punctuation, translating the collection.

For the next set, we translated the query. In figures 4.18 and 4.19 we can see a clear difference between the results obtained removing punctuation and not removing it.

Figures 4.16, 4.17, 4.18 and 4.19 show that while removing the punctuation didn’t improve the results for collection translation, it clearly improved the results for query translation.
Figure 4.17. Removing punctuation, translating the collection.

Figure 4.18. Removing punctuation, translating the query.
Figure 4.19. Removing punctuation, translating the query.
CHAPTER 5
DISCUSSION AND CONCLUSIONS

In experiment 1 we can see that the results showed clearly that an increase in the size of the parallel corpus results in better precision.

Given the relative size of the corpus, we can say that there was a significant increase in precision from 25k lines to the complete Bible, 27k lines. This can be an effect of words relevant for the experiment being included in the last books of the Bible.

In experiment 2 we can see that using paraphrasing to construct bigger corpus increases the precision for information retrieval purposes. The increase of precision in the experiment was consistent with the amount of parallel data that was used.

In experiment 3 we contrasted the translation of the query and the translation of the collection of documents. The translation of the collection proved to have a higher precision in all cases. To see if this was due to the quality of the parallel data, we contrasted the translation of the query and the collection when a different Quechua dialect was used. In this case, both the translation of the query and the collection returned comparatively similar results, indicating that choosing the query or the collection makes impact when the quality of the aligned data is higher.

In experiment 4 we analyzed the translation of the text using a different dialect from the one used originally. We translated the collection, and later the query.

For the collection translation, using the same dialect gave better precision than using a different dialect in all cases. We found the same difference at different amounts of information.

One interesting result was the translation of the query using different dialects. Both performed relatively similarly, under different amounts of parallel data.
In experiment 5 we shown that removing the symbols made a noticeable impact when doing query translation. For collection translation, there was almost no difference in the precision.

5.1. Contributions

For this work, we had to develop all the NLP resources for MT for Quechua Cuzco and Quechua Bolivia from the beginning.

We aligned this text with the English Bibles, and finally designed the experiments to measure the effect of different variables when performing IR.

We designed the experiments in order to compare different variables and see how they affect the results of SMT for IR purposes.

5.2. Conclusions

Based on this results, we have the following conclusions.

(i) How much does the amount of parallel data affect the precision of information retrieval?

All increases of additional parallel text increased the precision of the results. The best results corresponded to the parallel with more lines, with a significant difference over the other sets.

(ii) Is it possible to increase the precision of the information retrieval results using paraphrasing? We showed that paraphrasing had a direct impact on the results, increasing the precision. The best performing system was the corpus with 4 Bibles.

(iii) Is there a significant difference if we translate the query or the collection of documents? How is this affected by the other factors? Translating the collection performed better. While translating the query, we noticed that removing punctuation caused increased precision, while removing the punctuation didn’t make any impact for collection translation.
(iv) Is it possible to use Information Retrieval across different dialects of the same language? How much is the precision affected by this? While the results were inferior for the different dialect, they also made us believe that it is possible to use the same system for a different dialect. Collection and query translation had similar results in this case.
APPENDIX

TIME COLLECTION FILES
APPENDIX TIME COLLECTION FILES

Documents in the Time collection are identified by a number, which is associated with the query relevance document. This number range from 17 to 563.

In our experiment, we identified the documents using a correlative number, which has a mapping to the Time collection document.

Columns labeled ‘E’ contain the document number used in this experiment, while columns labeled ‘C’ contain the document number assigned in the times collection.
BIBLIOGRAPHY


