Query Translation and Disambiguation Approaches for Cross-Lingual Information Retrieval

Ph.D. Thesis

Vijay Kumar Sharma
(ID No. 2014RCP9541)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

December, 2019

Query Translation and Disambiguation Approaches for Cross-Lingual Information Retrieval

Submitted in

fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Vijay Kumar Sharma

ID: 2015RCP9541

Under the Supervision of

Dr. Namita Mittal

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

December, 2019

CERTIFICATE

This is to certify that the thesis entitled, **"Query Translation and Disambiguation Approaches for Cross-Lingual Information Retrieval"** being submitted by **Vijay Kumar Sharma (2014RCP9541)** is a bonafide research work carried out under my supervision and guidance in fulfillment of the requirement for the award of the degree of **Doctor of Philosophy** in the Department of Computer Science & Engineering, Malaviya National Institute of Technology Jaipur, India. The matter embodied in this thesis is original and has not been submitted to any other University or Institute for the award of any other degree.

Place: Jaipur
Date:

Dr. Namita Mittal
(Supervisor)
Associate Professor
Department of Computer Science and Engineering
MNIT Jaipur, India

DECLARATION

I, **Vijay Kumar Sharma**, declare that this thesis titled,”**Query Translation and Disambiguation Approaches for Cross-Lingual Information Retrieval**” and the work presented in it, are my own, I confirm that:

- This work was done wholly or mainly while in candidature for a research degree t this university.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always give. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself, jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Vijay Kumar Sharma
(2014RCP9541)

Date:
Place: MNIT Jaipur

ACKNOWLEDGMENT

Firstly, I would like to express my sincere gratitude to my advisor Dr. Namita Mittal for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I am grateful to Dr. Namita Mittal for enlightening me the first glance of research. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Girdhari Singh, Dr. E. S. Pilli, Dr. Santosh Vipparthi, and Dr. A. P. Mazumdar, not only for their insightful comments and encouragement, but also for their complicated set of questions which incented me to widen my research from various perspectives.

I would like to thank my co-scholars and friends Chandra Prakash, Kunwar Pal, Mithlesh Arya, Megha Sharma, Rohit Batwara, Virendra Meghwal, Pawan, Tanvi, and Avani for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. I would like to thank my senior Dr. Basant Agrawal for reviewing my thesis chapters. Also, I would like to thank my friends from other department in Malaviya National Institute of Technology Jaipur.

Last but not the least, I would like to thank my family: words can not express how grateful i am to my parents, brothers and sister for all of the sacrifices they have made on my behalf & for supporting me spiritually while writing this thesis and my life in general.

Abstract

The Web is a huge source of multi-lingual information. Internet users across the world are not familiar with all languages. It may also happen that the user required information is not available in the user's language but it is available in some other language. In such cases, the traditional information retrieval system could not retrieve the required information because it deals with only one language. There is a need for such a system that can deal with two languages, i.e., one is source or user's language and another is target document's language. Cross-Lingual Information Retrieval (CLIR) provides a bridge between the source and target language where a user can query in his regional language or get the required information in the target document's language.

Today's world is the world of the internet. The CLIR research is useful for internet users related to the application domains like agriculture, medical, tourism, research and development domains. The CLIR approaches depend on the language pair, training datasets and translation techniques. The translation techniques are categorized into conventional and trending. The manual dictionary and probabilistic dictionary are conventional translation techniques. Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) are trending translation techniques. A lot of research has been done in CLIR for foreign languages and many CLIR tools have been developed but for Indian languages, CLIR research is still in trend. The proposed CLIR research analyzes and explores the issues, challenges and opportunities for Hindi-English language.

In CLIR, a query is passed through four phases, i.e., translation, transliteration, disambiguation and expansion. Based on that, the techniques used in CLIR approaches are categorized into translation, disambiguation, transliteration and expansion models. In the literature, dictionary, parallel corpora, SMT, NMT, Wikipedia and web-based translation models are studied. Word co-occurrence and Point-wise Mutual Information (PMI) are the popular traditional disambiguation methods. Untranslated words are considered as the Out Of Vocabulary (OOV) words. Such words are transliterated by using transliteration generation and mining techniques. Query expansion is used to enhance IR effectiveness.

The proposed CLIR research depends on the language pair and training datasets. A lot of research results and training datasets are available for the foreign language but for Hindi-English language, limited researches are available in the literature and the latest datasets are FIRE 2010 and FIRE 2011 which are freely available.

The manual and probabilistic dictionary based CLIR approach are the basic approaches but we did not find state-of-art research results for FIRE 2010 and FIRE 2011 dataset. So we prepare experimental setup and get the research result for the manual and probabilistic dictionary based CLIR approach to avail the baseline results. Both the manual and probabilistic dictionary based CLIR suffers from the OOV word issue. To eliminate this issue, the size of the parallel corpus is increased. Now, it becomes necessary to retrain the IBM model to get the probabilistic dictionary. Therefore a term frequency model is proposed which uses example sentences and cosine similarity to translate the query words.

In the proposed research, previously unidentified translation and disambiguation issues related to the Hindi-English language pair are identified. Based on that, three research objectives are formulated that are related to translation, disambiguation and web-based translation resources. Hindi is a morphologically rich language. The state-of-art approaches use either the transliteration generation or transliteration mining technique to translate the morphologically variant or OOV words. Since transliteration generation and mining techniques are not able to fix the morphological irregularities, hence, four novel morphological variant solutions are proposed. Stop-words in Hindi and English languages are different and their effectiveness is null in perspective to information retrieval. But in perspective of CLIR, these stop-words are somehow related. These stop-words need to be refined. The proposed research deals with the translation mis-mapped and non-confident translation issues which are never been identified. The OOV word translation is the biggest challenge. Such words are even skipped by SMT. A novel solution is proposed for the OOV word issue.

In the literature, the OOV words are transliterated that is not a good solution because a word may be either a dictionary word or a named entity. The named entity recognition helps to recognize the named entities so that the transliteration technique can be followed for the named entities. Disambiguation performs a significant role to select the best translation. Previous CLIR approaches use either the maximum probabilistic score or maximum word association score (word co-occurrence or PMI). Weightage to both of the maximum probabilistic score and maximum association score may achieve better results. SMT provides static translation while web-based approaches provide the dynamic translation. Wikipedia is a popular encyclopedia. Wikipedia title and inter-wiki link attributes are used for translation. The proposed research addresses the Wikipedia issues of partially matched and ambiguous words which are never been handled in the literature. Other web lexical resources, i.e., Hindi WordNet, Indo WordNet, ConceptNet, and Online dictionaries are used for translation. These resources are never been used for Hindi-English CLIR.

The proposed research is evaluated with FIRE 2010 and 2011 datasets. Research effectiveness is measured by the recall, mean average precision, precision@5 and precision @10. The main findings of the proposed research are as follows.

1. Four morphological variant solutions and refined stop-word lists are incorporated in the translation induction algorithm which enhances the performance of CLIR over the probabilistic dictionary-based approach.

2. The solution of translation mis-mapped and non-confident translation issues are incorporated in the semantic morphological variant selection approach which semantically selects the word and enhances the performance of CLIR.

3. A context-based translation algorithm translates the OOV words which are even skipped by SMT. This algorithm performs better than SMT due to the translation of OOV words.

4. Bi-lingual word embedding skip-gram model does not perform well for the Hindi-English language pair due to the different sentence structure and the different number of vocabulary and stop-words in Hindi and English language.

5. Named entity based disambiguation did not perform well because nowadays word's translation or transliteration does not depend on the named entity tag. It depends on the word's popularity.

6. Maximum of average of words average probability and association score based disambiguation approach performs better than both of the individual maximum probability and maximum association score.

7. Wikipedia based CLIR addresses the partially matched word and ambiguous word issue, due to that, its performance is better than SMT for FIRE 2010 and approximately equal to SMT for FIRE 2011.

8. Other web based lexical resources are analyzed and experimented for exploring the oppurtunities for Hindi-English pair. Since Indo WorNet and ConceptNet are not lexically rich hence their performance is poor. Online dictionaries return too many irrelevant translations.

Contents

List of Tables

List of Figures

List of Abbreviations

IR - Information Retrieval
CLIR - Cross-Lingual Information Retrieval
DT - Dictionary-based Translation
CT - Corpora-based Translation
SMT - Statistical Machine Translation
NMT - Neural Machine Translation
PMI - Point-wise Mutual Information
RNN - Recurrent Neural Network
MD - Manually-constructed Dictionary
PD - Probabilistic Dictionary
WTD - Word Translation Disambiguation
STS - Statistical Term Similarity
FIRE - Forum for Information Retrieval Evaluation
OOV - Out Of Vocabulary
OOVTTM - Out Of Vocabulary Term Transliteration Mining
WSD - Word Sense Disambiguation
NER - Named Entity Recognition
LCSR - Longest Common Sub-sequence Ratio
TF-IDF - Term Frequency - Inverse Document Frequency
CSS - Cosine Similarity Score
EDS - Euclidean Distance Score
WE - Word Embeddings
NNM - Neural Network Model
RNN - Recurrent Neural Network
VSM - Vector Space Model
CFILT - Centre For Indian Language Technologies
HWN - Hindi WordNet
IWN - Indo WordNet
CON - COnceptNet
MAP - Mean Average Precision
BWESG - Bilingual Word Embedding Skip-Gram
CBTA-OOV - Context-Based Translation Algorithm for the Out Of Vocabulary

Chapter 1

INTRODUCTION

1.1 CLIR Introduction

Information Retrieval (IR) is a logical mechanism where a user can store, search, and retrieve relevant information from the target documents which should be in the same language as the query has. Today, the web has enriched by the multi-lingual content and the people across the world may not fully utilize the multi-lingual content because a user may feel the internet friendly with his/her regional language only. He/She may or may not query in the English language. According to global internet usage statistics, the non-English users and multi-lingual content are drastically increasing on the web as shown in Figure 1.1. The non-English users are unable to formulate their queries in English which generates an interruption to world communication.

In India, Hindi is the most popularly used language among the multiple local languages as shown in Figure 1.2. So, there is a need for such a platform where a user can query in his regional language and retrieve the relevant documents either in his regional or in other languages. IR technique which provides flexibility for the users to query in their regional language regardless of the target documents language is called Cross-Lingual Information Retrieval (CLIR) (Nagarathinam & Saraswathi, 2011). This motivates us to research in CLIR.

CLIR incorporates a translation approach which is either a query translation or a document translation approach. Query translation approach is preferred over the

[1] https://en.wikipedia.org/wiki/Global Internet usage

[2] http://www.mxmindia.com/2014/02/local-language-content-could-push-internet-users-drastically-says-iamai

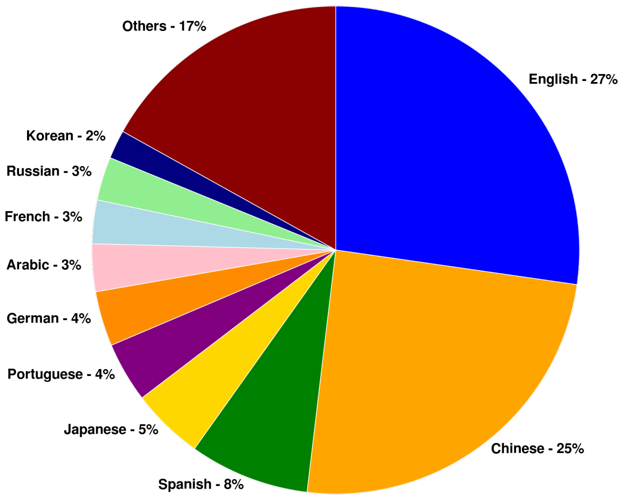


Figure 1.1: World's popular languages used over the Internet[1]

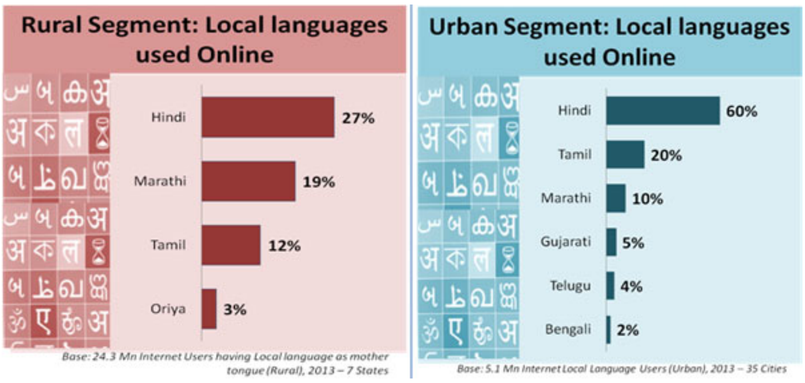


Figure 1.2: Indian local languages used over the Internet[2]

document translation due to the huge time and space consumed in document translation approach (Nasharuddin & Abdullah, 2010). Dictionary-based Translation (DT) and Corpora-based Translation (CT) are the state-of-art query translation approaches (Karimi, Scholer, & Turpin, 2011). Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) are the trending translation approaches which are also trained over the parallel corpus (Wu et al., 2016).

An easiest CLIR approach is the DT approach if a bilingual dictionary is available for that language pair. Since a bilingual dictionary has a limited vocabulary coverage, hence, many words are not translated by the bilingual dictionary. In the bilingual dictionary, a source language word has multiple translations which lead to the word translation disambiguation issue. The bilingual dictionary suffers from the dictionary coverage and word translation disambiguation issues (Pingali & Varma, 2006). A parallel or comparable bilingual corpus is used for query translation (Zhou, Truran, Brailsford, Wade, & Ashman, 2012). In parallel corpus, source and target language documents are aligned either at word level or sentence level. In comparable corpus, source and target language documents are aligned either at paragraph level or document level which convey the same meaning. Most of the CLIR approaches suffer from the non-availability of the parallel corpus and bilingual dictionary in the required language pair (Bharadwaj & Varma, 2011). If a parallel or comparable corpus is available for a language pair then a probabilistic dictionary can be constructed by applying an IBM model over the parallel corpus. A probabilistic dictionary has more vocabulary coverage than the manually constructed dictionary, however, many words are not found in the probabilistic dictionary such words are called Out Of Vocabulary (OOV) word and it is difficult to translate the OOV words (Zhou et al., 2012). Creating a parallel or comparable corpus from the raw corpora is another research issue. Parallel or comparable corpus-based approaches suffer from the corpus size, relevancy (well-matched training documents), granularity (finely grained corpora either at sentence level or document level) issues (Bradford & Pozniak, 2014).

The SMT is trained on a parallel corpus. Since the user queries are often too short that the SMT is unable to capture syntactic and semantic structure. SMT generates at most one translation for each query word. The accuracy of CLIR approach may be enhanced if synonyms of translated words or multiple translations are included in the translated query. The inclusion of multiple translations for each query word may also mix the noise in the translated query, due to that, the accuracy of the CLIR approach may be degraded. So, selection of N-best translations becomes a research issue (Ture & Lin, 2014). Word co-occurrence statistics, Pointwise Mutual Information (PMI) techniques are used to select N-best translations. An irrelevant translation is selected if PMI is applied on a document aligned corpus whereas a few translations are selected if PMI is applied on a sentence-aligned corpus, hence, there is a need for selection of parallel or comparable corpus of best size and quality. Neural networks impart a significant role in the field of data mining as it achieves surpassing results. The NMT model is trained and evaluated for the various foreign languages. A recurrent neural network based encoder-decoder architecture is trained over the parallel corpus to get the translations. Where, the author in (Kunchukuttan, Mehta, & Bhattacharyya, 2018) says that the SMT achieves higher BLEU score (a unit to measure translation accuracy) in comparison to NMT for Hindi to English translation.

A source language word is associated with the multiple translations in both of the manual dictionary and probabilistic dictionary, due to that, word translation disambiguation issue arises (Nagarathinam & Saraswathi, 2011; Zhou et al., 2012).

Since the translations in the probabilistic dictionary are associated with the probabilistic score, hence, the maximum probabilistic score is used to select the best translation. In a manual dictionary, the translations do not have any probabilistic score, therefore, a translation disambiguation approach is needed to select the best translation. Word co-occurrence, PMI, WordNet path length (Monz & Dorr, 2005), and statistical term similarity (Adriani, 2000) are the conventional translation disambiguation approaches. Statistical term similarity computes an optimal translation but it takes a high computation cost which increases exponentially with the sentence length. Today, Wikipedia is a huge source of comparable corpus. Internet users across the world can edit the Wikipedia knowledge base. Wikipedia becomes very useful for linguistic research due to its structure and content where an article has a unique title and links to the same-titled article in different languages called inter-wiki links (Schönhofen, Benczúr, Biro, & Csalogány, 2007).

CLIR research is promoted with the introduction of many workshops and forums like Cross-Language Evaluation Forum (CLEF) for European languages since 2000, NII Test Collection for IR System (NTCIR) for Japanese and other Asian languages, Forum for Information Retrieval Evaluation (FIRE) for Indian languages since 2008. An Indian government project *Development of Cross-Lingual Information Access system (CLIA)* is being developed by 11 Indian institutes. The objective of this project is to provide the query results in Hindi, English, and in the query language.

Recently, most of the IR and web search systems explicitly mention the problems of real-world data on the web. Especially in the case of Indian languages, string variation tends to occur a lot due to the larger size of the alphabet, lack of standardization, and transliteration variants. Some mapping rules are used to handle such data (Pingali & Varma, 2005). Most of the Indian language content on the web is not searchable due to multiple encodings of web pages. To overcome this problem, a transcoding method is exploited to convert non-UTF-8 encoding into UTF-8 encoding (Pingali, Jagarlamudi, & Varma, 2006). Conventional approaches, available tools and techniques, issues and challenges, research gaps and objectives, thesis organization are discussed in the subsequent sections.

Generally, CLIR is used where user queries are in different language than the target documents language, apart from that, the significance of CLIR research towards the different application domains are as follows.

1. In the agriculture domain, a farmer can search the required information in his regional languages which may or may not be available in his regional language.

2. In the medical domain, people are not aware of the medical terms so they can search in their regional languages.

3. In the tourism domain, people travel to the different-different countries but they don't know the language of that country and their survival become difficult in these countries. CLIR provides a way to get the information by using their regional language.

4. In the research and development, advanced countries follow their own languages. Their research about the technology, defense, arms, etc. are published in their own languages. It becomes difficult to get their research, therefore, CLIR provides a way by which one can access their research which is in a different language.

1.2 Conventional Approaches

Various conventional CLIR approaches are discussed in Table 1.1.

Table 1.1: Conventional CLIR approaches

S.no.	Approaches	Description & Issues	References
1	Bi-lingual dictionary	Contains translation pairs and suffers from the dictionary quality and coverage issue.	(Nagarathinam & Saraswathi, 2011; Nasharuddin & Abdullah, 2010)
2	Parallel corpus	Contains bilingual text which is either sentence aligned or word aligned. Corpus-based approach performs better than the dictionary but it suffers from the unavailability of parallel corpus issue.	(Nagarathinam & Saraswathi, 2011; Nasharuddin & Abdullah, 2010)

3	Machine translation	Machine translation system is computationally expensive for document translation whereas it does not translate the queries properly because queries are too short.	(Boretz & Adam, 2009).
4	Transliteration	Out of Vocabulary words are transliterated by either string matching or phonetic mapping. String matching is useful for the languages which share common alphabets but transliteration variant may be an issue. phonetic mapping is useful for the languages which contain dissimilar alphabets but missing sound may be an issue.	(Nagarathinam & Saraswathi, 2011; Makin, Pandey, Pingali, & Varma, 2007)
5	Co-occurrence method	Term-term co-occurrence method is used for translation disambiguation. It uses a monolingual corpus which is not available for a wide range of languages.	(Chinnakotla, Ranadive, Damani, & Bhattacharyya, 2007; Yuan & Yu, 2007).
6	Ontology	An explicit specification for a conceptualization, the ontological knowledge and its link to the dictionaries is used for CLIR.	(Yu, Zheng, Zhao, Li, & Yu, 2006; Monti, Monteleone, Di Buono, & Marano, 2013)
7	Wikipedia	It is an open access multilingual encyclopedia with the total of six million articles in 250 languages. Inter-wiki link is used for the translation purpose.	(Su, Lin, & Wu, 2007; Sorg & Cimiano, 2012)
8	Google translation	Google translation provides an online translation service which is biased towards named entities.	(Xiaoning et al., 2008)

9	Universal networking language	In universal networking language, concept nodes (Universal words) are connected in a hypergraph.	(Samantaray, 2012)
10	Web-based translation	The parallel and comparable web documents and search result documents are used for translation and disambiguation.	(J. Zhang, Sun, & Min, 2005; Nie, Simard, Isabelle, & Durand, 1999; Lu, Chien, & Lee, 2004)
11	Word sense disambiguation	Word sense disambiguation is the ability to identify the appropriate sense of the words.	(Navigli, 2009)
12	Named entity recognition	A word's category, i.e., person name, organizations, locations, etc is predicted.	NER (Prasad & Fousiya, 2015)
13	Lemmatization	A word is simplified to its uninflected form in lemmatization while different grammatical variants of a word are reduced to a common short form called a stem by removing the word endings in stemming.	(Nasharuddin & Abdullah, 2010)

1.3 Tools and Techniques

A brief study of six CLIR tools is represented in Table 1.2. These tools use a bilingual dictionary for the translation purpose due to the less computation time. An issue of user-assisted translation is tried to resolve in MIRACLE, MULTI LEX EXPLORER, and MULTI SEARCHER (Ahmed & Nürnberger, 2012).

Table 1.2: A brief study of CLIR tools

S.no.	Tools	Translation technique	Languages	Limitation
1	MULINEX	Bilingual dictionary, back translation	French, Germen, English	Homonymy & synonymy, user assisted translation
2	KEIZAI	Bilingual dictionary, parallel corpora	English, Japanese, Korean	Homonymy & synonymy, user assisted translation
3	UCLIR	Bilingual dictionary, machine translation	Arabic languages	Irrelevant translation & user assisted translation
4	MIRACLE	Bilingual dictionary	English and other languages	Resource unavailability, synonymy & homonymy
5	MULTILEX EXPLORER	WordNet & web search engine	Multi lingual	Wordnet not available for all languages
6	MULTI SEARCHER	Bilingual dictionary, parallel corpora, mutual information & entity recognition	Multi lingual	Non-availability of parallel corpora

1.4 Issues and Challenges

Various issues and challenges which are recognized during the literature survey and experimentation are discussed in Table 1.3.

Table 1.3: CLIR issues and challenges

Issues and Challenges	**Definition**	**Example**
Homonymy (Chinnakotla et al., 2007; Gao et al., 2001)	A word may have two or more different meanings	"Rock" means "a genre of music" or "a stone"
Polysemy (Chinnakotla et al., 2007; Gao et al., 2001)	A word may have multiple related meanings	"Hospital" may be a place for lodging guest or a place for treating the ill
Word inflection (Gaillard, Boualem, & Collin, 2010)	A word may have different grammatical forms	go, went, and gone are different forms of word "go"
Multi-word translation (Gaillard et al., 2010; Pingali & Varma, 2007)	Phrase gives different meaning then the words of phrase	"Couch potato" used for someone who watches too much television
Lack of resources (Bharadwaj & Varma, 2011)	Unavailability of resources for experimentation	Dictionary, parallel corpora, machine translation, character encoding
Out of vocabulary terms or dictionary coverage (Erdmann, Nakayama, Hara, & Nishio, 2009; Pingali & Varma, 2006; Zhou et al., 2012)	Words which are not available in dictionary or corpus	"H1N1 Malaysia" is a newly added term for influenza disease
Transliterable entity identification (Rao & Sobha, 2010; Saravanan, Udupa, & Kumaran, 2010)	Identify which words needs to be translated or transliterated	प्रतिभा, लाल

Transliteration variants (Karimi et al., 2011; Mathur & Saxena, 2014)	Different dialects in a language leads to transliteration variants	Vidhyalay & vidhalay
Named entity annotated data (Bhagavatula, GSK, & Varma, 2012)	Not enough annotated data available to identify out of vocabulary terms	
Corpus size, quality, relevancy, granularity (Bradford & Pozniak, 2014)	Corpus data needs to be matched with application data & fine-grained corpora	
Domain dependency (Shakery & Zhai, 2013)	Training data should be general	
Wikipedia issues (Schönhofen et al., 2007)	Article unavailability & wrong inter-wiki link	
Selecting n-best translation (Ture & Lin, 2014)	Selecting optimal number of translation for each query word	
Computational complexity (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Moen & Marsi, 2013)	Efficient techniques take a high computational cost	CL-LSI, Neural language model

1.5 Research Gaps

Following research gaps are identified during the literature survey.

1. Translation accuracy of probabilistic dictionary depends on the corpus size. A large parallel corpus is not available for Hindi-English language (Mikolov, Le, & Sutskever, 2013). Hindi is a morphologically rich language, due to that, many words remain untranslated. Although the morphological variants of such words are available in the parallel corpus but the actual word is skipped without performing translation due to the morphological irregularities (Saravanan et al., 2010; Shishtla, Ganesh, Subramaniam, & Varma, 2009).

2. A source language word has multiple target language translations. The right translation depends on the context which is computed by association based disambiguation approaches, i.e., word co-occurrence, PMI, WordNet path length, etc., which is not sufficient (Monz & Dorr, 2005). Statistical term similarity provides accurate translation but at the high computation cost which is not acceptable (Adriani, 2000).

3. Machine translator selects only one best translation. The CLIR performance may be enhanced with the selection of multiple translation or synonyms of the translation. Sometimes, the performance of CLIR may be degraded, hence, there is a need to define a translation term selection criteria (Federico & Bertoldi, 2002).

4. On-line translation resources are widely used for foreign language. Wikipedia is a good translation resource due to the title and inter-wiki links attributes which make Wikipedia capable to do the translation. Wikipedia and other web-based translation resources are needed to be explored for Hindi-English (Redkar, Singh, Joshi, Ghosh, & Bhattacharyya, 2015; Sorg & Cimiano, 2012).

1.6 Research Objectives

Following research objectives are formulated based on the above research gaps.

1. To develop efficient and effective word translation approaches.
   (a) Analysis of various statistical translation approaches (manual dictionary, probabilistic dictionary, SMT, NMT, CL-LSI, CL-LDA).
   (b) Developing an efficient and effective translation algorithm which includes morphological variants solutions to address the translation mis-mapped and non-confident translation issues.
   (c) Developing a translation approach for the out of vocabulary words.
2. To develop word translation disambiguation approaches.
   (a) Analysis of various translation disambiguation approaches (word co-occurrence, PMI, statistical term similarity, named entity recognition).
   (b) Developing the maximum of words average probability and association score based disambiguation approach.
3. To explore web resource based translation resources.
   (a) Developing an online Wikipedia based translation approach.
   (b) Exploring other web-based lexical resources for translation.

1.7 Thesis Contribution

In this thesis, our contributions are given as follows.

1. Proposing the manual dictionary and term frequency model based CLIR approaches.
2. Proposing a translation induction algorithm based CLIR approach to handle morphological irregularities.
3. Proposing a semantic morphological variant selection algorithm to address translation mis-mapped and non-confident translation issues.

4. Proposing a context-based translation algorithm for the out of vocabulary words.
5. Proposing the bilingual word vector and named entity recognition based disambiguation approaches.
6. Proposing a maximum of average of words average probability and association score based disambiguation approach.
7. Proposing a Wikipedia based translation approach to address the Wikipedia issues.
8. Proposing web-based lexical resources based translation approaches, i.e., Hindi WordNet, Indo WordNet, Concept Net, and online dictionaries.

1.8 Thesis Organization

This chapter represents the basic introduction of CLIR, motivation, conventional translation and disambiguation approaches, available CLIR tools and techniques, issues and challenges, research gaps, and research objectives. Thesis organization for the subsequent chapters is given as follows.

1. Chapter - 2 represents the literature survey about the dictionary and corpus-based translation approaches, SMT, NMT, word translation disambiguation approaches, and web-based translation approaches.
2. Chapter - 3 discusses the proposed translation approaches which are based on the manual dictionary, term frequency model, translation induction algorithm, semantic morphological variant selection, and context-based translation for the out of vocabulary words.
3. Chapter - 4 discusses the proposed translation disambiguation approach, which is based on both of the probabilistic score and association score. PMI and word embedding techniques are used to measure the association score.
4. Chapter - 5 explores the web-resources based translation approaches, i.e., Wikipedia, Hindi WordNet, Indo WordNet, ConceptNet, online dictionaries.
5. Chapter -6 represents the conclusion and future work for the thesis.

Chapter 2

LITERATURE SURVEY

CLIR incorporates a translation approach followed by mono-lingual information retrieval. In the literature, three types of translation approaches are discussed, namely, query translation, documents translation, and dual translation. The user queries are translated into target documents language in query translation approach. The target documents are translated into user query language in documents translation approach. In dual translation, the user queries and target documents are mapped into a common language or a dual semantic space (Karimi et al., 2011). The document translation approach consumes a lot of computation time and space, so the query translation approach is preferred.

A translation approach for CLIR includes query translation, disambiguation, transliteration, and expansion models. In the state-of-art, three main approaches are discussed for query translation, i.e., dictionary-based translation, corpora-based translation, and statistical machine translation (Karimi et al., 2011). The previous survey on CLIR is represented in Table 2.1, where each row represents the research paper followed by the CLIR approach and comments. The entries corresponding to the CLIR approach are 'Y' (Yes), if the paper contains a survey about the CLIR approach, else 'N' (No). The table contains the entries about the TransLation Models (TLM), Disambiguation Models (DM), TransliTeration Models (TTM), and Expansion Models (EM). In this chapter, the query translation, disambiguation, transliteration, and expansion models are discussed in sections 1, 2, 3 and 4 respectively.

Table 2.1: Study of previous CLIR literature survey papers

Articles	**TLM**	**DM**	**TT**	**EM**	**Comments**
(Karimi et al., 2011)	N	N	Y	N	Phonetic and spelling based transliteration models are covered
(Nasharuddin & Abdullah, 2010)	Y	N	N	N	DT, CT, and SMT are covered
(Nagarathinam & Saraswathi, 2011)	Y	Y	N	N	DT, CT, and Wikipedia based named entity recognition approaches are covered
(Zhou et al., 2012)	Y	N	N	Y	DT, CT, SMT, transitive and dual translation models are discussed
(Bajpai & Verma, 2014)	Y	N	N	N	Traditional translation models are discussed in Indian language perspective
(V. K. Sharma & Mittal, 2016a)	Y	Y	N	N	DT, CT, SMT, Wikipedia, and named entity recognition models are discussed

2.1 Translation Models

This section contains a detailed study of the conventional and trending translation models.

2.1.1 Dictionary-based Translation

The resource-scarce languages suffer from an issue of non-availability of a dictionary and it is cumbersome to create a bilingual dictionary manually. A word can have multiple translations in a bilingual dictionary. A word translation disambiguation technique is required to select the best one out of multiple translations

(Ballesteros & Croft, 1996; Duque, Martinez-Romo, & Araujo, 2015; Pingali & Varma, 2006, 2007).

Words and phrases are the commonly used translation units. The construction of phrasal dictionary is a cumbersome task (Gao et al., 2001). Word's domain knowledge is required for word translation disambiguation, hence, the HowNet[1] and ontologies are employed with the bilingual dictionary to acquire the domain knowledge (Yu et al., 2006). English is a very rich kind of language, which has a powerful feature of spelling standardization but Hindi doesn't have this feature, for example, a word "प्रयास" (prayas) in Hindi language has multiple variants "प्रयासरत" (prayasrat), "प्रयासशील" (prayassheel). Hence, the following approximate string matching techniques are used to map such words in the dictionary (Makin et al., 2007).

- Jaro-Winkler Similarity: The maximum score is assigned to a pair s, t which shares a common prefix of maximum length.

$$Jaro-Winkler(s,t) = Jaro(s,t) + \left(\frac{P}{10} * (1.0 - Jaro(s,t))\right) \tag{2.1}$$

$$Jaro(s,t) = \frac{1}{3}\left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s|}\right) \tag{2.2}$$

  P: length of common prefix; s, t: input strings; s': characters in s that are common with t; t': characters in t that are common with s; $T_{s',t'}$: number of transpositions of characters in s' relative to t'.

- Levenstein Distance: It is a string similarity measure which calculates the minimum cost needed to convert a string s into another string t. The Levenstein distance between two string a & b of length $|a|$ & $|b|$ is given by

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & if(min(i,j)=0)) \\ min\begin{cases} lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_i)} \\ lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \end{cases} & otherwise \end{cases} \tag{2.3}$$

  $lev_{a,b}(i,j)$ is the distance between the first i characters of a and first j characters of b, which is equal to 0 when a_i=b_j else equal to 1.

- Longest Common Sub-sequence Ratio (LCSR): It computes the ratio of the longest common sub-sequence of pair s, t to the longest string amongst the

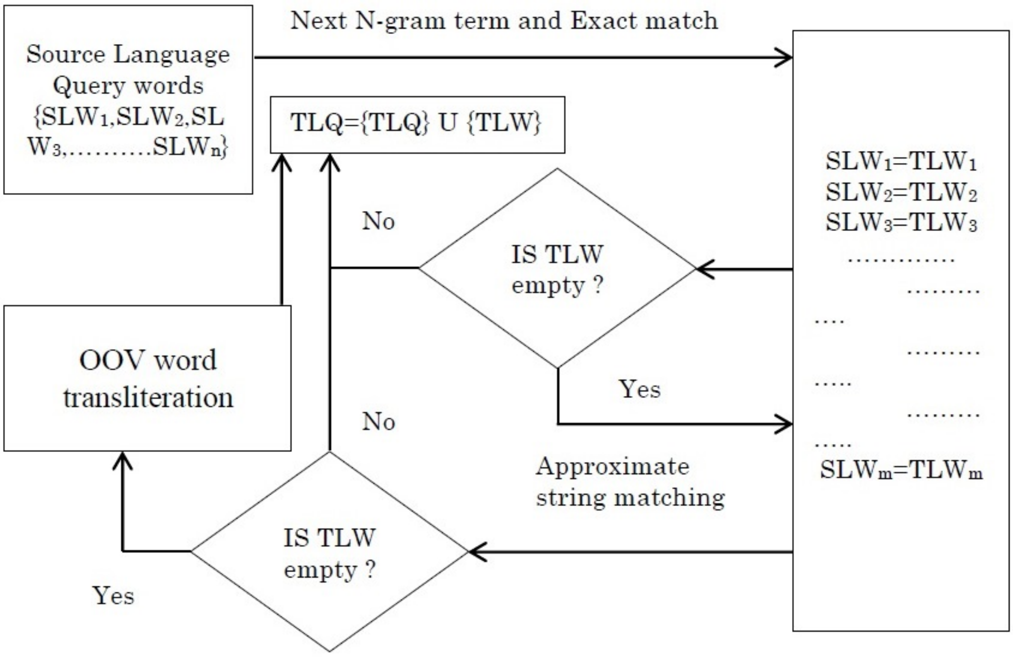
[1] http://www.keenage.com/html/c_index.html

two. The LCSR is defined as

$$LCSR(s,t) = \frac{|LCS(s,t)|}{max(|s|,|t|)} \tag{2.4}$$

LCS(s,t) is the Longest Common Sub-sequence between the string s and t.

The dictionary-based approach is represented in Figure 2.1. N-gram terms are created from the source language query words. These n-gram terms are searched into the dictionary, if found, then these terms are replaced by the corresponding translations. Words which are exactly not matched in to the dictionary are translated with the help of approximate string matching technique. In which, a query word is mapped to a source language dictionary word which has the maximum string matching score. Many words which are not translated by the dictionary are considered as the out of vocabulary words. These out of vocabulary words are either the named entity or newly identified terms (Chinnakotla et al., 2007; Sethuramalingam & Varma, 2008). It is very difficult to identify the named entities, for example, a word "प्रतिभा" (pratibha) may be a dictionary word (with the translations genius, and talent) or a named entity term (with the transliteration pratibha).


Next N-gram term and Exact match
Source Language Query words {SLW1,SLW2,SLW3,..........SLWn}
TLQ={TLQ} U {TLW}
SLW1=TLW1
SLW2=TLW2
SLW3=TLW3
SLWm=TLWm
No
IS TLW empty ?
Yes
OOV word transliteration
No
Approximate string matching
IS TLW empty ?
Yes

SLW: Source Language Word, TLW: Target Language Word, TLQ: Target Language Query

Figure 2.1: Dictionary based query translation

The remaining out of vocabulary words are transliterated by using the rule-based approach where an out of vocabulary word is converted into its romanized form. The dictionary-based approaches are not efficient due to the out of vocabulary words issue.

2.1.2 Corpora-based Translation

Parallel and comparable corpora are widely used to create a dictionary and to train the SMT. The resource-scarce languages suffer from the lack of availability of parallel and comparable corpus at any level (document, sentence or word level). It is arduous to construct such a corpus but if a parallel corpus is available then corpora-based approaches would be very efficient and effective. The effectiveness of corpora-based approaches depends on the corpus size and quality (Ganguly, Leveling, & Jones, 2012; Jagarlamudi & Kumaran, 2007; Shakery & Zhai, 2013). Since building a document aligned parallel corpus is easier than the sentence aligned, hence, the parallel sentence extraction mechanism from the document aligned parallel corpus is discussed for the resource-scarce languages (Chu, Nakazawa, & Kurohashi, 2016). Various corpora-based approaches are discussed in the subsections.

Probabilistic Dictionary based Translation

A probabilistic dictionary is used to translate the user queries (Ganesh, Harsha, Pingali, & Verma, 2008; Mahapatra, Mohan, Khapra, & Bhattacharyya, 2010; Nie et al., 1999; Saravanan et al., 2010). The IBM (Larkey, Connell, & Abduljaleel, 2003), and pearson's correlation coefficient (Shakery & Zhai, 2013) models are used to construct the probabilistic dictionary from the parallel corpus, which are briefly discussed below.

- IBM Model: A source language sentence s=(s_1,s_2,. . . .s_n) of length n and its corresponding target language sentence t=(t_1,t_2,. . . .t_m) of length m are given. The translation probability of a source language word s_i to a target language word t_j, with an alignment a : j → i is given as

  $$p(t,a|s) = \prod_{j=1}^{m} tp(t_j|S_{a(j)}) \tag{2.5}$$

  t_p represents translation probability of a target language word against a source language word.

- Pearson's Correlation Coefficient: Each word is represented by its frequency vector. The bilingual word pairs are scored based on the similarity between their frequency vectors. Let a word a in language L1 and word b in language L2 having the normalized frequency vectors (a_1,a_2,...a_n) and (b_1,b_2,...b_n) where $a_i = \frac{c(a,d_i)}{\sum_{j=1}^{n} c(a,d_j)}$, $b_i = \frac{c(b,d_i)}{\sum_{j=1}^{n} c(b,d_j)}$, and $c(a, d_i)$ is the count of word a in document d_i. The similarity between word a and b is computed by using $S(a, b)$.

$$S(a,b) = \frac{\sum_{i=1}^{n} a_i b_i - \frac{1}{n}\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i}{\sqrt{\left(\sum_{i=1}^{n} a_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)^2\right)\left(\sum_{i=1}^{n} b_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} b_i\right)^2\right)}} \tag{2.6}$$

A generalized N-gram cross term retrieval model is used for probabilistic information retrieval (Zhao, Huang, & Ye, 2014). In the probabilistic dictionary, a source language word has multiple target language translations associated with the probabilistic score which leads to a word translation disambiguation issue.

Cross-Lingual Latent Semantic Indexing (CL-LSI)

CL-LSI uses a parallel corpus and establishes a relationship among the words in a dual semantic space (Bradford & Pozniak, 2014; Zhou et al., 2012). Each source language document and its corresponding target language document are considered as a single document. A Term Frequency-Inverse Document Frequency (TF-IDF) matrix a is created, where each row corresponds to a term and each column corresponds to a document. An element $a_{m,n}$ of the matrix a represents the TF-IDF score of the m^{th} term corresponds to the n^{th} document. Further, singular value decomposition is applied to the matrix a which divides the matrix a into the matrices U, V & S, such that U and V are the orthogonal matrices, and S has the non-zero values only for the diagonal.

$$A = USV^T \tag{2.7}$$

The k largest values of S are deleted together with the corresponding columns in U and V to reduce the dimensionality. The proximity of any two words is measured by using cosine similarity score.

The performance of CL-LSI depends on the corpora size and quality as the TF-IDF matrix of a huge parallel corpus is also huge. The CL-LSI uses the singular value decomposition for dimensionality reduction, due to that, the singular value decomposition becomes inefficient with the large matrices. It takes a large amount of space at the time of matrix storage and a huge amount of time in dimensionality reduction phase, so the CL-LSI is not an efficient approach.

Cross Language Latent Dirichlet Allocation (CL-LDA)

A parallel corpus is utilized to construct a CL-LDA model (Negi, 2011; Vulić, De Smet, & Moens, 2013). A latent variable concept is used in the CL-LDA which represents a document as a mixture of topics. A dual-language document $d_i = \{w_{s1}, ..., w_{sn}, \quad w_{t1}, ..., w_{tm}\}$ is constructed from the parallel document as follows:

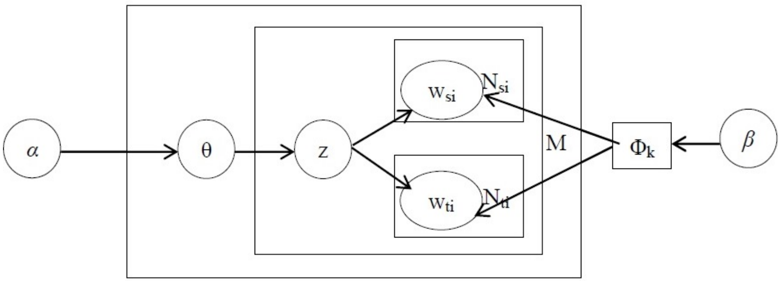


Figure 2.2: A CL-LDA Model

(i) Select a multinomial distribution $\theta_d \sim dir(\alpha)$ with a Dirichlet parameter α.

(ii) Choose a topic $z \in \{1, 2, ...K\}$ from the multinomial distribution θ_d , which determines topic assignment for the dual-language document.

(iii) Select a word token w from multinomial distribution $\phi_z \sim dir(\beta)$.

An CL-LDA model is represented in Figure 2.2, where k represents the number of topics, M represents the number of documents, N_{si} & N_{ti} are the number of tokens and w_{si} & w_{ti} are the words in the source & target language. Two sets of the probability distribution are acquired for each of the languages. The first set consists per-topic word probability distribution.

$$p(w_i|z_k) = \frac{n_k^{(w_i)} + \beta}{\sum_{j=1}^{|w^s|} n_k^{(w_j)} + |w^s|\beta} \tag{2.8}$$

$n_k^{(w_i)}$denotes the number of times when the topic z_k is assigned to the word w_i from the source vocabulary w^s. The sum $\sum_{j=1}^{|w^s|} n_k^{(w_j)}$ is the total number of words assigned to the topic z_k, and $|w^s|$ is the total number of distinct words in the source vocabulary. The second set consists per-document topic probability distribution.

$$p(z_k|d_j) = \frac{n_j^{(k)} + \alpha}{\sum_{j=1}^{k} n_j^{(l)} + K\alpha} \tag{2.9}$$

$n_j^{(k)}$ denotes the number of times when a word in the document d_j is assigned to the topic z_k. These two types of probabilities are obtained from CL-LDA model. These probabilities are used for target document retrieval and for bi-lingual lexicon construction.

- Target document retrieval: The probability of generating a source language query from a target language document is given by

  $$P(q|d_j) = P(q_1, ..., q_m|d_j)$$
  $$P(q|d_j) = \prod_{i=1}^{m} P(q_i|d_j)$$

  $$P(q_i|d_j) = (1 - \delta_1)\prod_{k=1}^{K} P(q_i|z_k^s)P(z_k^t|d_j) + \delta_1 P(q_i|Ref) \tag{2.10}$$

  δ_1 represents the interpolation parameter and $P(q_i|Ref)$ is the maximum likelihood estimate of the query word q_i in a monolingual source language reference collection. $P(q_i|z_k^s)$ is computed for all source language topics k = 1,2,...,K on query documents, and $P(z_k^t|d_j)$ is computed for all target language topics k = 1,2,...,K on target test documents.

- Bilingual lexicon construction: A bilingual lexicon is constructed by using Cue and Ti method. Since these methods use a limited topic space while extracting lexicon entries, hence, these methods are computationally feasible. The Cue method's of similarity between a source language word w_s and a target language word w_t is given as

  $$Sim(w_s, w_t) = \sum_{j=1}^{K} P(w_t|z_j)P(z_j|w_s) \tag{2.11}$$

  In Ti method, a Term Frequency-Inverse Topical Frequency (TF-ITF) matrix is constructed for source language words w_s and target language words w_t against k = 1,...,K topics. The TF scores of source language word w_s for the source topic z_k is computed as follows.

$$TF_{s,k} = \frac{n_{k,s}^{(w_s)}}{\sum_{w_j \in w^s} n_{k,s}^{(w_j)}}$$

ITF score measures the importance of the source word w_s against all source topics. Higher importance is given to the rare words. The ITF for the source word w_s is computed as

$$ITF_s = log \frac{K}{1 + |k : n_{k,s}^{(w_s)} > 0|}$$

The final TF-ITF score of a source word w_s and topic z_k is computed as $TF - IDF_{s,k} = TF_{s,k}.ITF_s$. A k-dimensional vector is constructed for all source language words that are represented by SV^s and for all target words that are represented by TV^t. The similarity between w_s and w_t is computed by the cosine similarity score .

The CL-LDA model is very effective and outperforms the CL-LSI model but it is very complex and can't be solved by the exact inferences.

Word Embedding

A parallel corpus is used for creating the word embeddings in a dual semantic space. The similar words in the different language may have approximately nearer word vector, hence, the query translation can be done by using the proximity measurements between the source and target language word vector. A neural network model and its variants are used to construct the word embeddings. It learns multi-dimensional representation which is used to estimate the probability distribution of word sequence for a word (Klementiev, Titov, & Bhattarai, 2012). The neural network model based learned representations captures the syntactic and semantic properties of contextual words. A sequence of words $w_{t-n+1:t} = (w_{t-n+1}, ..., w_{t-1},\ w_t); w_i \in V, i(1...T)$; is given as input to the neural network model. The probability distribution over the next word w_t is estimated as follows:

(i) Prepare a shared representation vector $c = (c_1^T, ..., c_1^T)^T$ which is used to map each of the context words $w_i, i \in [t-1, \ldots, t-n+1]$ to its distributional representation c_{w_t}.

(ii) Concatenate all of the word representation of context $w_{t-n+1:t-1}$ with preserving the order $(c^T_{w_{t-n+1}}, c^T_{w_{t-2}}, ..., c^T_{w_{t-1}})$.

(iii) A linear transformation followed by a logistic function is applied to the concatenated embeddings at the hidden layer.

(iv) The softmax function is applied to the output layer.

The model captures local context. A d-dimensional distributed vector is induced for words in the vocabulary. A shared embedding needs to be built across the language for the cross-language distributed vector representation, which is a three-step process:

(i) A document's lexical representation is mapped into a latent semantic space.

(ii) Canonical correlation analysis maps the semantic representation of queries and documents into a shared semantic space.

(iii) Finally, query-relevant documents are retrieved using a distance metric (Kim, Nam, & Gurevych, 2012).

Canonical correlation analysis is used for semantic transformation among the pair of documents $\{(d_{s1}, d_{t1}\ ), (d_{s2}, d_{t2}), ..., (d_{sn}, d_{tn})\} \in D$ where d_s and d_t are the source language and target language document. A function $f(x)$ maps a document into its semantic space. The canonical correlation analysis requires such basis vectors w_s and w_t that maximize cross-lingual correlation.

$$\rho = max_{w_s,w_t} \frac{\langle w_s f(d_s), w_t f(d_t)\rangle}{\|w_s f(d_s)\|\ \|w_t f(d_t)\|} \tag{2.12}$$

Word embeddings are learned by using the continuous bag-of-word, skip-gram, and log bilinear regression based recurrent neural network models (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, et al., 2013; Pennington, Socher, & Manning, 2014). Further, a query relevant document is retrieved by using language modeling, bi-lingual word embedding skip-gram model (V. K. Sharma & Mittal, 2016d).

- Language Modeling: It includes three sub-techniques of source language word transformation, i.e., direct term sampling, transformation via document sampling, and transformation via collection sampling (Ganguly, Roy, Mitra, & Jones, 2015). The query relevant documents are retrieved with a combination of three transformations. A term generation process from a document is given as follows.

$$P(t|d) = \lambda P(t|d) \\ + \alpha \sum_{t \in d} P(t,t'|d)P(t') + \beta \sum_{t' \in N_t} P(t,t'|C)P(t') + (1-\lambda-\alpha-\beta)P(t|C) \tag{2.13}$$

It is generalized version of standard language modeling, where $P(t,t'|d)$ and $P(t,t'|C)$ is the transformation probability via document sampling and collection sampling, i.e.,

$$P(t,t'|d) = \frac{sim(t,t')}{\sum(N_t)} \frac{tf(t',d)}{|d|} \qquad P(t,t'|C) = \frac{sim(t,t')}{\sum(N_t)} \frac{cf(t')}{cs}$$

t and t' are the source and target language terms, d represents a document, $tf(t',d)$ is the term frequency of term t' in document d, and $cf(t')$ is the collection frequency of term t'.

- Bi-lingual Word Embedding Skip-Gram Model: A dual space document corpus $C = (d_1^s, d_1^t), ..., (d_n^s, d_n^t)$ is constructed by random shuffling of all the words of parallel documents (Vulić & Moens, 2015). Further, a skip-gram model is trained on these bi-lingual documents and the word embeddings are learned for source and target language words. Different semantic compositions are used to construct query and document embeddings, one of them, is given as:

$$\overrightarrow{d} = \overrightarrow{w_1} \times \overrightarrow{w_2} \times ... \times \overrightarrow{w_n}$$
$$\overrightarrow{Q} = \overrightarrow{q_1} \times \overrightarrow{q_2} \times ... \times \overrightarrow{q_m}$$

The similarity between query Q and document d is calculated by using cosine similarity score as shown in Equation 2.14.

$$CSS = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2.14}$$

Parallel/Comparable Sentence/Fragments Extraction

The parallel corpus based CLIR approaches are very efficient and effective but the resource-scarce languages don't have such corpora. The parallel sentence or fragments extraction techniques from a document aligned comparable corpus are used in case of unavailability of a sentence-aligned parallel corpus. The LDA topic modeling technique is applied to construct a topic aligned corpus from a document aligned comparable corpus. The topic aligned corpus is almost equivalent to sentence aligned corpus, further an IBM word alignment model can be applied to construct a probabilistic dictionary (Liu, Duh, & Matsumoto, 2015). A parallel sentence identifier is prepared which is based on novel filtering with three novel feature sets. A word alignment model is used to locate parallel fragment candidates and a lexicon based filter is used to validate truly parallel fragments (Chu et al., 2016). A comparable text corpus is developed with the help of best resolution power keys which are extracted from one document collection by using the relative average term frequency value (Talvensaari, Laurikkala, Järvelin, Juhola, & Keskustalo, 2007).

2.1.3 Statistical Machine Translation

The SMT tool Moses[2] is trained over the parallel corpus and uses phonetic transliteration technique to translate the out of vocabulary words (Gupta, Sinha, & Jain, 2011; Jagarlamudi & Kumaran, 2007). Since the user queries are often very short and don't have the proper syntactic structure, hence, the queries are not properly translated. Each query word is replaced by only one best translation while the inclusion of multiple correct senses of a word could increase the CLIR performance. The Google translator is biased in favor of the named entities (Xiaoning et al., 2008). The SMT uses grammar based and decoder based approaches for query translation. The grammar based approach is more effective than the decoder based approach. The hierarchical translation grammar yields more effectiveness than flat translation grammar. They represent different ways of applying context disambiguation and preserving translation diversity (Ture & Lin, 2014). The online translators, i.e., Google, Yahoo, and Bing provide their services in various languages (Hosseinzadeh Vahid, Arora, Liu, & Jones, 2015).

An open source & language-independent machine translation toolkit Moses is trained on a parallel corpus (Koehn, 2009; Jagarlamudi & Kumaran, 2007), where an IBM model is used to learn a word alignment table. The Hindi and English language sentences are given as $h = \{h_1, h_2, ..., h_m\}$ of length m, and $e = \{e_1, e_2, ..., e_n\}$ of length n. An alignment function $a : j \to i$ for an English word e_j

[2] http://www.statmt.org/moses/

to a Hindi language word h_i is given as

$$p(e, a|h) = \frac{\epsilon}{(m+1)^n} \prod_{j=1}^{n} t(e_j | h_{a(j)}) \tag{2.15}$$

ϵ represents the normalization constant and $t(e_j|h_{a(j)})$ represents the translation probability. Different variation of IBM model and hidden markov model are used for the word alignment, in GIZA++[3], an IBM Model 5 and hidden markov model are used. Phrasal translation technique enhances the power of machine translation (Green, Cer, & Manning, 2014) which includes two steps, i.e., extraction of phrase pairs and scoring the phrase pairs. The best target language translation e_{best} with the highest translation probability is identified at the decoding stage.

$$e_{best} = argmax_e \; p(e|h)$$

$$\begin{gathered} e_{best} = argmax_e \prod_{i=1}^{l} \phi(\bar{h_i}, \bar{e_i}) \\ d(start_i - end_{i-1} - 1) \; p_{LM}(E) \end{gathered} \tag{2.16}$$

$\phi(\bar{h_i}, \bar{e_i})$ represents the translation probability, $d(start_i - end_{i-1} - 1)$ represents the reordering component, and $p_{LM}(E)$ represents a N-gram language model to generate a fluent target language translation. Language model follows n^{th} order markov chain property.

$$\begin{aligned} p(w_1 w_2 w_3 ... w_n) = p(w_1) p(w_2|w_1) p(w_3|w_2 w_1) \\ ......p(w_n|w_{n-1} w_{n-2} ... w_1) \end{aligned}$$

$$p(w_1 w_2 ... w_n) = \prod_i p(w_i | w_1 w_2 ... w_{i-1}) \tag{2.17}$$

2.1.4 Neural Machine Translation

A long short term memory encoder-decoder based NMT[4] model using residual and attention connection is trained and experimented for WMT[5] datasets in English, German, and French languages (Wu et al., 2016). A convolution neural network encoder-decoder machine translation model and its variants are trained for foreign languages (Dakwale & Monz, 2017; Gehring, Auli, Grangier, & Dauphin, 2016;

[3] https://github.com/moses-smt/giza-pp/blob/master/GIZA%2B%2B-v2/README

[4] https://github.com/tensorflow/nmt

[5] http://www.statmt.org/wmt14/

Gehring, Auli, Grangier, Yarats, & Dauphin, 2017; Meng et al., 2015). It is difficult to train the NMT models for the resource-scarce languages like Hindi. The SMT achieves better results compare to the NMT model (Kunchukuttan et al., 2018). An attention-based standard NMT model is used for cross-lingual pronoun detection task in four foreign languages, i.e., English, French, German, and Spanish (Jean, Lauly, Firat, & Cho, 2017). A selective decoding based translation model is used for cross-lingual information extraction in English-Chinese languages (S. Zhang, Duh, & Van Durme, 2017).

Neural networks impart a significant role in the field of data mining as it achieves surpassing results. The NMT model is trained and evaluated for the various foreign languages. A recurrent neural network based encoder-decoder architecture is trained on a parallel corpus to learn the conditional distribution where the conditional probability of generating a target language sentence against a source language sentence is to be maximized. In recurrent neural network based encoder-decoder architecture, a source language sentence is encoded into a set of vectors and this encoded set of vectors is decoded into the target language sentence (Dakwale & Monz, 2017; Du et al., 2016; Wu et al., 2016). For example, a sentence pair $\{X = (x_1, x_2, ..., x_M), Y = (y_1, y_2, ..., y_N)\}$ of size M and N is taken as input and the encoder simply encodes the sentence X into a set of vectors as given in Equation 2.18.

$$\begin{aligned} Ex_1, Ex_2, ..., Ex_M = \\ Encoder\_RNN(x_1, x_2, ..., x_M) \end{aligned} \tag{2.18}$$

The conditional probability to generate the next target language word is learned by the chain rule.

$$P(Y|X) = \prod_{i=1}^{n} P(y_i|y_0, y_1, ..., y_{i-1}; x_1, x_2, ..., x_M) \tag{2.19}$$

y_0 is a special starting symbol of target language sentence. Attention mechanism enhances the capabilities of recurrent neural network based encoder-decoder where a direct short-cut connection is established between the source and target language sentence by using an alignment matrix (Bahdanau, Cho, & Bengio, 2014; Wu et al., 2016).

2.1.5 Web-based Translation

Web provides promising results for a search query. It is also used for collecting parallel or comparable sentences (J. Zhang et al., 2005; V. K. Sharma & Mittal, 2016a). The source language user query is reframed by using target language vocabulary and searched on the web. The search engines return vast results that contain title and query biased summary in each result. Possible translations are expected in either title or query biased summary, so the intersection of substrings of different title pairs, summary pairs, and title - summary pairs are computed. A ranking function is used to select the best translation which depend on the substring frequency, inverse translation frequency, and top retrieved results.

$$P(c_i|w_s) = \partial \left(\frac{TF(c_i)}{Max(TF(c_j))} ITF(c_i) \right) + (1 - \partial) \frac{1}{(Rank(c_i) + \beta)} \tag{2.20}$$

c_i is a candidate translation, w_s is a source language query word, $TF(c_i)$ is term frequency of a candidate translation, $ITF(c_i)$ is inverse translation frequency of a candidate translation, and α, β are the adjusting factors. The online dictionaries are exploited to segment Japanese queries and to obtain all possible English senses. An EWC (ESA-Wikipedia, WordNet path length, Collocation index) measure is used to select the most related meaning from the translation choices (Klyuev & Haralambous, 2012). The Web-based translation uses existing keyword-based search engines which have a very less probability of returning the search results in the target language.

2.1.6 Wikipedia-based Translation

Wikipedia is an online knowledgebase. It is available in 294 languages and very useful resource for natural language processing research. Its structure makes it amenable to cross-lingual research. A Wikipedia article is associated with multiple attributes, i.e., title, abstract, inter-wiki links, redirect page, anchor text, infobox, forward/ backward links, category, sub-sections (Bharadwaj & Varma, 2011; Erdmann et al., 2009; Schönhofen et al., 2007; V. K. Sharma & Mittal, 2016c). The researchers use the title and inter-wiki links for query translations. The source language user query words are searched in Wikipedia. The titles are extracted from the resultant Wikipedia articles. Every Wikipedia article has the inter-wiki links which provide the same-titled articles in other languages, so the title of the target language article is extracted by using inter-wiki links. The hyperlinks, redirect page, anchor text, and category attributes are used for word sense disambiguation. The Wikipedia API libraries are available to use on-line Wikipedia knowledge base and the Wikipedia dumps are also available to use off-line Wikipedia knowledge base. It is also used for building a parallel or comparable corpus and Wikipedia based dictionary (Gaillard et al., 2010; Su et al., 2007).

Cross-Lingual Explicit Semantic Analysis (CL-ESA) indexes the user queries and target language documents with Wikipedia articles which are considered as the explicit concepts (Cimiano, Schultz, Sizov, Sorg, & Staab, 2009; Egozi, Markovitch, & Gabrilovich, 2011; Sorg & Cimiano, 2012). A word to concept lexicon is required for CL-ESA but for the resource-scarce languages, such lexicons are not available. So, the researchers use Wikipedia articles as explicit concepts. The source language user queries and target language articles are TF-IDF indexed against the source language and target language Wikipedia articles. The Inter-wiki database is used to replace the column keys in the indexing, where the source language articles are replaced by target language articles in the index file of the source language queries. Further, cosine similarity score is used to retrieve the top-k query relevant target language documents.

2.2 Disambiguation Models

Source language user query words have multiple target language translations. It doesn't matter what translation model is followed, it is very difficult to select the appropriate translation. Generally, a source language word is translated in favor of maximum probability scorer translation in the probabilistic dictionary based translation approach (Larkey et al., 2003; V. K. Sharma & Mittal, 2016a; Udupa, Jagarlamudi, & Saravanan, 2008). Various translation disambiguation models are discussed in the subsections.

2.2.1 Maximum Probability

A probabilistic dictionary contains translation pairs associated with the probability score. A source language word is translated in favor of maximum probability scorer translation among the multiple translations (Larkey et al., 2003; V. K. Sharma & Mittal, 2016b; Udupa et al., 2008).

2.2.2 Query Segmentation

A user query can be more accurately translated if it is segmented into some units before the translation. The noun phrases and dependency triplets of the queries are identified based on a fact that the larger translation units lead to a more specific model which achieves better translation (Gao & Nie, 2006). The query words and segmented query phrases are considered as the units for query representation (Wu et al., 2016).

2.2.3 Word Co-occurrence Statistics

A target language raw corpus is required to compute the word co-occurrence statistics. Let the translations of the source language words w_{s1}, and w_{s2} are $\{w_{t11}, w_{t12}, ..., w_{t1k}\}$ and $\{w_{t21}, w_{t22}, ..., w_{t2k}\}$ respectively. Appropriate translations are selected based on the probability of co-occurrence of two words (Gao et al., 2001; Yu et al., 2006) that is given by Equation 2.21.

$$P(w_{t1}, w_{t2}) = f(w_{t1}w_{t2}) \tag{2.21}$$

2.2.4 Point-wise Mutual Information

The co-occurrence statistics doesn't include the word's individual occurrence (Chinnakotla et al., 2007), while the PMI includes the word's individual probability along with the co-occurrence probability, and it is defined as:

$$PMI(w_{t1}, w_{t2}) = \frac{p(w_{t1}w_{t2})}{p(w_{t1})p(w_{t2})}$$

$$PMI(w_{t1}, w_{t2}) = \frac{f(w_{t1}w_{t2})/count(w)}{(f(w_{t1})/count(w))(f(w_{t2})/count(w))}$$

$$PMI(w_{t1}, w_{t2}) = \frac{f(w_{t1}w_{t2})count(w)}{f(w_{t1})f(w_{t2})} \tag{2.22}$$

count(w) represents the total number of words in the target language raw corpus.

2.2.5 Statistical Term Similarity

Word co-occurrence statistics and PMI select the optimal translation at word level whereas the statistical term similarity selects the optimal translation at the query level. The cost to compute such an optimal set of translations at the query level is very high. Let a source language query has the words $\{s_1, s_2, ..., s_n\}$ and the query words have the translation sets $\{\{t_{1,i}, t_{1,i+1}, ...\}, ..., \{t_{n,i}, t_{n,i+1}, ...\}\}$. The statistical term similarity algorithm selects the optimal translation set from all the possible combinations of the translation sets (Adriani, 2000). The drawback of this technique is that the complexity of this algorithm exponentially increases with the sentence length.

2.2.6 Multi-Word Expression

Unigram translations suffer from the ambiguities. A multi-word expression or phrase is a set of words which occurs frequently in the documents or sentences. Translation of a multi-word expression is less ambiguous than the unigram translation. Following techniques are discussed here to identify a multi-word expression.

Alignment/Frequency based method

Sentence aligned parallel corpus is used in alignment or frequency based method. The source and target language sentences are part of speech tagged then rule-based alignment is applied to extract the multi-word expression (de Caseli, Ramisch, Nunes, & Villavicencio, 2010; Sinha, 2011). The N-gram method with frequency in the sentence-aligned parallel corpus is used to identify the multi-word expressions (Hewavitharana & Vogel, 2013; Tsvetkov & Wintner, 2012) but this technique would be failed when the languages differ by the sentence structure.

PMI/Co-occurrence statistics

A set of multi-words is extracted from a large training corpus (Singh, Bhingardive, Patel, & Bhattacharyya, 2015; Sinha, 2011). This technique depends on corpora size and quality, where a threshold needs to be defined to classify the word pairs as the multi-word expressions.

WordNet/Word Embedding based Method

Synonyms and antonyms features are used to identify the multi-word expressions in a large corpus. Three rules are discussed which depends on WordNet and word embeddings (Singh et al., 2015).

(i) WordNet based: if $w_2 \in \{w'|w' = Is\_Synonyms\_or\_Antonyms(w_1)\}$ then w_1w_2 is an multi-word expression, where $Is\_Synonyms\_or\_Antonyms(w_1)$ function returns the synonyms and antonyms of word w_1.

(ii) Word Embedding based: if $w_2 \in \{w'|w' = Is\_a\_Neighbour(w_1)\}$ then w_1w_2 is an multi-word expression, where $Is\_a\_Neighbour(w_1)$ function returns top 20 nearest word of w_1.

(iii) WordNet and Word Embedding with Exact Match: if $\{w'|w' = Is\_Synonyms\_or\_Antonyms(w_1))\} \bigcap \{w'|w' = Is\_a\_Neighbour(w_2))\} \neq \phi$ then w_1w_2 is an multi-word expression. Where $Is\_Synonyms\_or\_Antonyms(w_1)$ function returns output strings from WordNet and $Is\_a\_Neighbour(w_2)$ function returns output strings from the word embeddings.

2.2.7 Latent Word Context Model

The LDA model is used to build the latent word context model (Brosseau-Villeneuve, Nie, & Kando, 2014), where each word is represented by the local context which is based on word features. A decaying function assigns the weights to the contextual words. The larger distant contextual words have very low weight. The context words are limited to the distance of 20. At least 500 documents are required to build the word's context and 1,00,000 documents are sufficient for training of the model. Feature sparseness (multiple words may have similar context representation) is the main issue in this model.

2.2.8 Named Entity Recognition

User queries contain either dictionary terms or named entity terms, but it is difficult to differentiate them. A dictionary term needs translation while the named entity term needs transliteration, but for the term which is both the dictionary as well as named entity term, it becomes difficult to decide whether the term needs a translation or transliteration (Rao & Sobha, 2010; V. K. Sharma & Mittal, 2016a). The named entity recognition is a solution. If a word is recognized as a named entity then it will be transliterated. The named entity recognition models are discussed in the next subsections.

Rule based Model

Hand coded rules, lists or a set of dictionaries, and such other documents are employed to recognize the named entities. A lot of human effort is required to construct such rules. It is highly dependent on the language and has poor performance. It is categorized into linguistic and list lookup techniques. The linguistic technique needs a linguist or language expert to generate hand-crafted rules which are based on syntactic, grammatical and orthographic features. The lists, i.e., person names, location names, organization names, gazetteers and other named entity relevant lists are used in list lookup technique (Prasad & Fousiya, 2015). These techniques are not so accurate but for the resource-scarce languages like Hindi, these techniques can be applied.

Association rule mining based model uses 30% minimum support and 80% minimum confidence (Jain, Yadav, & Tayal, 2014). The model defines three types of rules which are (i) dictionary rule: $\langle term_2 \rangle \Rightarrow name\_class2$, (ii) bi-gram rule: $\langle term_1, term_2 \rangle \Rightarrow name\_cla- \ ss2$, and (iii) feature rule: $\langle term_1, feature_2 \rangle \Rightarrow name\_class2$. This technique specifies limited rules so it can't be adopted.

Machine Learning Models

- Hidden Markov Model: It follows the markov chain property, i.e., the probability of occurrence of a particular state depends on the just previous state. Given a sequence of Words (W) and the Name Classes (NC), The most probable sequence of name classes NC for the words W can be obtained by the conditional probability.

$$P(NC|W) = P(W, NC)/P(W) \tag{2.23}$$

  A named class NC is assigned to a word W where $P(NC/W)$ is highest among the other name classes (Prasad & Fousiya, 2015).

- Maximum Entropy Model: A set of features and training corpora are given as input to train the maximum entropy model (Prasad & Fousiya, 2015). The most probable tags corresponding to each word are learned by this model. A user can choose the highest class conditional probability value. A different set of features, i.e., orthographic features (like capitalization, digits, decimal), affixes, left and right context, named entity specific trigger words, gazetteer features, part-of-speech and morphological features, etc. are generally used for named entity recognition. In the case of the Indian language, word features (suffixes, digits, special characters), context features, dictionary features, named entity list feature, etc. are used for named entity recognition (Chinnakotla et al., 2007).

- Support Vector Machine and Conditional Random Field: A named entity annotated training data is utilized in both of the support vector machine and conditional random field (Prasad & Fousiya, 2015; Krishnarao, Gahlot, Srinet, & Kushwaha, 2009; Shishtla et al., 2009). The word window, prefix information, suffix information, length of the word, sentence start information, two consecutive digits, four consecutive digits, word class, and brief class features are used in both of these learning algorithms. The support vector machine finds an optimal hyperplane which performs binary classification with maximal margin. In the case of named entity recognition, support vector machine can be used for classification of a word as an example of a particular class or all the other classes. A set of (X_i, Y_i) points are given to support vector machine and it finds an optimal hyperplane defined as:

$$\vec{w}.\vec{x_i} + b \geq 1, \quad if \quad y_i = 1$$
$$or$$
$$\vec{w}.\vec{x_i} + b \leq -1, \quad if \quad y_i = -1$$

It can be combined to get an optimization problem:

$$Minimize\, \|\vec{w}\|\, subject \quad to \quad y_i(\vec{w}.\vec{x_i} + b) \geq 1, for \quad i = 1, 2, ..., n \tag{2.24}$$

$\vec{w}$ is the normal vector to the hyperplane. An entire observation sequence is given and conditional random field is built as a single exponential model which determines the joint probability of sequences of labels. Conditional random field is a graphical model where possible labels $Y = Y_i; i$ belongs to the set of vertices. (X, Y) is a conditioned on X if and only if it satisfies

$p(y_i|x, y_q, i \neq q) = p(y_i|x, y_q, i \sim q)$, where $i \sim q$ represents that i and q are neighbors in the graph. The conditioned random field outperforms the support vector machine.

Wikipedia based Model

Named entity recognition is a difficult task for the resource-scarce language where sufficient gazetteers and annotated corpora are not available, hence, the English Wikipedia is used to bootstrap the named entities for other languages (Bhagavatula et al., 2012). This method contains two steps: (i) clustering of highly similar Wikipedia articles, and (ii) names entities in English articles are mapped to other language terms by using inter-wiki linked articles. The Wikipedia links can also be used to construct the named entity annotated data.

2.3 Transliteration Models

The out of vocabulary words are translated by using a transliteration mechanism. A rule-based method follows the source to target character level alignment rules. Phonetic and spelling based methods are based on the similar sounding property which state that the word's pronunciation will be same in the different languages (Karimi et al., 2011; Mathur & Saxena, 2014). The phonetic based method identifies phonemes of source language words which are mapped into the target language character representation. In spelling based methods, a group of source language characters is mapped into a group of target language characters. The spelling based methods are preferred due to the character's missing sound issue in the phonetic based method (Karimi et al., 2011). A source language word may have multiple correct transliterations, like a Hindi language word "गौतम" (gautam) have multiple transliterations, 'gautam', 'gautham', 'gowtam', 'gowtham'. It is very difficult to select the correct and most suitable transliteration.

2.3.1 Transliteration Generation with Hidden Markov Model and Conditional Random Field

Transliteration generation model incorporates two phases: (i) inducing character alignment over the word-aligned bi-lingual corpus, and (ii) statistical models (hidden markov model, conditional random field) are used to generate the target language transliteration (Ganesh et al., 2008; Larkey et al., 2003). The hidden markov model is used to maximize source and target language word pairs, further, the character-level alignments are set to the maximum posterior prediction of the model. A target language character is obtained by comparing each source language character to a word. The conditional random field method is used to generate target language transliteration after training the model on some word aligned pairs. It uses forward viterbi and backward search whose combination produces the exact n-best transliteration. The conditional random field is better than the hidden markov model (Larkey et al., 2003).

2.3.2 Transliteration Mining

Large parallel corpus is used for transliteration mining where a target language word is mined against a source language word but the resource-scarce languages don't have such large corpora (Saravanan et al., 2010; V. K. Sharma & Mittal, 2016a). A source language word w_s is transliterated into target language word w_{tt} by using rule-based approach. Further, the LCSR between w_{tt} and all words of the raw corpus are calculated and top-n transliterations $w_{ti} : i = 1...n$ with the maximum LCSR (Udupa, Saravanan, Bakalov, & Bhole, 2009) are selected. A highly robust and less rigid named entity phonetic matching translation model uses the similarity at phoneme level (Lam, Chan, & Huang, 2007). A holistic parallelized graph alignment approach is used for named entity transliteration which uses both the transliteration similarity and mono-lingual occurrences (You, Hwang, Song, Jiang, & Nie, 2012).

Compressed word format algorithm mines more accurate transliteration than the transliteration generation (Janarthanam, Subramaniam, & Nallasamy, 2008; Sethuramalingam & Varma, 2008). A source language word is transliterated into the target language using a rule-based method. Further, the compressed word format algorithm generates the word's minimal form which is compared across the indexed list of target language words. The minimum edit distance technique is used to select the top-n best transliteration.

2.4 Expansion Models

Query expansion is widely used to improve the retrieval effectiveness. The queries are expanded before the translation or after the translation called pre-translation or post-translation query expansion. Selection of the number of query expansion terms is another issue. The performance of the CLIR system increases or decreases based on the number of expansion terms. The query expansion techniques are discussed in the following sub-sections.

2.4.1 Pseudo-Relevance Feedback

Top-K expansion words are selected from initially retrieved top-N documents based on the rocchio relevance feedback algorithm (Sanderson, 2010).

$$\vec{Q_m} = \left(\alpha.\vec{Q_0}\right) + \left(\beta.\frac{1}{|D_r|}.\sum_{\vec{D_j}\in D_r}\vec{D_j}\right) - \left(\gamma.\frac{1}{|D_nr|}.\sum_{\vec{D_k}\in D_nr}\vec{D_k}\right) \tag{2.25}$$

Q_0 : initial query vector, Q_m : modified query vector, D_r : set of relevant documents, D_{nr} : set of non-relevant documents; D_j and D_k are the document vector; α, β, γ are the associated weights.

2.4.2 WordNet based expansion

WordNet is a lexical database which is available in multiple languages and returns lexical features like synonyms and antonyms. User queries are expanded by including the synonyms which are based on the different WordNet distance measures that are path_similarity, lch_similarity, wup_similarity, res_similarity, jcn_similarity, and lin_similairty. Wikipedia is also used for query expansion (Gan & Tu, 2014).

2.4.3 Word-Embedding based expansion

A word embedding model is trained on a raw corpus (Diaz, Mitra, & Craswell, 2016). In the CLIR scenario, queries are expanded before translation or after translation. A source language and a target language raw corpus is required to train the word embeddings. The top-n most similar words are selected based on the cosine similarity score as the query expansion terms.

2.4.4 Selection of N-number of Translations

The inclusion of more than one target language translation either increases or decreases the CLIR performance (V. K. Sharma & Mittal, 2016a; Federico & Bertoldi, 2002). It increases the system performance for a certain value of N after that system performance continuously decreases. So, it is very difficult to state whether the selection of N-number of translations would increase or decrease the system performance and if increase then what would be the value of N.

2.5 Summary

The bilingual dictionary is a widely used translation approach in CLIR (Rao & Sobha, 2010; V. K. Sharma & Mittal, 2018b; Yu et al., 2006). It has limited vocabulary coverage, hence, suffers from the out of vocabulary word issue. Since a bilingual dictionary contains multiple translations against a source language word so it needs a word translation disambiguation technique. The researchers follow either inclusion of the first sense or all senses of a source language word (Ballesteros & Croft, 1996; Pingali & Varma, 2006, 2007; Sethuramalingam & Varma, 2008). The inclusion of all senses of a word mixes the noise in the translated query. The dependency phrases, noun-phrases, and word co-occurrence techniques are widely used to disambiguate the target language translations (Chinnakotla et al., 2007; Federico & Bertoldi, 2002; Gao & Nie, 2006; Gao et al., 2001; Yu et al., 2006). The noun-phrases and dependency-phrases are not recognized due to the unavailability of dependency parser for resource-scarce languages like Hindi. The cognate identification technique depends on languages. It can be applied to the languages which contain similar character sets (Makin et al., 2007). Statistical term similarity method is very effective but not efficient because it involves all possible translations of all the words at the query level due to that it takes a huge amount of time to disambiguate the query (Adriani, 2000).

Parallel corpus is a popular translation resource. A probabilistic dictionary is created by training the IBM model on the available sentence-aligned parallel corpus. The probabilistic dictionary contains multiple translations for a source language word. All translations are associated with the probabilistic score. The best translation is selected in favor of the maximum probabilistic score (Larkey et al., 2003; Nie et al., 1999; Talvensaari et al., 2007; Udupa et al., 2009). The SMT returns at most one translation for a query word (Chen & Gey, 2003; Jagarlamudi &

Kumaran, 2007) while the researchers include top-4 translations for a word in the translated query to enhance the CLIR performance (Saravanan et al., 2010; Udupa et al., 2008). Pearson's correlation coefficient is also used to generate the translation probabilities for the probabilistic dictionary and word co-occurrence statistics is used to select the best translation (Shakery & Zhai, 2013). The probabilistic dictionary also suffers from the out of vocabulary word issue. New example sentences need to be added in the parallel corpus to alleviate the out of vocabulary word issue that need retraining of IBM model to update the probabilistic dictionary. The online translation systems (Google and Bing) are biased in favor of the named entities (Gupta et al., 2011; Hosseinzadeh Vahid et al., 2015; Xiaoning et al., 2008). The online dictionaries and EWC measures are used for translation and disambiguation (Klyuev & Haralambous, 2012). Grammar based and decoder based machine translation models are discussed to translate the queries (Ture & Lin, 2014).

The web search results are utilized to extract the target language translations (J. Zhang et al., 2005). The online dictionaries, Hindi WordNet, Indo WordNet, Wikipedia online knowledge base and offline dumps are analyzed to compute the target language translations (Schönhofen et al., 2007; V. K. Sharma & Mittal, 2016c, 2018a). The online dictionaries are better than the offline dictionaries but they also suffer from the out of vocabulary and word translation disambiguation issue. The WordNet and IndoWordNet are the poor lexical resources in perspective of translation. The Wikipedia dumps are considered as the comparable corpora because of the availability of the same-titled articles in multiple languages which are linked by inter-wiki links. In the case of Hindi-English, the Wikipedia dumps are the partial comparable corpora because the numbers of available Hindi articles are about 26% of the total number of available English Wikipedia articles. Many Hindi Wikipedia articles have only the 'title' attribute field. These articles don't have even a single line description.

The cross-lingual latent semantic indexing (Bradford & Pozniak, 2014), cross-lingual latent dirichlet allocation (Vulić et al., 2013) and cross-lingual explicit semantic analysis (Cimiano et al., 2009; Sorg & Cimiano, 2012) techniques map the source language queries and target language documents in dual semantic space. These techniques are very effective but not efficient. The CL-LSI needs a large storage space to store the TF-IDF matrix, further, the singular value decomposition is applied on the TF-IDF matrix which again needs large storage space with the high computation cost. It is very difficult to inference the results from the complex CL-LDA model. The CL-ESA model needs a word to concept mapping which is not available for the resource-scarce languages. The researchers use the Wikipedia article as the explicit concepts for the resource-scarce languages.

Word embeddings are prepared after training the deep neural network (Kim et al., 2012; Ganguly et al., 2015). The continuous bag-of-words, skip-gram (Vulić & Moens, 2015), and log bilinear regression models are used to learn the word embeddings. The bilingual word embedding skip-gram model is used to learn the bilingual word embeddings, in which, the sentences of the parallel corpus are merged and shuffled, further, skip-gram model is applied (Vulić & Moens, 2015). These bilingual word embeddings don't satisfy their characteristics due to the differences in source and target language sentence structure and the number of vocabularies. So, a probabilistic dictionary is also included with bilingual word embedding skip-gram model to bounds a source language word's translation against the top-k target language words; where k is a constant value (V. K. Sharma & Mittal, 2016d). A year-wise comparison chart of CLIR approaches for the Indian and foreign languages is presented in Table 2.2.

Taxonomy for the Table

BDTRR : Bi − lingual Dictionary based Topical Relevance retrieval
BWESG : Bi − lingual Word Embedding Skip Gram
CCA : Canonical Correlation Analysis
CL : Cross − Language
CTM : CLIR with TM
DBM : Decoder Based Model
DBN : Deep Belief Net
Desc : Description
DpTr : Dependency Translation
DRI : Direct Random Indexing
ESA : Explicit Semantic Analysis
GBM : Grammar Based Model
GLM : Generalized Language Model
GTTRR : Google Translation Topical Relevance Retrieval
HRBD : Human Readable Bi − lingual Dictionary
IDRI : Indirect Random Indexing
IWL : Inter Wiki Link
LDA : Latent Dirichlet Allocation
LM : Language Modeling
LSI : Latent Semantic Indexing
MAP : Mean Average Precision
MRD : Machine Readable Dictionary
MSOD : Mecab System &Online Dictionary
NE : Named Entity
NOT : Number Of Translations
NP : Noun Phrase
NPAD : News Paper Adhoc Dataset
PC : Parallel Corpus
PCC : Pearsons Correlation Coef ficient
QE : Query Expansions
RATF : Relative Average Term Frequency
RF : Relevance Feedback
RIML : Relevant Important Meaning will be Lost
TD : Title − Description
TM : Transliteration Mining
TP : Translation Probability
TRNS : TRaNSlation
WP : WikiPedia
WPDI : WikiPedia Dictionary
WPDU : WikiPedia Dumps

Author	**Languages**	**Data sets**	**Translation**	**Disambiguation**	**Transliteration, NOT, QE**	**Result (MAP)**	**Comments**
(Ballesteros & Croft, 1996)	Spanish-English	TREC-7	MRD	In the favor of first sense	RF with Pre & Post translation	0.1242	RIML, OOV
(Nie et al., 1999)	French-English	TREC-6 & 7	PD	Max. Probability	-	0.2166 0.3401	Unavailability of Parallel corpus
(Adriani, 2000)	Indonesian, English	TREC	MRD	Statistical term similarity	Top –n terms are included	0.2288	Computation cost is very high
(Gao et al., 2001)	English-Chinese	TREC	MRD	Multi-word NP Identification, Co-occurrence	-	0.3891, 75.40%	Unavailability of complete phrase dictionary
(Federico & Bertoldi, 2002)	Italian-English	CLEF 2000, 2001	MRD, LM	Co-occurrence	1, 5, 10 translatioms and RF	0.4379 0.4260 0.4226	Effect of N-best translations not computed with only title tag

(Larkey et al., 2003)	English-Hindi	BBC documents	PD	Max. Probability	Character n-gram trans literation and RF with Pre & Post translation	0.4285	Unavailability of Enough resources
(Chen & Gey, 2003)	English, Germen, French, Dutch, Italian, Spanish	CLEF 2003	PD, L&H MT	Max. Probability	-	0.4340 0.4003 Italian-English	Results are not computed with only ⟨title⟩ field
(J. Zhang et al., 2005)	English-Chinese	NTCIR-4	Web based	Intersection of title pairs	-	0.1582	Depends on existing search engine
(Gao & Nie, 2006)	English-Chinese	TREC-9	NP & DpTr	Co-occurrence	-	0.3303	NP depends on well syntactic queries
(Pingali & Varma, 2006)	Hindi, Telugu to English	CLEF 2006	HRBD	Include all senses of a word	Phoneme transliteration and RF	12.32% 8.09%	Multiple sense inclusion decrease the performance

(Yu et al., 2006)	Chinese-English	NTCIR-4	HRBD	Co-occurrence	HowNet, Ontology	0.2652	Ontology makes it domain dependent
(Chinnakotla et al., 2007)	Hindi, Marathi to English	CLEF-2007	HRBD	Co-occurrence	transliteration mining	0.2366 0.2163	Co-occurrence depends on document corpus quality, Spelling Variation
(Jagarlamudi & Kumaran, 2007)	Hindi-English	CLEF-2007	PD, LM	top-n translation	Phonetic transliteration	0.1994 with (TD)	MAP affected by PC size, so need to increase the PC size
(Makin et al., 2007)	Hindi-Telugu	BBC Hindi & nav bharat	HRBD, string matching	Cognate identification	Rule based transliteration	0.2771	can be applied where language share vocabularies
(Pingali & Varma, 2007)	Hindi, Telugu to English	CLEF 2007	HRBD, TF_IDF index	Include all senses	-	0.1560 0.2155	Phrasal translation is not includes
(Schönhofen et al., 2007)	Hungarian, German, English	CLEF 2007	IWL of WP	Hyperlink and bi-gram statistics	-	-	Wikipedia itself is not sufficient

(Yu et al., 2006)	Chinese, Japanese, Korean	NTCIR-6	HRBD, WPDI	-	-	0.0992	limited Vocabulary coverage
(Talvensaari et al., 2007)	Swedish, English, German	CLEF collection	RATF, Parallel corpus	-	RF	0.252	Depends on corpora size and quality
(Janarthanam et al., 2008)	English-Tamil	-	-	-	CWF transliteration mining	-	CWF performance is better
(Sethuramalingam & Varma, 2008)	Hindi- English	FIRE 2008	HRBD	Include all senses of a word	CWF transliteration mining	0.0907 (H-E), 0.1538 (E-H)	CWF needs a large corpus of unique words
(Udupa et al., 2008)	Hindi- English	FIRE 2008	PD	Max probability	Transliteration mining	0.4526 (81%)	PD depends on corpora size & quality
(Xiaoning et al., 2008)	Chinese-English	NTCIR-7	Google TRNS	-	RF	0.3889	Google biased in favor of Named entity terms
(Cimiano et al., 2009)	English, German, French	Multext, JRC-Acquis	ESA, LSI, LDA	-	-	.83 .56 .71 .55 .11 .08	CL-ESA outperforms CL-LSI and CL-LDA

(Udupa et al., 2009)	Hindi- English	CLEF 2006 2007	PD by IBM model	Include 4 translation of every word	TM from top-n result of first iteration	CTM .2527 .3389	Including 4 translation decrease the system performance
(Rao & Sobha, 2010)	Tamil- English	FIRE 2010	MRD	-	Transliteration of NE	0.3980	Need of named entity disambiguation
(Saravanan et al., 2010)	Hindi, Tamil, English	FIRE 2010	PD	Include 4 translation of a word	Transliteration Generation & Mining	0.3388 0.2816	TM increases the system performance
(Gupta et al., 2011)	Hindi- English	FIRE 2010	MT Google	-	-	0.2299 (MT) 0.3578 (Google)	Google biased in favor of NE
(Ganguly et al., 2012)	English-Bengali	FIRE 2010	BDTRR, GT-TRR	-	Google transliteration	0.1588 0.1652	Poor result due to topic relevance between query and target documents

(Klyuev & Haralambous, 2012)	Japanese-English	NTCIR	MSOD Google	EWC measure	-	0.2644 0.3017	Effectiveness of EWC measure is not discussed
(Sorg & Cimiano, 2012)	English, German, French, Spanish	JRC-Acquis, Multext (J,M)	CL-ESA CAT-ESA, TREE-ESA	-	-	.33(M), .28(J) (CL-ESA)	Wikipedia dumps needs a data refining process
(Kim et al., 2012)	English, German	NPAD, WPDU	DBN, CCA	-	RF	0.2916	DBN model have limited lexicon size
(Moen & Marsi, 2013)	German, English	CLEF	DRI, IDRI	-	-	0.0667 0.0176	Computationally efficient but not effective
(Shakery & Zhai, 2013)	Arabic-English	TREC 2002	PCC based TP	Top-k and Word Co-occurrence	RF	0.2617 (75.9%)	Queries and comparable corpora vocabulary should be from different domain

(Vulić et al., 2013)	Dutch- English	CLEF 2001-03	LDA & LDA Lex,	-	-	0.2995, 0.2083	Variants of LDA model are discussed but they all depends on PC size & quality
(Bradford & Pozniak, 2014)	English, Italian, Spanish	Reuters 21578	PC created by MT & CL-LSI	-	-	0.8692, 0.8586, 0.8703	Machine translation capabilities is not accurate for all languages
(Ture & Lin, 2014)	Arabic, Chinese, French	TREC NTCIR CLEF	GBM and DBM of MT	-	-	A-.293, C-.182, F- 0.297	Source language side phrases are not handled
(Ganguly et al., 2015)	-	TREC 6,7,8	WE with GLM	-	-	0.2287, 0.1958, 0.2503	WE not tested for cross lingual platform

(Hosseinzadeh Vahid et al., 2015)	Hindi, English, Spanish, Italian, German, Swedish	CLEF 2000, CL!NSS 2012-13	Google (G) Bing (B)	-	NE transliteration	.3578 .3673	MAP were not evaluated for CL!NSS dataset and these standard translator biased in favor of NE
(Vulić & Moens, 2015)	Dutch- English	CLEF 2001-03	BWESG , WE with LM	Document & query Semantic Composition	-	0.286 0.317 0.222	Shared semantic space creation would be reflected by language properties

Table 2.2: Comprehensive study of the CLIR approaches [see taxonomy for the short forms]

Chapter 3

Translation of the Semantically Selected Morphological Variants and Out Of Vocabulary Words

In this chapter, a manual dictionary based translation approach is proposed in section 3.1. Since the manual dictionary based approach suffers from the poor vocabulary coverage, hence, a term frequency model based translation approach is proposed in section 3.2 which uses a set of parallel sentences from parallel corpus and cosine similarity to compute target language translation. The proposed term frequency model suffers from the morphological irregularities, hence, a rule-based approach which resolves the morphological irregularities is incorporated in the proposed translation induction algorithm in section 3.3. This algorithm incorporates four morphological variant solutions to fix morphological irregularities and refined stop-word list. The above-mentioned approaches suffer from the Out Of Vocabulary (OOV) word translation issue even though SMT and NMT which are trained in section 3.4, also suffer from the OOV word translation issue. The proposed translation induction algorithm is based on the syntactic rules. It suffers from the translation mis-mapped and non-confident translation issues. Therefore, a semantic morphological variant selection algorithm is proposed in section 3.5 where continuous bag-of-word based word embeddings are used to semantically select the alternate word for morphological variant. OOV word translation is the biggest challenge. A context-based translation algorithm for the out of vocabulary words is proposed in section 3.6 to address the OOV word translation issue.

3.1 Manual Dictionary based

Manual Dictionary (MD) is a manually constructed dictionary where a source language word is associated with multiple translations. MD is used for translation purpose due to the fast computation which is done by a simple look-up technique. A bilingual dictionary is used for query translation from Hindi and Tamil to the English language where OOV words are transliterated by using probabilistic approach (Pingali & Varma, 2006, 2007).

The proposed MD based CLIR approach follows four steps, i.e., (i) tokenization and multi-word terms creation using n-gram technique, after that, stop-words elimination, (ii) query terms mapping to the dictionary and translation extraction, (iii) the OOV terms are transliterated by the transliteration mining technique, and (iv) Vector Space Model (VSM) is used for retrieving target documents. The proposed approach is represented in Figure 3.1.

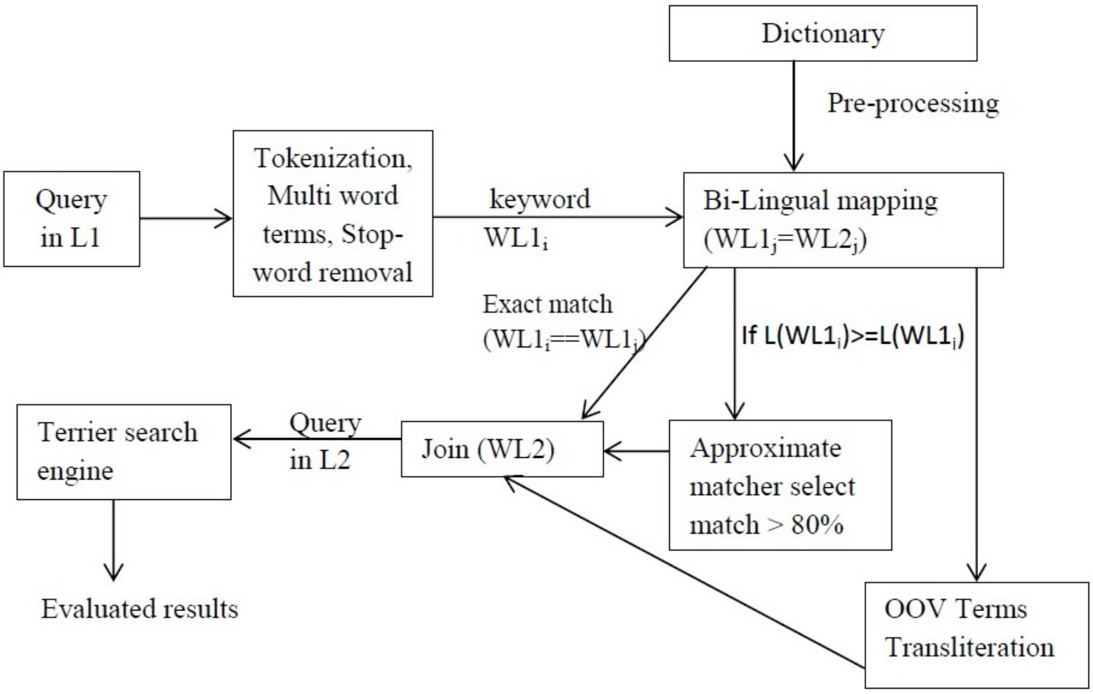


Figure 3.1: Manual dictionary based translation approach

A query string is tokenized and divided into multi-word terms before stop word removal. For example, the term "के विरुद्‌ध" translation is "against" but if stop-word is eliminated before multi-word term creation then the term "विरुद्‌ध" returns "opposite" and "repugnant" translations which are less accurate than "against". Stop words are removed in the case of unigram. A bunch of multi-word terms $WL1\{t_1, t_2, ..., t_n\}$ is prepared after query pre-processing.

Table 3.1: Transliteration mining for the out of vocabulary words

Query Word (q_i)	**Transliterated Word** (tq_i)	**Word** (w_i **from sw**)
मीणा	meena	meena
शंकर	shnkr	shankar
भारत	bhart	bharat

A query term $WL1_i$ is searched into a bilingual dictionary, if it is exactly matched then it is replaced by the corresponding translation $WL2_j$. For example, query word "समुदाय" returns translations "fold" and "tribe" from the dictionary. If the query terms $WL1_i$ is not translated by exact matching then it can be translated by using partial matching. In partial matching, if the length of query term $WL1_i \geq$ the length of bilingual dictionary word $WL1_j$ then a longest common sub-sequence $WL1_{LCS}$ is computed, after that, percentage match of $WL1_{LCS}$ in $WL1_i$ is also calculated. Further, the source language word from the bilingual dictionary which scores the highest percentage match is selected. The highest percentage match should be greater than the empirically defined threshold, i.e. 80%. For example, query term "प्रतिनिधी" matches with a dictionary word "प्रतिनिधि".

The query terms which are not translated by both of the exact and partial matchings are considered as the OOV term and such terms are transliterated by the Out Of Vocabulary Term Transliteration Mining (OOVTTM) technique. OOV terms are either named entities or newly added terms. An OOV term q_i is transformed into a roman format tq_i using a rule-based approach. A set of unique words S_w is collected from the randomly selected target documents. The romanized query term tq_i is mapped to the words in S_w. If tq_i is exactly matched to a word w_i in S_w then original query term q_i is replaced by w_i. Else first and last character of both of the transliterated query term tq_i and word w_i from set S_w are compared if they matched then all the word E_w from set S_w are extracted. Further, each word w_i of E_w is compared to tq_i and the word w_i with the minimum Euclidean distance is selected. Some examples of OOVTTM are shown in Table 3.1.

Terrier[1] search engine is utilized for indexing, retrieval, and evaluation where VSM is utilized to retrieve target documents. FIRE[2] 2010 dataset which contains 50 Hindi language queries and 1,25,638 English language target documents is used to evaluate the proposed MD based CLIR approach. A query has $< title >$, $< desc >$ and $< narr >$ tag, among them, only $< title >$ tag is used for the experiments.

[1] http://terrier.org/

[2] http://fire.irsi.res.in/fire/home

Table 3.2: Hindi – English dictionary based CLIR results evaluated for FIRE 2010 (LCS: Longest Common Sub-sequence, OOVTTM : Out Of Vocabulary Term Transliteration Mining)

S.no.	Experiment	Recall	MAP
1	Shabdanjali + Unigram + LCS	0.5629	0.0581
2	Shandanjali + Unigram + OOVTTM	0.5735	0.0971
3	Shabdanjali + N-gram + OOVTTTM	0.5906	0.1172
4	Shandanjali & English Hindi mapping + N-gram + OOVTTM	0.5464	0.0535
5	Shandanjali & English Hindi mapping + N-gram + OOVTTM & Rule based transliteration	0.5510	0.0579

Shabdanjali[3] and English-Hindi mapping[4] are used as the bilingual dictionary. N-grams up to tri-grams are constructed and searched into a bilingual dictionary, if tri-grams are not found then bi-grams are searched and then unigrams. OOV terms are transliterated by using the OOVTTM technique where a unique set of words S_w is prepared from the randomly selected 139 target documents.

The CLIR approach is evaluated by using recall and Mean Average Precision (MAP). The recall is the fraction of relevant documents that are retrieved. The MAP is the mean of the average precision score of each query. Precision is the fraction of retrieved documents that are relevant to the query. The experiment results are shown in Table 3.2. The proposed approach achieves a maximum 0.1172 MAP with only $< title >$ tag.

Although, the proposed experiment results for FIRE 2010 achieves better MAP 0.1172 than the MAP 0.0907 achieved by state of art approach for FIRE 2008 dataset in (Sethuramalingam & Varma, 2008). The proposed experiment results can not be compared to the state of art experiment results as the state of art results are not available with the same dataset and same language pair. The proposed MD based CLIR approach is considered as the baseline for further experiments.

As shown in experiment 1^{st} and 2^{nd}, The proposed approach achieves a better MAP due to the OOVTTM technique. For example, transliteration of the term **नागालैंड** (nagalaind) by rule-based transliteration is “nagalaind” while the correct transliteration is “nagaland” which is obtained from the set S_w. The OOVTTM technique increases the probability of correct translation and reduces the transliteration mining time due to the selected set E_w from S_w. As shown in 3^{rd} experiment, n-gram term translation enhances the CLIR performance. For example, a word

[3] http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

[4] http://www.cfilt.iitb.ac.in/Downloads.html

Table 3.3: Words translations achieved by the manual dictionary

Query word	अनारा	गुप्ता	मामले	राज्यों
Dictionary word	अनार	गुप्त	मामला	राज्य
Translated word	pomegranate	Secret, covert, undercover	case	state

Table 3.4: Examples of translation diversity

S.no.	Word	Shabdanjali translation	(Shabdanjali + English-Hindi) mapping translation
1	समुदाय	fold, tribe	fold, tribe, category, plurality, battalion, pack, community
2	विध्वंस	annihilation collapse	annihilation, collapse, destruction, wipeout, demolition

हवाई अड्डा (hawai adda) is translated as "airbase" which can not be obtained by unigrams.

The partially matched terms may return wrong translations as shown in Table 3.3. Such terms are either the non-dictionary words or named entities. Multiple dictionaries are added for dictionary enrichment purpose but the MAP is decreased due to the translation diversity as shown in Table 3.4. Rule-based transliteration could not significantly improve the MAP because it may wrongly transliterate some words.

Experiment results show that the inclusion of multi-word terms translation and OOVTTM increase the MAP. A maximum of 0.1172 MAP is achieved with Shabdanjali, N-gram, and OOVTTM. The size of the set S_w needs to be increased for better performance of OOVTTM. Since named entity terms are not correctly translated by the dictionary, hence, there is a need to distinguish between named entities and dictionary term. Including more dictionaries for the dictionary enrichment purpose lead to translation diversification issue.

3.2 Term Frequency Model based

A Probabilistic Dictionary (PD) is constructed by training of GIZA++[5] over the parallel corpus. In PD, a word has multiple translations associated with the probabilistic score. Sentence word overlap and WordNet similarity are used to select the best translation (Mahapatra et al., 2010). The OOV words are transliterated

[5] http://www.statmt.org/moses/giza/GIZA++.html

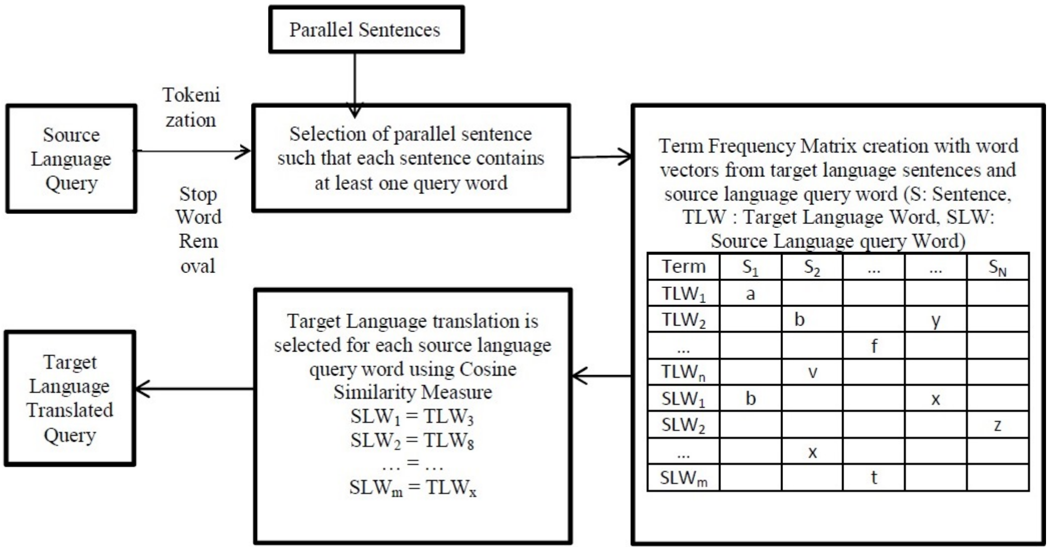
by using the transliteration generation and mining technique (Ganesh et al., 2008; Saravanan et al., 2010). The maximum probabilistic score and Point-wise Mutual Information (PMI) are generally used to select the best translation. Since the PMI score at sentence level is very low, hence, the maximum probabilistic score is used to select the best translation.

GIZA++ training takes much time to construct a PD. The latent semantic indexing technique applies the complex singular value decomposition method on a huge term-frequency matrix of the given parallel corpus. The whole process takes a high computation cost. The proposed approach provides an intermediate solution which reduces the computation cost where a small term frequency matrix is used to extract translation instead of creating PD by GIZA++. The proposed Term Frequency Model (TFM) is represented in Figure 3.2.

User queries are tokenized and stop words are removed to reduce noise. A set of parallel sentences from the parallel corpus is extracted for each query word such that, at least one query word should be present in every sentence. These extracted parallel sentences are merged such that each sentence S_i contains source and target language sentence. Further, a term frequency matrix is constructed as shown in Figure 3.2, which includes word vectors from both of the target language words from the extracted parallel sentences and query words. In word vectors, target language word and source language query word entries with the corresponding sentences are 1, if the target language word and source language query word are exactly matched in the sentences. If source language query word is not exactly matched, then all source language words from the extracted parallel sentences which have the length range between 70% to 130% length of source language query word are extracted. The Longest Common Subsequence Ratio (LCSR) between source language query word and all the extracted words are calculated. The LCSR between two string a, and b is computed by Equation 3.1.

$$LCSR(a,b) = \frac{|LCS(a,b)|}{Maximum(|a|,|b|)} \tag{3.1}$$

LCS(a,b) returns the Longest Common Sub-sequence between the strings a and b. If a word scores more than 75% then the source language query word entry with the corresponding sentence is 1. Further, the Cosine Similarity Score (CSS) is calculated between each source language word and all target language words. A target language word with the maximum CSS is selected as the best translation. The CSS between two given vectors $A = a_1, a_2, ..., a_N$ and $B = b_1, b_2, ..., b_N$ is


Parallel Sentences
Source Language Query
Tokeni zation
Stop Word Rem oval
Selection of parallel sentence such that each sentence contains at least one query word
Term Frequency Matrix creation with word vectors from target language sentences and source language query word (S: Sentence, TLW : Target Language Word, SLW: Source Language query Word)
Target Language Translated Query
Target Language translation is selected for each source language query word using Cosine Similarity Measure
SLW1 = TLW3
SLW2 = TLW8
… = …
SLWm = TLWx

Term	S_1	S_2	…	…	S_N
TLW_1	a				
TLW_2		b		y	
…			f		
TLW_n		v			
SLW_1	b			x	
SLW_2					z
…		x			
SLW_m			t		

Figure 3.2: Term frequency model based translation approach

calculated by the Equation 3.2.

$$CSS = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{3.2}$$

The proposed TFM is evaluated with FIRE[6] 2010 and 2011 datasets. Dataset statistics is represented in Table 3.5. A query includes $< title >$, $< desc >$ and $< narr >$ tag, among them, only $< title >$ tag is considered in our experiments. A Hindi-English parallel corpus HindiEnCorp[7] is utilized in both of the PD and TFM. VSM is used for indexing and retrieval. Recall and MAP which are the evaluation measures are used to evaluate the proposed TFM. The recall is the fraction of relevant documents that are retrieved and precision is the fraction of retrieved documents that are relevant to the query. MAP for a set of queries is the mean of the average precision score of each query where average precision is calculated by Equation 3.3.

$$\begin{aligned} Average\ Precision = \\ \frac{\sum_{k=1}^{n}(p(k) \times rel(k))}{Number\ of\ relevant\ documents} \end{aligned} \tag{3.3}$$

k is the rank in the sequence of retrieved documents, n is the number of retrieved

[6] http://fire.irsi.res.in/fire/home

[7] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0

Table 3.5: FIRE dataset statistics

Dataset Characteristic	**FIRE 2010**		**FIRE 2011**	
	Query	**Document**	**Query**	**Document**
Number of queries/sentence/documents	50	125586	50	392577
Average length (Number of Tokens) of query/sentence/document	6	264	3	245

documents, $p(k)$ is the precision at rank k, $rel(k)$ is equal to 1 if the document at rank k is relevant, otherwise 0. Experiment results for PD and TFM are presented in Table 3.6.

Table 3.6: Experiment results for the probabilistic dictionary and term frequency model based translation approach

Approach	FIRE 2010		FIRE 2011	
	Recall	MAP	Recall	MAP
MD (baseline)	0.5906	0.1172	0.4879	0.0893
PD (baseline)	0.7488	0.2267	0.6791	0.1672
TFM	0.7519	0.2637	0.6754	0.1623

The proposed TFM achieves better MAP than the baseline PD and MD based approach. The construction of PD from the parallel corpus takes much time during GIZA++ training whereas The proposed TFM does not need huge corpus, instead of, it selects only 250 to 500 sentences per query word. So, two benefits are achieved with the TFM over PD, i.e., TFM does not need huge corpus due to that the computation cost is reduced and it eliminates the big disadvantage of the latent semantic indexing which takes a high computation cost to process the huge matrix that is generated from the huge parallel corpus. The number of parallel sentences for each query word is decided based on the empirically defined thresholds which are 250 for FIRE 2010 and 500 for FIRE 2011 as shown in Figure 3.3.

The proposed TFM achieves better MAP in comparison to PD and it takes fewer computations in comparison to the latent semantic indexing. The graph which is shown in Figure 3.3 states that the MAP is approximately equal for every selection of sentences above 90 sentences. However, a maximum of 0.2637 MAP for FIRE 2010 is achieved with 250 sentences and a maximum of 0.1623 MAP for FIRE 2011 is achieved with 500 sentences. FIRE 2011 average query length is shorter than the FIRE 2010, hence, TFM performs approximately equal to PD approach for FIRE 2011. In the case of FIRE 2010, the TFM achieves better MAP than

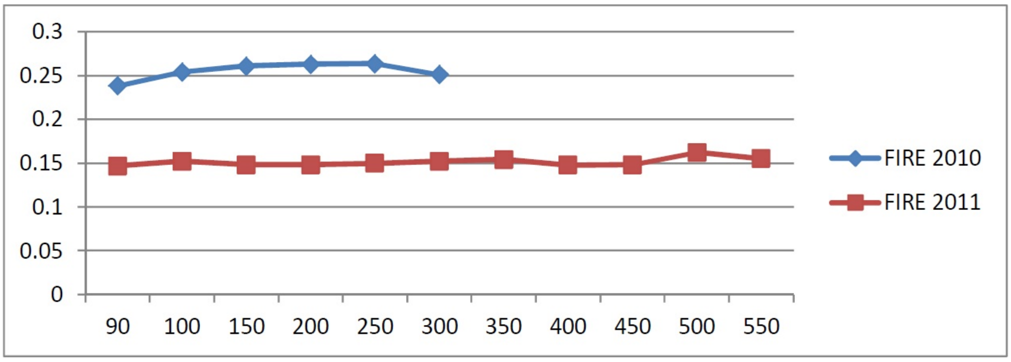


Figure 3.3: Mean average precision scores (Y-axis) against the number of selected parallel sentences (X-axis)

the PD approach.

3.3 Translation Induction Algorithm based

Manual dictionary, probabilistic dictionary, and parallel corpus-based approaches translate the exactly matched query words by a simple look-up technique and partial matched query words with the help of approximate string matching technique (Makin et al., 2007). The OOV words are transliterated by using the transliteration generation and mining technique (Saravanan et al., 2010; Ganesh et al., 2008). These techniques are not able to fix the morphological irregularities like nukta character, infrequent words, multiple morphological variants. A morphological variant word has many forms which all are not available in the parallel corpus but at least one of them may be available. Such morphological variants are considered as one of the type of the OOV word (Akhtar, Gupta, Vajpayee, Srivastava, & Shrivastava, 2017; Gujral, Khayrallah, & Koehn, 2016; Huck, Tamchyna, Bojar, & Fraser, 2017). Stop-words are the frequently occurring words which do not carry any significant information, hence, the stop-words are eliminated to enhance the IR effectiveness (El-Khair, 2006). Some stop-words carry significant information which may improve IR effectiveness. State-of-art SMT and NMT are unable to translate some morphological variants because they are trained on the parallel corpus which has limited vocabularies (Kunchukuttan et al., 2018).

A Translation Induction Algorithm (TIA) is proposed which incorporates the refined stop-words list and morphological variants solutions (V. Sharma & Mittal, 2019). Significant stop-words are eliminated from the standard stop-words lists in order to produce new refined stop-words lists for both of the source and target languages. Morphological variants solution are added to fix the morphological irregularities. Further, the contextual parallel sentences are exploited to compute the best translation. The old parallel corpus "HindiEnCorp" and the newly developed parallel corpus by CFILT lab at IIT Bombay "IITBCorpus" are tested for Hindi-English CLIR. The HindiEnCorp performs better in perspective of CLIR due to its better organization than the IITBCorpus.

Generally, the stop-words are removed from the queries but some source and target language stop words have multiple meaningful target and source language translations respectively which may convey significant information. The examples of such stop-words are represented in Table 3.7. These significant stop-words need to be eliminated from the Hindi and English standard stop-words lists. Such significant stop-words are listed in Table 3.8. Refined Stop-Words (RSW) lists for Hindi and English are produced after the elimination of the significant stop-words, further, these refined stop-words are removed from the queries.

Table 3.7: List of stop-words and their translations

Word	Translations
Against	खिलाफ (Khilaf), विपरीत (Viruddh), विरुद्ध (Vipareet), प्रतिकूल (Pratikool)
During	दौरान (Dauran), की अवधि में (Ki avadhi me), कालावधि तक (Kalavadhi Tak), पर्यन्त (Paryant)
बिल्कुल (Bilkul)	All, Completely, Perfectly, Quite
पूरा (Poora)	Complete, Finished, Total, Overall, Through

Table 3.8: List of meaningful stop-words for Hindi and English

Hindi Stop-Words	English Stop-words
बिल्कुल (bilkul), निहायत (nihayat), वर्ग (varg), रखें (rakhen), काफी (kaffi), निचे (niche), पहले (pahle), अंदर (andar), भीतर (bheetar), पूरा (poora), गया (gaya), बनी (bani), बही (bahi), बीच (bich)	About, above, after, again, against, all, because, before, below, between, but, down, during, few, more, most, off, only, ought, out, over, own, some, than, through, too, under, up

Query words which are exactly mapped in the probabilistic dictionary are translated by a look-up technique. The LCSR string matching technique is used to translate such morphological variants which are not exactly mapped in the PD. At many instances, LCSR is unable to trace the morphological variants due to the morphological irregularities in the Hindi language. Therefore, the following Morphological Variants Solutions (MVS) are applied to trace the approximately nearer word of the query word.

- Equality of nukta character with the non-nukta character: LCSR is unable to detect the equality between the nukta and non-nukta characters like, सड़क (sadak), लड़ाई (ladai), परवेज़ (parvez), hence, an equality solution is applied where nukta characters are replaced by the non-nukta characters.

- Auto-correction of user query words: A query word is searched in the parallel corpus as it appears but its correctness is not verified. A word's popularity based auto-correctness solution is applied where a query word's frequency wf_i over the parallel corpus is computed and compared to the empirically defined threshold T. If wf_i is less than T then the nearest word's (using LCSR) frequency cwf_i over the parallel corpus is computed. If $cwf_i > wf_i$, then the query word is replaced by its nearest word. Examples of such words are shown in Table 3.9.

Table 3.9: Auto-corrected words

Query Word	**Frequency**	**Closest Word**	**Frequency**
मसजिद (Masjid)	4	मस्जिद (Masjid)	229
सियाचिन (Siachen)	2	सियाचीन (Siachen)	6
मुसलिम (Muslim)	3	मुस्लिम (Muslim)	947

Algorithm 1 Translation Induction Algorithm

Input: Source Language Query $SLQ[w_1, w_2, ..., w_m]$ and a Parallel Corpus PC
Output: Top-k Target Language Translation (TLT) for each query word

```
for (i=0 to len(SLQ)) do
    if (SLQ[i] $\notin$ RSW) then
        SLQ[i]=MVS(SLQ[i]);
        if (SLQ[i] $\notin$ PC) then
            maxscore=0, maxkey=" ";
            foreach word in PC do
                if (LCSR(word,SLQ[i])>maxscore) then
                    maxscore=LCSR(word,SLQ[i]);
                    maxkey=word;
                end
            end
            if (maxscore>0.75) then
                SLQ[i]=maxkey;
            end
        end
    end
end
```

```
Sorted_PC= SORT_LEN(PC);
for (i=0 to len(SLQ)) do
    TC=[];
    TGC[SLQ[i]]=0;                          //TGC: Tri Gram Counter
    BGC[SLQ[i]]=0;                          //BGC: Bi Gram Counter
    tri_gram[3] = N_Gram(SLQ, SLQ[i], 3);
    bi_gram[2] = N_Gram(SLQ, SLQ[i], 2);
    foreach sentence in PC do
        if (tri_gram[0]∈ sentence & tri_gram[1]∈ sentence & tri_gram[2]∈ sentence)
        then
            TGC[SLQ[i]]+=1;
            TC.add(sentence);
        else
            if (bi_gram[0]∈ sentence & bi_gram[1]∈ sentence) then
                BGC[SLQ[i]]+=1;
                TC.add(sentence);
            end
        end
    end
    if (TGC[SLQ[i]]+BGC[SLQ[i]]<t) then
        count=0;
        foreach sentence in Sorted_PC do
            if ((count<z) & (SLQ[i] ∈ sentence)) then
                count+=1;
                TC.add(sentence);
            end
        end
    end
    index[][]=TF_IDF(TC);
    maxcosine=0;
    maxword="";
    foreach term in index.keys() do
        if (Cosine_Similarity(index[term], index[SLQ[i]])>maxcosine)) then
            maxcosine=Cosine_Similarity(index[term], index[SLQ[i]]);
            maxword=term;
        end
    end
    print(maxword);
end
```

- Equality of chandra-bindu with म् (m) *and* न् (n): A query word with chandra-bindu is equivalent to many other words like, a word "अंबानी"(ambanee) has same LCSR 0.83 with these three words "अम्बानी"(ambanee), "अंबाजी"(am- bajee), and "अल्बानी"(albanee). If the chandra bindu is replaced by "म् " (m) then a correct word "अम्बानी" (ambanee) with the maximum LCSR is selected.

- Auto-selection of the nearest query word: LCSR is used to select the nearest word if the word is exactly not present in the parallel corpus. Such words may have multiple morphological variants with the same LCSR as shown in Table 3.10. The compressed word format algorithm (Janarthanam et al., 2008) is used to auto-select the nearest query word. So far, the compressed word format algorithm is being used for transliteration mining.

Table 3.10: Multiple nearer words with the same longest common sub-sequence ratio

Query Word	**Corpus Word**	**LCSR Score**
गुटखा (Gutkha)	गुड़टा (Guita)	0.8
	गुटखे (Gutkhe)	0.8
	गुरखा (Gurkha)	0.8

The proposed TIA is represented in Algorithm 1, where the query words are searched in the parallel corpus after applying the refined stop-words and morphological variants solutions. An LCSR string matching technique is applied to the words which are not exactly matched in the parallel corpus and such words in the query are replaced by the nearest words.

A Sorted Parallel Corpus (Sorted_PC) is prepared by sorting the parallel corpus based on the sentence length. Further, a set of parallel sentences is selected for each query word w_i from the parallel corpus in a contextual manner such that each sentence contains either all three words of tri-gram or both of the words of bi-gram, independent of words order. The function $N\_Grams()$ returns trigrams or bigrams. If the number of selected parallel sentences is less than a threshold t then the z number of unigram based parallel sentences of minimum length are also included. Term Frequency-Inverse Document Frequency (TF-IDF) indexing is applied to the selected parallel sentences, further, cosine similarity scores are calculated between the query word and all target language words of the selected parallel sentences. A target language word with the maximum cosine similarity score is selected as the best translation. In the proposed algorithm, a context-based selection of the parallel sentences returns the more relevant translation. Target language refined stop-words are removed while performing the target document retrieval.

FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5 are used to evaluate the TIA. Following experimental setups are prepared to analyze the proposed approach.

- PD based Approach: A PD is learned from the HindiEnCorp. A maximum probability scorer translation is selected as the best translation. LCSR is used to search the not-exactly-mapped query words where the empirically defined LCSR threshold is 0.75. A Refined stop-words list is used instead of the standard stop-word list to analyze the impact of refined stop-words. The best translation is chosen from the top-k translations where k = 5 is an empirically defined constant.
- TIA: The proposed TIA uses different thresholds, i.e., LCSR threshold = 0.75, t = 10, T = 5, and z=70.

FIRE 2010 and 2011 Hindi language queries are translated by PD and TIA, further, these translated queries are used to retrieve the target English language documents. Target language refined stop-words are eliminated from the target documents. TF-IDF and cosine similarity are used for indexing and retrieval respectively.

The baseline PD based approach selects the maximum probability scorer translation while the proposed TIA selects the context based translation. TIA also incorporates the refined stop-word list and morphological variants solutions. The MAP for the baseline PD and proposed TIA are represented in Table 3.11.

Table 3.11: Experiment results for the translation induction algorithm

Experi-mental Setups	**Standard stop-words**				**Refined stop-words**			
	FIRE 2010		**FIRE 2011**		**FIRE 2010**		**FIRE 2011**	
	Recall	**MAP**	**Recall**	**MAP**	**Recall**	**MAP**	**Recall**	**MAP**
PD (baseline)	0.7488	0.2267	0.6791	0.1672	0.7761	0.2547	0.6917	0.1727
TIA (proposed)	0.8319	0.2498	0.7195	0.1874	0.8685	0.2818	0.7395	0.1921

The proposed TIA incorporates morphological variants solutions, hence, it achieves 0.2818 and 0.1921 MAP for FIRE 2010 and 2011 which is better than 0.2547 and 0.1727 MAP achieved by PD respectively. Baseline PD and the proposed TIA are also tested with the refined stop-words list. The refined stop-words list improves the MAP compared to the standard stop-words list as shown in Table 3.11. The proposed TIA incorporates refined stop-words list and morphological variants solutions, further, the translations are computed based on the contextual words, therefore, TIA outperforms the PD based approach.

Table 3.12: The experiment results of monolingual information retrieval for showing the impact of refined stop-words list

Retrieval Model	FIRE Topic set	Language	Standard stop-words list			Refined stop-words list		
			T	TD	TDN	T	TD	TDN
BM25	2010	Hindi	0.3197	0.4106	0.4954	0.3251	0.4267	0.5081
		English	0.3798	0.4538	0.5205	0.3845	0.4667	0.5309
	2011	Hindi	0.1846	0.2532	0.2819	0.1879	0.2571	0.2830
		English	0.2215	0.3069	0.3238	0.2215	0.3087	0.3236

A separate experiment of monolingual information retrieval is also performed with the FIRE 2010 and 2011 Hindi and English topic sets (queries) respectively to analyze the impact of the refined stop-words lists. The topic set has three tags in each query, namely, $\langle Title \rangle$ (T), $\langle Desc \rangle$ (D), and $\langle Narr \rangle$ (N). These all have individually experimented with a standard stop-words list and refined stop-words list. These experiments are performed on three fields of the queries and evaluated by using the MAP. Experiment results are shown in Table 3.12. Refined stop-words list achieves better MAP for FIRE 2010 topic sets in both of the languages while it achieves approximately equal performance for FIRE 2011 topic set because FIRE 2010 topic set has more stop-words than the FIRE 2011 topic set and the average query length for FIRE 2010 & FIRE 2011 topic sets are 6 & 3 respectively.

User queries are translated by using the baseline PD and the proposed TIA. Standard stop-words list has some significant stop-words whose presence may improve the CLIR performance, so these significant stop-words are eliminated from the standard stop-words list. The newly generated refined stop-words list enhances the MAP compared to standard stop-words list. The proposed TIA incorporates the refined stop-words and morphological variants solutions, apart from that, the query words are translated based on the contextual words, therefore, the TIA outperforms the PD based approach.

3.4 Trending Translation Techniques

Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) are the trending translation techniques. Although the author in (Kunchukuttan et al., 2018) says that the SMT performs better than the NMT for Hindi to English translations, however, an experimental study on SMT and NMT is discussed in this section.

3.4.1 Statistical Machine Tranlation based

SMT employs four components, namely, word translation, phrasal translation, decoding, and language modeling (Koehn, 2009; Green et al., 2014).

Word Translation

An IBM model is used to generate the word alignment table from the sentence aligned parallel corpus. The Hindi and English language sentences are given as $h = \{h_1, h_2, ..., hm\}$ of length m, and $e = \{e_1, e_2, ..., en\}$ of length n. An alignment function $a : j \rightarrow i$ for an English word e_j to a Hindi language word h_i is given in Equation 3.4.

$$p(e, a|h) = \frac{\epsilon}{(m+1)^n} \prod_{j=1}^{n} t(e_j|h_{a(j)}) \tag{3.4}$$

where ϵ represents the normalization constant and $t(e_j|h_{a(j)})$ represents the translation probability.

Since a source language word is likely to be aligned with different target language words in different iterations, hence, an expectation maximization algorithm is used to eliminate this problem. The expectation maximization follows an expectation step where the probabilities of alignments are computed and a maximization step where the model is estimated from the data. The expectation maximization algorithm is continuously applied until the convergence.

In expectation step, the probability of alignment $p(a|e, h)$ is computed as

$$p(a|e, h) = \frac{p(e, a|h)}{p(e|h)} \tag{3.5}$$

$p(e, a|h)$ computed by using equation 3.4, and $p(e|h)$ is calculated as

$$p(e|h) = \frac{\epsilon}{(m+1)^n} \prod_{j=1}^{n} \sum_{i=1}^{m} t(e_j|h_i) \tag{3.6}$$

In maximization step, the sentence pairs (e, h) in which e is a translation of h are calculated.

$$c(e|h; e, h) = \sum_{a} p(a|e, h) \sum_{j=1}^{n} \delta(e, e_j)\delta(h, h_{a(j)}) \tag{3.7}$$

The different variation of IBM model and hidden markov model are used for word alignment. GIZA++ implements an IBM Model 5 and hidden markov model.

Phrasal Translation

The phrase model is not limited to only linguistic phrases which can be a noun phrase, verb phrase, prepositional phrase etc. It includes two steps, i.e., extraction of phrase pairs and scoring phrase pairs. The phrase pairs are extracted such that they should be consistent with the word alignment. A phrase pair $(\bar{e}, \bar{h})$ is consistent with an alignment A, if all words $h_1, h_2, ..., h_l$ in $\bar{h}$ and $e_1, e_2, ..., e_l$ in $\bar{e}$ have the same alignment points in A and vice versa.

$$\begin{aligned}(\bar{e}, \bar{h})\ \mathit{consistent\ with}\ A \Leftrightarrow\ & \forall e_i \in \bar{e} : (e_i, h_j) \in A \rightarrow h_j \in \bar{h} \\ & AND\ \forall h_j \in \bar{h} : (e_i, h_j) \in A \rightarrow e_i \in \bar{e} \\ & AND\ \exists\ e_i \in \bar{e}, h_j \in \bar{h} : (e_i, h_j) \in A\end{aligned}$$

A translation probability is assigned to each phrase pair by calculating the relative frequency

$$\phi(\bar{h}, \bar{e}) = \frac{count(\bar{e}, \bar{h})}{\sum_{h_i} count(\bar{e}, \bar{h_i})} \tag{3.8}$$

Decoding

The best target language translation e_{best} with the highest translation probability is identified at the decoding stage.

$$e_{best} = argmax_e\ p(e|h)$$

$$e_{best} = argmax_e \prod_{i=1}^{l} \phi(\bar{h_i}, \bar{e_i}) \; d(start_i - end_{i-1} - 1) \; p_{LM}(E) \tag{3.9}$$

where $\phi(\bar{h_i}, \bar{e_i})$ represents the translation probability, $d(start_i - end_{i-1} - 1)$ represents the reordering component, and $p_{LM}(E)$ represents a N-gram language model to generate a fluent target language translation.

Language Modeling

A N-gram language model is used to generate a fluent translation output. The language model follows n^{th} order markov chain property.

$$p(w_1 w_2 w_3 ... w_n) = p(w_1)p(w_2|w_1)p(w_3|w_2 w_1)......p(w_n|w_{n-1} w_{n-2} ... w_1)$$

$$p(w_1 w_2 ... w_n) = \prod_i p(w_i|w_1 w_2 ... w_{i-1}) \tag{3.10}$$

3.4.2 Neural Machine Translation based

A Recurrent Neural Network (RNN) based encoder-decoder architecture is trained on parallel corpus to learn the conditional distribution, where the conditional probability of generating a target language sentence against a source language sentence is to be maximized. In RNN based encoder-decoder architecture, a source language sentence is encoded into a set of vectors and this encoded set of vectors is decoded into the target language sentence (Kunchukuttan et al., 2018; Pennington et al., 2014; Vulić et al., 2013; Zhou et al., 2012). As shown in Figure 3.4, a sentence pair X=(x_1,x_2,...,x_M), Y=(y_1,y_2,...,y_N) of size M and N is taken as input. The encoder simply encodes the sentence X into a set of vectors, as given in Equation 3.11.

$$Ex_1, Ex_2, ..., Ex_M = Encoder\_RNN(x_1, x_2, ..., x_M) \tag{3.11}$$

The conditional probability to generate the next target language word, is learned by the chain rule.

$$P(Y|X) = \prod_{i=1}^{n} P(y_i|y_0, y_1, ..., y_{i-1}; x_1, x_2, ..., x_M) \tag{3.12}$$

[8] https://github.com/tensorflow/nmt

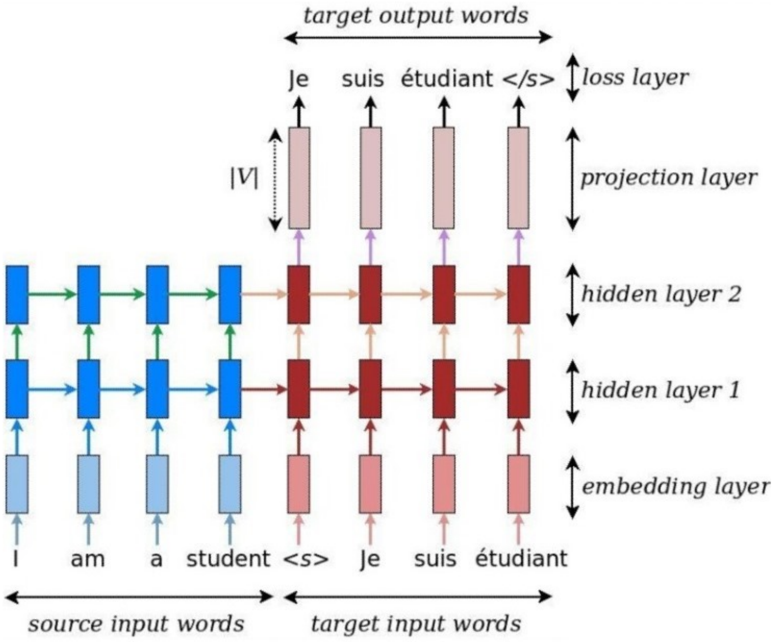


Figure 3.4: The recurrent neural network based encoder-decoder architecture[8]

y_0 is a special starting symbol of target language sentence which is represented by $\langle s \rangle$ in Figure 3.4.

Attention mechanism enhances the capabilities of RNN based encoder-decoder where a direct short-cut connection is established between the source and target language sentence by using an alignment matrix (Bahdanau et al., 2014). An attention-based RNN encoder-decoder is illustrated in Figur 3.5. The attention weights are computed by comparing the current target hidden state with all source states, as given in Equation 3.13.

$$\alpha_{ts} = \frac{exp(score(h_t, \bar{h_s}))}{\sum_{s'=1}^{S} exp(score(h_t, \bar{h_{s'}}))} \tag{3.13}$$

score compares the target hidden state with each of the source hidden states.

A context vector c_t is computed as the average of source states and the combination of c_t with the current target hidden state produces a final attention vector a_t. The computation of c_t and a_t is represented by Equation 3.14 and 3.15. The generated attention vector is fed as an input to the next step of conditional probability computation.

$$c_t = \sum_s \alpha_{ts} \bar{h_s} \tag{3.14}$$

[9] https://github.com/tensorflow/nmt

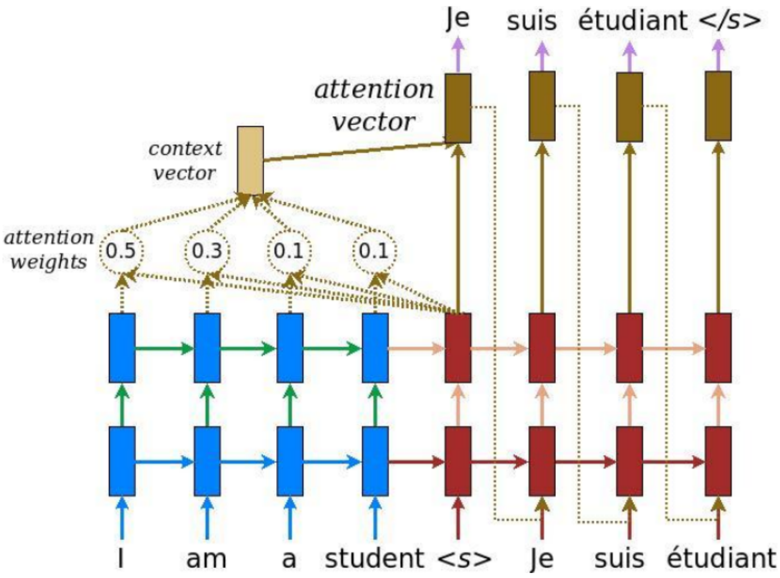


Figure 3.5: An attention based recurrent neural network encoder-decoder architecture[9]

$$a_t = f(c_t, h_t) \tag{3.15}$$

3.4.3 Experimental Results of the SMT and NMT based

The FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5 are used to evaluate the SMT and NMT based CLIR approach. The resources and dataset which are used to train SMT and NMT models are represented in Table 3.13. A parallel corpus HindiEnCorp which contains total 273886 parallel sentences and 20 words average sentence length is used to construct a probabilistic dictionary or to train the SMT. Seven SMT and NMT experimental setups are tuned and evaluated by using the common dev_set and test_set.

Table 3.13: Resources for the training of the SMT and NMT

Training_set	**Language Modeling**	**Dev_set**	**Test_set**
HindiEnCorp (273886 sentences) IITBCorpus (1,492,827 sentences)	HindiEnCorp IITBCorpus WMT News 2015 Corpus (3.3 GB)	WMT Dev_set (520 sentences)	WMT news test_set 2014 (2507 sentences), and FIRE 2008, 2010, 2011, and 2012 query set (each have 50 sentences)

Machine translation techniques are evaluated by using BLEU score which computes the n-gram overlap between the machine translation output and the referenced translation. It computes precision for n-grams of size 1 to 4, which is given as

$$precision = \frac{correct\ \ translation}{translation\ \ length}$$

BLEU score is computed for the entire corpus not for a single sentence (Koehn, 2009).

$$BLEU = min(i, \frac{output\ -\ length}{reference\ -\ length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} \tag{3.16}$$

CLIR performance is measured by using MAP. Three different SMT setups are trained to translate the user queries, which are given as follows:

- SMT_setup1: HindiEnCorp is used for both of the purposes of training and language modeling.
- SMT_setup2: A Hindi-English parallel corpus developed by IIT Bombay IITBCorpus[10] is used for both of the purposes of training and language modeling.
- SMT_setup3: IITBCorpus is used for training, while the WMT news corpus 2015 is used for language modeling.

Four different attention based RNN encoder-decoder models are trained with the different set of resources (Bahdanau et al., 2014), which are given as follows:

- NMT_setup1: A HindiEnCorp is used for training with the dropout value 0.0.
- NMT_setup2: A HindiEnCorp is used for training with the dropout value 0.2.
- NMT_setup3: An IITBCorpus is used for training with the dropout value 0.0.
- NMT_setup4: An IITBCorpus is used for training with the dropout value 0.2.

[10] http://www.cfilt.iitb.ac.in/iitb_parallel/

A byte pair encoding with 15,500 merge operation is used to learn the vocabulary (Sennrich, Haddow, & Birch, 2015). A *subword-nmt*[11] tool is used for obtaining the vocabulary. An open-source NMT[12] tool is used to train the attention based RNN encoder-decoder with the double hidden layer. Each layer has 512 units of embedding dimension and 20,000 training steps.

The SMT and NMT models are evaluated by using the BLEU score. These trained models are evaluated for five different test_sets. The experiment results are represented in Table 3.14. The News test_set 2014, FIRE 2008, 2010, and 2011 test sets are evaluated against the corresponding human translated text while FIRE 2012 test set is evaluated against the Google translated text because the human translated text for FIRE 2012 is not available.

Table 3.14: Experiment results for the SMT and NMT translation

Setups	**News test_set 2014**	**FIRE 2008**	**FIRE 2010**	**FIRE 2011**	**FIRE 2012**
SMT_setup1	7.05	10.76	4.48	8.13	17.11
SMT_setup2	9.70	11.72	6.75	6.53	17.59
SMT_setup3	8.95	11.45	5.13	8.77	17.75
NMT_setup1	3.56	07.43	5.35	6.92	09.87
NMT_setup2	3.46	06.80	4.58	7.42	10.80
NMT_setup3	3.20	09.34	4.54	8.40	10.70
NMT_setup4	2.40	10.47	5.26	9.56	12.98

The SMT setups achieve approximately similar BLEU score. The SMT setups perform better than the NMT setups. The SMT and NMT are expected to generate fluent translation output, so they can add unnecessary translations which actually increase the noise in translation. In the perspective of CLIR, the translation should be more accurate instead of fluent. These SMT and NMT setups are evaluated for CLIR by using the recall and MAP. Experiment results are represented in Table 3.15.

SMT_setup1 performs better than the SMT_setup2 and SMT_setup3 in perspective of CLIR. SMT_setup1 is trained on the HindiEnCorp which is smaller than the IITBCorpus. The IITBCorpus is used in SMT_setup2 and SMT_ setup3. Although the IITBCorpus is a superset of HindiEnCorp but it is not so well-organized and mixes the noise in the translation, hence, it achieves poor performance in perspective of CLIR. The SMT_setup3 uses WMT news corpora 2015 for language modeling due to that SMT_setup3 performs a little better than the SMT_setup2.

[11] github.com/rsennrich/subword-nmt

[12] https://github.com/tensorflow/nmt

Table 3.15: Experiment results for the SMT and NMT based CLIR approaches

Experimental	**FIRE 2010**		**FIRE 2011**	
setups	Recall	MAP	Recall	MAP
SMT_setup1	0.8284	0.2832	0.7084	0.1885
SMT_setup2	0.7718	0.2175	0.6602	0.1608
SMT_setup3	0.7978	0.2237	0.6602	0.1767
NMT_setup1	0.5835	0.0681	0.4864	0.0906
NMT_setup2	0.5696	0.0967	0.5110	0.1010
NMT_setup3	0.5789	0.0887	0.4969	0.0942
NMT_setup4	0.5406	0.1158	0.5324	0.1092

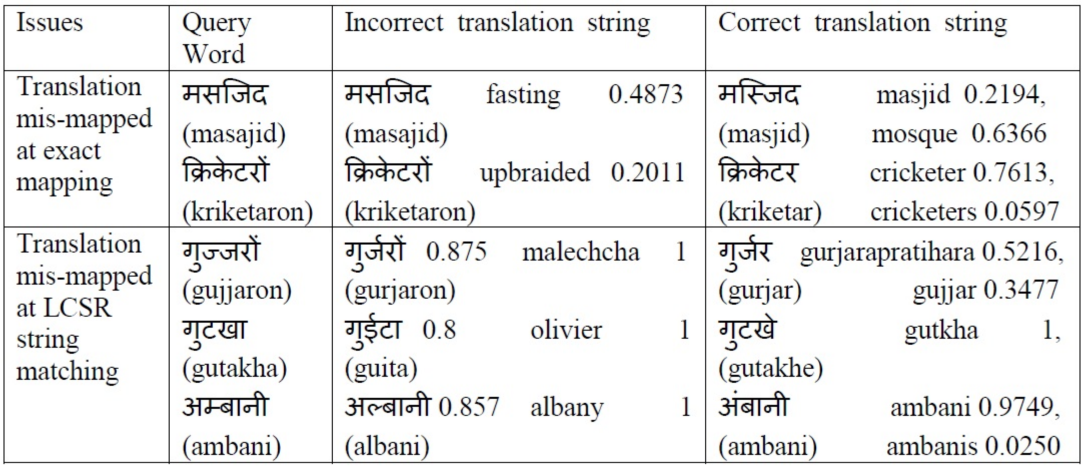
The NMT_setup4 (IITBCorpus, dropout value 0.2) achieves better MAP for both of the FIRE 2010 and 2011 datasets compared to other NMT setups. The experiment results show that the SMT performs better than the NMT because SMT uses the word alignment matrix while NMT uses attention mechanism. HindiEnCorp is smaller than the IITBCorpus, but it is well organized than the IITBCorpus. The SMT_setup1 uses a HindiEnCorp.

Our prepared SMT setup achieves better MAP 0.2832 than the state-of-art machine translation system i.e. 0.2299 achieved in (Gupta et al., 2011) and the baseline PD where MAP is 0.2267. This prepared SMT setup is used as the baseline for further experiments.

3.5 Semantic Morphological Variant Selection

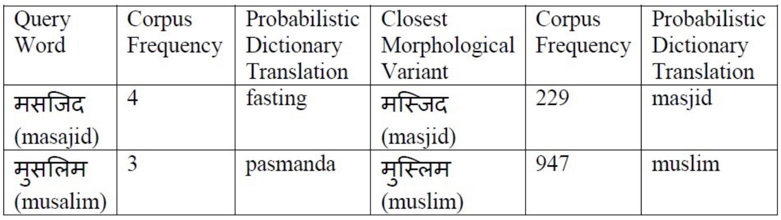
In order to produce a probabilistic dictionary, an IBM model is trained on a parallel corpus (Sorg & Cimiano, 2012; Zhou et al., 2012). A query word translation is extracted by exact mapping from a PD. The LCSR string matching technique is used to map those words which are not exactly mapped due to the morphological irregularities in Hindi language (Makin et al., 2007). A PD may not returns the correct translation string for both of the exact mapping and LCSR string matching technique which is called a translation mis-mapped issue. Examples of the translation mis-mapped issue are shown in Figure 3.6, where an exactly mapped word returns an incorrect translation string and in LCSR string matching either the source language word or the translation is incorrect.

In CLIR, a query word suffers from the translation mis-mapped issue at the exact mapping and LCSR string matching due to the low corpus frequency of the word



Issues	Query Word	Incorrect translation string	Correct translation string
Translation mis-mapped at exact mapping	मसजिद (masajid) क्रिकेटरों (kriketaron)	मसजिद fasting 0.4873 (masajid) क्रिकेटरों upbraided 0.2011 (kriketaron)	मस्जिद masjid 0.2194, (masjid) mosque 0.6366 क्रिकेटर cricketer 0.7613, (kriketar) cricketers 0.0597
Translation mis-mapped at LCSR string matching	गुज्जरों (gujjaron) गुटखा (gutakha) अम्बानी (ambani)	गुर्जरों 0.875 malechcha 1 (gurjaron) गुईटा 0.8 olivier 1 (guita) अल्बानी 0.857 albany 1 (albani)	गुर्जर gurjarapratihara 0.5216, (gurjar) gujjar 0.3477 गुटखे gutkha 1, (gutakhe) अंबानी ambani 0.9749, (ambani) ambanis 0.0250

Figure 3.6: Examples of the translation mis-mapped issue

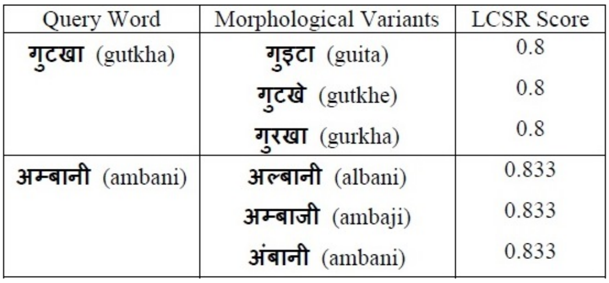


Query Word	Corpus Frequency	Probabilistic Dictionary Translation	Closest Morphological Variant	Corpus Frequency	Probabilistic Dictionary Translation
मसजिद (masajid)	4	fasting	मस्जिद (masjid)	229	masjid
मुसलिम (musalim)	3	pasmanda	मुस्लिम (muslim)	947	muslim

Figure 3.7: More relevant translations are appeared with the higher frequency

and multiple morphological variants respectively. An exactly mapped word with the higher corpus frequency has more relevant translation than the low corpus frequency as shown in Figure 3.7.

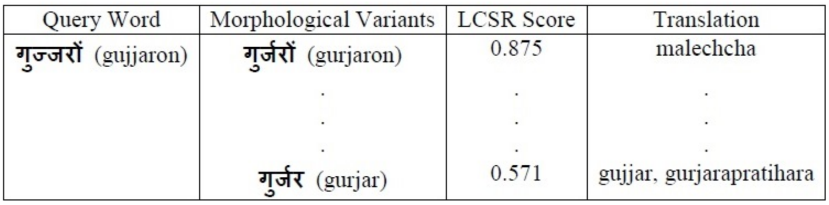
The LCSR may return multiple mapped morphological variants due to the same LCSR as shown in Figure 3.8.



Query Word	Morphological Variants	LCSR Score
गुटखा (gutkha)	गुइटा (guita) गुटखे (gutkhe) गुरखा (gurkha)	0.8 0.8 0.8
अम्बानी (ambani)	अल्बानी (albani) अम्बाजी (ambaji) अंबानी (ambani)	0.833 0.833 0.833

Figure 3.8: Multiple morphological variants due to the same LCSR score

A source language word which has at most one translation in the dictionary is considered as a non-confident translation while the word which has more than one translation is considered as a confident translation, as shown in Figure 3.6. A non-confident translation may be correct or incorrect. So, there is a need to increase the range of extracted translations for that query word which enhances the probability of correct translation. But, it will not work when there is a big difference between the LCSR of the correct translation and selected translation as shown in Figure 3.9.



Query Word	Morphological Variants	LCSR Score	Translation
गुज्जरों (gujjaron)	गुर्जरों (gurjaron)	0.875	malechcha
	.	.	.
	.	.	.
	.	.	.
	गुर्जर (gurjar)	0.571	gujjar, gurjarapratihara

Figure 3.9: Big difference between the LCSR scores of the correct translation and selected translations on the basis of maximum LCSR score

The words which suffer from the translation mis-mapped or non-confident translation issue, syntactically map to a morphological variant in the PD which may return an incorrect translation. Therefore, a Semantic Morphological Variant Selection (SMVS) algorithm is proposed in Algorithm 2 to select a morphological variant with the correct translation.

Algorithm 2 Semantic Morphological Variant Selection Algorithm

Input: An affected source language query word w_x which is either suffering from translation mis-mapped issue or a non-confident query word, Word Embedding (WE) for the source language, Proababilistic Dictionary (PD)

Output: Most similar source language words (morphological variants) translations up to five

```
translations=[];
t_count=0;
m_variants= [];
m_variants= WE_similar($w_x$, 20);
for (i=0 to len(m_variants)) do
    if (LCSR(m_variants[i], $w_x$)>0.50) then
        if (m_variants[i] in PD) then
            translations.append(PD[m_variants[i]]);
            t_count+=1;
            if (t_count==5) then
                break;
            end
        end
    end
end
```

The proposed algorithm includes the translation of more source language words which are semantically and syntactically nearer to the affected word (which is either suffering from translation mis-mapped issue or a non-confident query word).

Semantic and syntactic nearness are computed based on the word embedding and longest common subsequence ratio. Word embeddings are prepared by training of a Continuous Bag Of Word (CBOW) based recurrent neural network model on a large source language monolingual corpus. Then, top-20 similar words of an affected source language query word w_x are extracted by using $WE\_similar()$. Further, these semantically extracted words are syntactically verified by using LCSR which is calculated between an affected word and all semantically extracted words. If any semantic word score more than the LCSR threshold 0.50 and its translation is present in the PD, then its translation is included for that affected word. The threshold value 0.50 is marked as half-confidence. The LCSR which lies beyond the half-confidence returns the translation of a significant source language word. This threshold value can be varied from application to application. This algorithm returns up to the five semantically and syntactically verified source language words translations instead of the one syntactically selected word translation which is computed by using the LCSR.

FIRE 2010 and 2011 ad-hoc datasets, statistics are represented in Table 3.5 are used to evaluate the proposed algorithm. A query includes $< title >$, $< desc >$ and $< narr >$ tags, among them, only $< title >$ tag is considered for the evaluation. A Hindi-English parallel corpus HindiEnCorp is exploited to produce the PD. A large Hindi language raw corpus (approx 10GB) which is a combination of Hindi Wikipedia articles[13] and Bojar Hindi MonoCorp[14], is used to generate the Hindi language word embeddings. Following evaluation measures are used to evaluate the proposed approach.

- Recall: It is the fraction of relevant documents that are retrieved as shown in Equation 3.17.

$$Recall = \frac{|\{relevant\ documents\} \bigcap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{3.17}$$

- MAP: It is the mean of the average precision score of each query. Precision is the fraction of retrieved documents that are relevant to the query. Average precision of the query is calculated by the Equation 3.3.
- Precision@5 (P@5): It is an average precision computed at the top five target documents.

[13] https://dumps.wikimedia.org/backup-index.html

[14] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-6260-A

- Precision@10 (P@10): It is an average precision computed at the top ten target documents.

Experimental setups are prepared by using SMT and PD. A HindiEnCorp parallel corpus is used to train the SMT and to produce the PD (Koehn, 2009). In PD, a Hindi language word has multiple translations which are associated with the probabilistic score. Generally, translations are chosen in favor of the maximum probabilistic score. A PD with exact mapping and LCSR string matching technique is used to map the query words in the dictionary but it is not sufficient due to the translation mis-mapped and non-confident translation issue. Hence, the proposed SMVS algorithm is also included.

Table 3.16: Experimental results for the semantic morphological variant selection algorithm

Approach	FIRE 2010				FIRE 2011			
	Recall	**P@5**	**P@10**	**MAP**	**Recall**	**P@5**	**P@10**	**MAP**
SMT (baseline)	0.8284	0.2840	0.2300	0.2832	0.7084	0.2440	0.2320	0.1885
PD_LCSR (baseline)	0.8192	0.2520	0.2320	0.2645	0.6787	0.2120	0.2320	0.1667
PD_LCSR_ SMVS	0.8545	0.3000	0.2580	0.3032	0.7040	0.2240	0.2540	0.1865

SMT and PD with LCSR (PD_LCSR) are considered as the baselines to evaluate the proposed SMVS (PD_LCSR_SMVS) algorithm. The experiment results of the baselines and the proposed algorithm for FIRE 2010 and 2011 datasets are represented in Table 3.16. Impact of SMVS algorithm is represented in Figure 3.10. SMVS increases the recall, P@5, P@10, & MAP for FIRE 2010, and P@10 for FIRE 2011, whereas the MAP and recall remain approximately equivalent to SMT for FIRE 2011. So, it can be concluded that SMVS algorithm enhances the CLIR performance compared to the SMT and PD with LCSR string matching.

Source language queries are translated into the target language in order to perform cross-lingual information retrieval. A manual dictionary or a probabilistic dictionary is used to translate the user queries. Extraction of a correct translation string from a dictionary is sometimes not possible due to the translation mis-mapped and non-confident translation issues. The proposed SMVS algorithm is compared to the baselines SMT and PD_LCSR. It is concluded that the SMVS performs better than the SMT and PD_LCSR.

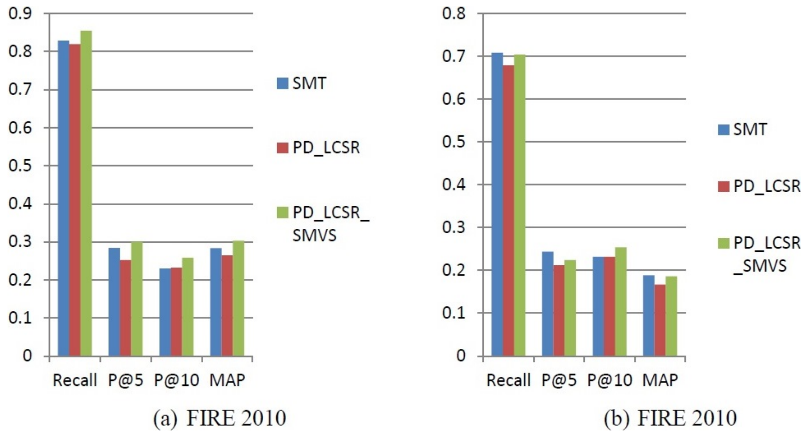


Figure 3.10: Impact of SMVS, evaluated by Recall, P@5, P@10, and MAP for (a) FIRE 2010 and (b) FIRE 2011

3.6 Context-Based Translation Algorithm for the Out Of Vocabulary Words

CLIR incorporates a translation technique which is based on either a dictionary or a parallel corpus. The trending translation techniques SMT and NMT are also trained over the parallel corpus. The efficiency of SMT and NMT depends on the parallel corpus of good quality and size which is difficult to obtain for the resource-scarce languages (Egozi et al., 2011; Nasharuddin & Abdullah, 2010; Vulić et al., 2013; Zhou et al., 2012). A parallel corpus is a set of mutually translated sentences which is not so absolute that could translate all words. In (Kunchukuttan et al., 2018) author says that the SMT achieves higher BLEU score than the NMT for Hindi to English translation. The SMT skips some words without performing translation and such words are called OOV words. The OOV word translation is the biggest challenge (Nagarathinam & Saraswathi, 2011; V. K. Sharma & Mittal, 2016a) as shown by the examples in Figure 3.11, where word “paairesee” and “aaepeeail” are not translated by the SMT.

In the literature, a phrase based graph propagation algorithm computes the translations from a large source language monolingual corpus and a parallel corpus where the translations are propagated directly from the labeled node to the OOV node or via unlabeled node (Razmara, Siahbani, Haffari, & Sarkar, 2013). The algorithm does not make use of the large target language corpus, so the translations are limited to only the parallel corpus vocabulary. The size of dictionaries and phrase tables are extended by utilizing a large monolingual corpus and a small bi-lingual parallel corpus where a transformation matrix is learned from the parallel corpus. The OOV words translations are computed with the help of

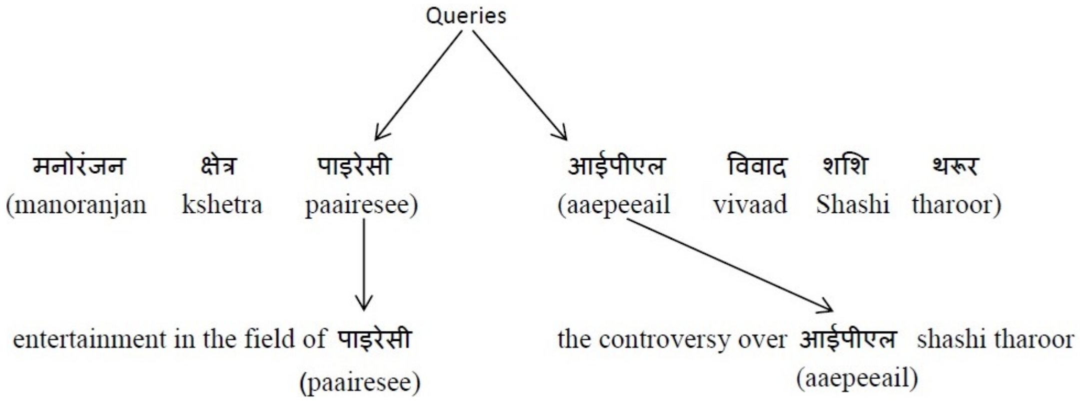


Figure 3.11: Examples of words that are not translated due to their unavailability in the parallel corpus

the transformation matrix. These translations are from the parallel corpus which does not provide any surety that these translations are accurate, not synonyms or morphological variants (Mikolov, Le, & Sutskever, 2013). A Japanese to English machine translation system handles the OOV words, with the condition that, at least one of the orthographic variant of that OOV word needs to be present in the parallel corpus (Luo & Lepage, 2015). SMT capability is also extended by adding either the translation pairs or artificial sentences for the unseen morphological variants (Akhtar et al., 2017; Huck et al., 2017; Pilehvar & Collier, 2017). The above-discussed methods compute the translation for the OOV morphological variants and missing words by using a transformation matrix learned from the parallel corpus which provides approximate translation while the proposed algorithm computes the translation based on the highly relevant contextual words, therefore, the proposed approach will translate the OOV words more effectively and accurately.

A Context-Based Translation Algorithm for the Out Of Vocabulary (CBTA-OOV) which incorporates a Similarity Computation (SC) and a Similarity Association (SA) component, is proposed to translate the OOV words. The SC component returns the PMI and Word Embedding (WE) based contextual words and the SA component selects the highly relevant contextual words. The proposed algorithm exploits two large unlabeled & unrelated mono-lingual corpora (in source and target language), and a small bi-lingual parallel corpus.

The CBOW and Skip-Gram based RNN models are used to learn the WE from the large corpora. The CBOW predicts a target word based on the contextual word whereas the Skip-Gram predicts the contextual words based on the current word (Mikolov, Chen, et al., 2013). Skip-gram with hierarchical softmax function is used to improve the vector representation for the word and phrases (Mikolov, Sutskever, et al., 2013). A Log-Bilinear Regression (LBR) model which efficiently uses the statistical information by using the co-occurrence matrix, is also used to learn the word vectors (Pennington et al., 2014). Attention mechanism enhances the

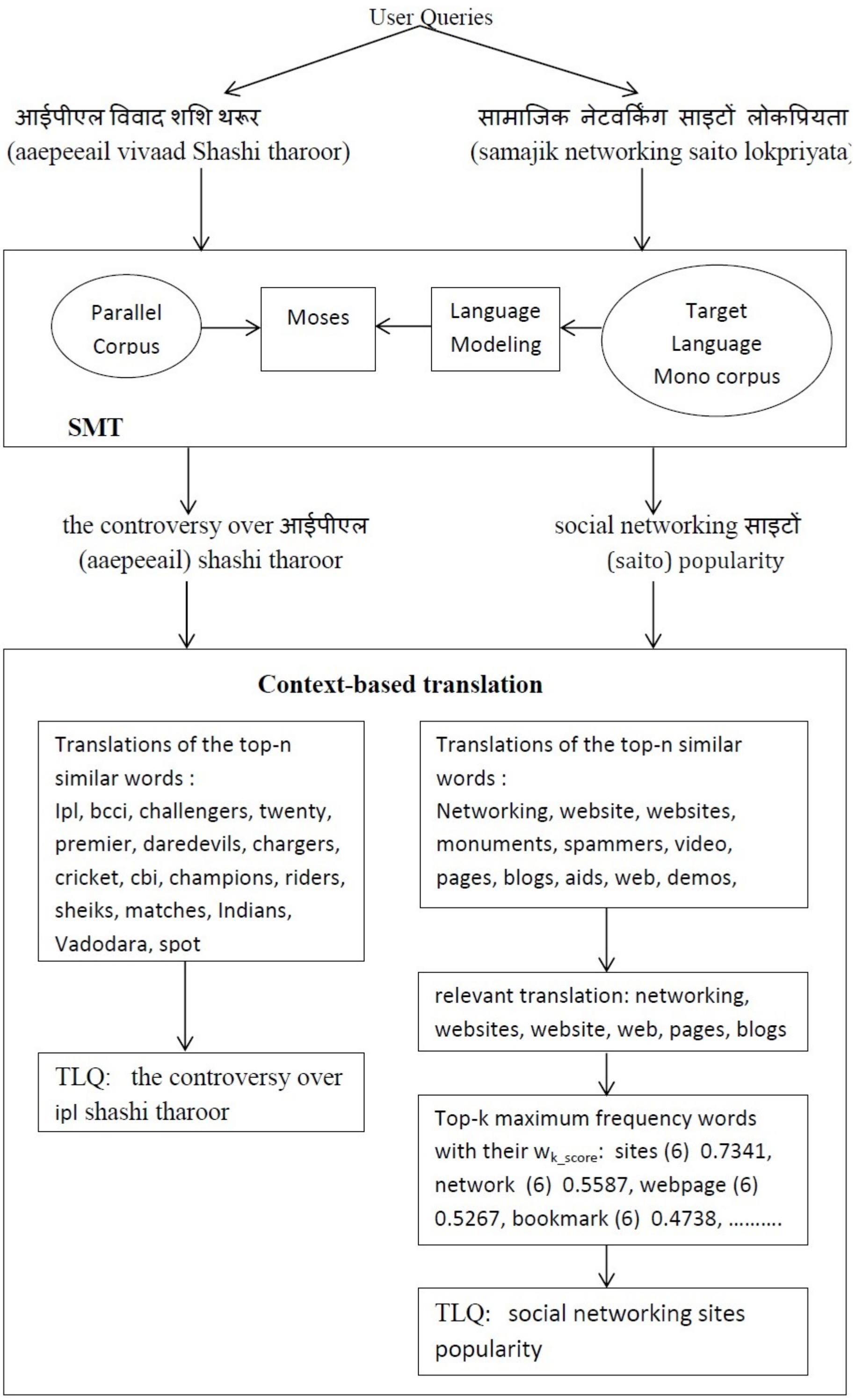

User Queries
आईपीएल विवाद शशि थरूर
(aaepeeail vivaad Shashi tharoor)
सामाजिक नेटवर्किंग साइटों लोकप्रियता
(samajik networking saito lokpriyata)
Parallel Corpus
Moses
Language Modeling
Target Language Mono corpus
SMT
the controversy over आईपीएल (aaepeeail) shashi tharoor
social networking साइटों (saito) popularity
Context-based translation
Translations of the top-n similar words : Ipl, bcci, challengers, twenty, premier, daredevils, chargers, cricket, cbi, champions, riders, sheiks, matches, Indians, Vadodara, spot
TLQ: the controversy over ipl shashi tharoor
Translations of the top-n similar words : Networking, website, websites, monuments, spammers, video, pages, blogs, aids, web, demos,
relevant translation: networking, websites, website, web, pages, blogs
Top-k maximum frequency words with their w_k_score: sites (6) 0.7341, network (6) 0.5587, webpage (6) 0.5267, bookmark (6) 0.4738, ...
TLQ: social networking sites popularity

Figure 3.12: Examples showing the OOV words translation with the help of contextual words

capabilities of RNN based encoder-decoder where a direct short-cut connection is established between the source and target language sentences by using the context vectors (Bahdanau et al., 2014). CLIR incorporates SMT which suffers from the OOV words issue due to the non-availability of the OOV words in the parallel corpus, therefore, the OOV words are skipped by the SMT without performing the translation. The OOV words may be translated with the help of contextual words as shown by query examples in Figure 3.12.

The query “aaepeeail vivaad shashi tharoor” is translated by SMT where word “aaepeeail” is skipped which remains in the source language form and that word is further translated with the help of top-n contextual words. The translation of the highest similar word of the OOV word “aaepeeail” is the required translation that is “ipl”. The query “samajik networking saito lokpriyata” is a little bit different than the previous query, here, the translation of the highest similar word of the OOV word “saito” does not return the required translation. In this situation, translations of the top-n similar words of the OOV word are computed and then the highly relevant translations are selected, further, top-m similar words of the highly relevant translations are collected in a container. Unique word frequencies are calculated from the container and a similarity score is assigned to each word based on other translated query words. The word which has the highest frequency and a maximum similarity score is selected as the best translation for the OOV word.

A context-based translation algorithm based on the above-discussed idea is proposed in Algorithm 3. The proposed algorithm incorporates a similarity computation & similarity association components and utilizes two large mono-lingual (source & target language) raw corpora along with the PD. The SC component returns contextual words by using either PMI or WE.

Algorithm 3 A Context-Based Translation Algorithm for the Out Of Vocabulary words

```
Input: A Target Language Query $TLQ[w_1, w_2, ..., w_w]$ obtained from the SMT where $w$ represents a word, a Probabilistic Dictionary (PD) produced by training an IBM model on the parallel corpus, A Source Language Mono-lingual Corpus (SLMC), and A Target Language Mono-Lingual Corpus (TLMC).
Output: A modified Target Language Query TLQ where OOV words are replaced by the best translation
Prepare Similarity Computation (SC) components SC_SLMC from SLMC and SC_TLMC from TLMC by using Equation 3.18 and 3.19;
```

```
for each OOV word $w_i$ in TLQ do
    flag=0;
    similar_words = SC_SLMC_most_similar($w_i, n$);
    for each word $w_j$ in similar_words do
        lcsr=$|LCS(w_i,\ w_j)|/Maximum(|w_i|,\ |w_j|)$; // LCS returns Longest Common Sub-sequence string
        if (lcsr $\geq$ 0.75 and $w_j$ in PD) then
            flag=1;
            TLQ[$w_i$] = PD[$w_j$];
        end
    end
    if (flag==0) then
        temp=[ ]; TLT=[][];
        for each word $w_j$ in similar_word do
            if ($w_j$ in PD) then
                temp.add(PD[$w_j$]);
            end
        end
        HR_class_translation = SA(temp); //SA: Similarity Association is discussed in the sub-sequent paragraph
        for each word $w_k$ in HR_class_translation do
            TR[ ] = SC_TLMC_most_similar($w_k, m$);
        end
        compute unique words frequency from TR[ ];
        Extract top-k maximum frequency word;
        for each word $w_k$ in top-k maximum frequency word do
            $w_{k_{score}}$=0;
            for each word $w_l$ in TLQ do
                if ($w_l \notin$ OOV words and $w_l \notin$ Stop-words) then
                    $w_{k_{score}}$=$w_{k_{score}}$+$SC\_TLMC\_similarity(w_k, w_l)$;
                end
            end
            TLT[$w_i$][$w_k$]=$w_{k_{score}}$;
        end
        TLQ[$w_i$] = max(TLT[$w_i$].iteritems(), key=operator.itemgetter(1))[0]; // key $w_k$ with the maximum value $w_{k_{score}}$ from TLT[$w_i$], is assigned to the TLQ
    end
end
```

In PMI based SC, two bi-gram lexicons are constructed. One is from Source Language Monolingual Corpora (SLMC) and the other one is from Target Language Monolingual Corpora (TLMC). These lexicons contain entries like $\{(w_i, w_j)\ PMI(w_i, w_j)\}$, where (w_i, w_j) is a word pair and $PMI(w_i, w_j)$ is the PMI score of the word pair (Razmara et al., 2013). The PMI score between the word w_i and w_j is computed by Equation 3.18.

$$PMI(w_i, w_j) \ = \ log_2 \frac{P(w_i, w_j)}{P(w_i)\, P(w_j)} \tag{3.18}$$

The OOV words are translated by using the PMI based contextual words which are considered as similar words. WE are prepared by using three popular deep learning architectures, i.e., CBOW, Skip-Gram (Mikolov, Chen, et al., 2013), and LBR models (Pennington et al., 2014) where each word of SLMC and TLMC is represented by a unique vector. A cosine similarity measure is used to compute the similarity between the word vectors A=$\{a_1, a_2, ..., a_n\}$ and B=$\{b_1, b_2, ..., b_n\}$, as shown in Equation 3.19.

$$Cosine\ Similarity(A, B) \ = \ \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\ \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{3.19}$$

The OOV morphological variants and missing words which are not translated by SMT are processed by the algorithm. Where, top-n similar words of the OOV word are extracted from the SC component of SLMC. Further, LCSR with the threshold 0.75[15] is used to select the highest similar word which must be a key in the PD. The translation of the highest similar word will be substituted at the place of OOV word in the target language query as shown in Figure 3.12, where “ipl” is substituted. The LCSR between two strings is computed by Equation 3.1.

The first segment of the algorithm translates almost all OOV morphological variants. The remaining OOV words (missing words which are not present in the parallel corpus) which are not translated by the first segment (flag remains 0 in the algorithm in that case) are translated by the second segment. In that case, target language translations are extracted from the PD for all top-n similar words where a word which is not found in the PD, is ignored. PD returns translations associated with the probability score but it does not provide any surety that the returned translations are correct or relevant to the OOV word. Irrelevant trans-

[15] The LCSR threshold is chosen experimentally, a number of experiment has been done, among them, in most of the cases the threshold 0.75 returns appropriate words

lations mix unwanted contextual words, due to that, the algorithm skips more accurate translation. Therefore, these irrelevant translations are eliminated by using the similarity association component.

In SA, the similarity is computed among the translations of the top-n similar words. Word pairs which achieve similarity scores above a threshold T are considered as the highly relevant words. Such word pairs should satisfy the Equation 3.20.

$$SC\_TLMC\left[w_i, w_j\right]_{i,j=1}^{n} \geq T; i \neq j \tag{3.20}$$

Table 3.17: Extraction of highly relevant words by using semantic association component

Translations of top-n similar words	Similarity scores among the translation	HR Class words
Networking, Websites,	Websites, website (0.6445)	Websites (10)
Monuments, Spammers,	Websites, Networking (0.6327)	website (10)
Website, Blogs, Web,	Networking, Website (0.6338)	Networking (10)
Pages, Aids, Video	....	....

Further, unique words are collected from the highly relevant word pairs and their frequencies $freq(w_i)$ are calculated, where $freq(w_i)$ represents the number of words which have a similarity score with w_i more than the threshold T. These words are grouped into two clusters, i.e., High Relevant (HR) and Low Relevant (LR), by using the k-means clustering algorithm with the Manhattan distance (Loohach & Garg, 2012) over the attribute $freq(w_i)$. HR class contains the high-frequency words and LR class contains the low-frequency words. The SA component returns the HR class words as shown in Table 3.17, where translations of top-n similar words of the OOV word "saito" are computed. The similarity scores among the translations are computed by using Equation 3.20 and these pairs (second column entries represent word pairs with a similarity score in the bracket) are collected in a container. Further, unique word frequencies are calculated and HR class words are computed based on that frequency (third column entries represent words with the frequency in the bracket).

After obtaining the HR class words, top-m similar words of HR class words are computed from SC of TLMC and unique words frequencies are calculated from the collection of top-m similar words. Then top-k maximum frequency words are extracted and a translation score is assigned to each word where the translation score is the addition of the similarity scores between the other translated words of

TLQ and that word. A word with the maximum translation score is substituted at the place of the OOV word in the target language query like the translation "sites" is replaced at the place of "साइटों" in Figure 3.12.

In our experiments, three SMT setups with the different training datasets, are trained to achieve the baseline results. Training datasets details are shown in Table 3.18. SMT_setup1 uses HindiEnCorp and SMT_setup2 uses IITBCorpus, for both of the purposes of training and language modeling whereas SMT_setup3 uses IITBCorpus for the training and WMT[16] news corpus 2015 for language modeling. A common WMT dev dataset is used for the purpose of validation.

Table 3.18: Resources which are used to train the SMT

Experiments	**Training datasets**	**Language Modeling**	**Dev datasets**
SMT_setup1	HindiEnCorp (2,73,886 sentences)	HindiEnCorp	WMT Dev_set (520 sentences)
SMT_setup2	IITBCorpus (1,492,827 sentences)	IITBCorpus	
SMT_setup3		WMT News 2015 Corpus (3.3 GB)	

The proposed algorithm translates the OOV words which are not translated by the SMT, with the help of a large Hindi language raw corpus (SLMC, approx 10GB) which is a combination of Hindi Wikipedia articles & Bojar Hindi MonoCorp, and a large English language raw corpus (TLMC, approx 12GB) which is collected from the English Wikipedia articles[17]. SC components are prepared for source and target language from SLMC and TLMC respectively. PMI based SC returns a Bi-gram Lexicon (Bi_Lex) and WE based SC returns the Word Vectors (W_Vec). W_Vec are constructed by using the Skip-Gram Vector (SG_Vec), Continuous Bag-Of-Word Vector (CBOW_Vec), and Log Bi-linear Regression Vector (LBR_Vec) models. These W_Vec models are trained with a context window of size s=10. Experimental parameters n, m, and k are tested for (20, 50), (500, 1000, 2000), and (10, 20, 50) respectively. SA component eliminates irrelevant translation based on a threshold T which is taken empirically 0.1 in the case of Bi_Lex while 0.50 in the case of W_Vec. The proposed CBTA-OOV has experimented with the different SC component and different n, m, k values. Therefore, experiments are named in the form of '**CBTA-OOV_SC_n_m_k**' where SC has four variants, i.e., Bi_Lex, SG_Vec, CBOW_Vec, and LBR_Vec.

[16] http://www.statmt.org/wmt15/translation-task.html

[17] https://dumps.wikimedia.org/backup-index.html

FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5 are used to evaluate the baseline SMT and the proposed CBTA-OOV based CLIR approach. Terrier open source search engine is used for indexing and retrieval, where TF-IDF is used for indexing and cosine similarity is used for retrieval. The CLIR system is evaluated by using recall, Precision@5 (P@5), Precision@10 (P@10), and MAP as discussed in Section 3.5.

Three SMT setups are trained and evaluated on FIRE ad-hoc dataset for Hindi-English CLIR and the best one is used as the baseline. Evaluation measures for these three SMT setups are presented in Table 3.19. SMT_setup1 performs better than the SMT_setup2 and SMT_setup3 in perspective of CLIR. SMT_setup1 is trained on the HindiEnCorp which is smaller than the IITBCorpus, which is used in SMT_setup2 and SMT_ setup3. IITBCorpus is a superset of HindiEnCorp, but it is not so well-organized and mixes the noise in the translation, hence it achieves poor performance compare to HindiEnCorp. SMT_setup1 is considered as the baseline experiment against the proposed approach.

Table 3.19: Baseline results for the SMT setups in perspective of Hindi-English CLIR

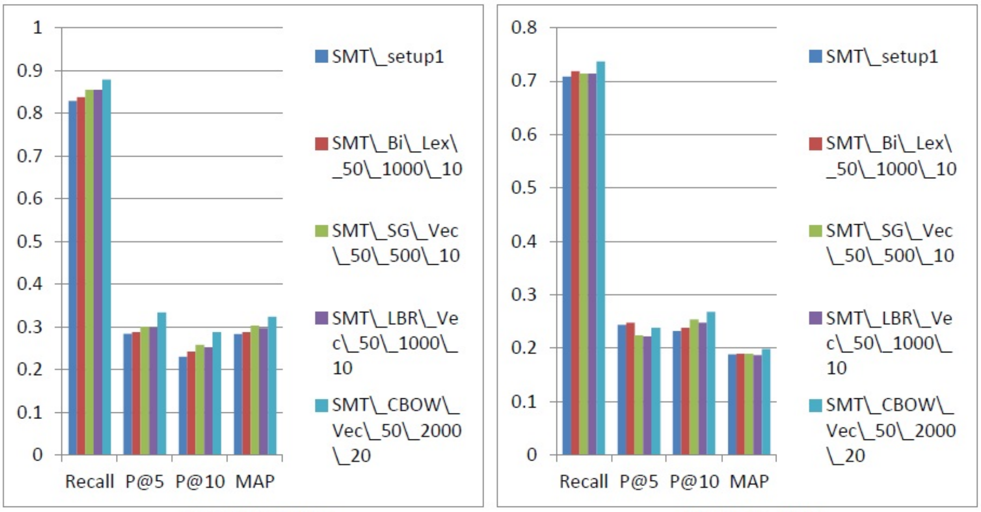
Setups	**Dataset**	**Recall**	**P@5**	**P@10**	**MAP**
SMT_setup1	FIRE 2010	0.8284	0.2840	0.2300	0.2832
	FIRE 2011	0.7084	0.2440	0.2320	0.1885
SMT_setup2	FIRE2010	0.7718	0.2220	0.1660	0.2175
	FIRE 2011	0.6602	0.1640	0.1220	0.1608
SMT_setup3	FIRE 2010	0.7978	0.2280	0.1740	0.2237
	FIRE 2011	0.6602	0.1780	0.1320	0.1767

The proposed algorithm in which the SMT_setup1 is used as the baseline is evaluated for FIRE 2010 and 2011 datasets. Experiment results for the proposed approach are represented in Table 3.20.

A comparison between the proposed approach and the baseline SMT_setup1 for both of the FIRE 2010 & 2011 datasets is represented in Figure 3.13. The proposed CBTA-OOV performs better than the baseline SMT as the recall, P@5, P@10, and MAP for SMT_CBOW_Vec_50_2000_20 are improved up to 6.04%, 17.60%, 25.21%, 14.37% in comparison to SMT for FIRE 2010, & recall, P@10, and MAP for SMT_CBOW_Vec_50_2000_20 are improved up to 3.96%, 15.51%, 5.46% in comparison to SMT for FIRE 2011. Bi_Lex model returns PMI based bi-grams and PMI is a co-occurrence method, hence, Bi_Lex does not return semantically similar words. LBR_Vec model generates the word vectors which are based on the co-occurrence matrix and SG_Vec predicts contextual words based on the cur-

Table 3.20: Experiment results for the proposed CBTA-OOV algorithm in perspective of Hindi-English CLIR

Experiments	**Dataset**	**Recall**	**P@5**	**P@10**	**MAP**
CBTA-OOV_	FIRE 2010	0.8379	0.2880	0.2420	0.2872
Bi_Lex_50_1000_10	FIRE 2011	0.7184	0.2480	0.2380	0.1896
CBTA-OOV_	FIRE 2010	0.8545	0.3000	0.2580	0.3032
SG_Vec_50_500_10	FIRE 2011	0.7140	0.2240	0.2540	0.1896
CBTA-OOV_	FIRE 2010	0.8545	0.2980	0.2520	0.2969
LBR_Vec_50_1000_10	FIRE 2011	0.7140	0.2220	0.2480	0.1885
CBTA-OOV_	FIRE 2010	0.8785	0.3220	0.2760	0.3196
CBOW_Vec_20_1000_10	FIRE 2011	0.7365	0.2340	0.2620	0.1946
CBTA-OOV_	FIRE 2010	0.8785	0.3340	0.2880	**0.3239**
CBOW_Vec_50_2000_20	FIRE 2011	0.7365	0.2380	0.2680	**0.1988**



(a) FIRE 2010 (b) FIRE 2011

Figure 3.13: Comparison of evaluation measures between the proposed approach and baseline SMT_setup1 for (a) FIRE 2010, and (b) FIRE 2011

rent word. SG_Vec and LBR_Vec SC components perform approximately equal. CBOW_Vec SC model outperforms the other models due to the difference among the model architecture as the CBOW_Vec predicts target word based on the contextual words. Experiments are performed with the different values of n, m, k, amongst them, only significant experiment results are reported in Table 3.20. Values of n, m, k carry a significant information as the SMT_CBOW_Vec_50_2000_20 achieves better results compared to SMT_CBOW_Vec_20_1000_10.

The proposed algorithm reduces the number of OOV words to improve evaluation measures. A comparison among the baseline and proposed algorithm is represented in Table 3.21 and by a chart in Figure 3.14 which shows that the CBOW_Vec SC model effectively reduces more number of OOV words in comparison to the other

Table 3.21: Out of vocabulary word's statistics for the FIRE topic sets after applying the baseline SMT and the proposed CBTA-OOV

Approaches	FIRE 2010 (Total: 245)	FIRE 2011 (Total: 173)
SMT	17 (06.93%)	7 (04.04)
SMT_Bi_Lex	15 (06.12%)	5 (02.89%)
SMT_SG_Vec	7 (02.85%)	5 (02.89%)
SMT_LBR_Vec	9 (03.67%)	7 (04.04%)
SMT_CBOW_Vec	2 (00.81%)	3 (01.73%)

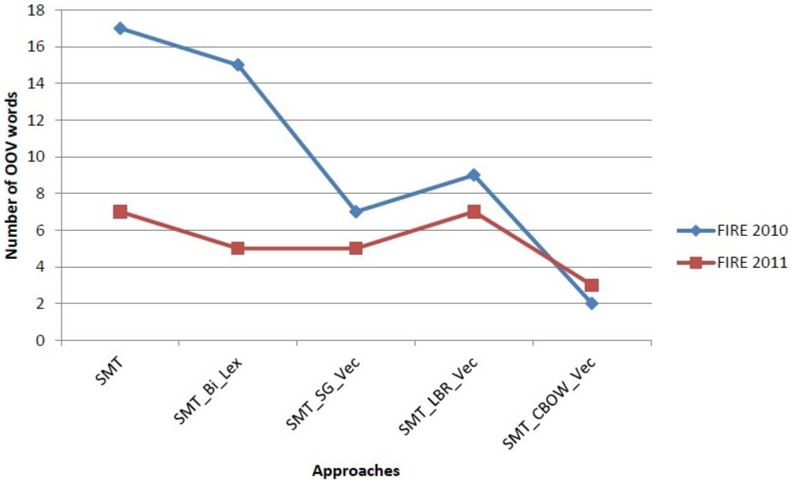


Figure 3.14: The number of out of vocabulary words remained after applying the proposed approach and baseline SMT_setup1 for (a) FIRE 2010, and (b) FIRE 2011 dataset

models.

Performance analysis of the proposed CBTA-OOV algorithm is done based on the Intermediate States (IS) which are given as follows.

- **IS1**: Number of correct or relevant translations n_f from the translation of top-n similar words where $n_f \leq n$. State IS1 is represented by a tuple $\{n, n_f\}$.
- **IS2**: Number of top-m similar words of translation of n_f words in which the target language translation of the OOV word is present. Tested values of m are 500, 1000, and 2000, if the required translation is even not present in m=2000 then the IS2 value will be NP (Not Present).

- **IS3**: Number of top-k target language translations in which the required translation is present. Tested values of k are 10, 20, 50, and if the required translation is even not present in k=50 then the value of IS3 will be NP (Not Present).

The performance analysis for the different SC modules based on these three intermediate states is represented in Table 3.22. The *IS1* has entries in the form of $\{n,n_f\}$ where n_f is the number of correct or relevant translations from the n words. The SA component eliminates irrelevant translations based on a threshold T, so the value of n_f is very less than the value of n. Defining a threshold T for PMI is an arduous task because it returns a wide range of co-occurrence scores and a large set of contextual words due to the bi-grams. The WE based SC is trained with the context window of size s, where s may be 5, 10, 20, etc. So the computed similar words by WE based SC, are more appropriate. Generally, two words are considered similar if their similarity score is more than 50%, so a threshold 0.50 is defined for WE based SC.

NP entry for *IS2* shows that the required translation is not present in even top-2000 similar words for the n_f translations. Although the required translation may be present in top-2000 similar words for n_{ff} translations where $n_{ff} < n_f$, but in that case, the translation will not appear in top-k maximum frequency words where $k = 50$ as shown in Table 3.22. If $IS2$ has an entry NP then the $IS3$ also has the same entry NP. If $IS2$ and $IS3$ has some numeric value, then the translation is found for all the n_f translations.

In PMI, the threshold T may be appropriate for some words but not for all words, so various translations in n_f are irrelevant. The required translation is not present in top-m similar words and top-k translations where $m = 2000$ and $k = 50$. So the PMI gives poor performance due to the two reasons, (i) it is based on bi-grams so it includes a large set of similar words, and (ii) defining a threshold T is not appropriate as an important word pair may have a low PMI score. The CBOW_Vec follows an architecture where the next word is predicted based on a context window of size s. So, the CBOW_Vec model performance is better than others as it returns exact translation for 5 words among the example set of 10 words while SG_Vec and LBR_Vec return for only two words.

Table 3.22: Performance Analysis of the Proposed CBTA-OOV Algorithm

SC	IS	**सीबीआई**	**निठारी**	**गुज्जरों**	**पाइरेसी**	**साइटों**	**चेचेन**	**आईपीएल**	**नरेगा**	**सर्वाइकल**	**आईफोन**
		cbi	nithari	gujjar	piracy	sighto	chechen	ipl	nrega	cervical	iphone
PMI	IS1	20,4	20,7	20,3	20,0	20,3	20,0	20,4	20,6	20,2	20,3
	IS2	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP
	IS3	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP
SG	IS1	50,2	50,4	50,4	50,3	50,3	50,4	50,4	50,2	50,3	50,6
_Vec	IS2	1000	NP	NP	NP	NP	NP	NP	NP	NP	1000
	IS3	50	NP	NP	NP	NP	NP	NP	NP	NP	50
CBOW	IS1	50,7	50,2	50,2	50,3	50,5	50,6	50,4	50,5	50,3	50,7
_Vec	IS2	1000	NP	NP	NP	2000	500	2000	NP	NP	1000
	IS3	20	NP	NP	NP	20	10	10	NP	NP	10
LBR	IS1	50,17	50,8	50,2	50,0	50,6	50,3	50,10	50,9	50,0	50,10
_Vec	IS2	NP	NP	NP	NP	2000	NP	NP	NP	NP	2000
	IS3	NP	NP	NP	NP	50	NP	NP	NP	NP	50

The target language translation depends on SC model along with the value of n_f, m, k. In SG_Vec, the value of n_f is low that is n_f=2, the translation is present in m=2000, and k is also high that is 50. In CBOW_Vec, n_f values are average that is 4 to 7, m values varied among 500, 1000 and 2000, and k values are 10 and 20. In LBR_Vec, the n_f values are average, but m and k values are high that is m=2000 and k=50. It is not necessary that if n_f values are high then the values of m and k should be low as shown in Table 3.22, LBR_Vec model has n_f=17 but the corresponding values of m and k are NP. CBOW_Vec model has a consistency that if n_f has average values and m is varied from 500 to 2000, then the target language translation will be present in top-k translations where k will be 10 and 20. The internal states of the algorithm do not depend on the languages. The proposed algorithm will also work for other language pairs as efficiently as it works for Hindi-English, only corresponding resources need to be changed.

The proposed CBTA-OOV approach which is varied based on three SC models is evaluated for FIRE ad-hoc dataset in Table 3.20. The CBOW_Vec model outperforms the other models as it achieves the evaluation measures 0.8785 recall, 0.3340 P@5, 0.2880 P@10, and 0.3239 MAP for FIRE 2010, & 0.7365 recall, 0.2380 P@5, 0.2680 P@10, and 0.1988 MAP for FIRE 2011 which are better than evaluation measures achieved by baseline SMT setup, i.e., 0.8284 recall, 0.2840 P@5, 0.2300 P@10, and 0.2832 MAP for FIRE 2010, & 0.7084 recall, 0.2440 P@5, 0.2320 P@10, and 0.1885 MAP for FIRE 2011. An analysis for the number of OOV words is also represented in Table 3.21. The CBOW_Vec model reduces the OOV words more effectively which are 2 (0.81%) for FIRE 2010 and 3 (1.73%) for FIRE 2011, in comparison to baseline SMT setup, i.e., 17 (6.93%) for FIRE 2010 and 7 (4.04%) for FIRE 2011.

3.7 Summary

In this chapter, manual dictionary based approach translates the source language query words by exact matching or LCSR based partial matching. The OOV words are translated by compressed word format transliteration mining technique. The manual dictionary based approach where Shabdanjali or English Hindi mapping is used as the dictionary is evaluated for FIRE 2010 dataset. The manual dictionary based approach with Shabdanjali dictionary achieves maximum MAP, i.e., 0.1172. The term frequency model addresses the issue of dictionary coverage and retraining of IBM model (probabilistic dictionary), hence, a set of query-relevant parallel sentences are exploited to compute the target language translation. The proposed term frequency model achieves better recall and MAP, i.e., 0.7519 and 0.2637 than the probabilistic dictionary, i.e., 0.7488 and 0.2267 respectively in case of FIRE 2010. In case of FIRE 2011, term frequency model achieves approximately equal evaluation measure to probabilistic dictionary due to the short length queries.

Refined stop-words lists are prepared for both of the source and target language and four morphological variants solutions are proposed to address the issue of morphological irregularities. The proposed translation induction algorithm incorporates the refined stop-words lists, morphological variants solutions, and the target language translations are computed based on the selected contextual parallel sentences for the query. The proposed translation induction algorithm achieves better MAP, i.e., 0.2818 and 0.1921 for FIRE 2010 and 2011 dataset, than the probabilistic dictionary based approach, i.e., 0.2547 and 0.1727 for FIRE 2010 and 2011 respectively. The experiments are also performed to test the performance of refined stop-word lists which performs better than the standard stop-word lists. The trending SMT and NMT are trained with the different dataset and parameter where it is found that the SMT performs better than the NMT.

Translation induction algorithm doesn't resolve all types of irregularities and suffers from the translation mis-mapped and non-confident translation issues. Therefore, the SMVS algorithm is proposed which selects the syntactically and semantically verified morphological variant. SMT is trained on an available HindiEnCorp. The OOV words are not translated even by the SMT, therefore, a CBTA-OOV algorithm which incorporates SC and SA components is proposed to translate the OOV words. SC component is prepared by using PMI and WE where WE are learned by the CBOW, Skip-Gram, and LBR models. SC exploits two large unlabeled and unrelated mono-lingual corpora (in source and target language) to prepare PMI and WE. A small bi-lingual parallel corpus is used to induce the translation for the OOV word. The proposed CBTA-OOV algorithm with CBOW model achieves better results than the SMT and CBTA-OOV with other models.

Chapter 4

Words Average Probability and Association Score based Disambiguation

A source language word is associated with the multiple translations which lead to Word Translation Disambiguation (WTD) issue. In this chapter, a bi-lingual word vector based translation disambiguation approach is proposed in section 4.1. Since Word Embedding (WE) provides context-based word vectors, hence, this bilingual word vector approach may disambiguate the words translation. There are some words which are both of the dictionary words and named entities. So it becomes difficult to decide that such words either need translation or transliteration. Therefore, a Named Entity Recognition (NER) based disambiguation approach is proposed in section 4.2. Generally, maximum probabilistic score based and association score based disambiguation approaches like word co-occurrence, PMI, WordNet path length, etc. are used to select the best translation. Both of these approaches may not provide the best translation in all situations, hence, a maximum of words average probability and association score based disambiguation approach is proposed in section 4.4. The association scores are calculated by using PMI or WE.

4.1 Hybrid Bilingual Word Vector Model based

Nowadays, distributed word vector representation has a trend in various natural language processing task. These word vectors are utilized to track similar contextual words. In this section, The effectiveness of bilingual word vectors for cross-lingual (Hindi-English) scenario is analyzed. The bilingual word vectors are learned by training the skip-gram model over the sentence-aligned parallel corpus. Most of the literature is focused on a single language (Ganguly et al., 2015; Pennington et al., 2014) where semantically similar words are recognized for a given word. Recently, the skip-gram model is extended from a single language to cross-lingual (Klementiev et al., 2012; Mikolov, Sutskever, et al., 2013). A Bilingual Word Embedding Skip-Gram (BWESG) model is trained over the combined and shuffled parallel sentences which actually creates a dual semantic space (Vulić & Moens, 2015).

In our approach, initially, BWESG model is used to learn the bilingual word vectors. Since the similar contextual words are having approximately nearer word vectors, hence, cosine similarity scores are calculated between each source language word and all target language words. A target language word with the maximum cosine similarity score is assigned to a source language word. This approach returns many target language words because these target language words all have similar word vectors to the source language word. So, the wrong target language word is assigned to a source language word. This approach does not provide good translations due to the different sentence lengths and sentence structure & different number of vocabulary words and stop-words across the languages. Therefore, a hybrid model is proposed, where the cosine similarity score is calculated between a source language word and top-k target language words which are extracted from the probabilistic dictionary.

User queries are translated into the target language using a hybrid model which incorporates BWESG along with the IBM model. Further, the vector space model is used to retrieve relevant documents from the target documents. A brief introduction of BWESG model (Klementiev et al., 2012; Vulić & Moens, 2015) and IBM model (Manning, Manning, & Schütze, 1999) is presented in the following sub-sections and then the proposed hybrid model is discussed.

BWESG Model

A combined bilingual corpus is prepared by merging parallel documents of a parallel corpus such that the newly prepared combined bilingual document corpus has the documents in the form of $\{d_1, d_2, ..., d_N\}$, where each document d_i is a merged document from both of the source and target language documents (d_s, d_t). The words from d_s and d_t are shuffled such that the final words in d_i are $\{w_{s1}, w_{t1}, w_{s2}, w_{t2}, ..., w_{sn}, w_{tm}\}$. The skip-gram model is trained over the combined bilingual document corpus. In the skip-gram model, the probability of predicting contextual words for the current word is to be maximized. The probability of predicting the context word v for the current word w is defined by softmax function as given in Equation 4.1.

$$p(v|w) = \frac{1}{1 + exp(-\vec{w}.\vec{v})} \tag{4.1}$$

The BWESG model learns the word embeddings for both of the source and target language words over *dim* embedding dimension. A *dim* dimensional vector for word w is:

$$\vec{w} = [f_{w,1}, f_{w,2}, ..., f_{w,dim}]$$

$f_{w,k}$ denotes the $k^t h$ inter-lingual feature. Further, semantic similarity is computed both of the monolingually or cross-lingually.

IBM Model

A parallel sentence pair (s, t) where $s = (s_1, s_2, ..., s_n)$ of length n and $t = (t_1, t_2, ..., t_m)$ of length m is given. A translation probability for each source language word s_i to a target language word t_j with an alignment $a : j \rightarrow i$ is given in Equation 4.2.

$$p(t, a|s) = \prod_{j=1}^{m} tp(t_j|s_{a(j)}) \tag{4.2}$$

t_p represents the translation probability of the target language words against a source language word.

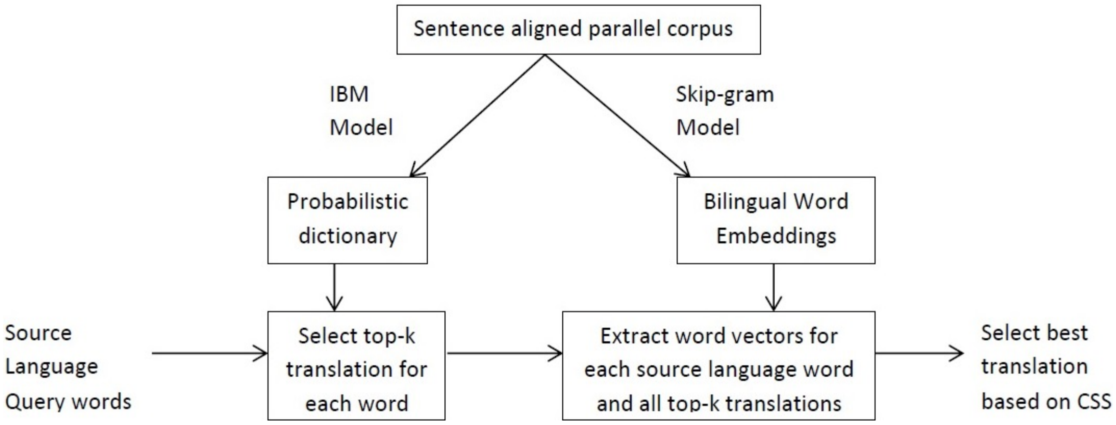

Sentence aligned parallel corpus
IBM Model
Skip-gram Model
Probabilistic dictionary
Bilingual Word Embeddings
Source Language Query words
Select top-k translation for each word
Extract word vectors for each source language word and all top-k translations
Select best translation based on CSS

Figure 4.1: Hybrid model based on BWESG and IBM model

Hybrid Model

Hybrid model is a combined model of IBM and BWESG model as shown in Figure 4.1. A probabilistic dictionary is constructed by using an IBM model. Top-k translations $(t_1, t_2, ..., t_k)$ are extracted from the probabilistic dictionary against a source language word s_i. Word vectors are extracted from the skip-gram learned bilingual word vectors, i.e., $\{s_{i,1}, s_{i,2}, ..., s_{i,dim}\}$ for each source language word s_i and $\{t_{j,1}, t_{j,2}, ..., t_{j,dim}\}$ for each target language word t_j. The maximum Cosine Similarity Score (CSS) and minimum Euclidean Distance Score (EDS) are used to select the best translation from the top-k translations. The CSS and EDS are calculated by using the equation 4.3 and 4.4.

$$CSS = \frac{\sum_{k=1}^{dim} s_{i,k}.t_{j,k}}{\sqrt{\sum_{k=1}^{dim} s_{i,k}^2}\sqrt{\sum_{k=1}^{dim} t_{j,k}^2}} \tag{4.3}$$

$$EDS = \sqrt{\sum_{k=1}^{dim}(s_{i,k} - t_{j,k})^2} \tag{4.4}$$

The proposed hybrid model is evaluated with FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5. A Hindi-English parallel corpus HindiEnCorp[1] is used to create word vectors and PD. The proposed hybrid model is used to translate the source language query string into the target language. VSM is used for indexing and retrieval of target documents. The recall and Mean Average Precision (MAP) which are discussed in Section 3.5 are used to evaluate the proposed hybrid model. Experiment results of the proposed hybrid model and the PD are

[1] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0

Table 4.1: Experiment results for Hybrid model

Experiment	**FIRE 2010**		**FIRE 2011**	
	Recall	**MAP**	**Recall**	**MAP**
Hybrid Model + Top-5 + CSS	0.5727	0.1223	0.4643	0.1053
Hybrid Model + Top-10 + CSS	0.4747	0.1059	0.4060	0.0758
Hybrid Model + Top-5 + EDS	0.6034	0.1382	0.5418	0.1085
Hybrid Model + Top-10 + EDS	0.5329	0.1144	0.4889	0.0866
PD (baseline)	0.7488	0.2267	0.6791	0.1672

represented in Table 4.1. Top-5 and top-10 translations are extracted from the PD against a source language word. CSS and EDS are used to select the best translation.

BWESG model gives a poor MAP due to different sentence structure and the different number of vocabulary & stop-words across the language, hence, a hybrid model is proposed which incorporates an IBM model. Here, IBM model limits the translation up to either top-5 or top-10. CSS and EDS are calculated between source language word vector and top-k target language word vectors. The target language word which has either the maximum CSS or the minimum EDS is selected as the best translation. Hybrid model with top-5 translations gives a better MAP than the top-10 due to the less number of target language translations. EDS gives a better MAP than the CSS. It is concluded from the experimental results which are reported in Table 4.1 that the hybrid model with top-5 translations and EDS gives better MAP.

Word vectors learning for recognizing similar contextual words perform well for the single language but across the languages, it recognizes non-contextual words. The hybrid model limits the target language word vectors by selecting only top-k translations. So, the hybrid model with top-5 and EDS performs better than the BWESG model. Although, the experimental result analysis shows that the PD based approach performs better than the hybrid model but our objective is not to show the superiority of the PD based approach over the word vectors. Our objective is to investigate the cross-lingual word vectors learning with respect to the Hindi-English CLIR.

4.2 Named Entity Recognition based

Machine translation is in the evolving stage for the Indian languages, where either a translation or transliteration technique is used to translate a word or phrase. Identifying whether a word needs a translation or transliteration technique, is still a challenge. Named entity recognition may be helpful to disambiguate the words in favor of either translation or transliteration due to the property of similar pronunciation of Named Entity (NE) terms across the languages. The term frequency model is used for translation along with the NER based disambiguation model.

NER is a task of recognizing whether a term is a NE, i.e., person name, location, and organization or not. A lot of research is being done for foreign and Indian languages, but the challenges are still not resolved. An issue of improper term translation or transliteration, i.e., whether a term needs either a translation technique or transliteration technique, is addressed in this section. Most of the previous machine translation systems suffer from poor quality translations due to the improper term translation or transliteration issue.

The proposed NER based translation disambiguation model is evaluated for CLIR where a parallel corpus-based term frequency model (V. K. Sharma & Mittal, 2016b) is used for translation. The proposed NER based disambiguation model incorporates two phases which are as follows:

- The named entity annotated data is collected and prepared from the different sources, and gazetteer list. The NER system is trained with some linguistic patterns.
- Is NER based disambiguation model pertinent to resolve the improper term translation or transliteration issue?

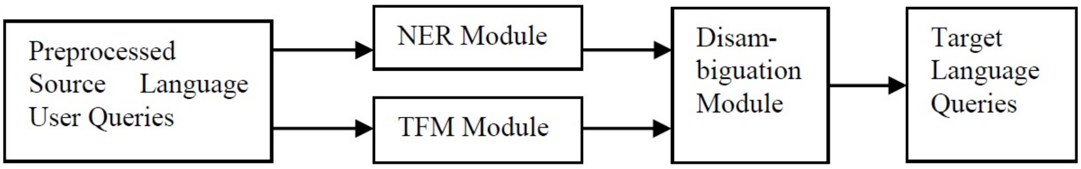
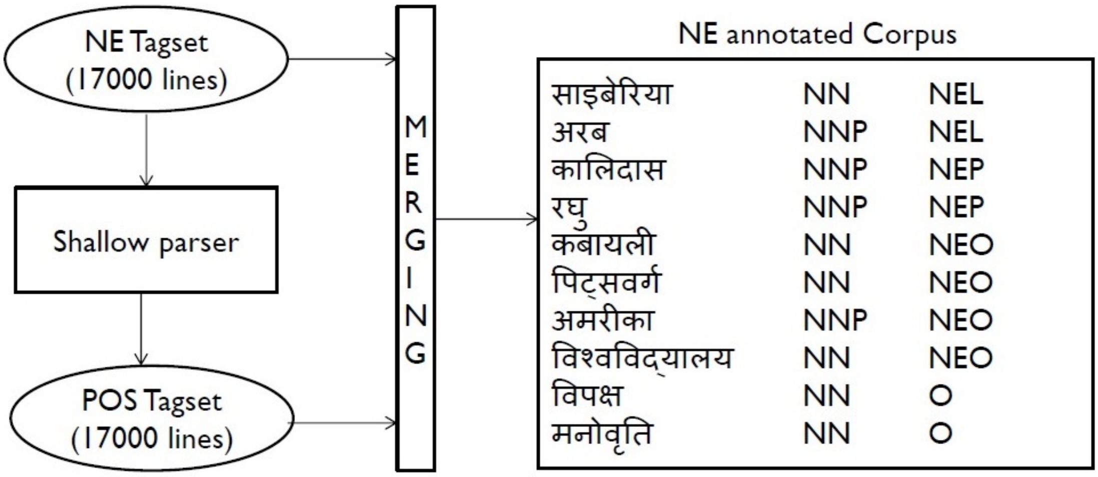


Figure 4.2: Named entity recognition based disambiguation approach

An experiment is also done to analyze the impact of multiple term translation which is shown in Section 4.3. User queries include three types of terms, i.e., stop words, the terms which need translation, and the terms which need transliteration. The proposed approach is represented in Figure 4.2, where stop words are eliminated in the preprocessing step and the remaining terms are passed through the NER and TFM module.

Named Entity Recognition

The conditional random field is better than the other machine learning algorithms (Krishnarao et al., 2009; Prasad & Fousiya, 2015), hence, the Stanford NER[1] (SNER) which uses conditional random field learning, is used to train the NER system. A lot of NE annotated data which is not available for the Hindi language is required to train the SNER, therefore, the NE annotated dataset and gazetteer lists need to be prepared for the training of SNER.



साइबेरिया	NN	NEL
अरब	NNP	NEL
कालिदास	NNP	NEP
रघु	NNP	NEP
कबायली	NN	NEO
पिट्सवर्ग	NN	NEO
अमरीका	NNP	NEO
विश्वविद्यालय	NN	NEO
विपक्ष	NN	O
मनोवृति	NN	O

Figure 4.3: Named entity annotated data preparation

An available NE tagged dataset[2] comprise of approx 17000 sentences is parsed by Shallow parser[3] developed by IIIT Hyderabad to obtain the part-of-speech tags. Further, NE tags and part-of-speech tags are combined to generate a NE annotated dataset for training the SNER system as shown in Figure 4.3. Any standard gazetteer list for NER is not available, hence, a gazetteer list is prepared by collecting various Indian named entity terms from the web. The named entity terms and their sources are listed in Table 4.2. A testing word is categorized into four categories, i.e., Person Name (NEP), Location (NEL), Organization (NEO) and non-NE terms (NOP). Various stop-word phrases are analyzed, among them, six phrases are recognized as patterns which are like *Word1 Stop-word Word2*. If any word in the recognized pattern is a NE then the other word is also tagged

[1] http://nlp.stanford.edu/software/CRF-NER.shtml
[2] http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5
[3] http://ltrc.iiit.ac.in/analyzer/hindi/run.cgi

with the same NE tag. The recognized patterns are represented in Table 4.3. The query word translation is computed by using the term frequency model which is discussed in Section 3.2.

Table 4.2: Web links for the named entities

Named Entity	Sources
List of First Names	http://www.studentsoftheworld.info/penpals/stats.php3?Pays=IND http://babynames.extraprepare.com/ http://www.indiaexpress.com/specials/babynames/ http://www.babycenter.in/a25012573/most-popular-indian-girl-names http://www.babynames.org.uk/indian-boy-baby-names.htm http://www.newlyborn.org/most-popular-names/hindu-names.htm
List of Middle Names	http://www.cs.colostate.edu/ malaiya/middlenames http://www.indianchild.com/indian_middle_names.htm http://www.top-100-baby-names-search.com/girl-middle-names.html
List of Last Names	https://en.wikipedia.org/wiki/Category:Indian_family_names http://surnames.behindthename.com/names/usage/indian https://en.wiktionary.org/wiki/Appendix:Indian_surnames http://www.lowchensaustralia.com/names/popular-indian-names.htm http://blogs.transparent.com/hindi/common-surnames-in-india/ http://www.indianhindunames.com/indian-surnames-origin-meaning.htm http://indiachildnames.com/surname/
List of Locations in India	https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_India https://en.wikipedia.org/wiki/List_of_state_and_union_territory
List of Suffixes	http://www.irfca.org/docs/place-names.html (Locations)
List of Organization	https://en.wikipedia.org/wiki/Category:Organisations_based_in_India https://en.wikipedia.org/wiki/List_of_Indian_government_agencies

Table 4.3: Stop-word Phrases

S.no.	Stop-Word	Example Phrases
1	और	हरियाणा और दिल्ली, पांडिचेरी और मुम्बई
2	एवं	महाभारत एवं रामायण, हिंदी एवं उर्दू
3	तथा	विष्णुगुप्त तथा कौटिल्य, गंगा तथा ब्रह्मपुत्र
4	या	राजग या एनडीए, देवी या दुर्गा
5	व	जम्मू कश्मीर व उत्तराखंड, नारायणपुर व बीजापुर
6	अथवा	महाराष्ट्र अथवा उड़ीसा, सिलीगुड़ी अथवा कोलकाता

Disambiguation

Disambiguation module receives NE tag from NER module and top-n translations from TFM module. A NE word's transliteration is also present in top-n translations if the word's transliteration is available in a parallel corpus but that word's transliteration has a poor translation cosine similarity score. Therefore, a disambiguation algorithm is proposed in Algorithm 4 to select the proper translation or transliteration.

Algorithm 4 NER based disambiguation algorithm

Input: Words NER tags and Top-n translations

Output: Proper translation or transliteration

```
if (word's NER tag $\neq$ NOP) then
    R_word = word's romanized form
    R_trans = find LCSR between R_word and all Top-n translations and select
    the translation with maximum LCSR
    if (LCSR (R_trans) $\geq$ 0.60) then
        R_trans is a proper transliteration
    else
        select the maximum CSS scorer translation
    end
else
    select the maximum CSS scorer translation
end
```

The proposed algorithm receives a word's NER tag and top-n translations, if the word's NE tag is not *NOP* then the word is converted into its romanized form that is R_{word}. Further, LCSR between R_{word} and all top-n translations are computed and a translation with maximum LCSR is selected that is R_{trans}. If R_{trans} has the LCSR greater than 0.60 then the R_{trans} becomes the proper transliteration else a maximum CSS scorer translation is selected. In another case, when a word in not a NE then the maximum CSS scorer translation is selected. The LCSR between two strings is computed by using the Equation 3.1.

The proposed NER based disambiguation approach is investigated with FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5. A preprocessed source language query is passed through the NER and TFM module separately and their outcome, i.e., a NE tagged query and top-5 translations for each query word are passed through the disambiguation module. The resultant outcome of the proposed approach is the target language query. VSM is used to retrieve target documents. NER based disambiguation approach is evaluated by using recall and MAP, discussed in Section 3.5. The experiment results are shown in Table 4.4.

Table 4.4: Experiment results of the named entity recognition based disambiguation approach

Approach	FIRE 2010		FIRE 2011	
	Recall	MAP	Recall	MAP
PD (baseline)	0.7488	0.2267	0.6791	0.1672
TFM	0.7519	0.2637	0.6754	0.1623
TIA	0.8315	0.2818	0.7257	0.1911
TIA + NER	0.7993	0.2785	0.6967	0.1776

The inclusion of NER based disambiguation decreases the performance of CLIR because at many occurrences the translation versions are more popular than the transliteration. Therefore, the TIA with the proposed NER based disambiguation achieves less MAP than the TIA alone. Although the TIA with NER based disambiguation achieves better evaluation measures than the baseline PD but it is more adaptive to compare it with TIA alone. A considerable difference between the popularity of the term's translation and transliteration is represented in Table 4.5. NER alone is not sufficient to select the proper translation or transliteration because term's popularity decides whether it needs either translation or transliteration.

Table 4.5: Effectiveness of named entity recognition based disambiguation approach

Terms	NE Tag	Translite ration	Trans lation	Is NER effec tive? Y/N
भारत	NEP	Bharat	India	N
प्रतिभा	NEP	Pratibha	Talent	Y
नगर	NEL	Nagar	City	N
भारतीय सेना	NEO	Bhartiya Nausena	Indian Navy	N

NER based disambiguation is investigated to extricate an improper translation or transliteration issue. Indian languages do not have enough NE annotated data and Gazetteer list. The NE annotated data is prepared with the help of IIIT Hyderabad's NE corpus and shallow parser. Different web resources are used to prepare the gazetteer lists. SNER is trained on the prepared NE annotated data and gazetteer list. The proposed stop-words patterns are used to improve the NER system. The TFM module returns the top-n translations. Disambiguation module selects the proper translation and transliteration on the basis of the outcome of

NER and TFM module. The proposed NER based disambiguation achieves less MAP than the TFM.

4.3 Multiple Translation Term Selection based

A separate experiment is performed to analyze the impact of multiple translations term selection which is shown in Table 4.6. This experiment is performed with FIRE 2010 and 2011 datasets where Top-n translations are selected from the probabilistic dictionary for each query term on the basis of maximum probabilistic score.

Table 4.6: Effect of Top-n translation selection

Dataset			**MAP**		
	Top-1	**Top-2**	**Top-3**	**Top-4**	**Top-5**
FIRE 2010	0.2637	0.2604	0.2583	0.2317	0.2238
FIRE 2011	0.1623	0.1785	0.1460	0.1448	0.1403

FIRE 2010 queries have an average length about six words which is sufficient for searching, hence, the inclusion of Top-2 or more terms are always degraded the CLIR performance. FIRE 2011 queries have an average length about three words which is not sufficient. In this case, Top-2 translation term selection achieves better performance but Top-3 achieves low MAP because FIRE 2011 queries with Top-2 terms get the sufficient length, more than that, always degraded the CLIR performance.

4.4 Maximum of Average of Words Average Probability and Association Score based

In the probabilistic dictionary, a word has multiple translations associated with the probabilistic scores. The best translation is selected in favor of maximum probabilistic score (Zhou et al., 2012). Since the manual dictionary does not have any probabilistic score, hence, the word co-occurrence and PMI are used to select the best translation (Monz & Dorr, 2005). The NER is used to disambiguate the named entity terms from the dictionary terms (V. K. Sharma & Mittal, 2017). Rule-based NER needs a lot of grammatical knowledge & experience about a particular language while machine learning techniques need a lot of named entity annotated data (Das, Ganguly, & Garain, 2017; Karimi et al., 2011; Prasad & Fousiya, 2015). Statistical term similarity disambiguation computes an optimal translation but it takes a high computation cost which increases exponentially with the sentence length (Adriani, 2000). Word embedding represents the words in the form of vectors which are prepared such that the semantically similar words have approximately similar vectors, based on that fact, the bi-lingual word embeddings are trained and are used for the translation purpose (Klementiev et al., 2012; Zou, Socher, Cer, & Manning, 2013). Continuous bag of words and skip-gram models are trained to learn the high-quality vectors which capture precise syntactic and semantic words relationship (Ganguly et al., 2015; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Vulić & Moens, 2015). A log-bilinear global matrix factorization method with local context window is also used to produce the word embeddings (Pennington et al., 2014). Since WE represents a word in the form of the vector which is enriched with the syntactical and semantic features, hence, the WE may impart a significant role in word translation disambiguation. An example in Figure 4.4 shows the extracted translations from the manual dictionary and probabilistic dictionary.

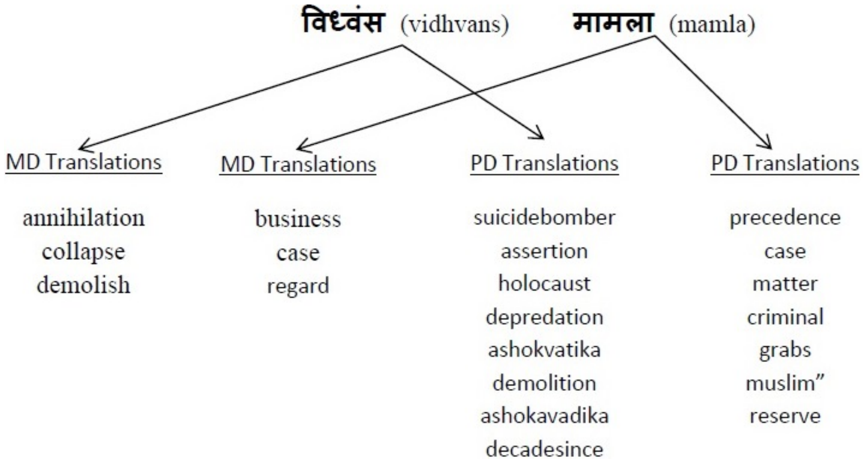


Figure 4.4: Extracted translations from the MD and PD

The MD does not return an appropriate translation pair in all of the nine combinations. In PD, the maximum probabilistic score based translation is always not a good idea because a more relevant translation may have a low probabilistic score. For example, the translation "demolition" has lower probabilistic score than the "holocaust" and the translation "holocaust case" is selected on the basis of maximum probabilistic score while the more relevant translation is "demolition case" which is selected on the basis of maximum association score between the words. In another example, the query समुदाय (samuday) संघर्ष (sangharsh) is translated as "mob riot" on the basis of maximum association score while the more relevant translation is “community conflict” which is selected on the basis of maximum probabilistic score. Therefore, it is difficult to say whether the more relevant translation will be selected based on the maximum probabilistic score or on the maximum association score.

CLIR follows either maximum probabilistic score or word association score based translation disambiguation but both of these techniques alone are not sufficient, therefore, a hybrid translation disambiguation approach is proposed which disambiguates based on both of the words association and average probabilistic score. This disambiguation approach is implemented at either word level and query level that means the translations are computed based on either two neighboring words or all query words respectively, which are discussed in Figure 4.5 and 4.6.

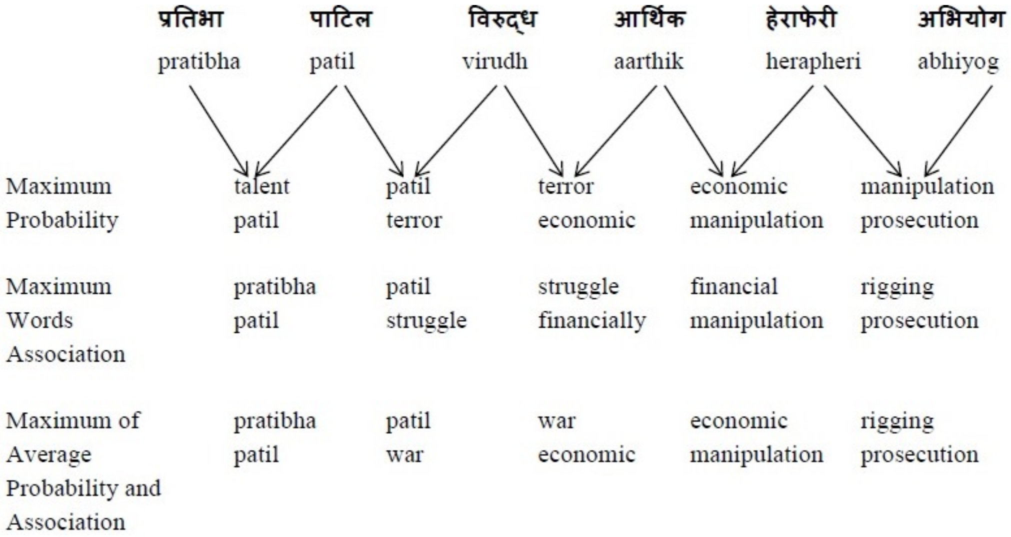


Figure 4.5: Translation disambiguation at word level by using maximum probability, maximum words association, and maximum of average of the average probability and association score

In Figure 4.5, maximum probabilistic score based technique computes irrelevant translation “talent” and the maximum words association score based technique computes less relevant translation “financial” while the maximum of average of the average probabilistic score and association score based technique computes the relevant translations which are “pratibha” and “economic” respectively.

In Figure 4.6, maximum words association score based technique computes less relevant translation “financially chicanery charge” while the maximum of average of the average probabilistic score and association score based technique computes more relevant translation “financially manipulation prosecution”.

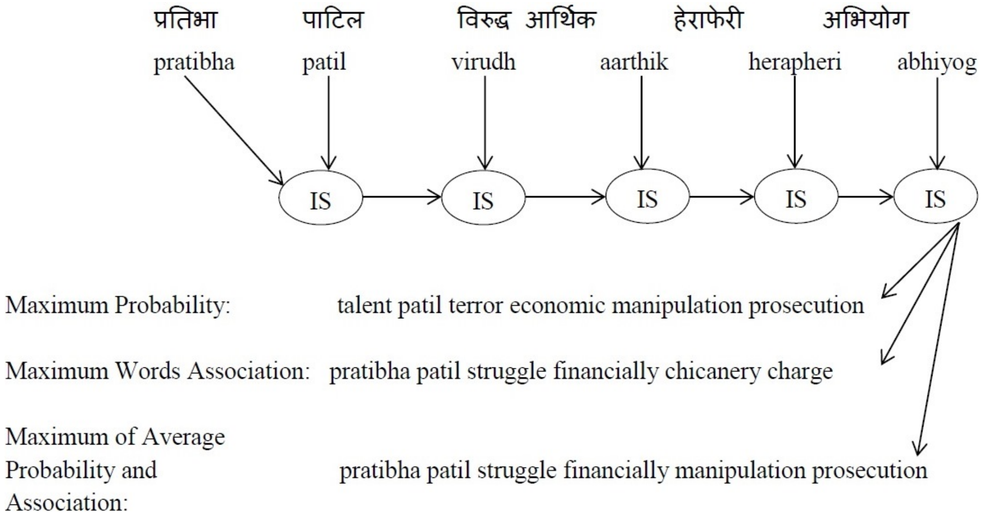

प्रतिभा
pratibha
पाटिल
patil
विरुद्ध
virudh
आर्थिक
aarthik
हेराफेरी
herapheri
अभियोग
abhiyog
IS
IS
IS
IS
IS
Maximum Probability:
talent patil terror economic manipulation prosecution
Maximum Words Association:
pratibha patil struggle financially chicanery charge
Maximum of Average Probability and Association:
pratibha patil struggle financially manipulation prosecution

Figure 4.6: Translation disambiguation at query level by using maximum probability, maximum words association, and maximum of average of the average probability and association score. IS: intermediate states

Words translations probabilistic scores are stored in the probabilistic dictionary and association scores are computed by using either PMI or WE. PMI method is applied to a large target language corpus in order to produce a bi-gram lexicon with the PMI score (Turney, 2004). WE with vectors size 200 are prepared by training a continuous bag of word based recurrent neural network model with a context window size 10 on a large target language corpus. Further, words association scores are computed by calculating the cosine similarity scores between the vectors. PMI based association score is used in the examples which are shown in Figure 4.5 and 4.6. The example of WE based association score at word level is discussed in Table 4.7, where the maximum of average of average probability and association score returns the translation pair "demolition case" which is more relevant than the other two.

Table 4.7: Translation disambiguation at word level where association scores are calculated by using word embedding

Translation Pair	**Average**	**Words**	**Average of Average**
Translation Pair	**Probability**	**Association**	**Probability and**
		Score	**Association Score**
Destroyed Matter	0.3646	0.6741	0.5058
Demolition Case	0.4563	0.5820	0.5191
Destruction Case	0.4783	0.3851	0.4317

Table 4.8: Experimental results for the maximum of words average probability and association score based disambiguation

Resource	String Matching Technique	Translation mis-mapped	Disambiguation Technique	Disambiguation Level	Word Association	FIRE 2010				FIRE 2011			
						Recall	P@5	P@10	MAP	Recall	P@5	P@10	MAP
SMT (B)						0.8284	0.2840	0.2300	0.2832	0.7084	0.2440	0.2320	0.1885
PD (B)	LCSR		max-prob			0.8192	0.2520	0.2320	0.2645	0.6787	0.2120	0.2320	0.1667
PD (B)	LCSR	SMVS	max-prob			0.8545	0.3000	0.2580	0.3032	0.7040	0.2240	0.2540	0.1865
PD	LCSR	SMVS	max-assoc	Word	PMI	0.8009	0.2760	0.2360	0.2598	0.6258	0.2040	0.2040	0.1517
PD	LCSR	SMVS	max-assoc	Query	PMI	0.7947	0.2320	0.2140	0.2282	0.6262	0.1760	0.1980	0.1480
PD	LCSR	SMVS	avg-prob & max-assoc	Word	PMI	0.8468	0.3200	0.2540	0.2860	0.6816	0.2480	0.2500	0.1772
PD	LCSR	SMVS	avg-prob & max-assoc	Query	PMI	0.7075	0.2120	0.1800	0.2133	0.5056	0.1480	0.1500	0.0964
PD	LCSR	SMVS	max-assoc	Word	WE	0.7702	0.2040	0.1760	0.2098	0.5407	0.1560	0.1480	0.1118
PD	LCSR	SMVS	max-assoc	Query	WE	0.7182	0.1800	0.1680	0.2030	0.5563	0.1520	0.1600	0.1313
PD	LCSR	SMVS	avg-prob & max-assoc	Word	WE	0.8744	0.3060	0.2600	0.2932	0.6707	0.2280	0.2540	0.1617
PD	LCSR	SMVS	avg-prob & max-assoc	Query	WE	0.6646	0.1600	0.1540	0.1778	0.5443	0.1600	0.1580	0.1229

B: Baseline; max-prob: maximum probability; max-assoc: maximum association;
avg-prob & max-assoc: maximum of average of average probability and association score

FIRE 2010 and 2011 ad-hoc datasets, statistics are represented in Table 3.5. A Hindi-English parallel corpus HindiEnCorp is exploited to produce the PD. A large Hindi language raw corpus (approx 10GB) which is a combination of Hindi Wikipedia articles[2] and Bojar Hindi MonoCorp[3], is used to generate the Hindi language WE. A large English language raw corpus (approx 12GB) which is collected from the English Wikipedia Articles[4], is used to produce a PMI based bi-gram lexicon and WE.

The Term Frequency - Inverse Document Frequency (TF-IDF) and cosine similarity techniques are used for target documents indexing and retrieval. The proposed approach is evaluated by using recall, MAP, P@5, and P@10, as discussed in Section 3.5. Experimental setups are prepared by using SMT and PD. In SMT, a HindiEnCorp parallel corpus is used to train it (Koehn, 2009). In PD, the same HindiEnCorp is used to produce a PD where a Hindi language word has multiple translations which are associated with the probabilistic score. Generally, translations are chosen in favor of the maximum probabilistic score.

A probabilistic dictionary with exact mapping and LCSR string matching technique is used to map the query words in the dictionary but it is not sufficient due to the translation mis-mapped issue. Additionally, the semantic morphological variant selection algorithm is also used which is proposed in Section 3.5.

Maximum word association score is used to select the best translation from the multiple translations where association scores are computed by using either PMI or WE. The proposed approach selects the best translation based on both of the word average probability and association scores at both of the word and query level. SMT and PD_LCSR are considered as the baselines to evaluate the proposed SMVS approach. SMT, PD_LCSR, and PD_LCSR_SMVS are considered as the baselines to evaluate the proposed disambiguation approach.

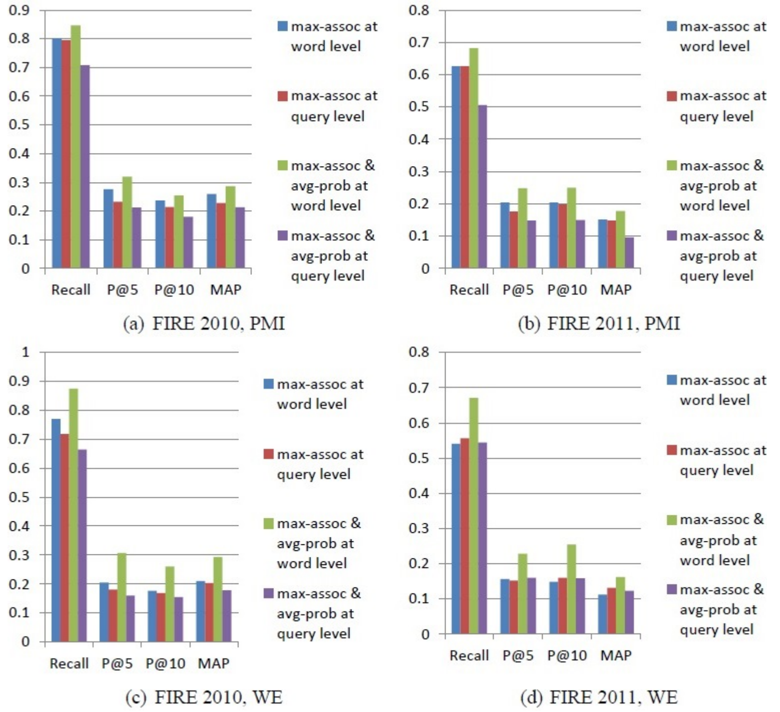
The proposed maximum of average of words average probability and association score based disambiguation approach is evaluated at the word level and query level for FIRE 2010 and 2011 datasets. Word association score is computed by using either PMI or WE. Experiment results are represented in Table 4.8. A comparison among these different disambiguation techniques is represented in Figure 4.7, where maximum of average of average probability and association score at word level based disambiguation technique enhances the CLIR performance for both of

[2] https://dumps.wikimedia.org/backup-index.html

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-6260-A

[4] https://dumps.wikimedia.org/backup-index.html

the FIRE 2010 and 2011 datasets by using PMI based association score as shown in Figure 4.7(a) & (b), and by using WE based association score as shown in Figure 4.7(c) & (d).



(a) FIRE 2010, PMI

(b) FIRE 2011, PMI

(c) FIRE 2010, WE

(d) FIRE 2011, WE

Figure 4.7: Impact of maximum of average of words average probability and association score based disambiguation approach, evaluated by Recall, P@5, P@10, and MAP for (a) FIRE 2010, PMI, (b) FIRE 2011, PMI, (c) FIRE 2010, WE, and (d) FIRE 2011, WE

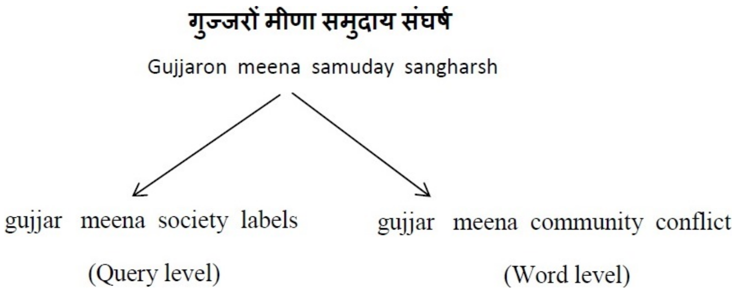
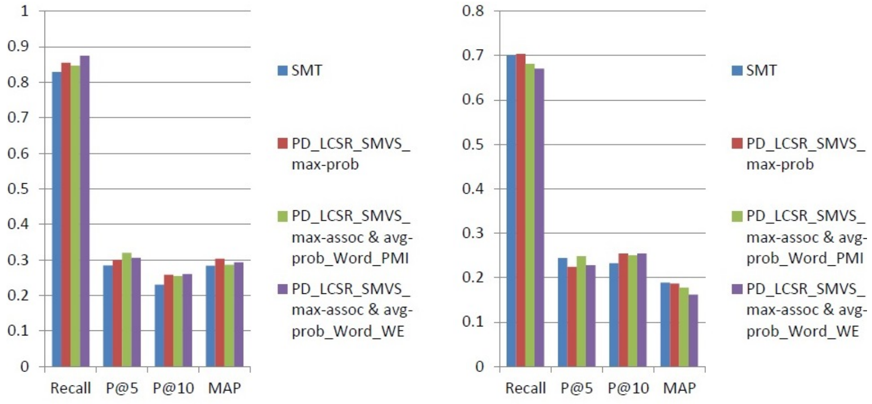

गुज्जरों मीणा समुदाय संघर्ष
Gujjaron meena samuday sangharsh
gujjar meena society labels
(Query level)
gujjar meena community conflict
(Word level)

Figure 4.8: An example showing the difference between the word level and query level maximum of average of average probability and association score bassed disambiguation approach

Maximum of average of average probability and association score at query level achieves low evaluation measures because it selects highly associated query words translations where the semantic relatedness between the first and last query word translation may or may not be strong. As shown in Figure 4.8, where "gujjar meena society labels" is a highly associated translation but its semantic relationship is not so strong, as the translation "gujjar meena community conflict" has.



(a) FIRE 2010

(b) FIRE 2011

Figure 4.9: Impact of the proposed maximum of average of words average probability and association score based disambiguation approach at word level for (a) FIRE 2010 (b)FIRE 2011

Maximum of average of average probability and association score based disambiguation at word level with PMI and WE is compared to the baseline SMT and the proposed PD_LCSR_SMVS_ max-prob which is shown in Figure 4.9. WE based disambiguation increases the recall for FIRE 2010, and P@5, P@10 for both of the FIRE 2010 and 2011. PMI based disambiguation increases the P@5 for both of the FIRE 2010 and 2011, and P@10 remains approximately similar to PD_LCSR_SMVS_max-prob. PMI based disambiguation is also compared to WE as shown in Figure 4.9 where P@10 remains same and P@5 increases with PMI for both of the FIRE 2010 and 2011 dataset. In addition to this, recall is increased and MAP remains the same for the FIRE 2010 in case of WE whereas recall and MAP are increased for FIRE 2011 in case of PMI. Maximum of average of average probability and association score at word level achieves approximately similar performance with both of the PMI and WE, and better performance compared to the baselines SMT and PD_LCSR_SMVS_max-prob.

Word translation disambiguation techniques which are based on either the maximum probability or the maximum word association score are not appropriate, therefore, a maximum of average of words average probability and association score based disambiguation approach is proposed where the association scores are calculated by using either PMI or WE. The proposed disambiguation approach is applied at the word level and query level, among them, the word level disambiguation approach performs better than the query level. The proposed disambiguation approach at the word level with both of the PMI and WE based association score achieves almost equal evaluation measures and better evaluation measures than the baselines SMT and the proposed PD_LCSR_SMVS_max-prob.

4.5 Summary

A source language word is associated with multiple translations in the dictionary which leads to word translation disambiguation issue. In the probabilistic dictionary, a maximum probability scorer translation is chosen as the best translation. Word embedding provides the context-based word vectors, hence, the bilingual word embedding may disambiguate the words translation. A hybrid model which incorporates both of the BWESG (based on the bilingual WE) and IBM model, is proposed to disambiguate the translation. The hybrid model with top-5 translations extracted from the PD and Euclidean distance similarity measure achieves good evaluation measures but not more than the probabilistic dictionary due to the different sentence structure and the different number of vocabulary & stop-words in source and target language. A query word may be a dictionary word or named entity. It becomes difficult to recognize whether the word is a dictionary word or named entity. The named entity recognition based disambiguation approach is proposed which recognize the word's named entity tag and translate/transliterate the word accordingly. Stanford NER tool is used to train the NER system. Named entity recognition based disambiguation approach with translation induction algorithm degrades the CLIR performance in comparison to translation induction algorithm because word's translation/transliteration depends on its popularity, not on its named entity tag. The CLIR performance may be enhanced after the inclusion of multiple translation terms or synonyms of the words translation. The selection of multiple translation terms depends on the query length.

Maximum probabilistic score or maximum association score like word co-occurrence, PMI, alone are not sufficient for disambiguation. A maximum of average of words average probability and association score based disambiguation approach which incorporates both of the probabilistic score and association score is proposed to disambiguate the queries more effectively. The disambiguation is performed at both of the word level or query level where the word level disambiguation approach performs better than the query level. The proposed disambiguation approach at the word level with PMI and CBOW based WE achieves almost equal evaluation measures, and better evaluation measures than the baselines SMT and the proposed PD_LCSR_SMVS_max-prob.

Chapter 5

WEB-RESOURCES BASED TRANSLATION

The number of multilingual content and its users are increasing on the web. Internet users perceive a multilingual web but they are not familiar with it, because a person likes to communicate in his regional language. In CLIR, a translation technique is used to translate the user queries into the target language. Conventional translation techniques are based on either a manual dictionary or a parallel corpus, and the trending Statistical Machine Translation[1] (SMT) and Neural Machine Translation[2] (NMT) techniques are trained on a parallel corpus. The SMT provides a static translation due to the limited vocabularies in the available parallel corpus, hence, SMT may not provide the translations for missing or unseen words. The web provides a dynamic interface where multiple users are updating information at the same time, hence, the web may provide the translations for missing or unseen words. Therefore, the web is effectively used for the technically developed languages like English, German, Spanish, Russian, Chinese, etc. In this chapter, different web resource based translation techniques are proposed and applied to the Hindi-English CLIR. The proposed approaches are compared with the SMT where the Wikipedia based approach achieves approximately similar mean average precision.

[1] http://www.statmt.org/moses/

[2] https://github.com/tensorflow/nmt

5.1 Wikipedia based

Conventional approaches depend on the quality and coverage of a resource. Manual dictionary and parallel corpus are static in nature which makes it impossible to translate the unseen or missing words of the queries. In the literature, the SMT and NMT (attention based recurrent neural network) are trained on the extended parallel corpus which is developed by CFILT lab, IIT Bombay India, where SMT achieves better BLEU score in comparison to NMT for Hindi to English translation (Kunchukuttan et al., 2018), hence, SMT is used as the baseline in our experiments. SMT is trained on parallel corpus with good quality and appropriate size but that corpus will always be static in nature while the web is dynamic in nature.

Wikipedia is an open-access online knowledge base which is editable by the users across the world. Wikipedia is available in 299 languages, out of them, 19 languages only have more than 5,00,000 articles[3]. Wikipedia is useful for linguistic research in resource-scarce languages like Hindi due to its content and structure. A Wikipedia article has a unique title and inter-wiki link attribute which provides same-titled articles in the different languages, due to that, Wikipedia is capable to do the translation. The user queries are translated by using the Wikipedia inter-wiki link attribute. Wikipedia hyperlinks are used for translation disambiguation (Schönhofen et al., 2007). Inter-wiki links are also utilized to construct a parallel or comparable corpus, other Wikipedia attributes like title, abstract, category, and infobox are used to build a Wikipedia mined bilingual dictionary. The queries are segmented and searched in the Wikipedia mined bilingual dictionary for English-French translation (Gaillard et al., 2010). Context information is gathered from the title, category, redirect title, in-links, out-links, and subsection attributes (Bharadwaj & Varma, 2011) which can be used for disambiguation purpose. The size of Wikipedia mined dictionary is increased by using the anchor text, redirect page, inter-wiki link, and forward/backward link attributes. Incorrect term translation pairs are filtered out based on the backward link attribute (Erdmann et al., 2009). Named entities are recognized with the help of Wikipedia where similar English Wikipedia articles are grouped and tagged by Stanford named entity recognizer, further, these English articles named entity tags are mapped to other language terms by using inter-wiki link attribute (Bhagavatula et al., 2012).

The proposed approach uses the title and inter-wiki link attributes for translation purpose. Many issues are identified during experimentation. Stop words may also

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias

mix the noise while searching the online Wikipedia as a user searches for a bi-gram "और मीणा" then Wikipedia API[4] returns a page with the title ‘‘किरोड़ी लाल मीणा", hence, stop words are needed to be eliminated before query processing. N-gram term variation occurs in many Wikipedia articles like a N-gram term "राम विलास पासवान" has the Wikipedia title "रामविलास पासवान" .

The proposed Wikipedia based approach is represented in Figure 5.1, where a query string is tokenized, stop-words are eliminated and multi-word terms are created using N-gram technique. The N-gram terms up to tetra-gram are translated using the title and inter-wiki link attributes of Wikipedia. Finally, the vector space model is used to retrieve the target documents.

Each N-gram term is searched on online Wikipedia knowledge base by using Wikipedia API which returns the titles of matching Wikipedia articles in the source language. If N-gram term matches with any title which has target language inter-wiki link then the title associated with target language inter-wiki linked article is extracted, else, merged N-gram is formulated by eliminating white spaces from the N-gram term. Online Wikipedia knowledge base is searched for the merged N-gram and the titles of all source language Wikipedia article are extracted. If merged N-gram matches to any title which has target language inter-wiki link then the title associated with the inter-wiki linked target language article is extracted, else, the titles of all Wikipedia article in the source language are extracted using N-gram term. Further, all the source language titles whose matching score with N-gram term is more than 80% are selected, if no title is selected then a maximally matched title is selected. If selected title is not more than one then the title of the inter-wiki linked articles becomes the desired translation, else all target language titles from all the inter-wiki linked articles are extracted and a maximum frequency words from all target language titles becomes the desired translation. This procedure is followed for tetra-gram and if we don't obtain any translation then the same procedure is followed for tri-gram, bi-gram, and unigram. A little modification needs to be followed in the case of unigram if the maximal matched title is selected then there is a possibility that the length of the selected title is more than one and in that case, the translation is extracted based on unigram position in source language title.

Terrier[5] search engine which supports many retrieval models like VSM, BM25, etc., is utilized for indexing, retrieval, and evaluation. VSM and BM25 are used

[4] https://pypi.python.org/pypi/wikipedia/
[5] http://terrier.org/

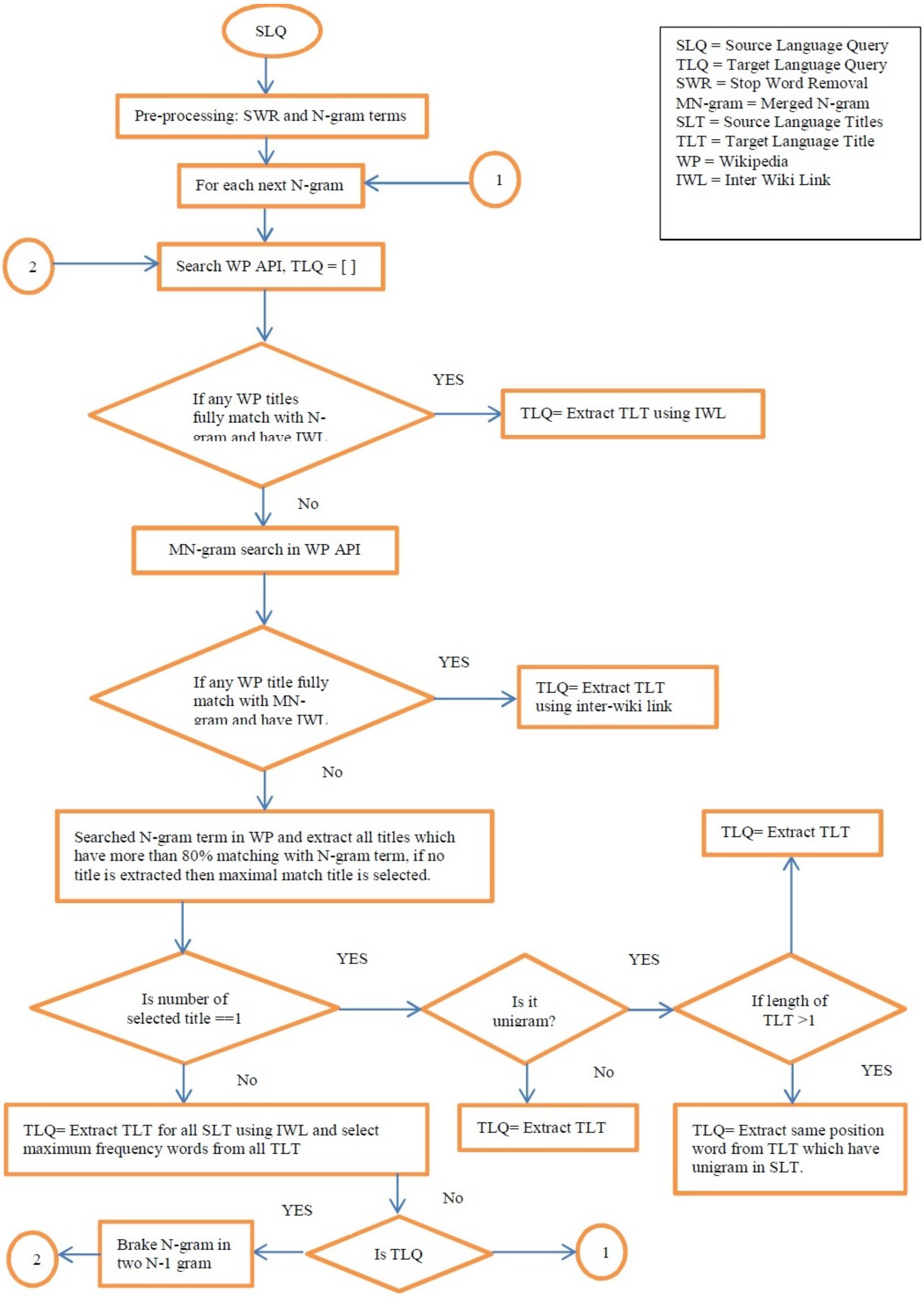

SLQ
Pre-processing: SWR and N-gram terms
For each next N-gram
1
2
Search WP API, TLQ = [ ]
SLQ = Source Language Query
TLQ = Target Language Query
SWR = Stop Word Removal
MN-gram = Merged N-gram
SLT = Source Language Titles
TLT = Target Language Title
WP = Wikipedia
IWL = Inter Wiki Link
If any WP titles fully match with N-gram and have IWL
YES
TLQ= Extract TLT using IWL
No
MN-gram search in WP API
If any WP title fully match with MN-gram and have IWL
YES
TLQ= Extract TLT using inter-wiki link
No
Searched N-gram term in WP and extract all titles which have more than 80% matching with N-gram term, if no title is extracted then maximal match title is selected.
TLQ= Extract TLT
Is number of selected title ==1
YES
Is it unigram?
YES
If length of TLT >1
No
No
YES
TLQ= Extract TLT for all SLT using IWL and select maximum frequency words from all TLT
TLQ= Extract TLT
TLQ= Extract same position word from TLT which have unigram in SLT.
YES
No
2
Brake N-gram in two N-1 gram
Is TLQ
1

Figure 5.1: Wikipedia API based query translation approach

in our experiments. The proposed approach is evaluated with FIRE[6] 2010 and 2011 datasets, statistics are represented in Table 3.5. A query includes $< title >$, $< desc >$ and $< narr >$ tag. The experiments are performed in two different ways with both of the datasets, one is with only $< title >$ tag and another is with both $< title >$ and $< desc >$ tag. Terrier search engine is utilized for indexing, retrieval, and evaluation. The recall and Mean Average Precision (MAP) are the evaluation measures, discussed in Section 3.5. The Proposed approach achieves good MAP without using any other resources for Hindi-English CLIR as shown in experiment results in Table 5.1.

Table 5.1: Wikipedia API based CLIR results

Experiment Results			Mono Lingual		Cross lingual	
			Fire 2010	Fire 2011	Fire 2010	Fire 2011
Vector Space Model	title	Recall	0.9709	0.8294	0.8315	0.5900
		MAP	0.3705	0.2688	0.1895	0.1096
	title & desc	Recall	0.9954	0.9434	0.9066	0.6914
		MAP	0.4597	0.3584	0.2685	0.1594
BM25 Model	title	Recall	0.9724	0.8290	0.8300	0.5925
		MAP	0.3714	0.2675	0.1899	0.1083
	title & desc	Recall	0.9969	0.9442	0.9081	0.6914
		MAP	0.4650	0.3559	0.2685	0.1601

The retrieval algorithms don't have any significant impact on the MAP as both of these achieve almost the same MAP. N-gram terms are used for phrasal translation which implicitly disambiguates the translations. Wikipedia knowledge base has fewer articles for the Hindi language as many query terms are not found. Approx 75% of Hindi Wikipedia articles have the inter-wiki link attribute and many of them are inter-wiki linked to wrong target language articles. These issues have a very bad impact on the MAP. The MAP for FIRE 2011 is lesser than FIRE 2010 because FIRE 2011 query length is shorter than to FIRE 2010 query length. A maximum of 0.1895 and 0.1096 MAP for FIRE 2010 and 2011 with the $< title >$ tag & 0.2685 and 0.1601 MAP for FIRE 2010 and 2011 with the $< title >$ and $<$ $desc >$ tag, is achieved by using the title and inter-wiki link attribute of Wikipedia.

The query terms which do not have any article or inter-wiki link are considered as Out Of Vocabulary (OOV) terms. Various Wikipedia issues, i.e., poor coverage of Hindi Wikipedia article, unavailability of inter-wiki links, wrong target language articles are identified during experimentation. Many N-gram terms, even, unigram also don't have any Wikipedia article. FIRE 2010 and 2011 datasets analysis

[6] http://fire.irsi.res.in/fire/home

with online Wikipedia knowledge base states that approx 75% of Hindi Wikipedia articles have English inter-wiki link. Wikipedia inter-wiki link statistics for Hindi-English language is shown in Table 5.2.

Table 5.2: Wikipedia Hindi-English inter-wiki link statistics

Datasets	Available Hindi article	Available English inter-wiki linked article	Percentage
Fire 2010	4185	3244	77.51%
Fire 2011	2460	1849	75.16%

In the above-discussed work, exactly matched query terms are translated by using the title and inter-wiki link attributes but partially matched terms are not handled. Apart from that, Wikipedia also suffers from the wrong inter-wiki link issue where a Wikipedia article may have a wrong inter-wiki link, for example, an article “स्टैनफोर्ड विश्वविद्‌यालय” (Stainphord vishvavidhyalaya) has a wrong English inter-wiki linked article “Pac-12 conference”. The ambiguous Wikipedia article issue, where Wikipedia may have multiple articles with the same title. The proposed Wikipedia based translation approach is extended by incorporating the solutions of the issues of partially matched terms, wrong inter-wiki links, and ambiguous articles. The proposed algorithm is represented in Algorithm 5

A larger translation unit which can be a phrase or multi-word term returns a better translation, hence, the N-gram tokens up to 4-gram are searched on Wikipedia by using Wikipedia.search() function which returns a set of relevant articles. The proposed algorithm incorporates three modules, i.e., Exactly Matched (EM), Partially Matched (PM), and disambiguation.

Algorithm 5 Wikipedia based translation algorithm

```
Input: A Source Language Query $SLQ[w_1, w_2, ..., w_n]$, a Source Language Stop-
       Word List (SLSWL), a Target Language Stop-Word List (TLSWL), and
       the Wikipedia on-line knowledgebase
Output: A Target Language Query (TLQ).

TLQ=[];
 english_titles={};
 dis={};
 // the remaining modules of the algorithm are shown on the next pages //
```

```
Function Disambiguation() is
    foreach (key in dis.keys()) do
        titleList=[]; targetWords= {};
        foreach (page ∈ dis[key].disambiguation()) do
            count=0;
            for (l=0 to Length(TLQ)) do
                if (page.abstract().substr(TLQ[l])) then
                    count++;
                end
            end
            if (count ≥ 1) then
                titleList.add(page.title());
            end
        end
        merged_title = " ".join(titleList);
        words= merged_title.split(" ");
        foreach (word ∈ words) do
            if (word ∈ targetWords.keys()) then
                targetWords[word]+=1;
            else
                targetWords[word]=1;
            end
        end
        TLQ.add(targetWords.maximum());
        break;
    end
end
Function Exactly_Matched(docs[]: array, value: string) is
    flag = 0;
    for (k=0 to Length(docs)) do
        if (docs[k].title() == value and docs[k].IWL(en) and
        docs[k].IWL(en).disambiguation() == TRUE) then
            dis[value]=docs[k].IWL(en).title();
            flag=1; break;

        else
            if (docs[k].IWL(en) and docs[k].title() == value) then
                if (docs[k].IWL(en).title().IWL(hi) and
                docs[k].IWL(en).title().IWL(hi).title() == value) then
                    TLQ.add(docs[k].IWL(en).title());    // reverse verification
                    flag = 1;
                    break;
                end
            end
        end
    end
    if flag==1 then
        return 1;
    else
        return 0;
    end
end
```

```
Function Partially_Matched(docs[]: array, value: string) is
    English_articles=[];
    for (k=0 to Length(docs)) do
        if (docs[k].title().substr(value) or PMS(docs[k].title(), value) > 0.80) then
            tempTitle = docs[k].IWL(en).title() ∩ docs[k].IWL(en).abstract();
            English_articles.add(docs[k].IWL(en).title());
            if (tempTitle ∈ english_titles) then
                english_titles[tempTitle]+=1;
            else
                english_titles[tempTitle]=1;
            end
        end
    end
    temp = english_titles.max();
    if (english_titles[temp] > 1) then
        TLQ.add(temp); return 1;

    else
        flag = 0;
        foreach substring in english_titles do
            if (Wikipedia.search(en, substring) ∈ English_articles) then
                TLQ.add(substring); flag=1;
            end
        end
        if (flag==1) then
            return 1;
        else
            return 0;
        end
    end
end
StopWord_Removal(SLQ);
for (i=0 to len(SLQ)) do
    for (j=0 to 3) do
        token = N_Gram(SLQ, i, 3-j);
        articles = Wikipedia.search(hi, token);
        if (Exactly_Matched(articles, token) or Partially_Matched(articles, token))
        then
            break;
        else
            if (j==3 and token ∉ dis.keys()) then
                TLQ.add(token)
            end
        end
    end
end
if (dis.Length() ≥ 1) then
    Disambiguation();
end
```

EM module provides a translation for the token which is exactly matched to a title of the Wikipedia article and that article should have an inter-wiki link. It returns at most one article with the unique title, but for the ambiguous pages, it returns multiple articles with the same title which is then handled by the disambiguation module. The function $title()$ and $IWL()$ return the title and inter-wiki link of a Wikipedia article respectively while $disambiguation()$ returns the ambiguity status of an article. EM module eliminates the wrong inter-wiki linked articles by reverse verification, where the Hindi inter-wiki linked title of the English inter-wiki linked article is compared to the Hindi token, and if that Hindi inter-wiki linked title matches to the Hindi token then the title of the corresponding English inter-wiki linked article is included in the target language query. Ambiguous Wikipedia titles are disambiguated based on the target language query words.

Table 5.3: Translation achieved by the exactly matched module

Example Type	**Token**	**Hindi Wikipedia title**	**English inter-wiki linked title**
Unique Page	“मनमोहन सिंह” (manmohan singh)	“मनमोहन सिंह” (manmohan singh)	Manmohan Singh
Wrong inter-wiki link	“स्टैनफोर्ड विश्वविद्यालय” (stainphord vishvavidhyalaya)	“स्टैनफोर्ड विश्वविद्यालय” (stainphord vishvavidhyalaya)	Pac-12 Conference
Ambiguous Page	“हाईवे” (haaiwe)	“हाईवे” (haaiwe)	Highway (disambiguation)

Example tokens of these three types of issues are shown in Table 5.3. The token “manmohan singh” has a unique article which is correctly translated as “Manmohan Singh” whereas the token “stainphord vishvavidhyalaya” has a wrong inter-wiki link so it is eliminated by reverse verification, further, the translations are computed after dividing the token into unigrams, this scenario is represented in Figure 5.2, where the token “stainphord” does not have an exactly matched Wikipedia article so this token is translated by partially matched module and the token “vishvavidhyalaya” is correctly translated. The ambiguous token “haaiwe” is also correctly translated by using the disambiguation module which is represented in Figure 5.3.

Partially matched module computes the translation from the articles whose titles are partially matched with the token that means the token should be either a substring of the title or Partial Matching Score (PMS) of the token with the title should be greater than 0.8. PMS is computed by using an Equation 5.1, where

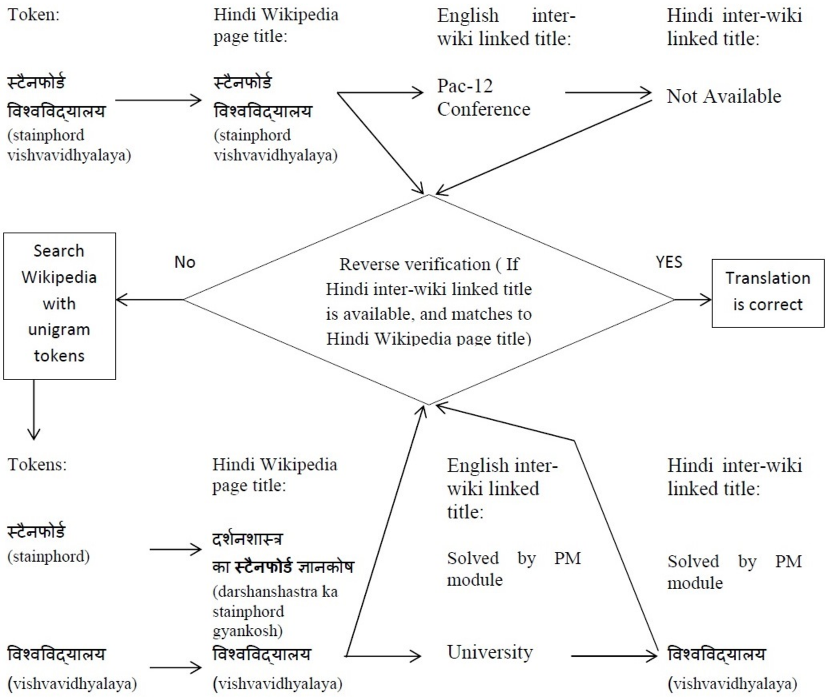

Token:
स्टैनफोर्ड विश्वविद्यालय
(stainphord vishvavidhyalaya)
Hindi Wikipedia page title:
स्टैनफोर्ड विश्वविद्यालय
(stainphord vishvavidhyalaya)
English inter-wiki linked title:
Pac-12 Conference
Hindi inter-wiki linked title:
Not Available
Search Wikipedia with unigram tokens
No
Reverse verification (If Hindi inter-wiki linked title is available, and matches to Hindi Wikipedia page title)
YES
Translation is correct
Tokens:
स्टैनफोर्ड
(stainphord)
विश्वविद्यालय
(vishvavidhyalaya)
Hindi Wikipedia page title:
दर्शनशास्त्र का स्टैनफोर्ड ज्ञानकोष
(darshanshastra ka stainphord gyankosh)
विश्वविद्यालय
(vishvavidhyalaya)
English inter-wiki linked title:
Solved by PM module
University
Hindi inter-wiki linked title:
Solved by PM module
विश्वविद्यालय
(vishvavidhyalaya)

Figure 5.2: An example of the wrong inter-wiki link issue, resolved with unigram tokens

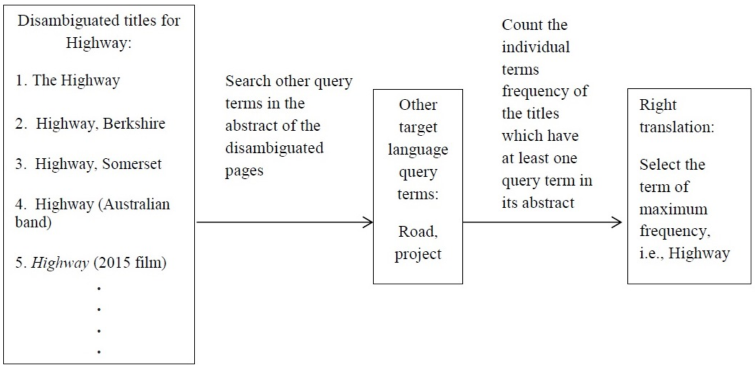

Disambiguated titles for Highway:
1. The Highway
2. Highway, Berkshire
3. Highway, Somerset
4. Highway (Australian band)
5. *Highway* (2015 film)
Search other query terms in the abstract of the disambiguated pages
Other target language query terms:
Road, project
Count the individual terms frequency of the titles which have at least one query term in its abstract
Right translation:
Select the term of maximum frequency, i.e., Highway

Figure 5.3: An example of the disambiguation issue, resolved by using the page abstract and other query terms

the function $LCS()$ returns the Longest Common Sub-sequence string.

$$PMS(token, title) = \frac{|LCS(token, title)|}{|token|} \tag{5.1}$$

Wikipedia returns either more than one article or at most one article, which should have an English inter-wiki link. The common substrings between the titles and abstracts of the inter-wiki linked articles are computed. The frequency of unique common substrings are calculated and the substring with the maximum frequency is selected as the translation. In the case of at most one article, the highest frequency is 1 and no translation will be selected then the right translation will be guessed from the English Wikipedia articles which are extracted for the substrings, where the extracted article should be the same article from which the English substring is produced. Table 5.4 and 5.5 demonstrates the examples of the partially matched module where ”bus” is selected with the maximum frequency and ”ram mandir” is selected by guessing from the English Wikipedia articles.

Table 5.4: Translation achieved by the partially matched module, where Wikipedia returns more than one English inter-wiki linked articles

Hindi Wikipedia articles for “बस” (Bus)	**English inter-wiki linked articles**	**Commom substrings between the titles and abstracts**	**Calculating frequency of common substrings**
चेन्नई मोफस्सिल बस टर्मिनस (Chennai mofassil bas tarminas)	Chennai Mofussil Bus Terminus	bus terminus, chennai, bus	bus terminus (1) inter state bus (1) chennai (1)
महाराणा प्रताप अन्तर्राज्यीय बस अड्डा, दिल्ली (Maharanha pratap antarrajyiy bas adda, dillee)	Maharana Pratap Inter State Bus Terminus	inter state bus, bus	mini satellite (1) satellite (1)
भारतीय लघु उपग्रह बस (Bharteey laghu upgrah bas)	Indian Mini Satellite bus	mini satellite	bus (4) inter state bus terminus (1)
स्वामी विवेकानंद अंतर्राज्यीय बस अड्डा (Swami vivekanand antarrajyiy bas adda)	Swami Vivekanand Inter State Bus Terminus	inter state bus terminus, bus	
उपग्रह बस (upagrah bas)	Satellite bus	bus, satellite	

If a token is neither exactly matched nor partially matched to any Wikipedia title then that token may be present in other Wikipedia attributes like abstract, category, content, image, hyperlinks, and an infobox. These attributes are used for the technically developed languages but not for Hindi-English, because a Hindi Wikipedia article may have lesser content than the English inter-wiki linked article like “श्री लंका क्रिकेट टीम” (shree lanka cricket team) and a query term may or may not have a Wikipedia article with a proper title like “गुरिल्ला” (gurilla). Such tokens are forwarded in their actual form to target language query.

A Hindi-English parallel corpus HindiEnCorp[7] is used to train the SMT, which is considered as the baseline technique. Terrier open source search engine is used for

[7] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0

Table 5.5: Translation achieved by the partially matched module, where Wikipedia returns at most one English inter-wiki linked article

Hindi Wikipedia articles for “राम मंदिर” (Ram mandir)	**English inter-wiki linked articles**	**Commom substrings between the title and abstract and its frequency**	**English Wikipedia article corresponding to English substring**
राम मंदिर रेलवे स्टेशन (ram mandir relve steshan)	Ram Mandir railway station	ram mandir (1), railway station (1), railway (1), station (1), ram (1)	Ram Mandir railway station, Train station, Rail transport, Station, Ram

indexing and retrieval, where Term Frequency-Inverse Document Frequency (TF-IDF) technique is used for indexing and cosine similarity is used for retrieval. The proposed approach is evaluated by using recall, Precision@5, Precision@10, and MAP, as discussed in Section 3.5. Experiment results for the proposed Wikipedia-based CLIR approach against the baseline SMT are represented in Table 5.6.

Table 5.6: Experiment results for the Wikipedia based approach against the baseline SMT

Resource	**Dataset**	**Recall**	**Precision@5**	**Precision@10**	**MAP**
SMT	FIRE2010	0.8284	0.2840	0.2300	0.2832
	FIRE2011	0.7084	0.2440	0.2320	0.1885
Wikipedia	FIRE2010	0.8820	0.2880	0.2380	0.2880
	FIRE2011	0.6807	0.2320	0.2280	0.1772

SMT is unable to translate some words like कॉरीडोर (koridor), आईफोन (aaiphone), डान्स (dans), which are easily translated by Wikipedia whereas some words which are not translated by Wikipedia are easily translated by SMT like एन्सेफलाइटिस (enseflaetis), प्रस्तावित (prastavit), कांड (kand). Therefore, it can be stated that SMT and Wikipedia-based CLIR approach achieve approximately equal evaluation measures.

The Wikipedia offline dumps can be used as the parallel corpus. The Wikipedia dumps for Hindi and English languages are the partial comparable corpus where source and target language documents quality and coverage are not in the congruent condition. The Hindi Wikipedia corpus size is around 2.73% of English Wikipedia corpus and the length (number of articles) of Hindi Wikipedia corpus is around 28.57% of English Wikipedia corpus. The Wikipedia dump statistics for Hindi and English language is shown in Table 5.7.

Table 5.7: Wikipedia dumps statistics for Hindi-English

Corpus	**Corpus length**	**Corpus size**
Hindi Wikipedia	40,000 (28.57%)	336 MB (2.73%)
English Wikipedia	1,40,000	12 GB

The English inter-wiki linked Wikipedia Hindi language articles are very few. The inter-wiki link statistics for Hindi Wikipedia dumps is shown in Table 5.8. Various Hindi Wikipedia articles have the only title. These articles don't have its description while the corresponding inter-wiki linked English article has a sufficient description. A combination of source and target language Wikipedia corpus becomes a partial comparable corpus due to the differences in (i) The number of articles, (ii) The document quality and coverage.

Table 5.8: Wikipedia Hindi-English inter-wiki link statistics for Wikipedia Dumps

Datasets	Available Hindi article	Available English inter-wiki linked article	Percentage
Wikipedia Hindi Dumps	140K	37K	26.42%

5.2 Web-based Lexical Resources

Manual dictionary and parallel corpus are the traditional translation resources which suffer from the OOV issues, so the researchers are moving towards web resources because these resources are dynamic in nature. In the literature, various web resources are investigated and utilized for foreign languages but not for the Hindi language. In this section, Hindi WordNet[8] (HWN), Indo WordNet[9] (IWN), COnceptNet[10] (CON), On-line dictionaries are investigated and applied to Hindi-English CLIR.

HWN and IWN are the lexical databases which provide word relations in 18 different Indian languages including the English language, where the relation “synsets“ returns the same words in the different languages (Bhattacharyya et al., 2017). CON provides a strong semantic network for only 10 core supported languages

[8] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

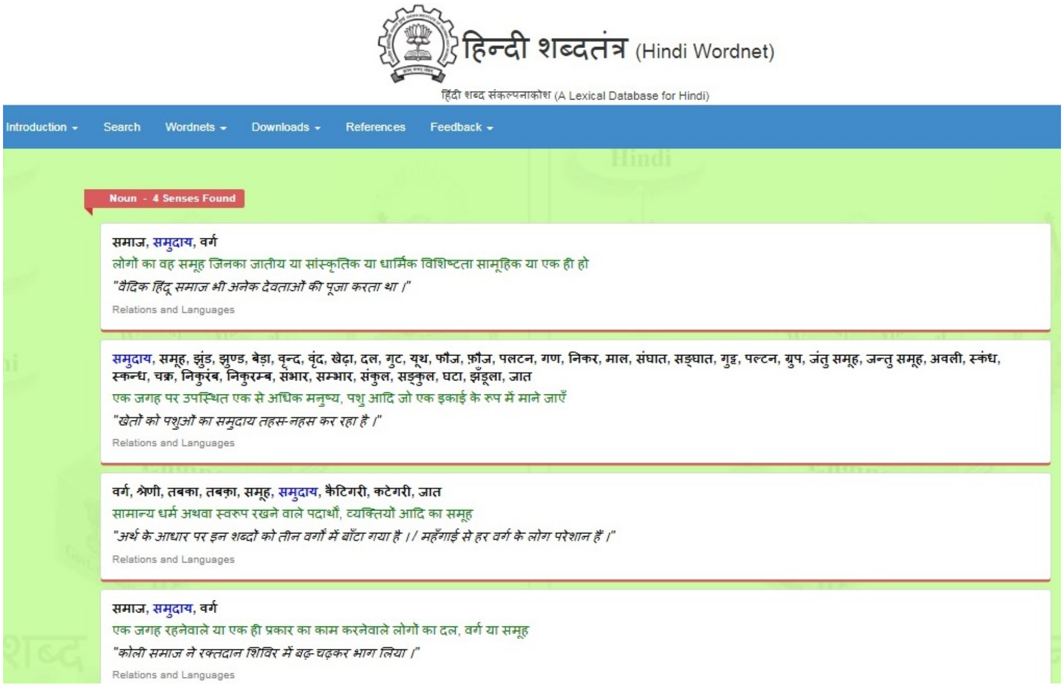
[9] http://www.cfilt.iitb.ac.in/indowordnet/

[10] https://github.com/commonsense/conceptnet5/wiki/Languages

and poor semantic network for other 77 moderate supported languages, where the relations “Related terms” and “Synonyms” return the English translations for a Hindi word. A number of online dictionaries are also available which provide a dynamic translation, therefore, the researchers have put their efforts towards the development of web-based translation approaches (Hosseinzadeh Vahid et al., 2015). A lot of research has been done for the technically developed languages like English, German, Spanish, Russian, Chinese, etc. but not for Hindi.

5.2.1 Hindi WordNet, Indo WordNet, and COnceptNet

HWN fetches syntactic and semantic relations amongst the Hindi words. IWN brings syntactic and semantic relations in English, Hindi, and 17 other Indian languages. It also has advanced features than the HWN. Both of the HWN and IWN are developed by the CFILT lab, IIT Bombay. HWN returns translations in the form of “English synset”, an example of which, is shown in Figure 5.4. In addition to the “English synset” IWN also returns “synonymy”, “gloss”, and “example sentence” attributes which can be useful for the disambiguation purpose, that is shown in Figure 5.5. Both of the HWN and IWN return the same set of English synset.


हिन्दी शब्दतंत्र (Hindi Wordnet)
हिंदी शब्द संकल्पनाकोश (A Lexical Database for Hindi)
Introduction
Search
Wordnets
Downloads
References
Feedback
Noun - 4 Senses Found

समाज, समुदाय, वर्ग
लोगों का वह समूह जिनका जातीय या सांस्कृतिक या धार्मिक विशिष्टता सामूहिक या एक ही हो
"वैदिक हिंदू समाज भी अनेक देवताओं की पूजा करता था ।"
Relations and Languages

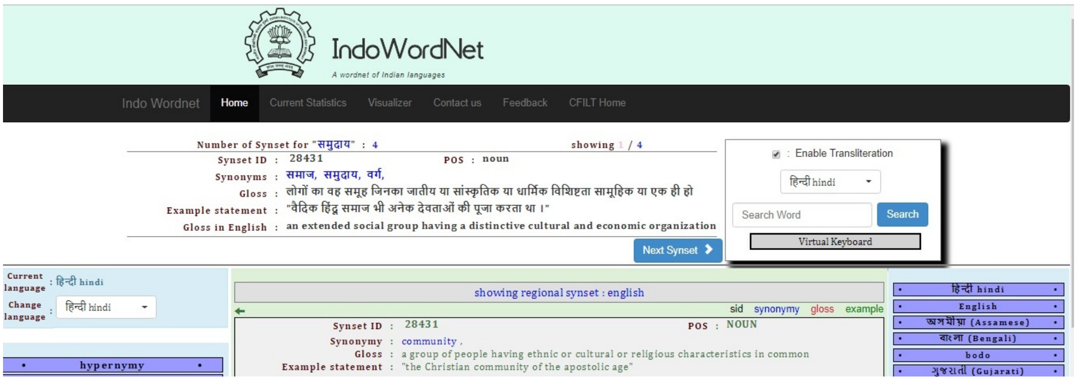
समुदाय, समूह, झुंड, झुण्ड, बेड़ा, वृन्द, वृंद, खेढ़ा, दल, गुट, यूथ, फौज, फ़ौज, पलटन, गण, निकर, माल, संघात, सङ्घात, गुट्ट, पल्टन, ग्रुप, जंतु समूह, जन्तु समूह, अवली, स्कंध, स्कन्ध, चक्र, निकुरंब, निकुरम्ब, संभार, सम्भार, संकुल, सङ्कुल, घटा, झँडूला, जात
एक जगह पर उपस्थित एक से अधिक मनुष्य, पशु आदि जो एक इकाई के रूप में माने जाएँ
"खेतों को पशुओं का समुदाय तहस-नहस कर रहा है ।"
Relations and Languages

वर्ग, श्रेणी, तबका, तबक़ा, समूह, समुदाय, कैटिगरी, कटेगरी, जात
सामान्य धर्म अथवा स्वरूप रखने वाले पदार्थों, व्यक्तियों आदि का समूह
"अर्थ के आधार पर इन शब्दों को तीन वर्गों में बाँटा गया है ।/ महँगाई से हर वर्ग के लोग परेशान हैं ।"
Relations and Languages

समाज, समुदाय, वर्ग
एक जगह रहनेवाले या एक ही प्रकार का काम करनेवाले लोगों का दल, वर्ग या समूह
"कोली समाज ने रक्तदान शिविर में बढ़-चढ़कर भाग लिया ।"
Relations and Languages

Figure 5.4: Hindi WordNet English synset attribute returns possible translations

CON contains concepts and relationship among the concepts where concepts and relationship are represented by nodes and arcs respectively. Concepts are the



Number of Synset for "समुदाय" : 4 showing 1 / 4

Synset ID : 28431 POS : noun
Synonyms : समाज, समुदाय, वर्ग,
Gloss : लोगों का वह समूह जिनका जातीय या सांस्कृतिक या धार्मिक विशिष्टता सामूहिक या एक ही हो
Example statement : "वैदिक हिंद्र समाज भी अनेक देवताओं की पूजा करता था ।"
Gloss in English : an extended social group having a distinctive cultural and economic organization

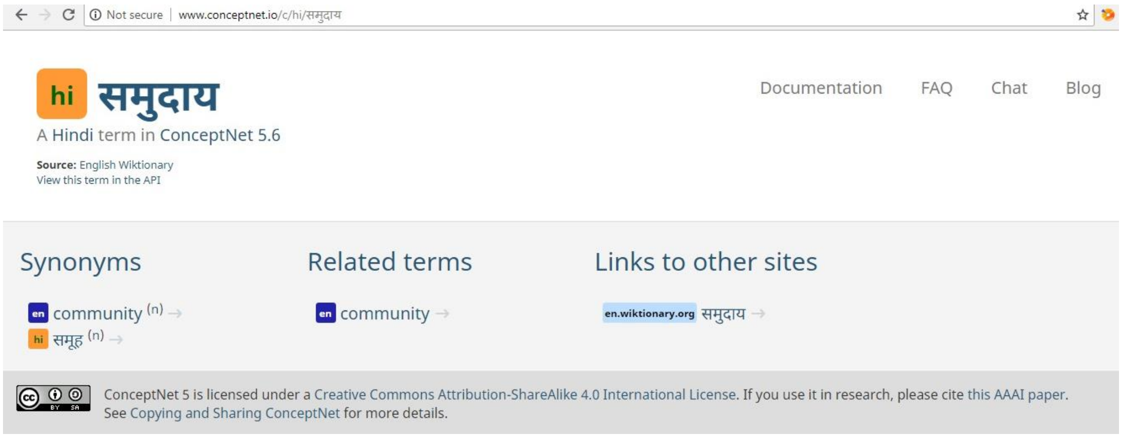
Current language : हिन्दी hindi
Change language : हिन्दी hindi

showing regional synset : english

Synset ID : 28431 POS : NOUN
Synonymy : community ,
Gloss : a group of people having ethnic or cultural or religious characteristics in common
Example statement : "the Christian community of the apostolic age"

Figure 5.5: Indo WordNet useful attributes other than the English synsets

words or short phrases. The relationship represents the relationship amongst the words or phrases. It is available in various languages including the Hindi language. CON attributes which return translations are "Related terms"and "Synonymy". An example of CON interface is shown in Figure 5.6.



hi समुदाय

A Hindi term in ConceptNet 5.6

Source: English Wiktionary
View this term in the API

Synonyms	Related terms	Links to other sites
en community (n) → hi समूह (n) →	en community →	en.wiktionary.org समुदाय →

Figure 5.6: ConceptNet interface where RelatedTo attribute returns translation

FIRE 2010 and 2011 dataset are used to investigate the web-based translation resources. These web-based resources are manually analyzed because evaluation matrices are not available. A comparison amongst the HWN, IWN, and CON is presented in Table 5.9 where a query is represented by a row. The subsequent columns show the entries for query words & senses on HWN (H), IWN(I) and CON (C). HWN returns only senses while the translations are computed from IWN because IWN contains Hindi and English synonym mutually. It is also possible that a word may have multiple HWN senses but it doesn's have a single sense on the IWN and IWN may return empty linked pages for the pages which are suggested by itself.

The CON attributes TranslationOf and RelatedTo return the target language translations. A Hindi query word is searched to Hindi version of CON where the number of CON attributes is 0 for most of the words and 1 for a few words.

Table 5.9: A comparative analysis among Hindi WordNet, Indo WordNet and ConceptNet

				FIRE 2010 Queries				
S.no.	Word	H,I,C	Word	H,I,C	Word	H,I,C	Word	H,I,C
1	गुज्र्जरों	2,0,0	मीणा	0,0,0	समुदाय	3,3,1	संघर्ष	1,0,1
2	राम	2,2,0	मंदिर	2,2,0	आडवाणी	0,0,0	सिंघल	0,0,0
				FIRE 2011 Queries				
3	स्वाइन	0,0,0	फ्लू	1,0,0	टीके	0,0,0		
4	माइकल	0,0,0	जैक्सन	0,0,1	अचानक	1,0,1	मृत्यु	1,2,1

In HWN & IWN, the query words are searched and their English synsets are extracted which are more than one for a single query word, therefore, the WordNet path_similarity[11] is used to select the best synset. Path_similarity is measured on the basis of the shortest path that connects the senses in hypernym/hyponym taxonomy. HWN & IWN based approach is described by using an example which is shown in Figure 5.7, where all English synsets are extracted for each query word and the query words which are not available in HWN & IWN are skipped, further, path_similarity returns a synset pair with the maximum similarity, for example, “jati community”and “community struggle”.

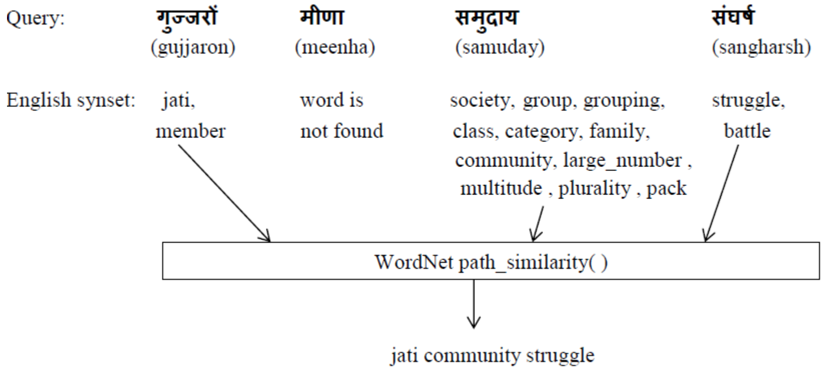


Figure 5.7: Query translation by using Hindi WordNet & Indo WordNet based approach

CON based translation approach is represented in Figure 5.8, where words translations are extracted by using the CON attributes. CON internally uses various resources including the multi-lingual WordNet, therefore, the WordNet path_similarity is used for translation disambiguation.

[11] http://www.nltk.org/howto/wordnet.html

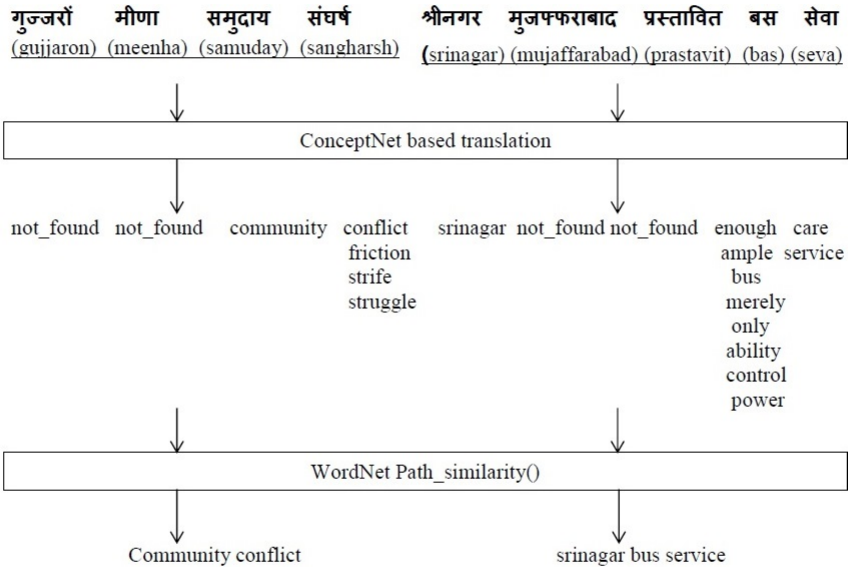

गुज्जरों मीणा समुदाय संघर्ष
(gujjaron) (meenha) (samuday) (sangharsh)
श्रीनगर मुजफ्फराबाद प्रस्तावित बस सेवा
(srinagar) (mujaffarabad) (prastavit) (bas) (seva)
ConceptNet based translation
not_found not_found community conflict friction strife struggle
srinagar not_found not_found enough ample bus merely only ability control power care service
WordNet Path_similarity()
Community conflict
srinagar bus service

Figure 5.8: Query translation by using ConceptNet based approach

5.2.2 Online Dictionaries

A number of dictionaries are available on the web, among them, five popular online dictionaries are used for translation, i.e., Universal[12], Shabdkosh[13], Raftar[14], Tamilcube[15], and Indiatyping[16]. A comparative study of these five dictionaries is shown in Table 5.10. A dictionary returns multiple translations, therefore, the WordNet path_similarity measure is used for the disambiguation purpose.

- Universal Word Based Dictionary (UWBD): It is developed by CFILT lab. The users are able to search words and phrases in Hindi and English languages. The dictionary contains 136109 words and provides word's grammatical, morphological and semantic relation. The UWBD interface provides two options for searching, i.e., exact match and maximal match. Exact match searching generally suffers from the dictionary coverage issue, so maximal match option is preferred which includes many irrelevant and wrong translation pairs leads to WTD issue.

- Shabdkosh Dictionary (SD): It is developed by a graduate student from IIT Delhi in 2003. Initially, it contained 15000 words but today, it contains tens of thousands of words. It suffers from the dictionary coverage issue and many irrelevant translations are included.

[12] http://www.cfilt.iitb.ac.in/ hdict/webinterface_user/dict_search_user.php
[13] http://www.shabdkosh.com/
[14] http://shabdkosh.raftaar.in/
[15] http://dictionary.tamilcube.com/
[16] http://indiatyping.com/index.php/hindi-dictionary/hindi-english-dictionary

- Raftar Dictionary (RD): Raftar is a Hindi search engine started in 2005. It also provides a Hindi-English dictionary. Hindi word is searched against all Hindi entries and transliterated Hindi word is searched against all English entries. All translation pairs which contain either Hindi word or transliterated Hindi word as a substring are extracted. It returns a few relevant translation pairs.
- Tamilcube Dictionary (TD): It is a Hindi to English and English to Hindi dictionary containing 200000 words and it is continuously growing. It returns all the Hindi word and phrases which contains query words as a substring, hence, many irrelevant translations are included in the translation set.
- Indiatyping Dictionary (ID): It is a web portal which provides a typing solution for all Indian languages. It also provides a Hindi-English translation service where a Hindi to English dictionary is used. It follows maximal match searching, so many irrelevant translations are included.

Table 5.10: A comparative study of the on-line dictionaries

Dictionary	**Creator**	**Length**	**Searching technique**	**Weak point**
Universal Word (UW)	CFILT, IIT Bombay	136109	exact and max match	inclusion of irrelevant, OOV issue, and wrong translations
Shabdkosh (SK)	Students IIT Delhi	100000	exact match	inclusion of irrelevant, OOV issue translations
Raftar (RF)	raftar.in	-	exact match	few relevant translations found, Roman transliterated searching
Tamilcube (TC)	tamilcube. com	200000	max match	inclusion of irrelevant translations
IndiaTyping (IT)	indiatyping .com		exact and max match	inclusion of irrelevant translations

Five online dictionaries which are briefly discussed above are investigated. A comparative analysis with respect to the translation is represented in Table 5.11 where the column next to the query word's represent a five digit entry. Each digit shows the number of available translations in the online dictionary, in the case of more than five translations, the $ symbol is used. The digits are stored according to the order of dictionaries that is [UW, SK, RF, TC, IT]. Most of the entries are $ which leads to a WTD issue while the entries with digit 0 are for the named entities or newly added terms which lead to an OOV issue.

FIRE 2010 and 2011 datasets, statistics are represented in Table 3.5 are used to evaluate the effectiveness of web-based lexical resource based CLIR approaches.

Table 5.11: On-line dictionaries analysis in perspective of translation

FIRE 2010 Queries								
S.no.	Word	NOT	Word	NOT	Word	NOT	Word	NOT
1	रिश्वत	$$2$$	बदले	$10$0	संसद	$21$3	प्रश्न	$$$$$
2	गुटखा	00000	मालिकों	00000	अन्डरवर्ल्ड	00000	उलझाव	341$4
FIRE 2011 Queries								
3	अबू	$00$0	गरीब	$$$$$	जेल	$$$$$	अत्याचार	$$$$$
4	भोपाल	00001	गैस	$$1$$	दुर्घटना	$$$$$		

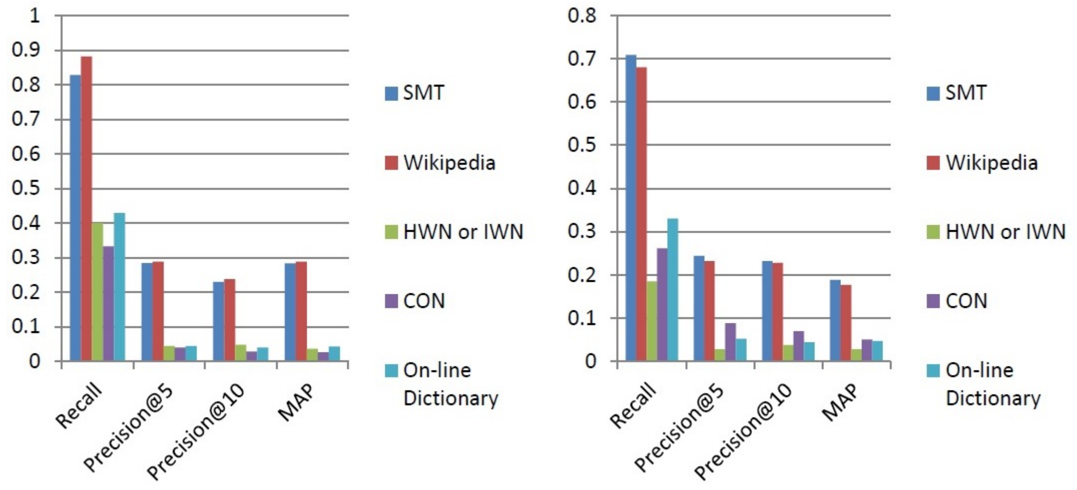
A Hindi-English parallel corpus HindiEnCorp is used to train the SMT which is considered as the baseline technique. Terrier open source search engine is used for indexing and retrieval, where TF-IDF is used for indexing and CSS is used for retrieval. The proposed CLIR approaches are evaluated by using recall, Precision@5, Precision@10, and MAP, discussed in Section 3.5.

Experiment results for the web-based lexical resource based CLIR approaches including Wikipedia and the baseline SMT are represented in Table 5.12. A comparison graph for FIRE 2010 and FIRE 2011 is represented in Figure 5.9.

Table 5.12: Experiment results for the web resource based translation approaches

Resource	**Dataset**	**Recall**	**Precision@5**	**Precision@10**	**MAP**
SMT	FIRE2010	0.8284	0.2840	0.2300	0.2832
	FIRE2011	0.7084	0.2440	0.2320	0.1885
Wikipedia	FIRE2010	0.8820	0.2880	0.2380	0.2880
	FIRE2011	0.6807	0.2320	0.2280	0.1772
HWN & IWN	FIRE2010	0.3996	0.0440	0.0480	0.0366
	FIRE2011	0.1850	0.0280	0.0380	0.0275
CON	FIRE2010	0.3323	0.0400	0.0280	0.0259
	FIRE2011	0.2614	0.0880	0.0700	0.0505
On-line	FIRE2010	0.4287	0.0440	0.0400	0.0430
Dictionary	FIRE2011	0.3306	0.0520	0.0440	0.0471

SMT and Wikipedia achieve approximately equal evaluation measures while HWN & IWN, CON, and online dictionaries achieve poor evaluation measures compared to the SMT and Wikipedia as shown in Figure 5.9. HWN & IWN and CON are not lexically rich and suffer from missing or unseen words more than the SMT and Wikipedia as shown in Table 5.13. Therefore, HWN & IWN and CON return poor evaluation measures. Online dictionary has less number of missing or unseen words compared to the HWN & IWN, and CON, but it returns too many irrelevant translations. Therefore, it performs approximately equal to HWN &



(a) FIRE 2010

(b) FIRE 2011

Figure 5.9: A comparison graph for the evaluation of web resource based translation approaches applied to (a) FIRE 2010, and (b) FIRE 2011

Table 5.13: Number of missing or unseen words where n represents total number of vocabulary

	FIRE2010 (n=257)	**FIRE2011 (n=178)**
SMT	14	10
Wikipedia	13	12
HWN & IWN	61	51
CON	128	58
On-line Dictionary	37	24

IWN, and CON. Named entity terms are translated by the Wikipedia but due to the unavailability, named entities are not translated by the HWN & IWN, CON, and online dictionaries. A sub-query translation is represented in Table 5.14, to show the effectiveness of different web resource based translation approaches.

SMT and Wikipedia achieve almost the same and correct translation while HWN & IWN, CON, and online dictionary do not return the correct translation. HWN & IWN and CON not only have limited translations for individual words but also return less relevant translations. Online dictionary returns a wide range of translations for an individual word which includes both the relevant and irrelevant translations. Less relevant and irrelevant translations generate noise in the target language query, due to which, online dictionary is unable to achieve a good MAP.

The state-of-art SMT uses a parallel corpus which is static in nature and has a limited number of vocabularies, hence, the SMT may not translate the missing or unseen words whereas the web is dynamic in nature, therefore, the web-based

Table 5.14: Different translations achieved for a sub-query

Translation Resource	रक्षा (raksha)	विभाग (vibhag)	रक्षा विभाग (raksha vibhag)
SMT	defense	department	department of defense
Wikipedia	defence	division, department	defence division department
HWN & IWN	protection	division, department, section, sectionalization, component part, portion, partitioning, segmentation, partition	protection part
CON	defence, protection, safety, support, defense	bureau, department	defence bureau
Online Dictionary	panoply, auspice, ward, patronage, back, preservation, care, charge, protection, conservancy, custody, refuge, defence, safeguard, defense, safety, defensive, salvation, egis, security, guard, shelter, insurance	partition, allotment, personnel, analysis, portfolio, block, portion, branch, realm, bureau, school, category, sector, compartment, unit, work unit, department	

translation techniques, Wikipedia, HWN & IWN, CON, and online dictionaries are proposed. Wikipedia based approach effectively translates the user queries with the help of exactly matched, partially matched, and disambiguation modules which are introduced to solve the issues of partially matched terms, wrong inter-wiki links, and ambiguous articles. A Hindi Wikipedia article may have lesser content compared to an English article and due to which, Wikipedia is unable to translate the words which are neither exactly matched nor partially matched to Wikipedia articles, however, Wikipedia achieves 0.2880 (FIRE 2010), 0.1772 (FIRE 2011) MAP which is approximately equal to 0.2832, 0.1885 achieved by the SMT, this is possible because Wikipedia computes translations for many missing or unseen words which are not translated by the SMT.

HWN & IWN and CON are not lexically rich. They are having more number of missing or unseen words compared to the SMT and Wikipedia. Online dictionaries return too many irrelevant translations where the best translation is selected by using the WordNet path similarity and that selected best translation pair is not correct due to the irrelevant translations. Apart from this, HWN & IWN, CON, and online dictionaries are unable to translate the named entities which are translated by the SMT and Wikipedia. HWN & IWN, CON are lexically poor and online dictionaries return too many irrelevant translations, hence, leading to a poor MAP. In the future, the efficiency may be enhanced by applying a filtering technique to remove the irrelevant or less relevant translations.

5.3 Summary

SMT provides static translations due to the limited vocabularies in the available parallel corpus while the web provides a dynamic interface and it is effectively used for the technically developed language like English, German, Spanish, Russian, Chinese, etc. Wikipedia and other web-based lexical resource based approaches are proposed for Hindi-English CLIR. Wikipedia *title* and *inter-wiki link* attributes are used for translation. The proposed Wikipedia based CLIR approach addresses the issues of partially matched terms, wrong inter-wiki links, and ambiguous Wikipedia article issue. The proposed Wikipedia based approach achieves approximately equal evaluation measures to SMT.

HWN and IWN attribute “Synset” & CON attributes “Related terms” and “Synonyms” returns the English translation for a Hindi word. WordNet path similarity is used to disambiguate the word’s translation. Five online dictionaries are used to extract the translation where WordNet path similarity is used to disambiguate the translation. HWN & IWN and CON are not lexically rich and suffer from missing or unseen words, therefore, HWN & IWN and CON achieve poor evaluation measures. Online dictionary has less number of missing or unseen words compared to the HWN & IWN, and CON, but it returns too many irrelevant translations, therefore, it performs approximately equal to HWN & IWN, and CON. Named entity terms are also easily translated by the Wikipedia but not by the HWN & IWN, CON, and online dictionaries because named entity terms are not available in the HWN & IWN, CON, and online dictionaries.

Chapter 6

CONCLUSION AND FUTURE WORK

CLIR provides the accessibility of relevant information in a language different than the query language. A CLIR approach incorporates a translation approach followed by monolingual information retrieval. There are two types of translation approaches, namely, query translation and documents translation. Query translation approach is preferred over the document translation due to a lot of computation time and space elapsed in document translation approach. Dictionary and corpus-based are the state-of-art translation approaches while the statistical machine translation and neural machine translation are the trending translation approaches which are also trained over the parallel corpus. CLIR approaches are suffered from the morphological variants (transliteration variants), dictionary coverage or out of vocabulary word, and word translation disambiguation issues. The performance of CLIR may be enhanced by including multiple translation or synonyms of translated terms, hence, there is a need to set criteria by which it becomes easy to select the multiple translation terms. A number of tools and techniques are available for the foreign language, all are based on the bilingual dictionary due to the fast computation. These tools suffered from out of vocabulary words, synonymy, homonymy, irrelevant translation, and user-assisted translation issues.

In the literature, a number of issues are recognized, i.e., morphological irregularities or transliteration variants, word translation disambiguation, multiple term translation selection, and out of vocabulary words. Online resources are also not much explored for the Hindi language. In this thesis, research objectives are formulated based on these research gaps or issues, i.e., (i) Development of a manual dictionary and term frequency model based CLIR approach, (ii) Development of translation induction algorithm based CLIR approach to handle morphological irregularities, (iii) Development of a semantic morphological variant selection algorithm to address translation mis-mapped and non-confident translation, (iv) Development of a context-based translation algorithm for the out of vocabulary word translation algorithm, (v) Development of the bilingual word vector and named entity recognition based disambiguation approaches, (vi) Development of a maximum of word's average probability and association score based disambiguation approach, (vii) Development a Wikipedia based translation approach, and (viii) Development of web-based lexical resources based translation approaches.

Manual dictionary based CLIR approach is the easiest approach. It takes a low computation cost. The out of vocabulary words are transliterated using the out of vocabulary term transliteration mining technique. A highest MAP 0.1172 for FIRE 2010 dataset is achieved with Shabdanjali dictionary which is better than the 0.0907 MAP achieved in (Sethuramalingam & Varma, 2008). The probabilistic dictionary based approach achieves better MAP, i.e., 0.2267 and 0.1672 for FIRE 2010 and FIRE 2011 respectively than the manual dictionary. The term frequency model is proposed due to the retraining of the IBM model. The term frequency model is based on the example sentences where a set of query-relevant parallel sentences are extracted from the parallel corpus and cosine similarity score is used to compute the best translation. The term frequency model achieves better MAP, i.e., 0.2637 and 0.1623 for FIRE 2010 and FIRE 2011 respectively than the probabilistic dictionary based approach.

Hindi is a morphologically rich language which suffers from the morphological irregularities like nukta character, infrequent words, and multiple morphological variants. To overcome these irregularities, four morphological variants solutions are proposed. The refined stop-words lists and morphological variants solutions are incorporated in the translation induction algorithm. Further, query translations are computed from the selected contextual parallel sentences for the query. Translation induction algorithm achieves better MAP, i.e., 0.2818 and 0.1921 for FIRE 2010 and FIRE 2011 respectively than the probabilistic dictionary and term frequency model. The SMT and NMT are the trending translation techniques which are trained over the two different parallel corpus, i.e., HindiEnCorp and IITBCorpus with the different tuning parameter where SMT performs better than the NMT.

A semantic morphological variant selection algorithm is proposed to overcome the translation mis-mapped and non-confident translation issue. The semantic morphological variant selection algorithm achieves 0.3032 MAP for FIRE 2010 which is better results than 0.2832 achieved by baseline SMT. In the case of FIRE 2011, the semantic morphological variant selection algorithm achieves approximately equal performance to SMT because FIRE 2011 queries average length is 3 which is just half of the FIRE 2010 queries average length that is 6. The out of vocabulary word translation is the biggest challenge. Therefore, a context-based translation algorithm for the out of vocabulary word is proposed. The proposed approach with the continuous bag of word model achieves highest evaluation measures than the other models. It achieves the evaluation measures 0.8785 recall, 0.3340 P@5, 0.2880 P@10, and 0.3239 MAP for FIRE 2010, & 0.7365 Recall, 0.2380 P@5, 0.2680 P@10, and 0.1988 MAP for FIRE 2011 which are better than evaluation measures achieved by the baseline SMT, i.e., 0.8284 recall, 0.2840 P@5, 0.2300 P@10, and 0.2832 MAP for FIRE 2010, & 0.7084 recall, 0.2440 P@5, 0.2320 P@10, and 0.1885 MAP for FIRE 2011. An analysis for the number of out of vocabulary words is also done where the continuous bag of word model reduces the out of vocabulary words more effectively which are 2 (0.81%) for FIRE 2010 and 3 (1.73%) for FIRE 2011, in comparison to baseline SMT, i.e., 17 (6.93%) for FIRE 2010 and 7 (4.04%) for FIRE 2011.

In the dictionary, a source language word is associated with multiple translations which lead to a word translation disambiguation issue. Maximum probabilistic score and maximum association score like word co-occurrence & PMI are the conventional disambiguation methods. Statistical term similarity is an optimal translation disambiguation technique but its computation cost is very high. Word embedding provides the contextual word vectors which are semantically nearer to each other. The BWESG model learns the bilingual word vectors, further, the cosine similarity score is used to compute the nearest target language word for a source language word. The BWESG does not provide the appropriate translation due to the different sentence structure and the different number of vocabulary & stop-words, hence, a hybrid model is proposed which incorporates the BWESG and IBM model where IBM model limits the number of translations. The hybrid model with top-5 translation and Euclidean distance score achieves 0.1382 and 0.1085 MAP for FIRE 2010 and FIRE 2011 respectively which are lesser than the probabilistic dictionary based, therefore, it can be stated that bilingual word vector based model does not improve the CLIR performance.

Identifying whether a word needs a translation or transliteration technique is still a challenge, therefore, a named entity recognition based disambiguation approach is proposed to disambiguate the named entity terms from the dictionary term. A Stanford NER tool which uses a conditional random field machine learning algorithm is trained over the prepared named entity annotated data. Further, the trained Stanford NER is used to tag the queries. The proposed approach translates a query word in the favor of transliteration if that query word is associated with a named entity tag and its transliteration should be present in the top-n translations. The proposed named entity recognition based disambiguation combined with the translation induction algorithm achieves 0.2785 and 0.1911 MAP for FIRE 2010 and FIRE 2011 respectively, which is lesser than the MAP achieved by the translation induction algorithm alone. The named entity recognition based disambiguation actually degrades the CLIR performance because the term popularity decides that whether a term needs to be translated or transliterated. The effectiveness of named entity recognition based disambiguation is also discussed which shows that the NER alone is not able to solve the issue of improper translation/transliteration. Multiple translation term selection criteria depend on the query length as the average query length of FIRE 2011 queries is three, hence, with the inclusion of two best translation these queries achieve sufficient length and CLIR performance is enhanced but the inclusion of three best translations degrades the CLIR performance. In case of FIRE 2010, the average length of queries is six which is a sufficient length, hence, the inclusion of one more translation degrades the CLIR performance.

Maximum probabilistic score and maximum association score based disambiguation approaches individually are not sufficient, therefore, a maximum of average of words average probability and association score based disambiguation approach is proposed where association scores are calculated using the PMI or WE. WE are prepared by using the continuous bag-of-word, skip-gram, and log bilinear regression based recurrent neural network model. The proposed disambiguation approach is implemented at both of the word and query levels, among them, the word level disambiguation technique performs better than the query level. The proposed disambiguation technique at the word level with both of the PMI and WE based association score achieves almost equal evaluation measures and better evaluation measures than the baselines SMT and the proposed semantic morphological variant selection algorithm

Web resource based translation techniques provide dynamic translations but these resources need to be explored for Hindi-English CLIR. Wikipedia is widely used for the technically developed language due to the Wikipedia title and inter-wiki link attributes but Hindi is resource-scarce language. The proposed Wikipedia based CLIR approach addresses the issues of partially matched terms, wrong inter-wiki links, and ambiguous Wikipedia article. Therefore, it incorporates exactly matched, partially matched, and disambiguation modules. The proposed Wikipedia based approach achieves 0.2880 and 0.1772 MAP for FIRE 2010 and FIRE 2011 which is approximately equal to 0.2832 and 0.1885 MAP for FIRE 2010 and FIRE 2011 achieved by the baseline SMT. The web-based lexical resources like Hindi WordNet, Indo WordNet, COnceptNet, and online dictionaries based CLIR approaches are also proposed and analyzed. HWN & IWN attribute "English Synset"and CON attribute "TranslationOf"& "RelatedTo"are used for translation. Since the HWN, IWN, and CON are not lexically rich for Hindi-English translation, hence, they perform poor than the SMT and Wikipedia based CLIR approach. In HWN, IWN, and CON, the number of missing or unseen words are more compared to the SMT and Wikipedia. Online dictionaries return too many irrelevant translations where the best translation is selected by using the WordNet path similarity and that selected best translation pair is not correct due to the irrelevant translations. Apart from this, HWN & IWN, CON, and online dictionaries are unable to translate the named entities which are translated by the SMT and Wikipedia. HWN & IWN, CON are lexically poor and online dictionaries return too many irrelevant translations which reduce the MAP.

In this thesis, we propose the following approaches.

1. The term frequency model to address the dictionary coverage and retraining of the IBM model issues.
2. The translation induction algorithm to address the morphological irregularities.
3. The semantic morphological variant selection algorithm to address the translation mis-mapped and non-confident translation issues.
4. The context-based translation algorithm for the out of vocabulary words to address the out of vocabulary word translation issue.
5. The bilingual word vector and named entity recognition based disambiguation approaches.
6. The maximum of average of words average probability and association score based disambiguation approach.

7. The Wikipedia based translation approach which addresses the Wikipedia issues.

8. Web-based lexical resources based translation approaches.

The proposed semantic morphological variant selection algorithm performs better than the proposed term frequency model, translation induction algorithm, and baseline SMT. The proposed context-based translation algorithm effectively reduces the number of out of vocabulary words, due to that, it performs better than the baseline SMT. The proposed maximum of average of words average probability and association score based disambiguation approach combined with the semantic morphological variant selection algorithm performs better than both of the maximum probability and association score based disambiguation approaches. The proposed Wikipedia based translation approach effectively translates both of the dictionaries and named entity terms. It achieves approximately equal performance to the baseline SMT.

In future, the size of the parallel corpus will need to enhance which will increase the dictionary size and will reduce the number of out of vocabulary words, consequently, the CLIR performance will be increased. As per the conclusion of NER based disambiguation, word popularity based disambiguation approach may achieve better results so a word popularity based disambiguation algorithm may be developed to translate or transliterate a query word. According to the analysis of online resources in the perspective of Hindi language, Indo WordNet and ConceptNet are the poor resources. The efficiency of online dictionaries may be enhanced by applying a filtering technique to remove the irrelevant or less relevant translations. Other attributes of Wikipedia will be analyzed.

References

Adriani, M. (2000). Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information retrieval*, *2*(1), 71–82.

Ahmed, F., & Nürnberger, A. (2012). Literature review of interactive cross language information retrieval tools. *International Arab Journal of Information Technology*, *9*(5), 479–486.

Akhtar, S. S., Gupta, A., Vajpayee, A., Srivastava, A., & Shrivastava, M. (2017). Unsupervised morphological expansion of small datasets for improving word embeddings. In *International conference on computational linguistics and intelligent text processing* (pp. 1–15). Budapest, Hungary.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations.*

Bajpai, P., & Verma, P. (2014). Cross language information retrieval: In indian language perspective. *International Journal of Research in Engineering and Technology*, *3*(10), 46–52.

Ballesteros, L., & Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. In *International conference on database and expert systems applications* (pp. 791–801). Zurich, Switzerland.

Bhagavatula, M., GSK, S., & Varma, V. (2012). Language-independent named entity identification using wikipedia. In *Proceedings of the first workshop on multilingual modeling* (pp. 11–17). Jeju, Korea.

Bharadwaj, R. G., & Varma, V. (2011). Language independent identification of parallel sentences using wikipedia. In *Proceedings of the 20th international conference companion on world wide web* (pp. 11–12). Hyderabad, India.

Bhattacharyya, P., et al. (2017). Indowordnets help in indian language machine translation. In *International conference on information networking.* Da Nang, Vietnam.

Boretz, A., & Adam, A. (2009). Apptek launches hybrid machine translation software. *SpeechTechMag. com*, *2*.

Bradford, R., & Pozniak, J. (2014). Combining modern machine translation software with lsi for cross-lingual information processing. In *Information technology: New generations (itng), 2014 11th international conference on* (pp. 65–72). Las Vegas, NV, USA.

Brosseau-Villeneuve, B., Nie, J.-Y., & Kando, N. (2014). Latent word context model for information retrieval. *Information retrieval*, *17*(1), 21–51.

Chen, A., & Gey, F. C. (2003). Combining query translation and document translation in cross-language retrieval. In *Workshop of the cross-language evaluation forum for european languages* (pp. 108–121). Trondheim, Norway.

Chinnakotla, M. K., Ranadive, S., Damani, O. P., & Bhattacharyya, P. (2007). Hindi to english and marathi to english cross language information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages* (pp. 111–118). Budapest, Hungary.

Chu, C., Nakazawa, T., & Kurohashi, S. (2016). Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese–japanese wikipedia. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *15*(2), 10.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., & Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *Ijcai* (Vol. 9, pp. 1513–1518).

Dakwale, P., & Monz, C. (2017). Convolutional over recurrent encoder for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 37–48.

Das, A., Ganguly, D., & Garain, U. (2017). Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *16*(3), 18.

de Caseli, H. M., Ramisch, C., Nunes, M. d. G. V., & Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *Language resources and evaluation*, *44*(1-2), 59–77.

Diaz, F., Mitra, B., & Craswell, N. (2016). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (p. 367-377). Berlin, Germany.

Du, J., Hou, H., Wu, J., Shen, Z., Li, J., & Wang, H. (2016). Key research of preprocessing on mongolian-chinese neural machine translation. In *2016 2nd international conference on artificial intelligence and industrial engineering (aiie 2016).*

Duque, A., Martinez-Romo, J., & Araujo, L. (2015). Choosing the best dictionary for cross-lingual word sense disambiguation. *Knowledge-Based Systems*, *81*, 65–75.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 8.

El-Khair, I. A. (2006). Effects of stop words elimination for arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences*, *4*(3), 119–133.

Erdmann, M., Nakayama, K., Hara, T., & Nishio, S. (2009). Improving the extraction of bilingual terminology from wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *5*(4), 31.

Federico, M., & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval* (pp. 167–174). Tampere, Finland.

Gaillard, B., Boualem, M., & Collin, O. (2010). Query translation using wikipedia-based resources for analysis and disambiguation. In *Proceedings of the 14th annual conference of the european association for machine translation (eamt 2010), saint-raphaël, france. european association for machine translation.*

Gan, L., & Tu, W. (2014). Improving query expansion using wikipedia. In *International conference on management of e-commerce and e-government (icmecg), 2014* (pp. 143–146). Shanghai, China.

Ganesh, S., Harsha, S., Pingali, P., & Verma, V. (2008). Statistical transliteration for cross language information retrieval using hmm alignment model and crf. In *Proceedings of the 2nd workshop on cross lingual information access* (pp. 12–47). Hyderabad, India.

Ganguly, D., Leveling, J., & Jones, G. (2012). Cross-lingual topical relevance models. In *Proceedings of the international conference on computational linguistics 2012* (pp. 927–942). Mumbai, India.

Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. (2015). Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 795–798). Santiago, Chile.

Gao, J., & Nie, J.-Y. (2006). A study of statistical models for query translation: finding a good unit of translation. In *Proceedings of the 29th annual inter-*

national acm sigir conference on research and development in information retrieval (pp. 194–201). Washington, USA.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval* (pp. 96–104). Louisiana, USA.

Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2016). A convolutional encoder model for neural machine translation. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (p. 123-135). Vancouver, Canada.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 1243–1252). Sydney, NSW, Australia.

Green, S., Cer, D., & Manning, C. (2014). Phrasal: A toolkit for new directions in statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 114–121). Baltimore, Maryland, USA.

Gujral, B., Khayrallah, H., & Koehn, P. (2016). Translation of unknown words in low resource languages. In *Proceedings of the conference of the association for machine translation in the americas (amta), association for computational linguistics (acl)* (pp. 1–14). Austin, Texas, USA.

Gupta, S. K., Sinha, A., & Jain, M. (2011). Cross lingual information retrieval with smt and query mining. *Advanced Computing*, *2*(5), 33.

Hewavitharana, S., & Vogel, S. (2013). Extracting parallel phrases from comparable data. In *Building and using comparable corpora* (pp. 191–204). Sofia, Bulgaria: Springer.

Hosseinzadeh Vahid, A., Arora, P., Liu, Q., & Jones, G. J. (2015). A comparative study of online translation services for cross language information retrieval. In *Proceedings of the 24th international conference on world wide web* (pp. 859–864). Florence, Italy.

Huck, M., Tamchyna, A., Bojar, O., & Fraser, A. (2017). Producing unseen morphological variants in statistical machine translation. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (Vol. 2, pp. 369–375). Valencia, Spain.

Jagarlamudi, J., & Kumaran, A. (2007). Cross-lingual information retrieval system for indian languages. In *Workshop of the cross-language evaluation forum for european languages* (pp. 80–87). Budapest, Hungary.

Jain, A., Yadav, D., & Tayal, D. K. (2014). Ner for hindi language using association rules. In *Data mining and intelligent computing (icdmic), 2014 international conference on* (pp. 1–5). New Delhi, India.

Janarthanam, S. C., Subramaniam, S., & Nallasamy, U. (2008). Named entity transliteration for cross-language information retrieval using compressed word format mapping algorithm. In *Proceedings of the 2nd acm workshop on improving non english web searching* (pp. 33–38). Napa Valley, California, USA.

Jean, S., Lauly, S., Firat, O., & Cho, K. (2017). Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the third workshop on discourse in machine translation* (pp. 54–57). Copenhagen, Denmark.

Karimi, S., Scholer, F., & Turpin, A. (2011). Machine transliteration survey. *ACM Computing Surveys (CSUR)*, *43*(3), 17.

Kim, J., Nam, J., & Gurevych, I. (2012). Learning semantics with deep belief network for cross-language information retrieval. In *Proceedings of the international conference on computational linguistics 2012: Posters* (pp. 579–588). Mumbai, India.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of the international conference on computational linguistics 2012* (pp. 1459–1474). Mumbai, India.

Klyuev, V., & Haralambous, Y. (2012). Query translation for clir: Ewc vs. google translate. In *Information science and technology (icist), 2012 international conference on* (pp. 707–711). Hubei, China.

Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

Krishnarao, A. A., Gahlot, H., Srinet, A., & Kushwaha, D. (2009). A comparative study of named entity recognition for hindi using sequential learning algorithms. In *Advance computing conference, 2009. iacc 2009. ieee international* (pp. 1164–1169). Patiala, India.

Kunchukuttan, A., Mehta, P., & Bhattacharyya, P. (2018). The iit bombay english-hindi parallel corpus. In *Proceedings of the eleventh international conference on language resources and evaluation.* Miyazaki, Japan: European Language Resource Association.

Lam, W., Chan, S.-K., & Huang, R. (2007). Named entity translation matching and learning: With application for mining unseen translations. *ACM Transactions on Information Systems (TOIS)*, *25*(1), 2.

Larkey, L. S., Connell, M. E., & Abduljaleel, N. (2003). Hindi clir in thirty days. *ACM Transactions on Asian Language Information Processing (TALIP)*, *2*(2), 130–142.

Liu, X., Duh, K., & Matsumoto, Y. (2015). Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian and Low-resource Language Information Processing*, *14*(3), 11.

Loohach, R., & Garg, K. (2012). Effect of distance functions on k-means clustering algorithm. *International Journal of Computer Application*, *49*(6), 7–9.

Lu, W.-H., Chien, L.-F., & Lee, H.-J. (2004). Anchor text mining for translation of web queries: A transitive translation approach. *ACM Transactions on Information Systems (TOIS)*, *22*(2), 242–269.

Luo, J., & Lepage, Y. (2015). Handling of out-of-vocabulary words in japanese-english machine translation by exploiting parallel corpus. *International Journal of Asian Language Processing*, *23*(1), 1–20.

Mahapatra, L., Mohan, M., Khapra, M. M., & Bhattacharyya, P. (2010). Owns: Cross-lingual word sense disambiguation using weighted overlap counts and wordnet based similarity measures. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 138–141). Los Angeles, California.

Makin, R., Pandey, N., Pingali, P., & Varma, V. (2007). Approximate string matching techniques for effective clir among indian languages. In *International workshop on fuzzy logic and applications* (pp. 430–437). Camogli, Italy.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

Mathur, S., & Saxena, V. P. (2014). Hybrid appraoch to english-hindi name entity transliteration. In *Electrical, electronics and computer science (sceecs), 2014 ieee students' conference on* (pp. 1–5).

Meng, F., Lu, Z., Wang, M., Li, H., Jiang, W., & Liu, Q. (2015). Encoding source language with convolutional neural network for machine translation. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (pp. 20–30). Beijing, China.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). Lake, Tahoe.

Moen, H., & Marsi, E. (2013). Cross-lingual random indexing for information retrieval. In *International conference on statistical language and speech processing* (pp. 164–175). Tarragona, Spain.

Monti, J., Monteleone, M., Di Buono, M. P., & Marano, F. (2013). Natural language processing and big data-an ontology-based approach for cross-lingual information retrieval. In *International conference on social computing (socialcom), 2013* (pp. 725–731). Alexandria, USA.

Monz, C., & Dorr, B. J. (2005). Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval* (pp. 520–527). Salvador, Brazil.

Nagarathinam, A., & Saraswathi, S. (2011). State of art: Cross lingual information retrieval system for indian languages. *International Journal of Computer Applications*, *35*(13), 15-21.

Nasharuddin, N. A., & Abdullah, M. T. (2010). Cross-lingual information retrieval: State-of-the-art. *Electronic Journal of Computer Science and Information Technology: eJCIST*, *2*(1), 1-5.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, *41*(2), 10.

Negi, S. (2011). Mining bilingual topic hierarchies from unaligned text. In *Proceedings of 5th international joint conference on natural language processing* (pp. 992–1000). Thailand, Chaina.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval* (pp. 74–81). California, USA.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Doha, Qatar.

Pilehvar, M. T., & Collier, N. H. (2017). Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of 15th conference of the european chapter of the acl* (p. 388-393). Valencia, Spain.

Pingali, P., Jagarlamudi, J., & Varma, V. (2006). Webkhoj: Indian language ir from multiple character encodings. In *Proceedings of the 15th international conference on world wide web* (pp. 801–809). Edinburgh, Scotland.

Pingali, P., & Varma, V. (2005). Word normalization in indian languages. In *Proceedings of the international conference on natural language processing* *(icon–2005)* (p. 107). Kanpur, India.

Pingali, P., & Varma, V. (2006). Hindi and telugu to english cross language information retrieval at clef 2006. In *Cross-langauge evaluation forum (working notes).* Alicante, Spain.

Pingali, P., & Varma, V. (2007). Iiit hyderabad at clef 2007-adhoc indian language clir task. In *Cross-language evaluation forum (working notes).* Budapest, Hungary.

Prasad, G., & Fousiya, K. (2015). Named entity recognition approaches: A study applied to english and hindi language. In *2015 international conference on circuit, power and computing technologies (iccpct)* (pp. 1–4). Nagercoil.

Rao, P. R., & Sobha, L. (2010). Au-kbc fire2010 submission-cross lingual information retrieval track: Tamil-english. In *In the working notes of forum for information retrieval evaluation (fire)* (pp. 1–5). DAIICT, Gandhi Nagar, Gujrat, India.

Razmara, M., Siahbani, M., Haffari, R., & Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 1105–1115). Sofia, Bulgaria.

Redkar, H., Singh, S., Joshi, N., Ghosh, A., & Bhattacharyya, P. (2015). Indowordnet dictionary: An online multilingual dictionary using indowordnet. In *Proceedings of the 12th international conference on natural language processing* (pp. 71–78). Trivendrum, India.

Samantaray, S. (2012). An intelligent concept based search engine with cross linguility support. In *7th ieee conference on industrial electronics and applications (iciea), 2012* (pp. 1441–1446). Singapore.

Sanderson, M. (2010). Manning christopher d., raghavan prabhakar, schütze hinrich, introduction to information retrieval, cambridge university press. 2008. isbn-13 978-0-521-86571-5, xxi+ 482 pages. *Natural Language Engineering*, *16*(1), 100–103.

Saravanan, K., Udupa, R., & Kumaran, A. (2010). Crosslingual information retrieval system enhanced with transliteration generation and mining. In *Forum for information retrieval evaluation (fire 2010).* Kolkata, India: Citeseer.

Schönhofen, P., Benczúr, A., Biro, I., & Csalogány, K. (2007). Cross-language retrieval with wikipedia. In *Workshop of the cross-language evaluation forum for european languages* (pp. 72–79). Budapest, Hungary.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (p. 1715-1725). Berlin, Germany.

Sethuramalingam, S., & Varma, V. (2008). Iiit hyderabad's clir experiments for fire-2008. In *The working notes of first workshop of forum for information retrieval evaluation (fire)* (p. 1-6). Kolkata, India.

Shakery, A., & Zhai, C. (2013). Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. *Information retrieval*, *16*(1), 1–29.

Sharma, V., & Mittal, N. (2019). Refined stop-words and morphological variants solutions applied to hindi-english cross-lingual information retrieval. *Journal of Intelligent & Fuzzy Systems*, *36*(3), 2219–2227.

Sharma, V. K., & Mittal, N. (2016a). Cross lingual information retrieval (clir): Review of tools, challenges and translation approaches. In *Information systems design and intelligent applications* (pp. 699–708). Visakhapatnam, India: Springer.

Sharma, V. K., & Mittal, N. (2016b). Exploiting parallel sentences and cosine similarity for identifying target language translation. *Procedia Computer Science*, *89*, 428–433.

Sharma, V. K., & Mittal, N. (2016c). Exploiting wikipedia api for hindi-english cross-language information retrieval. *Procedia Computer Science*, *89*, 434–440.

Sharma, V. K., & Mittal, N. (2016d). Exploring bilingual word vectors for hindi-english cross-language information retrieval. In *Proceedings of the international conference on informatics and analytics* (p. 28). Pondicherry, India.

Sharma, V. K., & Mittal, N. (2017). Named entity identification based translation disambiguation model. In *International conference on pattern recognition and machine intelligence* (pp. 365–372). Bangalore, India.

Sharma, V. K., & Mittal, N. (2018a). A comparative study of online resources for extracting target language translation. In *Recent findings in intelligent computing techniques* (pp. 95–101). Goa, India: Springer.

Sharma, V. K., & Mittal, N. (2018b). Cross-lingual information retrieval: A dictionary-based query translation approach. In *Advances in computer and computational sciences* (pp. 611–618). New Delhi, India: Springer.

Shishtla, P., Ganesh, V. S., Subramaniam, S., & Varma, V. (2009). A language-independent transliteration schema using character aligned models at news 2009. In *Proceedings of the 2009 named entities workshop: Shared task on transliteration* (pp. 40–43). Bangalore, India.

Singh, D., Bhingardive, S., Patel, K., & Bhattacharyya, P. (2015). Detection of multiword expressions for hindi language using word embeddings and wordnet-based features. In *Proceedings of the 12th international conference on natural language processing* (pp. 295–302). Trivendram, India.

Sinha, R. M. K. (2011). Stepwise mining of multi-word expressions in hindi. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world* (pp. 110–115). Portland, USA.

Sorg, P., & Cimiano, P. (2012). Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, *74*, 26–45.

Su, C.-Y., Lin, T.-C., & Wu, S.-H. (2007). Using wikipedia to translate oov term on mlir. In *Ntcir-6* (p. 109-115). Tokyo, Japan.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., & Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, *25*(1), 4.

Tsvetkov, Y., & Wintner, S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, *18*(4), 549–573.

Ture, F., & Lin, J. (2014). Exploiting representations from statistical machine translation for cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, *32*(4), 19.

Turney, P. D. (2004). Word sense disambiguation by web mining for word co-occurrence probabilities. In *Third international workshop on the evaluation of systems for the semantic analysis of text, association for computational linguistics.* Barcelona, Spain.

Udupa, R., Jagarlamudi, J., & Saravanan, K. (2008). Microsoft research india at fire2008: Hindi-english cross-language information retrieval. In *Working notes for forum for information retrieval evaluation (fire) workshop.* Kolkata, India.

Udupa, R., Saravanan, K., Bakalov, A., & Bhole, A. (2009). "they are out there, if you know where to look": Mining transliterations of oov query terms for cross-language information retrieval. In *European conference on information retrieval* (pp. 437–448). France.

Vulić, I., De Smet, W., & Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, *16*(3), 331–368.

Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 363–372). Santiago, Chile.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiaoning, H., Peidong, W., Haoliang, Q., Muyun, Y., Guohua, L., & Yong, X. (2008). Using google translation in cross-lingual information retrieval. In *Proceedings of ntcir-7 workshop meeting* (pp. 16–19). Tokyo, Japan.

You, G.-W., Hwang, S.-W., Song, Y.-I., Jiang, L., & Nie, Z. (2012). Efficient entity translation mining: A parallelized graph alignment approach. *ACM Transactions on Information Systems (TOIS)*, *30*(4), 25.

Yu, F., Zheng, D., Zhao, T., Li, S., & Yu, H. (2006). Chinese-english cross-lingual information retrieval based on domain ontology knowledge. In *International conference on computational intelligence and security, 2006* (Vol. 2, pp. 1460–1463). Guangzhou.

Yuan, S. A., & Yu, S. N. (2007). A new method for cross-language information retrieval by summing weights of graphs. In *Fourth international conference on fuzzy systems and knowledge discovery, 2007. fskd 2007.* (Vol. 2, pp. 326–330). Haikou, Hainan, China.

Zhang, J., Sun, L., & Min, J. (2005). Using the web corpus to translate the queries in cross-lingual information retrieval. In *Proceedings of 2005 ieee international conference on natural language processing and knowledge engineering, 2005. ieee nlp-ke'05.* (pp. 493–498). Wuhan, China.

Zhang, S., Duh, K., & Van Durme, B. (2017). Selective decoding for cross-lingual open information extraction. In *Proceedings of the eighth international joint conference on natural language processing (volume 1: Long papers)* (Vol. 1, pp. 832–842). Taipei, Taivan.

Zhao, J., Huang, J. X., & Ye, Z. (2014). Modeling term associations for probabilistic information retrieval. *ACM Transactions on Information Systems (TOIS)*, *32*(2), 7.

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, *45*(1), 1.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1393–1398). Seattle, Washington, USA.

Appendices

Journal Publications

J1 Sharma V, Mittal N. Refined stop-words and morphological variants solutions applied to Hindi-English cross-lingual information retrieval. Journal of Intelligent & Fuzzy Systems. 2019 Jan 1;36(3):2219-27.

J2 Sharma VK, Mittal N. An Improvement in Statistical Machine Translation in Perspective of Hindi-English Cross-Lingual Information Retrieval. Journal of Computación y Sistemas. 2018 Dec 30;22(4).

J3 Sharma VK, Mittal N. Exploiting parallel sentences and cosine similarity for identifying target language translation. Journal of Procedia Computer Science. 2016 Jan 1;89:428-33.

J4 Sharma VK, Mittal N. Exploiting Wikipedia API for Hindi-English cross-language information retrieval. Journal of Procedia Computer Science. 2016 Jan 1;89:434-40.

J5 Sharma VK, Mittal N. Context-based Translation for the Out Of Vocabulary Words Applied to Hindi-English Cross-Lingual Information Retrieval, Journal of IETE Technical Review (SCIE-indexed, Accepted with minor revision).

Conference Publications

C1 Sharma VK, Mittal N. A Comparative Study of Online Resources for Extracting Target Language Translation. InRecent Findings in Intelligent Computing Techniques 2018 (pp. 95-101). Springer, Singapore.

C2 Sharma VK, Mittal N. Named Entity Identification Based Translation Disambiguation Model. InInternational Conference on Pattern Recognition and Machine Intelligence 2017 Dec 5 (pp. 365-372). Springer, Cham.

C3 Sharma VK, Mittal N. Cross-lingual information retrieval: A dictionary-based query translation approach. InAdvances in Computer and Computational Sciences 2018 (pp. 611-618). Springer, Singapore.

C4 Sharma VK, Mittal N. Exploring bilingual word vectors for Hindi-English cross-language information retrieval. InProceedings of the International Conference on Informatics and Analytics 2016 Aug 25 (p. 28). ACM.

C5 Sharma VK, Mittal N. Cross lingual information retrieval (CLIR): review of tools, challenges and translation approaches. InInformation systems design and intelligent applications 2016 (pp. 699-708). Springer, New Delhi.

Bio-Data

Vijay Kumar Sharma is a Ph.D. scholar at the Department of Computer Science & Engineering in Malaviya National Institute of Technology Jaipur, India. His research areas are Natural Language Processing, Text Mining, and Information retrieval. He has published several research papers in reputed international conferences and journals which were indexed in ACM, Springer, and Elsevier.