Tribhuvan University

Institute of Science and Technology

Creation of Parallel Corpus from Comparable Corpus (For English-Nepali Language Pair)

Dissertation

Submitted to

Central Department of Computer Science and Information Technology
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements for the Degree of
Master of Science in Computer Science and Information Technology

Submitted By

Hari Prashad Pant

November, 2011

Tribhuvan University

Institute of Science and Technology

Creation of Parallel Corpus from Comparable Corpus
(For English-Nepali Language Pair)

Dissertation

Submitted to

Central Department of Computer Science and Information Technology
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements for the Degree of
Master of Science in Computer Science and Information Technology

Submitted By

Hari Prashad Pant

November, 2011

Supervisor	**Co-Supervisor**
Assoc. Prof. Dr. Subarna Shakya	**Mr. Bikash Balami**

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Date: - ……..…………….

Supervisor's Recommendation

We hereby recommend that the dissertation prepared under our supervision by **Mr. Hari Prashad Pant** entitled **"Creation of Parallel Corpus from Comparable Corpus (For English-Nepali Language Pair)"** be accepted as fulfilling in partial requirements for the degree of Master of Science in Computer Science and Information Technology. In my best knowledge this is an original work in computer science.

Dr. Subarna Shakya
Associate Professor
Department of Electronics and
Computer Engineering
Institute of Engineering (IOE), Tribhuvan University
Pulchowk, Lalitpur, Nepal
(Supervisor)

Mr. Bikash Balami
Lecturer
Central Department of Computer Science
and Information Technology (CDCSIT)
Tribhuvan University
Kirtipur, Kathmandu, Nepal
(Co-Supervisor)

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

LETTER OF APPROVAL

We certify that we have read this dissertation work and in our opinion it is satisfactory on the scope and quality as a dissertation in the partial fulfillment for the requirement of Master of Science in Computer Science and Information Technology.

Evaluation Committee

Dr. Tanka Nath Dhamala
Head of Department
Central Department of Computer Science
and Information Technology
Tribhuvan University
Kirtipur

Dr. Subarna Shakya
Associate Professor
Department of Electronics and
Computer Engineering
Institute of Engineering (IOE),
Pulchowk, Lalitpur, Nepal
(Supervisor)

(External Examiner)

(Internal Examiner)

Acknowledgement

Finally, I am in a situation where I would like to express my heartfelt gratitude to my supervisor **Associate Professor Dr. Subarna Shakya**, Assistant dean of the Institute of Engineering (IOE), Pulchowk for providing me invaluable suggestions, encouragement and strong guidelines throughout the research period. Without his cooperation and suggestions, this dissertation could not be completed. I would also like to thank respected Head of the Department of Central Department of Computer Science and Information Technology (CDCSIT), TU, Assoc. Prof. Dr. Tanka Nath Dhamala for his kind help, encouragement and constructive suggestions regarding the dissertation.

I am also grateful to my course instructor and dissertation co-supervisor Mr. Bikash Balami, CDCSIT, TU for the guidance he provided throughout this research work. Without his scholarly vision, guidance, invaluable co-supervision and continuous encouragement, this dissertation would never have come in this form. Working under his co-supervision is really an unforgettable and he is my consistent source of inspiration not only in the field of research but also in my life.

This dissertation would not have been possible without the advice and support of my respected teachers, colleagues, friends and family members. Therefore, I would also like to express my special gratitude to my honourable teachers Prof. Dr. Shashidhar Ram Joshi, Prof. Sudarsan Karanjit, Mr. Min Bahadur Khati, Mr. Bishnu Gautam, Mr. Dinesh Bajracharya, Mr. Navaraj Poudel, Mr. Arjun Singh Saud, Mr. Jagdish Bhatta, Mr. Yog Raj Joshi, Mr. Ashim Ghishing of CDCSIT, TU, for providing me such a broad knowledge and inspirations within the time period of two years, which definitely helped me to live independently in this competitive world of IT.

My special thanks go to my family, especially my respected Father and Mother, my brothers Birendra Raj Pant, Janak Raj Pant, my sister Laxmi Pant, my sister-in-law Ratna Devi Pant and my dearest and nearest niece Punam Pant for their continuous encouragement, motivation and support. I really feel myself to be a lucky person for being a part of such a great caring and loving family. It was really a crucial time for me in order to get the higher education but this dream comes true really because of my Dad. His dedication, every time motivation, responsibility towards the future of we all brothers and sister, mental, physical and economical support is a source of inspiration for me to achieve this dream. So, I am really proud of you my dearest and lovely Dad.

Lastly, but not the least, I wish to thank to all my colleagues and friends especially Mr. Pravakar Ghimire, Mr. Roshan Silwal, Mr. Dinesh Kuamar Khadka, Mr. Karn Bahadur Bohara, Mr. Tej Bahadur Shahi for their valuable suggestions in one or another ways to complete my dissertation.

I have done my best to complete this research work but there may be some errors in my work, so any suggestion regarding the mistake of this dissertation will be always welcomed.

Hari Prashad Pant
Kathmandu, Nepal

Abstract

Statistical Machine Translation system is a great need of different multilingual countries like Nepal. But, one of the major bottlenecks in the development of Statistical Machine Translation systems for different language pairs is the lack of bilingual parallel data used for training such systems. Such parallel data contains the more or less exact translation of some source language sentence to the target language sentence. This is what we call parallel corpus used for training the Statistical Machine Translation System. There are such parallel corpora available relatively for few language pairs, for few domains and in limited size. Constructing such useful parallel data manually for different language pairs, different domains, and of sufficiently large size and good quality is really costly both human and monetarily.

It is parallel corpora may be the scarce resource, but comparable corpora are the rich, diverse resource that are readily available in several domains and language pairs. These corpora consists of a set of documents in two different languages which are not the exact translations of each other but contain somewhat related and similar information on the same topic. Such texts in large quantities can be found on the Web, good examples are online news agencies like CNN, BBC, etc.

In this dissertation, a method is proposed, which lets us to exploit such diverse resource: comparable corpora in order to extract the parallel data from them in an automated manner. The proposed method first tries to tokenize the documents at paragraph level and then candidate target sentences for each source sentence are obtained by using the sentence-length based method. After that the best match among the candidate sentences is made based on the bilingual dictionary. It has been observed that the quality and the number of words present in the bilingual dictionary enhance the accuracy of the model for the creation of parallel corpus from the comparable corpora.

Objectives

The major objectives of the research proposed in this thesis are:

- To find the best alignments for the source (English) sentence to the target (Nepali) sentence from the given comparable documents.
- To reduce the complexity required to generate the parallel corpus for English-Nepali language pair that is very much essential component in the field of SMT (Statistical Machine Translation) and various other fields of Natural Language Processing (NLP). It is not difficult to find such parallel texts written on paper, but of course, they are useless in machine translation studies. Thus this thesis is concentrated on finding good quality parallel texts on digital environment and hope that it will save lots of time for collection of good quality parallel data in future studies.
- To achieve significant improvements in the SMT system by adding our extracted corpus to the already available human-translated corpora.

Dedicated

To

My family members

&

Specially to my father Mr. Bhawani Datt Pant

Table of Contents

List of Tables

Details **Pages**

List of Figures

Details **Pages**

List of Abbreviations

AER	–	Average Error Rate
AI	–	Artificial Intelligence
APF	–	Armed Police Force
BBC	–	British Broadcasting Corporation
CALL	–	Computer-assisted Language learning
CL	–	Computational Linguistics
CLIR	–	Cross Language Information Retrieval
CNN	–	Cable News Network
EBMT	–	Example-Based Machine Translation
EM	–	Expectation Maximization
GDI	–	Gross Domestic Income
GNI	–	Gross National Income
GNP	–	Gross National Product
IR	–	Information Retrieval
LDC	–	Linguistic Data Consortium
LL	–	Language Learning
ME	–	Maximum Entropy
MT	–	Machine Translation
NLP	–	Natural Language Processing
OCR	–	Optical Character Reader
POS	–	Part of Speech
RBMT	–	Rule-Based Machine translation
SL	–	Source Language
SMT	–	Statistical Machine Translation
TL	–	Target Language
UN	–	United Nations
WSD	–	Word Sense Disambiguation

CHAPTER 1

1. Introduction

It is the true fact that Nepal is a multilingual, linguistically dense and diverse country with rich resources of information, where there are peoples who speak different local language (or mother tongue languages) like Nepali, Doteli, Maithili, Bhojpuri, Newari, Hindi, Tharu, etc and also follow different cultures. So there are different native languages with different no. of speakers. Not only Nepal, in most of the countries of the world, there are various languages with thousands of native speakers are being spoken and there is no any single dominant language spoken by the public. In this scenario, it require various textual materials like as school books, user manuals of different products, official documents, etc to be published and/or written in different local languages so that all peoples can equally get the desired education, information and knowledge. Therefore, in such situation knowledge of different languages is essential. Not only knowledge of different languages but also the faster and efficient automated natural language processing is the most essential thing. This is the most dominant factor for the requirement of Natural Language Processing (NLP) and Computational Linguistics (CL) which helps to solve the problem in an automated manner, efficiently and with a great speed.

Thus, among the various applications or fields of NLP, machine translation (MT) is applicable for handling such critical situations. Among the machine translation systems, the Statistical Machine Translation (SMT) system is so much popular to handle this kind of problems though there is Example Based Machine Translation System (EBMT) [19] or Rule Based Machine Translation (RBMT) system as well but it requires much more knowledge of different rules of linguistics [9; 19]. For the SMT systems, which are almost data-driven techniques, we require a huge amount of parallel data in the form of training data to perform the translation in a correct and efficient manner. Such a collection of data which is specially used to train or test the SMT systems is given a special name called parallel corpus (parallel corpora in plural) [4; 23; 24].

The methods based on data-driven techniques take a more important place in the research field of NLP these days, which require enough parallel corpora with high quality [4]. Parallel corpora are the foundational resource in data-driven NLP, which has a direct impact on the effectiveness of various technologies such as statistical machine translation, cross-lingual information retrieval, and automatic lexicon generation [4]. But it is disappointed that there are only few parallel corpora publicly available today and most of them are very small in size, are for narrow areas (domains), cover only few limited language pairs and are really out-of-date meaning that not updated with current new terminologies or the vocabulary [4; 12; 24; 25; 26]. It is very much costly both in time and human labor to construct such parallel corpora of large scale and high quality which constraints the increase of parallel corpora. However, the rapid development of the Internet and the extreme communication between different countries gives hope that we can construct such parallel corpora automatically by extracting information from the web as well [4].

Bilingual corpora are not always readily available for Nepalese languages. However, the author in paper [2] gives an overview of some of the major primary resources and applications developed in the field of NLP for the Nepali language and their prospective for building advanced NLP applications. The paper also discusses the different approaches used by the current applications in NLP and their coverage including their limitations as well.

Though enough monolingual data is available for most language pairs, it is the parallel corpus, well suited for everyday life and domain adapted translations is the sparse resource [30]. That is why the proposed thesis tries to generate such parallel corpus automatically from the comparable corpus, so that it could be used in the field of SMT systems for the English/Nepali language pair and could generate the different foundational resources needed for the multilingual NLP in the context of our country Nepal.

1.1. Natural Language Processing

Natural Language Processing is state-of-art technology that is in huge expansion these days. Natural language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of Artificial Intelligence (AI). NLP is a discipline that aims to build computer systems that will be able to analyze, understand and generate human speech as well. Therefore, NLP application areas are: Speech Recognition (speech analysis), Speech Synthesis (speech generation), Document processing as information extraction & summarization, Machine Translation, Dialog systems (typed and spoken) and many more. [3; 9; 19]

1.2. Corpus

A corpus is a large set of texts, where we assume the texts are stored electronically, in a given file format and character encoding, without any formatting information, eventually provided with metadata and/or linguistic annotation, which is useful for linguistic research [28]. Such corpus may be of different types like monolingual corpus, bilingual corpus or multilingual corpus depending upon the number of languages used in them. A monolingual corpus is a corpus of texts in one language. A bilingual corpus is a collection of texts in two different languages where each of one is translation of other. A multilingual corpus is a corpus of texts in more than one language [19; 28]

1.2.1. Parallel Corpus

A parallel corpus consists of document pairs that are more or less exact translations of each other. It is a bilingual corpus consisting of texts organized in pairs which are translations of each other, i.e. they include the same information (parallel texts) and especially are aligned at the sentence level [1; 13; 24]. It is a collection of sentences in two different languages. Among those two languages, one is considered as the Source Language (SL) and another is the Target Language (TL). The sentences in the TL are translations of sentences in the SL. The problem of sentence alignment means finding out which TL sentence is the translation of which SL sentence. An aligned parallel corpus is a collection of such pairs of sentences [31].

As we know, parallel corpus can be symmetrical and asymmetrical. Symmetrical parallel corpus normally contains one-to-one sentence correspondences between source and target languages. Whereas, asymmetrical parallel corpus contains a large proportion of one-to-zero/zero-to-one, one-to-many/many-to-one sentence correspondences in the corpus as well [5].

The most well-known parallel corpora are Europarl, JRC-Acquis multilingual parallel corpus, the OPUS corpus, etc [1]. Other available multilingual parallel corpora are developed in the framework of projects of Multilingual Corpora for Cooperation (MLCC), the Integrated European language data Repository Area

(INTERA2) eContent, SEEERAnet and so on [12]. Most of the parallel corpora available for only few language pairs and among them most of the parallel corpora contain English as one of the two languages in them.

1.2.2. Comparable Corpus

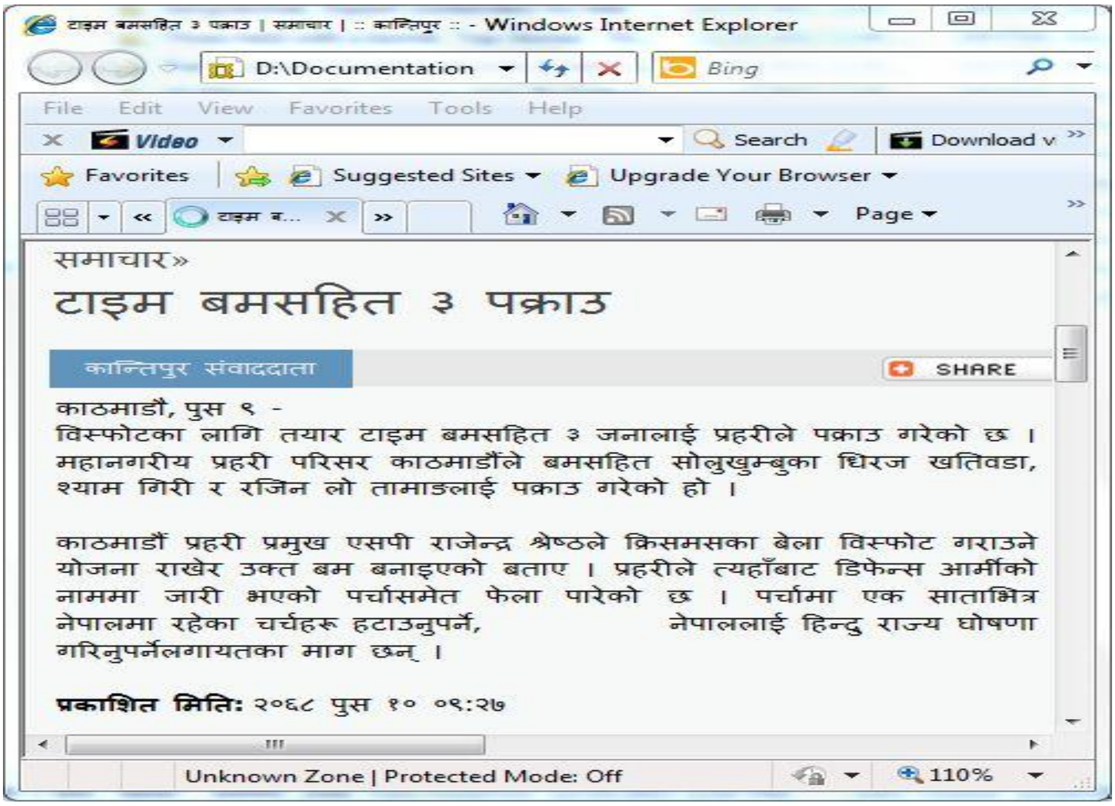
A comparable corpus is a bilingual corpus consisting of texts organized in pairs (comparable documents) which are only approximate translations of each other. In a comparable corpus, the document pairs are not exact translations but have similar vocabulary; that is texts in two different languages that are similar in content, convey overlapping information, but are not translations [1; 13]. The prototypical examples of comparable texts are two news articles in different languages which report on the same event. They are (most often) produced independently, but express overlapping content, and are therefore likely to contain some parallel data. Similarly the different publications which publish the same book in different languages, etc also can be considered as the sources of comparable corpora.

Potential sources of comparable corpora are multilingual news reporting agencies like Cable News Network (CNN), British Broadcasting Corporation (BBC), Al-Jazeera, etc or multilingual encyclopedias like Wikipedia, Encarta, etc [24; 25; 30]. Such comparable corpora are widely available from Linguistic Data Consortium (LDC[1]), in particular the Gigaword corpora, or over the web for many different languages and domains, e.g. Wikipedia [24; 25; 30]. Reliable identification of these pairs would enable the automatic creation of large and diverse parallel corpora [30]. However talking about Nepal, such comparable corpora can be taken from different news portals, news papers, books designed for some school curriculum that are published both in English and Nepali, user manuals of different products, etc. The degree of comparability of different documents varies, but we believe that the more comparable the corpora are, it is more useful for various NLP research task because it can be exploited to get the different desirable results like as parallel sentences, parallel documents, parallel words or sub-sentential data from it [32].

Figure 1.1 shows one example of comparable corpus. It shows the articles published on the Kantipur Publications Pvt. Ltd.'s website www.ekantipur.com for

[1] http://www.ldc.upenn.edu

both Kantipur Daily and The Kathmandu Post Daily on Dec-25, 2011. These articles report on arresting three peoples by the police along with time bomb.


टाइम बमसहित ३ पक्राउ | समाचार | :: कान्तिपुर :: - Windows Internet Explorer
D:\Documentation
Bing
File Edit View Favorites Tools Help
Video
Search
Download v
Favorites
Suggested Sites
Upgrade Your Browser
टाइम ब...
Page
SHARE
Unknown Zone | Protected Mode: Off
110%

समाचार»

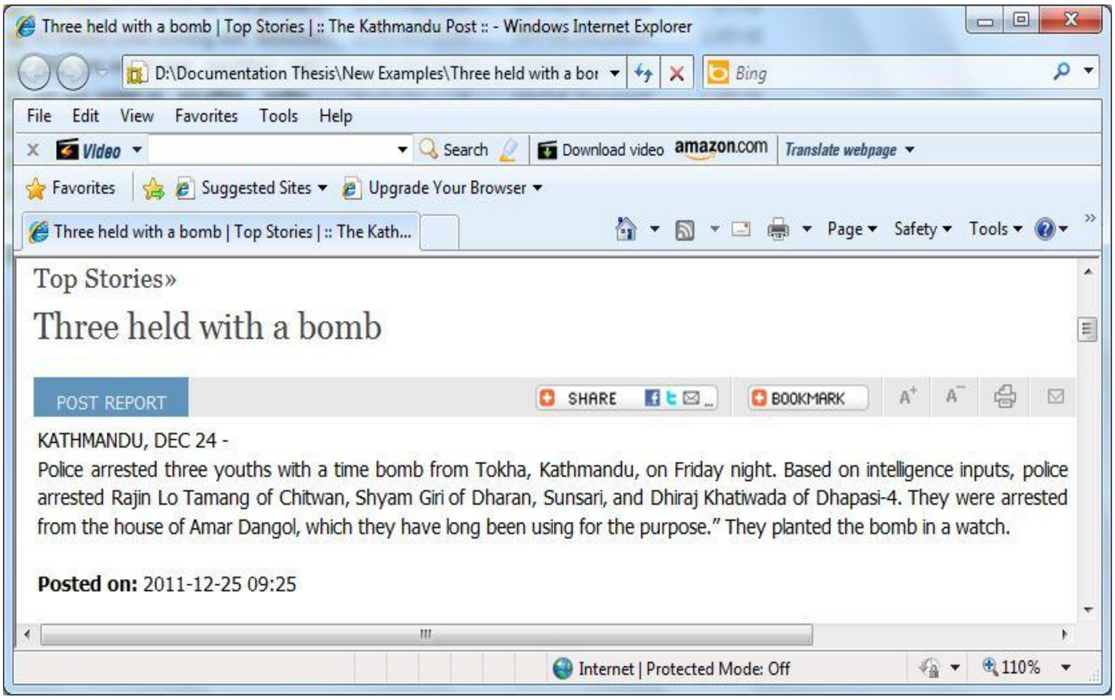
टाइम बमसहित ३ पक्राउ

कान्तिपुर संवाददाता

काठमाडौं, पुस ९ -
विस्फोटका लागि तयार टाइम बमसहित ३ जनालाई प्रहरीले पक्राउ गरेको छ । महानगरीय प्रहरी परिसर काठमाडौंले बमसहित सोलुखुम्बुका धिरज खतिवडा, श्याम गिरी र रजिन लो तामाङलाई पक्राउ गरेको हो ।

काठमाडौं प्रहरी प्रमुख एसपी राजेन्द्र श्रेष्ठले क्रिसमसका बेला विस्फोट गराउने योजना राखेर उक्त बम बनाइएको बताए । प्रहरीले त्यहाँबाट डिफेन्स आर्मीको नाममा जारी भएको पर्चासमेत फेला पारेको छ । पर्चामा एक साताभित्र नेपालमा रहेका चर्चहरू हटाउनुपर्ने, नेपाललाई हिन्दु राज्य घोषणा गरिनुपर्नेलगायतका माग छन् ।

प्रकाशित मिति: २०६८ पुस १० ०९:२७


Three held with a bomb | Top Stories | :: The Kathmandu Post :: - Windows Internet Explorer
D:\Documentation Thesis\New Examples\Three held with a bor
Bing
File Edit View Favorites Tools Help
Video
Search
Download video
amazon.com
Translate webpage
Favorites
Suggested Sites
Upgrade Your Browser
Three held with a bomb | Top Stories | :: The Kath...
Page Safety Tools
SHARE
BOOKMARK
Internet | Protected Mode: Off
110%

Top Stories»

Three held with a bomb

POST REPORT

KATHMANDU, DEC 24 -
Police arrested three youths with a time bomb from Tokha, Kathmandu, on Friday night. Based on intelligence inputs, police arrested Rajin Lo Tamang of Chitwan, Shyam Giri of Dharan, Sunsari, and Dhiraj Khatiwada of Dhapasi-4. They were arrested from the house of Amar Dangol, which they have long been using for the purpose." They planted the bomb in a watch.

Posted on: 2011-12-25 09:25

Figure 1.1: - Two comparable texts [News on same topic found in an online version of Kantipur Daily and The Kathmandu Post, Published Date – Sunday, Dec-25, 2011]

1.2.3. Domain-Specific Corpus

A domain-specific corpus (or in-domain corpus) is a corpus of texts from a given domain [28]. Such corpora provide very high amount of accuracy when the testing data comes from the same domain in which the system is trained but if used with some out domain data then the accuracy falls very down.

1.2.4. General-Domain Corpus

A general-domain corpus (or out-domain corpus) is a corpus of texts containing general language texts, i.e. texts from no any specific domain [28]. If such corpus is used for any SMT system, then the accuracy can be found in a quite good amount for any general domain test data, because the system is already trained with the general domain corpus. However, no such corpora are available even for the languages which are being spoken by most of the peoples of the world like for the languages English, Chinese, Arabic, Dutch, French, etc. So it is really difficult to have such corpora for the language which are popular among very few peoples and spoken by very few peoples like for the languages Nepali, Newari, Bhojpuri, Tharu, etc.

The corpora that are of interest for the work presented in this thesis are those which are not parallel, but do contain parallel data which can be of use for Statistical Machine Translation (SMT) after processing them. In order to obtain SMT training data, one must extract a set of parallel sentences from such comparable corpus, which serve as the parallel corpus needed for SMT systems [24]. The major task of this thesis is to find the good parallel data from the comparable corpus available in huge amount.

1.3. Motivation for the Dissertation

The lack of previous work on parallel texts extraction in an automated manner between English and Nepali is the most important motivation for making a research in this field. That is; the motivation behind this work is the lack of a very useful resource parallel text, which is essential and the most important for parallel translation. Parallel texts are useful essentially in any cross lingual NLP fields but they are most particularly important for SMT systems, where they provide the training data from which the system learns everything it knows about translation [1; 24]. Such collection of parallel text which is especially known as parallel corpus is a scarce resource and requires very much effort to generate manually for certain domain or for certain language pair. And to our knowledge, there are no any tools which generate the parallel corpus automatically for English-Nepali language pair that is why most of the parallel corpora which are being used for various tasks related to NLP, are generated manually. So, if it (parallel corpus) could be generated automatically, for the language pair English-Nepali, it would definitely help for various NLP related research works in Nepal. This fact really motivated me to perform the research in this particular topic.

The chart in figure 1.2 shows that the total number of speakers of different languages. The data about the different languages and the number of speakers of the language in different countries of the world can be found through Ethnologue[2].

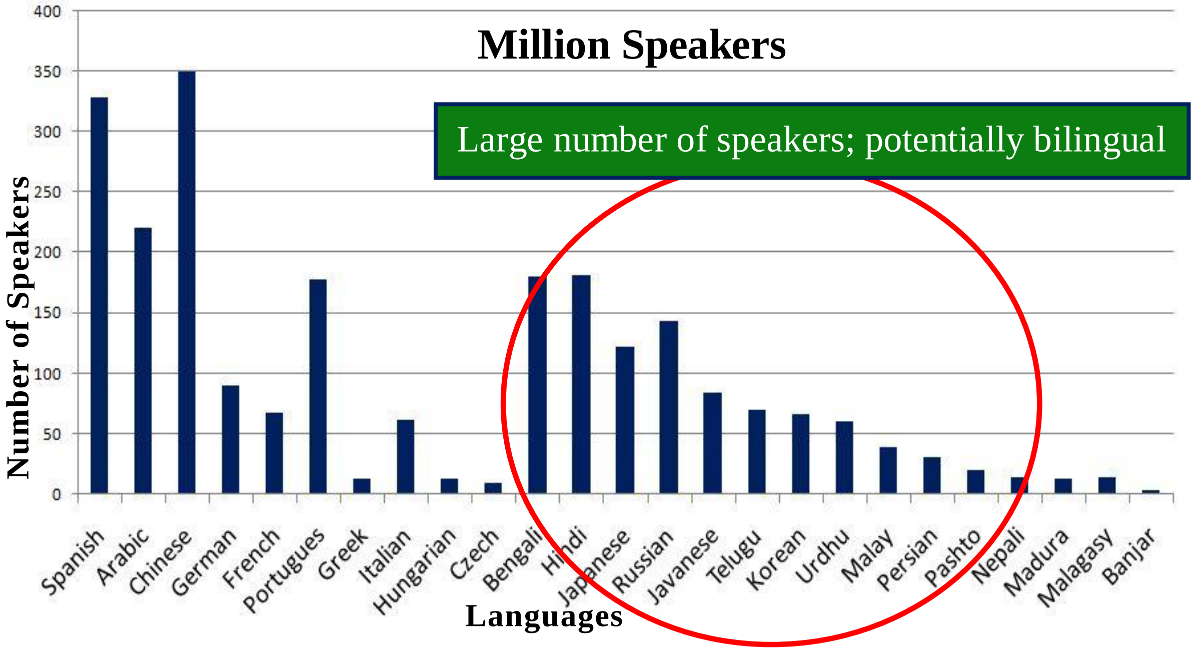


Figure 1.2: - Languages and Speakers [34; Ethnologue]

Similarly, the chart shown in figure 1.3 describes the total number of words of a language translated into English. It describes that even there are languages which are spoken by hundreds of millions of speakers don’t have parallel data available to the research community for the NLP related tasks especially SMT. It means that Nepali is one of the low-resource language for which there are no any parallel data available as well.

[2] http://www.ethnologue.com

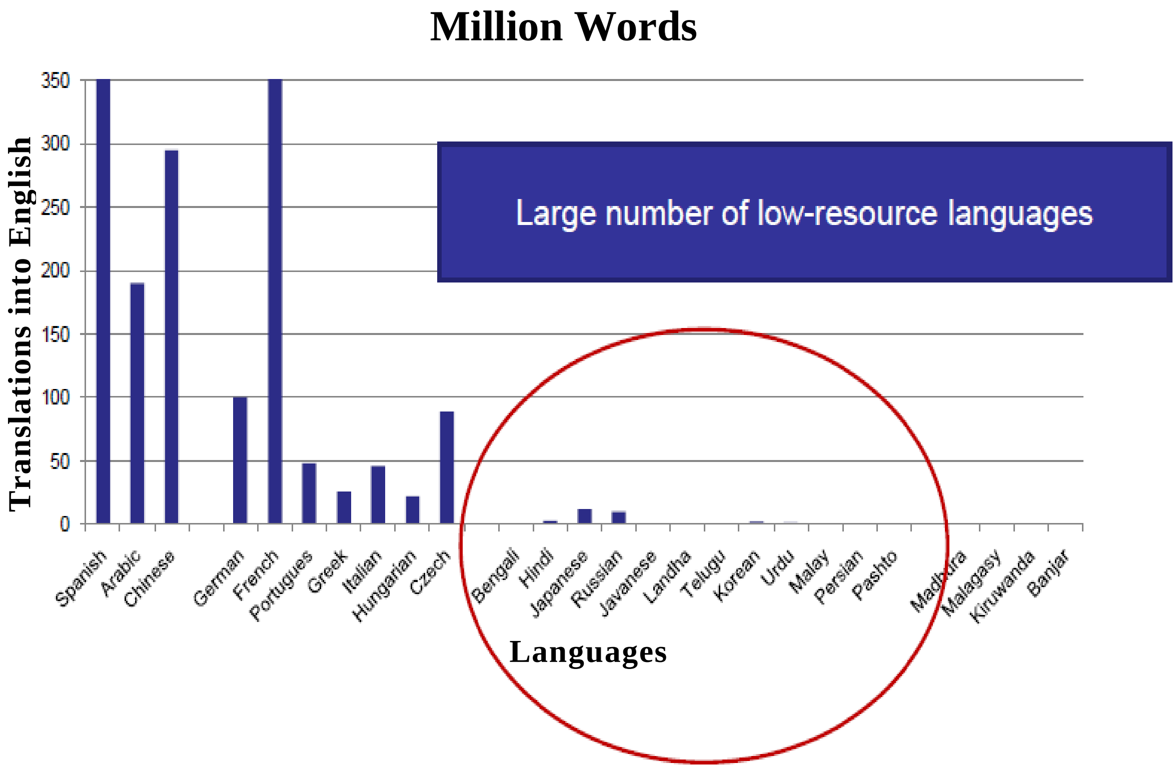


Figure 1.3: - Languages and Parallel Data Available [34]

Thus, the figure 1.2 and figure 1.3 shows the statistics about the languages which are being spoken by maximum number of speakers even don't have sufficient volume of parallel data. Also there are various languages which neither do have large number of speakers nor do have such parallel resources to the research community for their development.

1.4. Applications of parallel corpus

Sentence-level aligned bilingual parallel corpora hold a huge amount of linguistic information and this is the reason why they have several applications of high importance, especially in the field of NLP like Cross Language Information Retrieval (CLIR), Machine Translation (MT), Lexicography, Language Learning, Word Sense Disambiguation (WSD), etc. [18; 30; 35]. In order to be applicable or useful; those resources must be available in reasonable quantities, because most application methods are based on statistics rather than the exact rules of the linguistics. Unfortunately, such parallel corpora are often scarce resources: limited in size, language coverage and general domain data. The quality of the results depends a lot on the size of the corpora, which means robust tools are needed to build and process them [3; 4; 12; 19; 20; 24; 25; 26].

Some major application areas of parallel corpus are described briefly in the following section:

1.4.1. Machine Translation

The translation of some source language to target language automatically using computer is the machine translation. The major resource used for the automated translation is the parallel corpora, so they are used as a training corpus in various automated translation systems. Machine translation systems require bilingual lexica from which to get translations of terms. The well known data-driven approaches to machine translation are basically two: Example based machine translation (EBMT) [19] and statistical machine translation (SMT) [29; 30].

So far, the maximum use of parallel corpora is in the field of SMT, that is; the parallel corpora provide indispensable training data for SMT and the basis for SMT is basically word translations. The lack of bilingual parallel training data is one of the major bottlenecks in the development of SMT systems for most language pairs. Lately the use of bigram or trigram models [19] has been practiced because translations are not always one to one correspondence of words and also because such sequences of strings or phrases lead to sentences faster than the word to word translations [19].

1.4.2. Cross-language information retrieval

Cross-language information retrieval is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. Information Retrieval (IR) systems search and retrieve relevant documents based on a user query. Mono-lingual IR systems find documents only in the language of the query. For example a query in English that includes the term AIDS in English will not find possibly relevant information in other languages [18; 29]. Nowadays, Google like search engines have started embedding cross-language IR systems. It is the parallel corpora which are the most essential things for developing such IR systems. Information retrieval systems that retrieve documents from more than one language can use bilingual lexica by which query words are translated and the search is carried out in different languages. For example, a user may create their query in English but retrieve relevant documents written in Nepali.

Domain specific bilingual lexica, particularly, provide very useful support in getting the sense of words in a specified context. Such kind of bilingual dictionaries are simply generated by searching repeated co-occurrence [19; 29].

1.4.3 Word sense disambiguation

The task of word sense disambiguation (WSD) is to determine the correct meaning, or sense of a word in context, i.e. the process of determining the correct sense of a given word in context is called word sense disambiguation. Many words in natural languages have multiple meanings. It is important to identify the correct sense of a word before we take up translation, query-based information retrieval, information extraction, question answering, etc using that word. For example, the word *'bank'* can be *'any organization that provides various financial services, for example keeping or lending money'* as well as *'the side of river, canal, etc and the land near it'* [18].

Word sense disambiguation is the process of identifying which sense of a word is used in any given sentence, when the word has a number of distinct senses. It is a fundamental problem in natural language processing (NLP). Given a word-aligned parallel corpus, the different translations in a target language serve as the sense-tags of an ambiguous word in the source language. The outcome of word sense disambiguation of a source language word is the selection of a target word, which directly corresponds to word selection in machine translation. [19]

1.4.4. Computer-assisted language learning

Computer-assisted Language learning (CALL) refers to programs designed to help people learn foreign languages. CALL is an innovative approach of second language acquisition. Natural language processing has been enlisted in several ways in CALL: including carrying out bilingual text alignment so that a learner who encounters an unknown word in a second language can see how it was rendered in translation. Parallel texts can help the language learners to know the grammatical patterns of a language in advance. The grammar of two languages can be compared with the help of aligned texts so that the language learners can take benefit on language understanding. [19]

1.5. Thesis Organization

After giving the brief introduction of parallel and comparable corpora, their applications including the dire need of such parallel corpora in the context of a country like Nepal, the rest of the material presented in this thesis is structured as follows.

Chapter 2 gives a review on different types of alignments can be found on two comparable documents along with the definition of problem for the proposed study. The review of the literature related to the dissertation including the different related works and works done related to sentence alignment are described in detail in Chapter 3. This chapter provides the knowledge regarding the different methods used by different researchers to solve the problem of sentence alignment in both parallel corpora as well as comparable corpora.

The Chapter 4 provides an implementation overview of the different phases used to align the parallel sentences from the comparable corpus in order to generate the required parallel corpus. The empirical analysis of parallel sentence extraction model is made in Chapter 5 where the different test measures are taken in order to suggest the solution. The results of sentence alignment are shown using the bar diagrams along with different measures.

Finally, Chapter 6 focuses on the conclusion made and limitation and future works for the proposed model. Appendix A gives the testing performed on different comparable corpora taken for the proposed study with the source code for the implementation of the model in Appendix B.

CHAPTER 2

2. Background

Corpora are the term used on linguistics, which corresponds to a finite collection of texts in a specific language. A collection of text in a single language is called monolingual corpus, whereas the collection of documents in more than one language is called multilingual corpora. A bilingual corpus is a collection of texts in two different languages where each of one is translation of other. A parallel corpus is a collection of texts in different languages where one of them is the original text also called source language text and the other is the translation of original text called target language text. That is, parallel corpora are sentence aligned parallel texts between source language and the target language [1; 4; 24].

Parallel corpora are the scarce resources for most of the languages which are spoken by very few peoples in this world [4; 24]. Even for those languages which are being used frequently by majority of the people do not contain sufficient amount of general-domain parallel corpus. The domain-specific corpora as well are available in rare amount. Now, for English-Nepali, language pair this is further difficult to have such parallel corpora. According to the state of the art there are no methods that could enable the wholly automatic production of parallel corpora. The method we propose is based on statistical sentence alignment on comparable corpora based on both length-based and dictionary lookup approaches.

Parallel corpora are very useful resources for the NLP related tasks such as machine translation, linguistic studies, information retrieval systems development, lexicography, language learning, word sense disambiguation, etc [18; 30; 35]. In order to be useful, these resources must be available in reasonable quantities however they are not available easily for most of the languages with thousands/millions of native speakers; that is; these are the scarce resources not found in sufficient amount for various language pairs and are limited in size as well. The other problem with parallel corpora is that they cover only few domains and few languages [12; 24; 25]. But, the comparable corpora are readily available that, while not parallel in the strict sense, are closely related and convey the same information. Such parallel corpora can be generated through the alignment of source language sentences to the corresponding target language sentences of easily available comparable corpora. Thus the fundamental aspects of alignment are described in the following sections.

2.1 Alignment

Alignment is the task of identifying correspondences between the texts written in two different languages. Statistical Machine Translation is the data driven approach of finding the correspondence in two different languages. The aligned text will play the role of data in SMT. Hence text alignment plays an important role to make bilingual parallel corpora which will be very useful in SMT. Text alignment is done in different levels; it includes document alignment, paragraph alignment, sentence alignment, sub-sentence level alignment like word alignment, chunk alignment, phrase alignment, etc [19; 24; 28]. Document alignment means aligning the parallel documents from source language to target language, paragraph alignment means aligning the two documents at paragraph level. Sentential alignment (or sentence alignment) refers to alignment of sentences and sub-sentential alignment refers to alignment of sub-sentential elements, such as words, chunks, phrases, etc [24; 28].

2.1.1. Document alignment

Document alignment is the process of finding the document pair that is translation of one another from the collection of bilingual texts [19]. The two documents are declared to be aligned only when most of the sentences of the two documents are aligned properly between the two documents. So, sentence alignment is a declarative factor to identify whether the two documents are aligned or not.

2.1.2 Paragraph Alignment

Paragraph are often aligned sequentially, i.e. first paragraph of one language to first paragraph of another and so on. This might not be always true. Insertions, deletion, splitting and merging may appear on translating the paragraph of different language. Paragraph marker is used to separate the different paragraph on the document. Sometimes the use of the cognates and collocation is also used to recognize translation paragraphs. Aligned paragraph are further segmented into sentences [19].

2.1.3. Sentential alignment

Sentence alignment is the problem of determining which sentences are translations of each other. That is, the sentence alignment is the task of identifying correspondences between sentences in one language and sentences in the other language [16]. This task is a first step toward the more ambitious task finding correspondences among words. Prior to this step, sentence boundaries must be identified in both sides of the parallel documents, by using the process of sentence segmentation. A sentence-aligned parallel corpus is one of the two essential data resources required for training SMT systems whereas the other one is a TL monolingual corpus used for language modeling. In general, all possible alignment combinations are allowed: 1-1 when one sentence in one language fully corresponds to one sentence in the other language. 1-0 or 0-1 in case a sentence is not translated on the other side, or M-N when M>0 sentences on one side correspond to N>0 sentences on the other side [28]. Sentence alignment is not trivial because translators do not always translate one sentence in the input into one sentence in the output. Another problem is that of crossing dependencies, where the orders of sentences are changed in the translation.

Sentence alignment is usually applied on a parallel corpus where the parallel texts are assumed to be reliable translations of each other. In this thesis, this assumption cannot generally be made because the bilingual resources acquired here are comparable corpora, as the parallel texts in such corpora may not be the exact translations of each other (for example Wikipedia articles in multiple languages; they can but may not be accurate translations of each other, similarly the news published by different news agencies like CNN, BBC, etc. in different languages on the similar topic may not be exact translations of each other). So, the thesis tries to find the good parallel sentences among the different sentences present in the comparable corpora.

There are well established algorithms for aligning sentences across parallel corpora. Some are pure length based approaches, some are lexicon based, and some are a mixture of the two approaches

For sentence alignment, paragraph alignment is performed first, and then sentence within a paragraph are aligned. Paragraphs within a document can be aligned manually by inserting the paragraph marker within the document. The sentences within the paragraph can be aligned to any of the sentences in a cross level sentence alignment manner. The possible sentence alignment can be shown in the following figure 2.1:

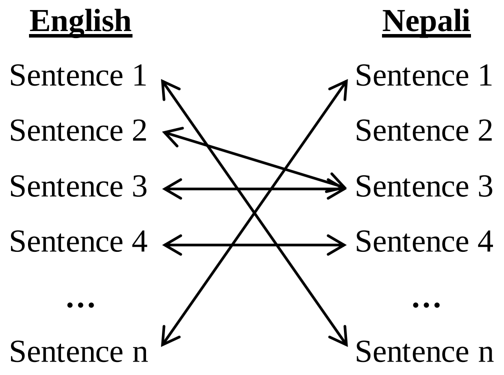


Figure 2.1: Possible Cross Level Sentence Alignment [5]

2.1.4. Sub-sentential alignment

The sub-sentential alignment is the task of identifying the parallel parts within the sentences that may be any word, chunk, phrase, etc. So the sub-sentential alignment can further be described as follows:

A. Word Alignment

The basic approach to sub-sentential alignment is word alignment. Word alignment is the identification of corresponding words in two source and target language sentences. It is a fundamental component of all modern SMT systems where it is used in order to extract a set of translation phrase pairs into a translation table. Word alignment is also employed in other NLP applications, such as translation lexicon induction and cross-lingual projection of linguistic information. Prior to word alignment, word boundaries must be identified in both sides of the parallel sentence by using the process of word tokenization. Word alignment is the natural language processing task of identifying translation relationships among the words in a bi-text, resulting in a bipartite graph between the two sides of the bi-text, with an arc between two words if and only if they are translations of one another.

Word alignment is typically done after sentence alignment of already identified pairs of sentences that are translations of one another. In general, all possible alignment combinations are allowed: 1-1 if one word on one side exactly corresponds to one word on the other side, M-N where M>0 words on one side correspond to N>0 words on the other side. 1-0 and 0-1 alignments are used when for a given word there is no translation equivalent on the other side of the parallel sentence (a word is deleted or inserted on the target side, respectively) [28]. The different words of a source sentence are aligned with the words at different places in the target sentence. This is clearly shown in figure 2.2:

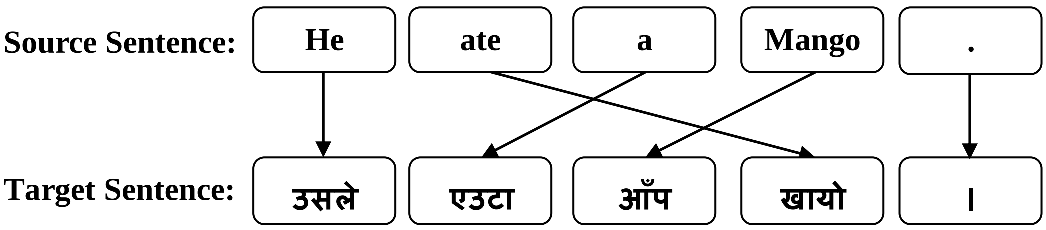


Figure 2.2: Word level alignment between English and a Nepali Sentence [34]

Here, in the above example, the alignment of words in both source (S) and target (T) sentences is given as:
Alignment: {S, T} = {1-1, 2-4, 3-2, 4-3}

B. *Chunk alignment*

Chunk alignment is the identification of corresponding chunks (syntactic constituents, such as noun phrases, verb phrases, etc.) in the two sentences. Chunking must be applied prior to this step. The assumption made during chunk alignment by using different methods is that the number of chunks in both sides of the parallel sentence is more or less the same and they can be aligned in a (more or less) 1-1 manner, although in general, 1-n chunk alignments are allowed too. Chunk alignment employed in SMT better captures local reordering and reduces the size of a translation table [19; 28].

2.2. Challenges of Parallel Corpus Creation

The problem of creation of parallel corpus actually is the problem of finding a source language sentence and its translation in the target language. In the thesis proposed, the problem of creation of parallel corpus is actually the problem of extracting the parallel sentences from the comparable corpora. This problem is ultimately the problem of aligning only the parallel sentences among the several sentences of the comparable corpus. Thus, the challenges of sentence alignment are also the challenges of parallel corpus creation.

The more we can enlarge a parallel bilingual corpus, the more we have made it effective and powerful. Providing such corpora demands special efforts both in seeking for as much already translated texts as possible and also in designing appropriate sentence alignment algorithms with as less time complexity as possible [38]. According to researchers in [17] the sentence alignment process has some important challenges, they are:

1. First of all, it is not the case that sentences always align one-to-one. Sometimes a sentence may be translated in more than one sentence in the other language or some part of a text may be deleted or some additional sentences may be added to the text so that we don’t have matches in the corresponding text. Even the existence of a small amount of such sentences results in remarkable deviations in the matching of sentence beads[3] in such situation. Because sentences do not always align one-to-one, the sentence alignment task is non-trivial [5]. The different sentence alignments may be treated as one-to-one, one-to-zero/zero-to-one, one-to-many/many-to-one or many-to-many and finding such alignments is really difficult.
2. Secondly, the structure of the different languages is not same. For example; an English sentence “I eat mango” has a translation into a Nepali sentence “म आँप खान्छु”. Here, English sentence has a basic structure of “Subject + Verb + Object” whereas Nepali sentence has a basic structure of “Subject + Object + Verb”. So, in real life, most of the texts have great inconsistencies with their translation such as the layout of texts, format differences, omission of some part of text and crossovers or inversions in text. Thus, the sentence alignment algorithms and programs must be devised in such a way to deal with such diverse situations and problems found in sentence alignment.
3. Finally, the accuracy always depends on the domain of the input text. For example an alignment program may give wonderful results when applied on a sports text but its success decline dramatically when applied on a scientific

[3] See section 3.2 for more details on beads

text. So, 100% accurate alignments are not possible even if the texts are “clean” and easy, which is affected by the input of the text.

The achievement of high accuracy with minimal consumption of computational resource is a common requirement for sentence alignment approaches. For a sentence alignment program to be called “ideal”, it should be fast, highly accurate and require no special knowledge about the corpus or the two languages [5; 17]. That is, the sentence alignment should also work in an unsupervised fashion as well as it should be language pair independent. By “unsupervised”, we denote methods that infer the alignment model directly from the data set to be aligned. “Language pair independence” refers to approaches that require no specific knowledge about the languages of the parallel texts to align [5]. In real world achieving all of these goals is a difficult task because of the following characteristics of real text [17]:

1) The paragraph boundaries in the real bilingual text are not represented in a similar manner in all cases. Which character is the paragraph separator, it is really difficult to identify in the bilingual text. For example in some texts have return character at the end of each line. So, it makes the case more difficult to determine if a newline means a new paragraph or not.
2) Sentence boundaries are also difficult to be identified properly because in many cases the same letter or symbol can be used as a sentence separator or not. That is, there is a variance in the ending and starting characters of the sentences at different places in the bilingual text. For example, a period (full stop) symbol is not only capable of detecting the sentence boundaries, for that we must also know the full meaning of the sentence. There are various terms like Mr., Ms., Prof., Dr., etc which also use period sign but doesn’t mean the end of sentences. In addition, the text may have many punctuation errors or symbols, pictures, etc. which make it impossible to determine the boundaries of sentences without knowing their meanings [17].
3) There might not be the case that the bilingual texts have exactly the same number of sentences and paragraphs. Some paragraphs or sentences may be merged into a larger paragraph or sentence because of the translator’s individual idea and so can’t say all the bilingual texts contain the same structure on sentence or paragraph representation.

4) Similarly, there might be different cases where the text may have crossing dependencies meaning that the order of sentences is changed in the translation.
5) In case of Nepali language, the different tools like Part of Speech (POS) tagger, Chunk tagger, stemmer, morphological analyzer, etc are not found in a fully functioning and freely available form, results the difficulties to find the accuracy in some extent to identify the parallel sentences. For a dictionary lookup approach, the dictionary mostly contains the root words and their translations, so if there is a stemmer, it can be used to find the root word from the given word and can be checked in the dictionary. It has the maximum probability to be found in the dictionary.
6) Also, the comparable data always could not be found in digital form. Large volume of such data can be found in the printed or hand written form. So we need some mechanisms to read such comparable data and present it in the digital form so that we could use it for our research task. Such major task can be done by using some automatic character recognizers like Optical Character Reader, etc. But, there is a chance that, some characters might not be read or identified properly by the OCR and lot more problems need to be faced. This definitely reduces the accuracy of the system. However, [8] is a research task related to handle such OCR errors in some extent.

2.3. Problem Definition

The parallel corpus is one of the most important resources in the research field of NLP especially for the SMT field. The major problem for this thesis is to extract the parallel sentences from the comparable corpus to generate a parallel corpus. The more we can enlarge a parallel bilingual corpus, the more we have made it effective and powerful. Providing such corpora demands special efforts both in seeking for as much already translated texts as possible and also in designing appropriate sentence alignment algorithms with as less time complexity as possible [38].

In this thesis, a method will be presented that enables automatic creation of parallel corpora by exploiting a rich, diverse, and readily available resource: comparable corpora for the language pair proposed. That is, the presented novel method will allow us to extract good-quality parallel data from such comparable collections for the English-Nepali language pair. The automatic extraction of bilingual parallel corpus from the comparable corpus mostly depends upon the parallel sentence alignment approaches.

We are given an English monolingual corpus and a Nepali monolingual corpus which are comparable to each other. We are assuming that both corpora contain the equal number of paragraphs in them. Suppose $EP_1, EP_2, ..., EP_n$ are the English paragraphs and $NP_1, NP_2, ..., NP_n$ are the Nepali paragraphs. Each such English and Nepali paragraphs may contain any number of sentences within it. In our proposed study let us suppose that English paragraph EP_1 contains *m* sentences as $ES_1, ES_2, ... ES_m$ and Nepali paragraph NP_1 contains *n* sentences as $NS_1, NS_2, ... NS_n$.

Now, for an English sentence ES_i $(1 \leq i \leq m)$ from an English paragraph EP_k, there are different similarity scores (probabilistic scores) to align with every Nepali sentence NS_j $(1 \leq j \leq n)$ of corresponding Nepali paragraph NP_k, where $1 \leq k \leq n$. Such alignment probabilities or similarity scores can be obtained by using a different strategy described as follows:

For an English sentence ES_i $(1 \leq i \leq m)$ from an English paragraph EP_k, we evaluate the similarity of sentence *ESi* with all the sentences *NSj* $(1 \leq j \leq n)$ of Nepali paragraph NP_k. Now, for each English word W_{Eng} in sentence ES_i, if one of its translations in the bilingual dictionary occurs in the sentence NS_j, the *translation count* will be added by 1 else it is kept as it is. Then the similarity of ES_i and NS_j is given by:

$$similarity(ES_i, NS_j) = \frac{translation \quad count}{\max\left(lengt\ h(ES_i), lengt\ h(NS_j)\right)} \text{ ------------- } 2.1$$

Where, *similarity*(ES_i, NS_j) is the sentence similarity of English sentence ES_i and the Nepali sentence NS_j, *length*(ES_i) and *length*(NS_j) are the count of words in ES_i and NS_j respectively, and the value of function *max(length(ESi),length(NSj))* is the bigger one between *length*(ES_i) and *length*(NS_j). The value of the *sentence similarity* obtained in this way is in between 0 and 1 and can be obtained different for different sentence pairs. The bigger the value is the more similar (parallel) the two sentences are [4].

Finally, the major problem of parallel sentence alignment is to find an alignment ***A***, such that it maximizes the probability over all possible alignments. This can be denoted as:

argmaxA P(A, S, T)

Where, *A* is the alignment, *P (A, S, T)* is the probability of being target sentence *T* as the alignment *A* of source sentence *S*. This probability is the similarity score obtained from equation 2.1. Thus, this is equivalent to

argmaxA similarity(ES_i , NS_j)

Thus, the alignment **A** of a source sentence and a target sentence which has the maximum similarity score is the best parallel sentence pair.

2.4. Stages in Parallel Corpus Creation

The basic steps used for finding the parallel sentences from the comparable corpus are shown in the architectural diagram of the proposed system in figure 2.3.

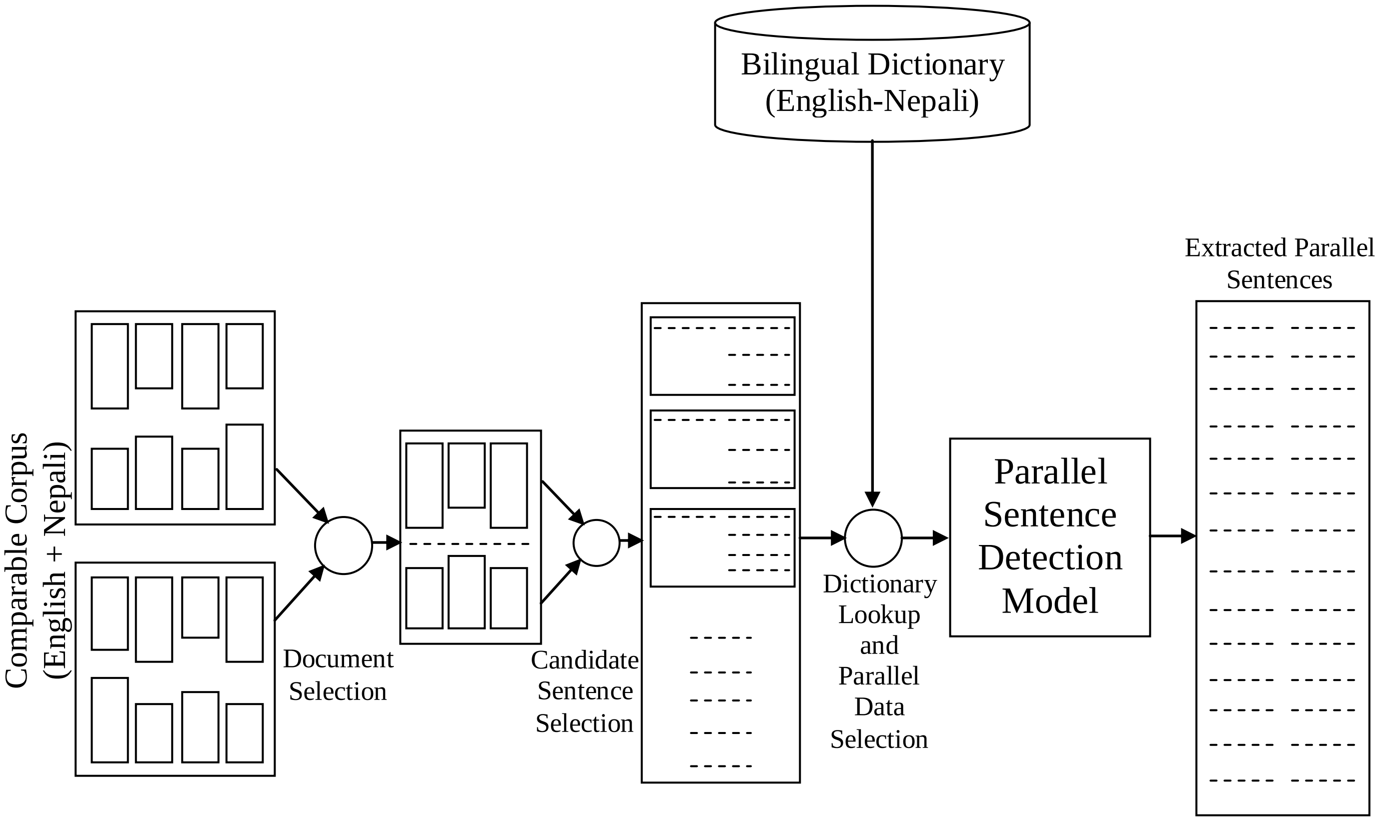

Bilingual Dictionary (English-Nepali)
Extracted Parallel Sentences
Comparable Corpus (English + Nepali)
Document Selection
Candidate Sentence Selection
Dictionary Lookup and Parallel Data Selection
Parallel Sentence Detection Model

Figure 2.3: The Architecture of the parallel sentence extraction system [24; 25]

The total work is divided into the following stages:

Document Selection

It includes selection of the two documents that are comparable to each other and contain some sentences about the same topic among a large no. of documents in the comparable corpus.

Candidate Sentence Selection

This step is used to select the candidate sentences among the selected two documents. As there are no one-to-one sentence alignments in the comparable documents, there may be some extra sentences that are not the translations of some other sentences present in the source or target corpus. So if we are able to extract some possible candidate sentences, it would reduce both time and computational cost to detect the correct alignments. Candidate sentences can also be determined based on the ratio of lengths of the two sentences. For our proposed model, the candidate sentences for a source sentence are the total number of sentences present in the corresponding paragraph of the target language.

Parallel Sentence Detection Model

It is used to compute the possible best translation of source language words to the target language words so that it will result in the best sentence-level alignment among the taken candidate sentences. For this, the bilingual dictionary is used to find the best translations.

2.5. Proposed Model for Parallel Corpus Creation

The fundamental idea about finding the solution to the problem of identification of parallel sentences from the comparable corpus as described in the section 2.3 is given in the proposed model shown in figure 2.4.

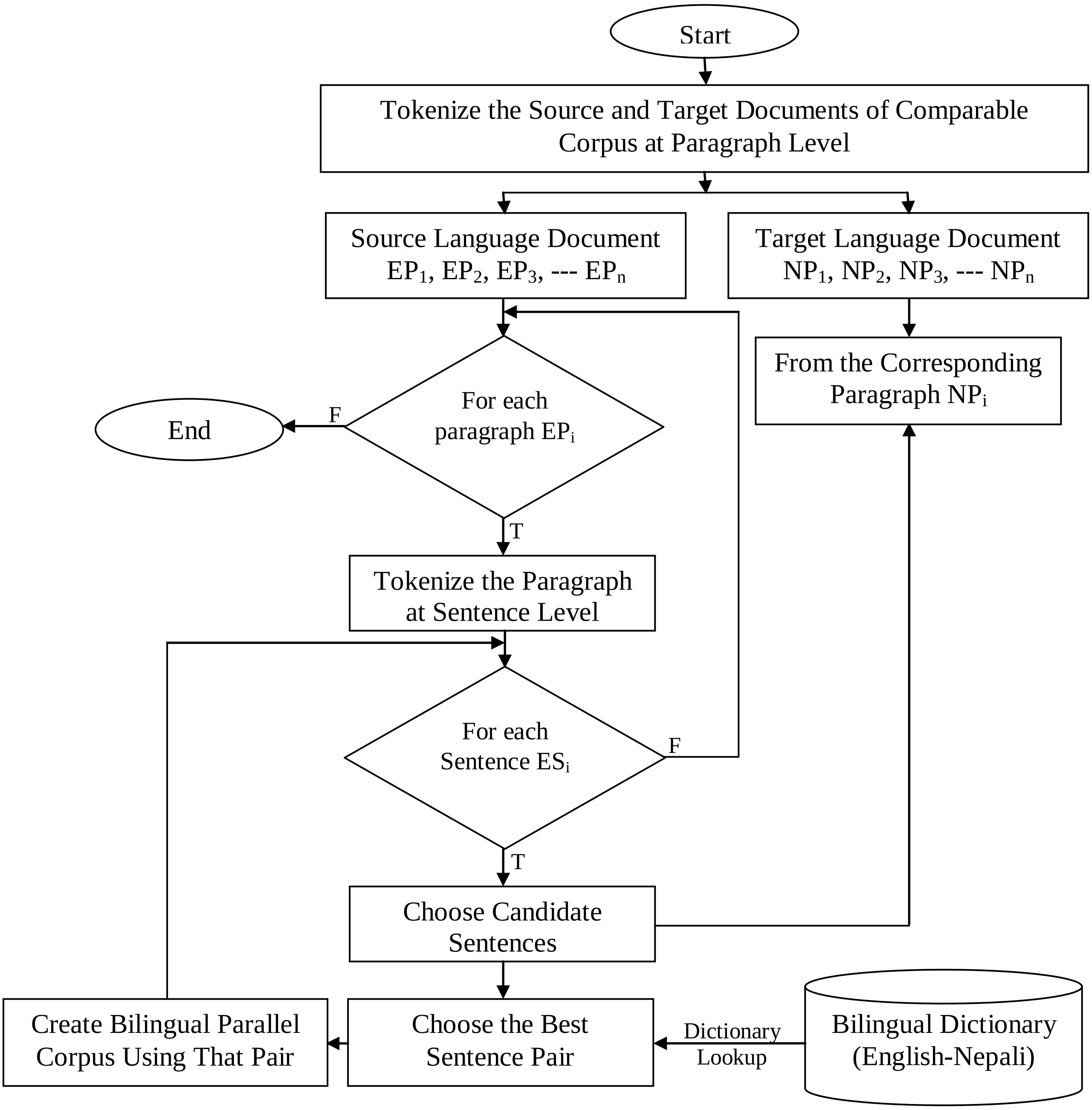

Start
Tokenize the Source and Target Documents of Comparable Corpus at Paragraph Level
Source Language Document EP$_1$, EP$_2$, EP$_3$, --- EP$_n$
Target Language Document NP$_1$, NP$_2$, NP$_3$, --- NP$_n$
From the Corresponding Paragraph NP$_i$
End
F
For each paragraph EP$_i$
T
Tokenize the Paragraph at Sentence Level
For each Sentence ES$_i$
F
T
Choose Candidate Sentences
Create Bilingual Parallel Corpus Using That Pair
Choose the Best Sentence Pair
Dictionary Lookup
Bilingual Dictionary (English-Nepali)

Figure 2.4: - Proposed model for parallel sentence extraction from comparable corpus

The best match among the candidate sentences is found by using the domain specific dictionary to calculate the highest score of parallelism between the sentences. After finding the best match between the source language sentence and the target language sentence, the problem of parallel corpus creation is to list the all aligned target sentences to the source sentences into two separate files.

CHAPTER THREE

3. Literature Review

3.1. Overview

For any statistical machine translation system, the size of the parallel corpus used for training the system is a major factor in its performance [32]. Such parallel data are readily available in large amount for some language pairs, such as Chinese-English, Arabic-English, French-English, etc, but for the most language pairs this is not the case. The size of parallel data not only influences the quality of translations, the domain of the parallel corpus also strongly influences the quality of translations produced from SMT systems [32]. Many parallel corpora are taken from the news domain, or from parliamentary proceedings [24; 25; 32]. Translation quality suffers when a system is not trained on any data from the domain it is tested on.

It is parallel corpora may be scarce resource, but comparable corpora, semi-parallel corpora are the rich resource that are readily available in several domains and language pairs. These corpora consist of a set of documents in two languages containing similar information. A detailed description of the types of non-parallel corpora is given in [13].

The basic aim of the research made in thesis is to extract parallel data from the comparable corpora where the parallel data can be any words, phrases, chunks, sentences, paragraphs and even the documents. But, the major goal is to find parallel sentences; this could be done by using different sentence-alignment algorithms. There are various methods developed for finding parallel sentences in parallel documents-to noisy parallel or comparable documents. Therefore, review of some of the sentence alignment algorithms is made in *section 3.2* and continues with the discussion of some approaches used for finding parallel sentences from comparable corpora in *section 3.3*.

3.2. Sentence Alignment Algorithms

Here in this section we consider various methods that are being used for aligning the bilingual corpora. Sentence alignment algorithms are mainly used to create a sentence level alignment between the two different documents. Most of the algorithms are able to align the sentences between parallel documents and some are able to align the sentences between non-parallel documents. It is a major idea to consider that the documents are a series of minimal translated units called *beads* [24]. A *(p, q)* bead, consists of *p* source language sentences and *q* target language sentences, where *p* and *q* can also be zero. Thus, a perfectly literal translation would be a sequence of *(1, 1)* beads; the existence of a source sentence with no any translations would be explained with a *(1, 0)* bead; the existence of a target sentence with no any source matching is represented with a *(0, 1)* bead. Similarly, one target sentence which translated two source ones would be represented by a *(2, 1)* bead and so on. Thus, the major goal of any sentence alignment algorithm is to find the best sequence of beads that generate the two documents under consideration [24].

There are many papers published at different places regarding the problem of sentence alignment and those papers propose different methods for aligning the sentences. But as far as the methodology they use is considered, all of these approaches or methods can be classified into following four categories:

3.2.1. Length-Based Approaches:

Length based approaches do not consider any semantic analysis of the different sentences in the process of sentence alignment. These approaches use the statistical methods for the task of sentence alignment between the different source and target language documents. These methods only consider the length of sentences while making decision for alignment. These approaches are based on the idea that long sentences will be translated into long sentences and short sentences into short ones [6; 15; 16]. The length based approach works remarkably well on language pairs with high length correlation, such as French and English. Its performance degrades quickly, however, when the length correlation breaks down, such as in the case of Chinese and English. These approaches are very simple and the length of different sentences is calculated in different manner in different approaches proposed by different researchers. Some of them measure the length on the basis of counting of the number of characters present in the sentences and some of them count the length on the basis of counting the number of words present in the sentences. That is, the length is either the number of characters or the number of words present in the sentences. Among the various length based algorithms Gale and Church Algorithm [15; 16] is the famous one. The Gale-and-Church Algorithm is basically dependent on the length of the sentence in terms of characters and the Brown's algorithm [6] is dependent on the length of the sentence in terms of words. These methods are very simple to align the sentences. Despite their simplicity, these methods work with a great accuracy. It means that these methods are simpler, highly accurate and efficient as well [17].

Some of the examples of length-based approaches are described in the following section:

A) **Brown et al., 1991 [6]:** This method aligns sentences with their translations in two parallel corpora based on a statistical technique. It is the method similar to Gale and Church (see section 3.2.1, part B), except that sentence lengths are computed in terms of words rather than characters and the comparison is made between the source sentences and target sentences [6; 22]. They have used the method to align several million sentences in the English-French Hansard[4] corpora. They have achieved accuracy in access of 99% in a randomly selected set of 1000 sentence pairs that are verified by hand. They have expected accuracy between 96% - 97% on other varieties of texts than they have tested [6; 17].

B) **Gale and Church, 1993 [16]:** The algorithm uses a simple statistical model of character lengths for aligning sentences of Source Language (SL) text and Target Language (TL) text. The algorithm uses sentence length measured in terms of number of characters to decide if some sentences in source language text are the alignment of some other sentences in the target language text [15; 16]. In fact, Gale and Church’s method is inspired from Brown’s method. The program uses the fact that longer sentences in SL tend to be translated into longer sentences in the TL and that shorter sentences tend to be translated into shorter sentences. The algorithm also uses the concept of Dynamic Programming which allows the system to consider all possible alignments and finding the minimum cost alignment effectively. The algorithm performs very well for the languages which are related like French and English. The method also aligns the sentences of parallel corpus. It has very good accuracy and gets a 4% error rate. It works best on 1:1 alignments and has high error rate on more difficult alignments [16; 17]. The method aligns the Canadian Hansard with such accuracy and the Hansard is the parallel data itself.

[4] The official written record of everything that is said in the parliament (in the British, Canadian, Australian, New Zealand or South African parliaments)

C) Wu, 1994 [37]: In this method Wu applies the Gale and Church's method to a corpus of parallel English and Cantonese (a version of Chinese) Text. Here the two languages were not similar meaning that they were unrelated; whereas Gale and Church have performed the experiment on related languages. But the results of the experiment are not much worse than Gale and Church's method which shows that the method can also be used on unrelated languages as well. The length based information was not sufficient to have much expected results, so to improve the accuracy; Wu uses *lexical cues*[5] in his method [17; 22; 37].

3.2.2. Location-Based Approaches:

These methods are also similar to length based approaches as they use statistical approach for finding the probable alignment between source and target language sentences. They use the fact that most of the times, beads of sentences in two different texts to be aligned, have similar positions. It means that if a sentence in the source text is in the middle of the text, its conjugate in the target text is probably in the middle of the text too. So, location of the bead is a major fact considered in this kind of location based methods. Some of the examples of location based approaches are now described in the following section:

A) Church, 1993 [8]: Church argues that length-based methods work well on clean text such as Canadian Hansard, where these methods have at least 96% accuracy but the accuracy of the methods may break down in various real-world situations where the input is noisy[6]. In such noisy inputs, the problem of finding paragraph boundaries, sentence boundaries is difficult due to noise to work well for sentence length-based approaches. So, Church's method is to make an alignment by using cognates[7] at the level of character sequences rather than at sentence/paragraph level [8; 17].

B) Aligning Bilingual Corpora Using Sentences Location Information [36]: - In this method the lexical information is not used for the main text but only

[5] A small corpus-specific bilingual lexicon

[6] due to OCR and/or unknown markup conventions

[7] words that are similar phonetically across languages

used for finding higher accuracy. The major property of this method is that it uses not only the length of sentences but also the length of texts, the length of upper and lower part of the candidate sentences, and some information like that to reinforce the effect of location of sentences in the text. For this reason, it can be said that it is an improved version of pure length-based method [17].

In this paper, authors describe a method for aligning bilingual corpora mainly based on the observation that the location of sentence pairs in two languages, are distributed in the texts similarly [17]. The Detailed description of the working of the method is found in [36]

3.2.3. Lexical (Lexicon-Based) Approaches:

These are the methods which take into account the lexical information about the texts for the purpose of sentence alignment. These methods try to overcome the weakness of the length based approaches by utilizing the lexical information from translation lexicons, and/or through the identification of cognates. These methods are based on the lexical resources such as bilingual lexicon (bilingual dictionary) or bilingual corpora. In most of these approaches, a bilingual corpus is used to find the best match of the content words in one text with their correspondences in the other text and use these matches as anchor points in the sentence alignment process. In some lexicon based methods, instead of content word pairs, cognates which are the words in language pairs that resemble each other phonetically are used for determining the beads of sentences. Thus, lexicon-based approaches give better alignment results than the sentence length-based approaches. However, the lexicon-based approach has the limitation of computational cost meaning that they take longer time than the length-based approaches. Some of the examples of lexicon-based approaches are given in the following section:

A) Kay & Roscheisen, 1993 [21]: They start their iterations by the assumption that the first and last sentences of the texts align [17; 21]. These are the initial anchors. Then, until most sentences are aligned:

1) Form an envelope of possible alignments.
2) Choose pairs of words that tend to co-occur in these potential partial alignments.

3) Find pairs of source and target sentences which contain many possible lexical correspondences. The most reliable of these pairs are used to induce a set of partial alignments which will be part of the final result.

B) Chen, 1993 [7]: S.F. Chen describes a fast algorithm for aligning sentences with their translations in a bilingual corpus. In this method, Chen constructs a simple statistical word-to-word translation model as he goes along the alignment process [7]. Here, the author tried to find an alignment that maximizes the probability of generating the corpus with this translation model. That is; the best alignment is the one that maximizes the likelihood of generating the corpus give the translation model. This best alignment is found by using the dynamic programming in language independent manner. The algorithm achieved an error rate of approximately 0.4% on Canadian Hansard data, which was a significant improvement over previous results. The method gave better accuracy that the sentence length-based method but the method was “tens of times slower than the Brown and Gale algorithms” [7, p.15]. It was perhaps first shown by Chen in [7] that word correspondence based models can be used to produce higher-accuracy sentence alignment than sentence length-based models alone.

3.2.4. Mixed (Hybrid) Approaches:

Now days, the sentence alignment is carried out by different researchers by using the combination of both length-based approaches and lexicon-based approaches. These methods use the length based approaches to find the most of the candidate sentences in the target document for a source sentence. Then, the lexicon-based approaches are used to find the best match among these provided by the length-based approach [1; 24; 25; 26]. Some of the examples of this method are mentioned below:

A) Moore 2002 [22]: Robert C. Moore presents a method for aligning sentences with their translations in a parallel bilingual corpus. The method adapts and combines the sentence length-based approaches and word correspondences based approaches to achieve high accuracy at a modest computational cost, and the method requires no knowledge of the language or the corpus beyond division into words and sentences [22].

Moore proposes a multi-pass search procedure where sentence length-based statistics are used in order to extract the training data for the IBM Model-1[8] translation tables. The acquired lexical statistics are then combined with the sentence-length based model in order to extract 1-to-1 correspondences with high accuracy. The method is highly accurate and is a language independent. The problem with the method is the slowness of the algorithm; as the extraction of a bilingual corpus from the texts at hand is not a straightforward and cheap operation [17; 22].

3.3. Methods for Finding Parallel Sentences from Comparable Documents

In recent years, there have been several approaches developed for obtaining parallel sentences from non-parallel, or comparable data. In most previous work on extraction of parallel sentences from comparable corpora, some coarse document-level similarity is used to determine which document pairs contain parallel sentences [32]. In this section, we discuss about various methods used for the extraction of parallel sentences from the comparable corpus:

3.3.1. Creation of Parallel Corpus from Comparable Corpus [1]: - In the paper [1], they present an approach for automatic creation of Hindi-Panjabi[9] parallel corpus from the comparable corpus. They used two-pass approach to align the two comparable documents at sentence level. The method described in the paper [1] uses various alignment parameters such as sentence length to linguistic parameters like syntactic level similarity. In the first pass they use the sentence length for selecting the candidate sentences. In the second pass they tried to find the best possible alignments of the sentences using the different tools available for the languages considered in [1].

3.3.2. Mining Parallel Text from the Web based on Sentence Alignment [4]: - In paper [4] the researchers brought a novel strategy to automatically fetch parallel text from the web. In the method presented, it first downloads the web pages from certain hosts and then the candidate web pages are prepared among such downloaded pages.

[8] Used to find the word level alignment

[9] Hindi and Panjabi are the most widely spoken languages in India.

The candidate web pages are then evaluated to find the best alignment of sentences within them. They also tried to evaluate the similarity of the two web pages based on the similarity of the aligned sentences. The experiments show the satisfactory performance on the multilingual web sites [4].

3.3.3. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment [32]: - The method proposed in [32] advances the state of the art in parallel sentence alignment by modeling the document level alignment. They have demonstrated that Wikipedia is a useful resource for mining parallel data. They retrieved the sheer volume of parallel data from Wikipedia which is a somewhat surprising result in the light of Wikipedia's construction. They proposed a novel approach which results in improved average precision while retaining simplicity and clarity in the models. They have shown that substantial gains can be achieved by using an induced word-level lexicon in combination with sentence extraction [32].

3.3.4. Adaptive parallel sentences mining from web bilingual news collection [38]: - This is an extension of various algorithms presented in section 3.2. In this method the authors combine both sentence length-based models and lexicon-based models under a maximum likelihood criterion. This method first tries to find the parallel document pairs and then sentence align them. The researchers compute the score of a document pair by defining a generative model of a target document as per given source document. They consider all document pairs as parallel whose score is above certain threshold. Now, the sentence-alignment is performed on the obtained parallel document pairs and the score of a sentence bead is computed as a combination of length information and IBM Model 1 lexical score. Zhao and Vogel evaluate the extracted sentences by showing that they improve the accuracy of automatically computed word alignments [24; 38].

3.3.5. Reliable measures for aligning Japanese-English news articles and sentences [33]: - It is also similar to the approach given by Zhao and Vogel in [38]. Here, they use the BM25[10] Score to find the parallel document pairs. After that, they align the obtained parallel document pairs at sentence-level by using a score based on

[10] **BM25** is a ranking function used by search engines to rank matching documents according to their relevance to a given search query

the lexical information i.e. the number of translated words in each sentence bead. Then, they use the sentence similarity scores to compute new document matching scores and the new document scores to compute new sentence similarity score and show that this gives the parallel data of higher quality. They evaluate their aligned sentences by analyzing randomly sampled alignments [24; 33].

3.3.6. Improving machine translation performance by exploiting non-parallel corpora [25]: - Munteanu & Marcu in [25] present a method for finding the parallel sentences from comparable, non-parallel corpora. In this method, they train a maximum entropy classifier such that, given a pair of sentences, can reliably determine whether or not the given pair of sentences are translations of each other. They have extracted parallel data from large Chinese, Arabic, and English non-parallel newspaper corpora by using the method [25]. They begin by selecting similar document pairs from two large monolingual corpora containing non-parallel documents. From the collected such pairs, they generate all possible sentence pairs and pass them through a word-overlap-based filter, which gives possible candidate sentence pairs. Now, such candidate sentence pairs are presented to a maximum entropy (ME) classifier that decides whether the sentences in each pair are the mutual translations of each other. The quality of extracted data is evaluated by showing that it improves the performance of a state-of-the-art statistical machine translation system. The method proposed in [25] is a language independent method and can be adapted for any language pairs.

3.3.7. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM [13]: - They pair each source document with all target documents whose similarity score is higher than a certain threshold value, where the similarity score is calculated by using cosine similarity measure. They work on very-non parallel corpora. They generate all possible sentence pairs from each document pairs and select the best ones based on a threshold value. Using the extracted sentences they learn a dictionary and iterate over with more sentence pairs [13].

CHAPTER 4

4. Implementation

4.1. Phases of Implementation

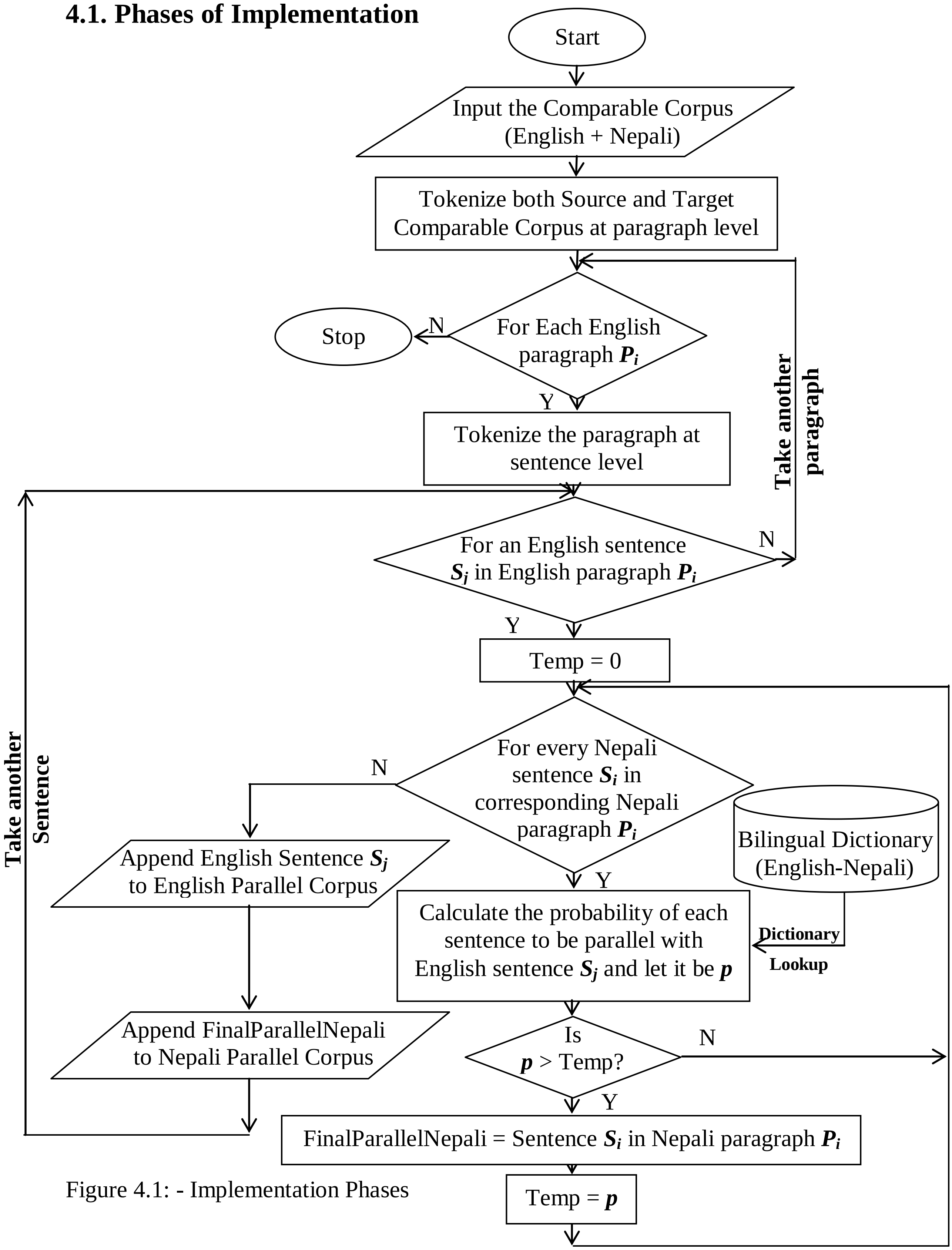

Start
Input the Comparable Corpus (English + Nepali)
Tokenize both Source and Target Comparable Corpus at paragraph level
For Each English paragraph P_i
N
Stop
Y
Take another paragraph
Tokenize the paragraph at sentence level
For an English sentence S_j in English paragraph P_i
N
Y
Temp = 0
Take another Sentence
For every Nepali sentence S_i in corresponding Nepali paragraph P_i
N
Y
Bilingual Dictionary (English-Nepali)
Append English Sentence S_j to English Parallel Corpus
Calculate the probability of each sentence to be parallel with English sentence S_j and let it be p
Dictionary Lookup
Append FinalParallelNepali to Nepali Parallel Corpus
Is p > Temp?
N
Y
FinalParallelNepali = Sentence S_i in Nepali paragraph P_i
Temp = p

Figure 4.1: - Implementation Phases

4.2. Description of Implementation Phases

The process shown in figure 4.1 is a novel approach used to find the scarce parallel data from the highly available form of comparable data. This model is a statistical model for finding the parallel corpus from the comparable corpus. In the proposed model, at first both the source and target comparable corpora are read. Then the source corpus i.e. the English corpus and the target corpus i.e. the Nepali corpus are broken down at paragraph levels. After this, the paragraph of source corpus is further segmented into different sentences by using the process or sentence segmentation and same task is also done for the paragraph of Nepali corpus. Furthermore, the sentences of source and target paragraphs are tokenized into words by using the process of word tokenization. This lets us to use the bilingual dictionary for finding the best sentence level alignments to generate the parallel corpus. Now, for each source sentence of a paragraph, we try to find the parallel sentence in the target language paragraph. For this, we first try to find the candidate sentences from the target paragraph. For this, all the sentences from target paragraph are treated as the candidate sentences. Now, among those sentences, the best parallel sentence is obtained by using the bilingual dictionary for both English and Nepali language pair. During the dictionary lookup to find the best parallel sentence, the individual words of source sentence are analyzed in the dictionary and the corresponding translation word is searched in the target language sentence.

It is supposed that the sentence(s) *ES* in English language and the sentence(s) *NS* in Nepali language are the candidate sentences from the corresponding paragraphs of the actual comparable corpora used in the experiment. Then a strategy is designed to evaluate the similarity of the sentences *ES* and *NS*. For each word W_{Eng} in the sentence *ES*, its translation word is identified from the bilingual dictionary and searched in the sentence *NS* to find the similarity score of the two sentences. Now, the translation probability (similarity score) of parallelism of that source sentence with all other candidate sentences is obtained. In this way, the translation probability for different Nepali sentences can be found different. Thus, the Nepali sentence *NS*, whose translation probability is found to be the highest among all the other candidate Nepali sentences, is treated as the best parallel Nepali sentence for the English sentence *ES*. Finally all such best pairs are written or collected in the form of parallel corpus.

CHAPTER 5

5. Testing and Analysis

In this chapter, the description of different data collection mechanisms used and the different sources that were used to collect the data related to the proposed study are discussed. Besides that the different testing performed on the collected data using the proposed model including the analysis measure for different parameters are also described here in this chapter. So, the *section 5.1* describes the different data collection related activities and *section 5.2* describes the different testing and verification of the collected data in order to obtain the expected goal.

5.1 Data Collection

The bilingual dictionary and test data are collected or developed to perform the parallel sentence alignment i.e. to extract the parallel sentences from the comparable documents (or comparable corpora). In this work, the different English-Nepali comparable texts are collected from different news websites, news papers, online magazines, and translated course books for secondary level education following the curriculum of Ministry of Education, Government of Nepal. All these data from such resources are collected for finding the major goal of this thesis, which is to find the parallel training data required for SMT, automatically from the different comparable data collected. Such collected data is either in the form of bilingual dictionary to find the best sentence alignments or in the form of testing data used to identify the accurate alignments of the model.

5.1.1. Bilingual Dictionary Structure

The *domain-specific bilingual dictionary* is collected for the proposed language pair, which is used to find the best translated words in Nepali for English words so that the best alignment between the sentences of the source and target language comparable corpus are identified. The bilingual dictionary used in this study is domain-specific in the sense that we have collected different comparable corpora and the dictionary is generated to handle the different words of those collected corpora. So, if such dictionary is found in a general domain, the performance of the proposed study will be improved so much.

The basic structure of the bilingual dictionary is that it contains a source (English) language word and its translation word in the target (Nepali) language. For the convenience the source word and target word are kept in two separate files, where the two files can be treated as the parallel corpus with the word level parallelism between the source and target file. The dictionary is generated with some restrictions, where the multi-word translations for a source word are kept in the dictionary by putting the first word of such multi-word translation in TL only such as for the word **produced = उत्पादन गरिएका**, the dictionary will only contain **produced = उत्पादन**. In case of multi-word source entry the source word and its translation both are not kept in the dictionary such as **due to = गर्दा**, **does not = होइन** are not kept in the dictionary. To read the dictionary, the file with SL words is read and its equivalent translation is identified from the file with TL words. The dictionary is used each and every time to identify the translation of a source word W_{Eng} of English Sentence (*ES*) and then the translated word is searched in the Nepali Sentence (*NS*) to identify the similarity score for that candidate Nepali sentence.

5.1.2. Test Data

Different possible comparable documents containing most of the parallel sentences are given as an input to the system which can be helpful to evaluate the accuracy of the actual parallel sentence alignment among the given non-parallel (comparable) documents. There are five different comparable corpora that we have collected to test the system. The different corpora contain different number of source and target sentence. In our collection, the Corpus 1 contains 30 source sentences and 28 target sentences. Similarly, Corpus 2 contains 20 source sentence and 19 target sentences, Corpus 3 contains 17 source sentences and 30 target sentences, Corpus 4 contains 23 source sentences and 25 target sentences whereas the Corpus 5 contains 12 source sentences and 12 target sentences. Therefore, the testing comparable corpora used in the research overall contain 102 source sentences and 114 target sentences. Thus, the more total number of comparable sentences exists in the comparable documents, the more generalized accuracy measures can be obtained from the model proposed.

Here, the input files contain English and Nepali test data respectively, both are then segmented into paragraph, each such paragraph is then segmented into a number of sentences and finally each such sentences are tokenized into tokens or words (delimited by space characters) to find the matched translations from the bilingual dictionary used for training the system.

5.2. Testing and Verification

The output will be tested and verified with respect to the manually taken comparable document pairs for the English (source) & Nepali (target) languages and translational information can be obtained from the manually generated domain specific bilingual dictionary for the language pair considered for this research task. Finally the empirical analysis is done to measure the accuracy of the model proposed and to suggest the solution. During the testing and analyzing period following measures are taken in order to verify the accuracy of the method or model. The measures taken are:

5.2.1. Precision

Precision is the number of correct results divided by the number of all results returned by aligner. Mathematically it is shown as follows:

$$\Pr ecision = \frac{Number\ of\ correctly\ aligned\ sentences}{Number\ of\ aligned\ sentences}$$

5.2.2. Recall

Recall is the number of correct results divided by the number of results that should have been aligned. Mathematically it is shown as follows:

$$\mathrm{Re}\,call = \frac{Number\ of\ correctly\ aligned\ sentences}{Total\ number\ of\ correctly\ aligned\ sentences\ in\ the\ comparable\ corpus}$$

5.2.3. F-Measure

The F-measure can be interpreted as a harmonic mean of precision and recall. It is defined as

$$F - measure = \frac{2\ \times (recall \times precision)}{(recall + precision)}$$

The Average Error Rate (AER) is calculated as

$$AER = 1 - (F - measure)$$

5.3. Input/output of program (Testing)

Comparable Corpus 1 (English)

[Social Studies Grade 9, Readmore Publishers and distributors,
Second Edition-2066 B.S., page no. 8-9]

<p> Development means a process of positive change. It is a dynamic process. People , family , community , town , etc keep on changing. Use of resources available in the society brings changes in quality of life of people. It enhances the living standard of the people living anywhere in the country. Development is linked with economic development. Economic development means achieving higher level economic condition. But, economic development alone does not indicate or reflect the overall development. According to the UN charter, “Development is related not only with the material needs of people but also with the improvement of social conditions”. Hence, development doesn't mean economic development only. It is also social, cultural and institutional growth. Change merely in specific people's life in no way reflects the development of a country. The country can advance ahead on the path of development only when a positive change takes place in the lives of all citizens.

<p> Some countries are developing at a rapid pace. These countries are called developed countries. Living standard of the people in the developed countries is very high. It has become half a century only since the developmental activities started in African, Asian and Latin American countries. Most of the countries in these continents are developing. People's living standard in developing countries hasn't been improved much due to the historical, geographical, cultural, social, political and economic reasons. Generally, those countries having a slow economic and technical growth rate and low per capita income are called developing countries. They are also called the least developed countries. All-round development is essential for improving the living standard in these countries. Every citizen of these countries should increase awareness and commitment to development.

<p> The total volume of all goods and services produced by a country in a year is called the Gross National Product (GNP). Similarly, the total value of all these products is called Gross National Income (GNI). Remittances from the foreign employment and foreign trade assistance are also included in Gross National Income. Gross National Income is considered these days as one of the main factors to measure

the economic development. When the Gross Domestic Income (GDI) of a region or country is divided by the population of that area or country, whatever number comes out is the per capita Income. Per capita income alone can't reflect the true state of development. A few rich people and their high income may show high per capita income, whereas a majority of people could be below the poverty line.

Comparable Corpus 1 (Nepali)

[सामाजिक अध्ययन - कक्षा ९, प्रकाशक: नेपाल सरकार, शिक्षा मन्त्रालय, पाठ्यक्रम विकास केन्द्र, सानोठिमी, भक्तपुर - २०६४ पेज न.२-३]

<p> विकास भनेको सकारात्मक परिवर्तन हो । मानिस , परिवार , समुदाय , सहर र गाउँहरू सधैँ परिवर्तनशील हुन्छन् । समाजमा उपलब्ध साधनहरू को प्रयोग बाट जनता को सुविधा मा वृद्धि हुन्छ । यसले कुनै पनि मुलुक मा बस्ने प्रत्येक व्यक्ति को जीवन मा सुधार ल्याउँछ । विकास लाई आर्थिक विकास सँग जोड्ने गरिन्छ । तल्लो आर्थिक अवस्था बाट माथिल्लो आर्थिक अवस्था मा पुग्नु नै आर्थिक विकास हो तर आर्थिक उन्नति ले मात्र विकास को अवस्था दर्साउन सक्दैन । संयुक्त राष्ट्रसंघ को बडापत्र मा उल्लेख भए अनुसार “ विकास ले मानिस को भौतिक चाहना सँग मात्र नभई सामाजिक अवस्था को सुधार सम्बन्धी कुराहरू सँग पनि सरोकार राख्दछ ”। तसर्थ , विकास भनेको आर्थिक विकास मात्र होइन । सामाजिक , सांस्कृतिक तथा संस्थागत वृद्धि पनि हो । कुनै खास व्यक्ति को जीवन मा मात्र परिवर्तन आउँदैमा देश को विकास हुन्छ भन्न सकिंदैन । जब राष्ट्र का सम्पूर्ण व्यक्तिहरू को जीवन मा सुधारात्मक परिवर्तन आउँछ तब देश विकास को पथ मा अगाडि बढ्न सक्छ । </p>

<p> विश्व का केही मुलुकहरू ले तीव्र रूपमा विकास गरिरहेका छन् । त्यस्ता मुलुकहरू लाई विकसित मुलुक भनिन्छ । ती मुलुकहरू मा बस्ने धेरै जसो मानिसहरू को जीवनस्तर उच्च छ । अफ्रिका , एसिया र दक्षिण अमेरिका का देशहरू ले विकास को थालनी गरेको आधा शताब्दी मात्र भएको छ । यी देशहरू विकासोन्मुख हुन् ।

ऐतिहासिक , भौगोलिक , साँस्कृतिक , सामाजिक , राजनीतिक तथा आर्थिक कारण ले गर्दा यी विकासोन्मुख देश को जीवनस्तर अझै उकास्न सकिएको छैन । सामान्यतया आर्थिक तथा प्राविधिक विकास दर कम भएको र प्रति व्यक्ति आम्दानी कम भएको मुलुक लाई विकासोन्मुख देश भनिन्छ । यिन लाई कम विकसित मुलुक पनि भनिन्छ । यस्ता मुलुकहरू को जीवन उकास्न चौतर्फी विकास गर्न आवश्यक हुन्छ । विकासोन्मुख देश को विकासका लागि सम्बन्धित देशका नागरिकहरू जागरुक हुनुपर्दछ । </p>

<p> कुनै मुलुक मा एक वर्ष भित्र उत्पादन गरिएका वस्तु र सबै सेवाको मुल्य लाई कुल राष्ट्रिय उत्पादन भनिन्छ । त्यस्ता सबै वस्तु र सेवाको मुल्य लाई कुल राष्ट्रिय आय भनिन्छ । कुल राष्ट्रिय आय मा वैदेशिक व्यापार सहायता र वैदेशिक रोजगार बाट प्राप्त रकम पनि जोडिन्छ । आजकल कुल राष्ट्रिय आय लाई आर्थिक विकास मापन गर्ने एक मुख्य आधार मानिन्छ । कुनै निश्चित क्षेत्रको कुल घरेलु आम्दानी लाई त्यस क्षेत्र को जनसङख्या ले भाग गर्दा आउने रकम नै प्रति व्यक्ति आय हो । प्रति व्यक्ति आय ले मात्र देश को विकास को यथार्थ अवस्था देखाउन सक्दैन । केहि धनि मानिसहरू को उच्च आय ले प्रति व्यक्ति आय बढी देखिएको पनि हुन सक्छ जहाँ अत्याधिक मानिस गरिबी को रेखामुनि पनि हुन सक्छन् ।   </p>

Output of the program as a parallel corpus 1:

S.N.	English Sentence	Nepali Sentence
1	Development means a process of positive change.	विकास भनेको सकारात्मक परिवर्तन हो ।
2	It is a dynamic process.	विकास भनेको सकारात्मक परिवर्तन हो ।
3	People, family, community, town, etc keep on changing.	मानिस , परिवार , समुदाय , सहर र

		गाउँहरू सधैँ परिवर्तनशील हुन्छन् ।
4	Use of resources available in the society brings changes in quality of life of people.	जब राष्ट्र का सम्पूर्ण व्यक्तिहरू को जीवन मा सुधारात्मक परिवर्तन आउँछ तब देश विकास को पथ मा अगाडि बढ्न सक्छ ।
5	It enhances the living standard of the people living anywhere in the country.	यसले कुनै पनि मुलुक मा बस्ने प्रत्येक व्यक्ति को जीवन मा सुधार ल्याउँछ ।
6	Development is linked with economic development.	विकास लाई आर्थिक विकास सँग जोड्ने गरिन्छ ।
7	Economic development means achieving higher level economic condition.	तल्लो आर्थिक अवस्था बाट माथिल्लो आर्थिक अवस्था मा पुग्नु नै आर्थिक विकास हो तर आर्थिक उन्नति ले मात्र विकास को अवस्था दर्साउन सक्दैन ।
8	But, economic development alone does not indicate or reflect the overall development.	तल्लो आर्थिक अवस्था बाट माथिल्लो आर्थिक अवस्था मा पुग्नु नै आर्थिक विकास हो तर आर्थिक उन्नति ले मात्र विकास को अवस्था दर्साउन सक्दैन ।
9	According to the UN charter, “Development is related not only with the material needs of people but also with the improvement of social conditions”.	संयुक्त राष्ट्रसंघ को बडापत्र मा उल्लेख भए अनुसार “ विकास ले मानिस को भौतिक चाहना सँग मात्र नभई सामाजिक अवस्था को सुधार सम्बन्धी कुराहरू सँग पनि सरोकार राख्दछ ”।

10	Hence, development doesn't mean economic development only.	तसर्थ , विकास भनेको आर्थिक विकास मात्र होइन ।
11	It is also social, cultural and institutional growth.	सामाजिक , सांस्कृतिक तथा संस्थागत वृद्धि पनि हो ।
12	Change merely in specific people's life in no way reflects the development of a country.	कुनै खास व्यक्ति को जीवन मा मात्र परिवर्तन आउँदैमा देश को विकास हुन्छ भन्न सकिंदैन ।
13	The country can advance ahead on the path of development only when a positive change takes place in the lives of all citizens.	जब राष्ट्र का सम्पूर्ण व्यक्तिहरू को जीवन मा सुधारात्मक परिवर्तन आउँछ तब देश विकास को पथ मा अगाडि बढ्न सक्छ ।
14	Some countries are developing at a rapid pace.	विश्व का केही मुलुकहरू ले तीव्र रूपमा विकास गरिरहेका छन् ।
15	These countries are called developed countries.	त्यस्ता मुलुकहरू लाई विकसित मुलुक भनिन्छ ।
16	Living standard of the people in the developed countries is very high.	ती मुलुकहरू मा बस्ने धेरै जसो मानिसहरू को जीवनस्तर उच्च छ ।
17	It has become half a century only since the developmental activities started in African, Asian and Latin American countries.	अफ्रिका , एसिया र दक्षिण अमेरिका का देशहरू ले विकास को थालनी गरेको आधा शताब्दी मात्र भएको छ ।
18	Most of the countries in these continents are developing.	ती मुलुकहरू मा बस्ने धेरै जसो मानिसहरू को जीवनस्तर उच्च छ ।

19	People's living standard in developing countries hasn't been improved much due to the historical, geographical, cultural, social, political and economic reasons.	ऐतिहासिक , भौगोलिक , साँस्कृतिक , सामाजिक , राजनीतिक तथा आर्थिक कारण ले गर्दा यी विकासोन्मुख देश को जीवनस्तर अझै उकास्न सकिएको छैन ।
20	Generally, those countries having a slow economic and technical growth rate and low per capita income are called developing countries.	सामान्यतया आर्थिक तथा प्राविधिक विकास दर कम भएको र प्रति व्यक्ति आम्दानी कम भएको मुलुक लाई विकासोन्मुख देश भनिन्छ ।
21	They are also called the least developed countries.	यिन लाई कम विकसित मुलुक पनि भनिन्छ ।
22	All-round development is essential for improving the living standard in these countries.	यस्ता मुलुकहरू को जीवन उकास्न चौतर्फी विकास गर्न आवश्यक हुन्छ ।
23	Every citizen of these countries should increase awareness and commitment to development.	ती मुलुकहरू मा बस्ने धेरै जसो मानिसहरू को जीवनस्तर उच्च छ ।
24	The total volume of all goods and services produced by a country in a year is called the Gross National Product.	कुनै मुलुक मा एक वर्ष भित्र उत्पादन गरिएका वस्तु र सबै सेवाको मुल्य लाई कुल राष्ट्रिय उत्पादन भनिन्छ ।
25	Similarly, the total value of all these products is called Gross National Income (GNI).	त्यस्ता सबै वस्तु र सेवाको मुल्य लाई कुल राष्ट्रिय आय भनिन्छ ।
26	Remittances from the foreign employment and foreign trade assistance are also included in Gross National Income.	कुल राष्ट्रिय आय मा वैदेशिक व्यापार सहायता र वैदेशिक रोजगार बाट प्राप्त रकम पनि जोडिन्छ ।

27	Gross National Income is considered these days as one of the main factors to measure the economic development.	आजकल कुल राष्ट्रिय आय लाई आर्थिक विकास मापन गर्ने एक मुख्य आधार मानिन्छ ।
28	When the Gross Domestic Income (GDI) of a region or country is divided by the population of that area or country, whatever number comes out is the per capita Income.	कुनै निश्चित क्षेत्रको कुल घरेलु आम्दानी लाई त्यस क्षेत्र को जनसङ्ख्या ले भाग गर्दा आउने रकम नै प्रति व्यक्ति आय हो ।
29	Per capita income alone can't reflect the true state of development.	प्रति व्यक्ति आय ले मात्र देश को विकास को यथार्थ अवस्था देखाउन सक्दैन ।
30	A few rich people and their high income may show high per capita income, whereas a majority of people could be below the poverty line.	केहि धनि मानिसहरू को उच्च आय ले प्रति व्यक्ति आय बढी देखिएको पनि हुन सक्छ जहाँ अत्याधिक मानिस गरिबी को रेखामुनि पनि हुन सक्छन् ।

Table 5.1: Aligned Parallel Sentences in Comparable Corpus 1 (Testing)

Here, in this corpus following measures are found:

Total No. of Sentences in Source Corpus = 30

Total No. of sentences in Target Corpus = 28

No. of aligned sentences = 30

No. of correctly aligned sentences = 26

Correct alignments present in the corpora = 29

Precision = (26 / 30) = 0.866667 = 86.67%

Recall = (26 / 29) = 0.896552 = 89.66%

F-measure = ((2*0.866667*0.896552) / (0.866667+0.896552))
= 0.881356 = 88.14%

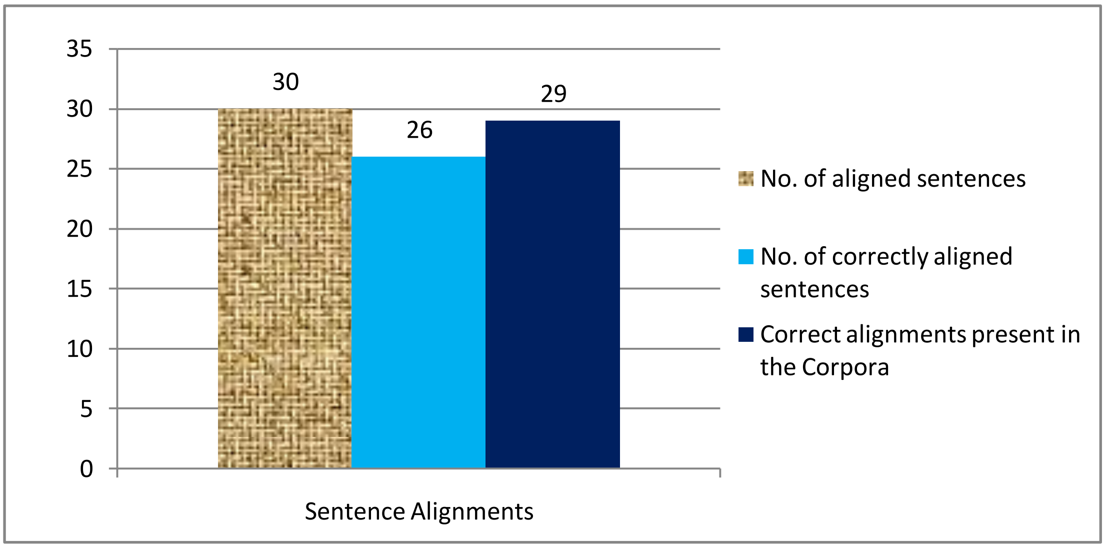


Figure 5.1: - Sentence alignment in Comparable Corpus 1

The testing on the other four comparable corpora including the measurements taken is shown in the Appendix A.

5.4. Analysis

The average precision and recall measure of the above given sentences is found by calculating the individual precision and recall measure of the given paragraph and finding the average of them.

Corpus/Measures	Corpus 1	Corpus 2	Corpus 3	Corpus 4	Corpus 5	Average
Precision	0.866667	0.85	0.941176	0.826087	0.916667	0.880119
Recall	0.896552	0.85	0.941176	0.863636	0.916667	0.893606
F-Measure	0.881356	0.85	0.941176	0.844444	0.916667	0.886729

Table 5.2: Final Analysis

Thus, final measures acquired are:

Precision = 0.88019

Recall = 0.893606

F-Measure = 0.886729

AER = 1- (F-Measure) = 1- 0.886729 = 0.113271

This result shows that the proposed model works well for the given input. In our analysis the recall is shown greater than precision it means that precision gives the comparison of alignment accuracy when it is compared to the result given by the model and Recall gives the accuracy of the alignment when compared with the actual alignment given by the human annotator. Besides that, this accuracy obtained or identified here may vary while testing a huge document. It means the accurate alignments for larger data sets may change. However, the overall accuracy depends on the number of words and the domain of the words present in the bilingual dictionary used in this proposed model. So, if we have a huge bilingual dictionary with sufficient amount of words from both source and target languages, the accuracy may be acquired in the range obtained in this research work.

CHAPTER 6

6. Conclusion and Future Works

6.1. Conclusion

Parallel texts are the useful resources and widely used in different fields of NLP can be found in sufficient amount in different comparable corpora that are found easily in huge amount. It is really a difficult task to generate such parallel corpora manually for the various NLP related projects, and is also time consuming as well as costlier. So, a mechanism of automatically generating such parallel text from the comparable corpora is a novel and effective approach for the language pairs proposed in this study. So far talking about the results obtained from this study, there is an accuracy of nearly up to 0.887 on the comparable corpora taken. The results of the experiment on the collected comparable documents show that the method has 88% precision and 89% recall. So the F-measure obtained for the proposed study is about 0.887 on the collected data sets. However, the accuracy and other measures taken in the research work may vary on the larger comparable corpora.

It seems that the study performed throughout this research lets us to generate good quality parallel corpus from the comparable corpus in very short time period. The performance of the proposed study is highly dependent on the quality of the bilingual dictionary used in the research work. The more advanced the dictionary is the more accurate the alignments are. The bilingual dictionary used in the study is made manually from the different testing comparable corpora used in the study. So, it is domain specific in some sense. However, if a dictionary is available in general domain, the method is capable of aligning the sentences from the general domain in a great extent with a great accuracy.

Thus, the method proposed in this study will be helpful to find the parallel data from the comparable data efficiently and accurately in a limited amount of time.

6.2. Limitations and Future Works

Since the model presented in this study contains a bilingual dictionary with limited words, so sometimes the words in the sentences of the testing corpus will not be found in the bilingual dictionary to find its translation in target language and will result in wrong alignments. The accuracy of the alignments is heavily dependent on the quality of the bilingual dictionary used in this model. The testing so far we have made and identified accuracy may vary while testing a huge comparable document as we have tested it on limited sentence sets.

It also has another limitation related to the dictionary used in the proposed study. Since, the dictionary may contain some of the words in either source or target language which might have more than one word in their translations but the dictionary used in this study handles only the single word translations. For example; **development** = **विकास**, **all = सबै**, **also = पनि**, **and = र**, etc. like single word entry in the dictionary are handled where as the word like **produced = उत्पादन गरिएका** , **responded = जवाफ दियो, developed** = **निर्माण गरिएको**, **due to = गर्दा**, **does not = होइन**, **twenty five** = **पच्चिस**, etc. multiword entries are not handled in this study.

There may be the situation where the sentence boundary is not identified correctly while considering the period or full stop (.) as a sentence boundary separator. For example in case of different words like Dr., Mr., Ms., Prof., etc which represent different short hand notations are not handled properly in the proposed study.

Here in this study, the parallel data is collected from the manually collected comparable documents but it could be much more effective if such comparable documents are also aligned automatically as a good comparable documents. Besides that the study is carried out by using a bilingual dictionary which is also generated manually. If such bilingual dictionaries could also be generated in an automated manner from the comparable corpora, it will improve the performance of the overall system. Thus, these areas could be taken as further studies.

References

[1] Arora, S., Tyagi, R., and Singla, S. R. Creation of Parallel Corpus from Comparable Corpus. In *Proceedings of ASCNT – 2010*, pp.77-83, CDAC, Noida, India.

[2] Bal, B. K. Towards Building Advanced Natural Language Applications – An Overview of the Existing Primary Resources and Applications in Nepali. In *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009,* pages 165–170, Suntec, Singapore, 6-7 August 2009.

[3] Balami, B. A Chunk Level Statistical Machine Translation. A *dissertation* submitted to Central Department of Computer Science and Information Technology (CDCSIT), Tribhuvan University, Kirtipur, Kathmandu, Nepal. April, 2010.

[4] Bo Li; Juan Liu; Huili Zhu. Mining Parallel Text from the Web based on Sentence Alignment. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation, 2007,* pages 285-292.

[5] Braune, F., and Fraser, A. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. *Coling 2010: 23rd International Conference on Computational Linguistics*, 23-27 August 2010, Beijing International Convention Center, Beijing, China, *Posters volume*; pp.81-89.

[6] Brown, P. F.; Lai, J. C.; and Mercer, R. L. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, pp. 169–176. ACL 1991.

[7] Chen, S.F. Aligning sentences in bilingual corpora using lexical information. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics,* pages 9–16, Columbus, Ohio, June 1993.

[8] Church, K.W. Char_align: A program for aligning parallel texts at the character level. In*Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 1–8. 1993

[9] D. Jurafsky, J. H. Martin, *Speech and Language Processing: An Introduction to Speech Recognition Natural Language Processing and Computational Linguistic*, (2006).

[10] D. Tufis¸, R. Ion, A. Ceaus¸u, and D. S¸ tefˇanescu. Improved lexical alignment by combining multiple reified alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pages 153–160, April.

[11] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy. Parallel corpora for medium density languages. *In Proceedings of the RANLP 2005*, pages 590-596.

[12] Eisele, A., & Xu, J. Improving machine translation performance using comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora LREC 2010*, pp. 35-41.

[13] Fung, P., and Cheung, P. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, 57–63, Barcelona, Spain, July, Association for Computational Linguistics.

[14] Fung, P., and McKeown, K. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic warping. In *First Conf. of the Association for Machine Translation in the Americas (AMTA 94)*, pages 81–88, Columbia, MD, October.

[15] Gale, W. A. and Church, K. W. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th ACL*, Berkeley, CA, pp. 177–184. ACL 1991.

[16] Gale, W. A., and Church, K. W. A program for aligning sentences in bilingual corpora. *Computational Linguistics,* 19(1), pages 75–102. 1993

[17] Güngör, A. M., and Taşcı, S. Turkish-English Sentence Alignment. Senior Project, Boğaziçi University, 2006

[18] Hla Hla Htay, G. Bharadwaja Kumar, Kavi Narayana Murthy. Constructing English-Myanmar Parallel Corpora. In *Proceedings of ICCA 2006: International Conference on Computer Applications,* pp 231-238, Yangon, Myanmar, 23-24 February 2006.

[19] Joshi, Y.R. A Chunk Alignment Model for Statistical Machine Translation on English-Nepali Parallel Corpus. A *dissertation* submitted to Central Department of Computer Science and Information Technology (CDCSIT), Tribhuvan University, Kirtipur, Kathmandu, Nepal. May, 2010.

[20] K. T´oth, R. Farkas, and A. Kocsor. Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm. *Acta Cybernetica*, vol. 18, no. 3, pp. 463–478, 2008.

[21] Kay, M., Röscheisen, M. Text-Translation Alignment. *Computational Linguistics*, 19(1) pages 121–142. 1993

[22] Moore, R. C. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 135–144. 2002

[23] Mukesh, G.S. Vatsa, Nikita Joshi, and Sumit Goswami. Statistical Machine Translation. *DESIDOC Journal of Library & Information Technology,* Vol. 30, No. 4, July 2010, pp. 25-32.

[24] Munteanu, D. S. Exploiting Comparable Corpora. A Dissertation Proposal Presented to the Faculty of the Graduate School, University of Southern California. December 2006.

[25] Munteanu, D. S., and Marcu, D. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics,* 31(4):477-504. 2005

[26] Munteanu, D. S., Fraser, A., and Marcu, D. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* pages 265–272, Boston, MA. 2004

[27] P. Resnik and N.A. Smith. The web as a parallel corpus. Computational Linguistics, 29(3):349-380. 2003

[28] Pavel Pecina, Antonio Toral, Gregor Thurmair, Andy Way, Carsten Schnober. Parallel technology tools and resources. *PANACEA Project,* 2010.

[29] Philip Resnik. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), University of Maryland, College Park, Maryland, June 1999.

[30] Sadaf Abdul-Rauf and Holger Schwenk. Exploiting Comparable Corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP,* pages 46-54, Suntec, Singapore, 6 August 2009.

[31] Singh, A.K. and Husain, S. Exploring Translation Similarities for Building a Better Sentence Aligner. In *Proceedings of IICAI.* 2007, 1852-1863.

[32] Smith, J.R., Quirk, C. and Toutanova, K. Extracting Parallel Sentences form Comparable Corpora using Document Level Alignment. *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL,* pages 403-411, Los Angeles, California, June 2010.

[33] Utiyama, M., and Isahara, H. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Morristown, NJ, USA, Association for Computational Linguistics. 2003

[34] Vamshi Ambati. Active Learning for Machine Translation in Sparse Data Scenarios. LTI Thesis Proposal. November, 2010.

[35] Vosoughpour, M., and Faili, H. Generating English-Persian Parallel Corpus Using an Automatic Anchor Finding Sentence Aligner. *In the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'10)* pp. 578-583.

[36] Weigang Li, Ting Liu, Zhen Wang and Sheng Li. Aligning Bilingual Corpora Using Sentences Location Information. In Proceedings of 3rd ACL SIGHAN Workshop, pp. 141-147. 1994

[37] Wu, D. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico (1994) 80–87

[38] Zhao, B., and Vogel, S. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the IEEE International Conference on Data Mining*, pages 745–748, Washington, DC, USA, IEEE Computer Society. 2002

Appendix A:

Input/output of the Program (Testing):

Comparable Corpus 2 (English)

[Social Studies Grade 9, Readmore Publishers and distributors,
Second Edition-2066 B.S., page no. 23]

<p> We have not been able to achieve desired improvement in the people's health because of widespread poverty , illiteracy and lack of awareness. The government has not been able to operate hospitals and manage doctors in all places. There are some places where there are no hospitals , some places have hospitals but no doctors and there are some places having doctors but no necessary tools , equipment and medicines. Doctors are not so interested to go and work in remote areas. They prefer to stay in cities and towns , where all facilities and privileges are available. Contrary to it , there are no such health services in the remote areas. Due to the lack of education and awareness , people prefer to go to witch-doctors (dhami jhankri) instead of seeing or going to doctors. People living in remote areas are poor and even can't pay for doctors' fee. The prescribed salary can't meet all the requirements of doctors. So , the doctors like to do part time jobs for additional income in the private hospitals , clinics , etc. But , the poor people in rural areas can't afford to pay for expensive services of private hospitals , nursing homes and clinics.

<p> The status of health of the people in countries having high life-expectancy is considered to be good. Here people are healthy due to the availability of nutritious food are attacked less by diseases , take regular exercise , enjoy good health care services , etc. Life-expectancy of people in Nepal is comparatively low compared to developed countries. Size of urban population in Nepal is increasing. However , their health condition is deteriorating day by day in the absence of nutritious foods , proper exercises and health services. There is a need of balanced diet , regular exercise , and some physical labour to maintain a healthy life. High child mortality rate is yet another indication of poor health condition in Nepal. Significant attempts are lately being initiated to establish and expand health posts all over the country. Health workers and volunteers are mobilized throughout the country.

Comparable Corpus 2 (Nepali)

[सामाजिक अध्ययन - कक्षा ९, प्रकाशक: नेपाल सरकार, शिक्षा मन्त्रालय, पाठ्यक्रम विकास केन्द्र, सानोठिमी, भक्तपुर - २०६४ पेज न.१४]

<p> हाम्रो देश मा व्याप्त गरिबी , अशिक्षा र जनचेतना को अभाव मा मानिसहरू को स्वास्थ्य मा अपेक्षाकृत सुधार हुन सकेको छैन । नेपाल सरकार ले सबै ठाउँ मा अस्पताल खोल्न र डाक्टर पठाउन सकेको छैन । कहीँ अस्पताल छन् त डाक्टर छैनन् , कहीँ अस्पताल र डाक्टर छन् भने आवश्यक उपकरण र औषधी छैनन् । भौगोलिक विकटता तथा पछौटेपन ले गर्दा डाक्टरहरू दुर्गम क्षेत्र मा गएर सेवा गर्न चाहँदैनन् । यिनीहरू प्राय : सहर मा नै बस्न चाहन्छन् । किनभने सहर मा सबै किसिम का सुविधाहरू प्राप्त छन् । यसको विपरीत गाउँ मा सुविधाहरू पनि छैनन् र जनचेतना को अभाव मा मानिसहरू पनि डाक्टर सँग सम्पर्क राख्न को सट्टा धामीझाँक्री को सल्लाह अनुसार काम गर्दछन् । गरिबी को कारण ले दुर्गम क्षेत्रका मानिसहरू डाक्टर लाई माग अनुसार शुल्क दिन सक्दैनन् । सरकार द्वारा निर्धारण गरेको तलब ले डाक्टर का सबै किसम का आवश्यकताहरू पूरा हुन सक्दैनन् । त्यसैले यीनीहरू सरकारी अस्पताल मा काम सके पछि अतिरिक्त समय मा निजी अस्पताल , क्लिनिक आदि मा गई अतिरिक्त आम्दानी गर्न चाहन्छन् तर गरिबी ले गर्दा दुर्गम का सबै मानिसहरू बढी शुल्क तिरेर निजी अस्पताल मा जान सक्दैनन् । </p>

<p> जुन् देश को औसत आयु बढी हुन्छ , त्यस देश का जनता को स्वास्थ्य को अवस्था राम्रो मानिन्छ । पर्याप्त पोषण , रोग को सङ्क्रमण को कमी , औषधी उपचार को सुविधा , व्यायाम आदि का कारण मानिस निरोगी हुन्छ र उसको औसत आयु पनि बढी हुन्छ । हाम्रो देश को जनताहरू को औसत आयु विकसित मुलुक का मानिसहरू को तुलना मा निकै कम रहेको छ । हाम्रो देश मा पनि सहरी जनसङ्ख्या निकै बढेर गएको छ । नियमित व्यायाम र सन्तुलित भोजन को अभाव ले गर्दा

उनीहरू को स्वास्थ्य दिन प्रति दिन बिग्रँदै गएको छ । त्यस का लागि सन्तुलित खाना , नियमित व्यायाम र शारीरिक श्रम गर्नु जरुरी छ । स्वास्थ्य को कमजोर अवस्था ले गर्दा नेपाल मा शिशु र बाल मृत्यु दर बढी भएको देखिन्छ । नेपाल मा हाल आधारभूत स्वास्थ्य सुविधा प्रदान गर्न गाउँ गाउँ मा उपस्वास्थ्य चौकी तथा स्वास्थ्य केन्द्रहरू सञ्चालित छन् । हरेक टोल र वडा मा स्वास्थ्य स्वयम्सेविकाहरू परिचालित छन् । </p>

Output of the program as a parallel corpus 2:

S.N.	English Sentence	Nepali Sentence
1	We have not been able to achieve desired improvement in the people’s health because of widespread poverty, illiteracy and lack of awareness.	हाम्रो देश मा व्याप्त गरिबी , अशिक्षा र जनचेतना को अभाव मा मानिसहरू को स्वास्थ्य मा अपेक्षाकृत सुधार हुन सकेको छैन ।
2	The government has not been able to operate hospitals and manage doctors in all places.	नेपाल सरकार ले सबै ठाउँ मा अस्पताल खोल्न र डाक्टर पठाउन सकेको छैन ।
3	There are some places where there are no hospitals, some places have hospitals but no doctors and there are some places having doctors but no necessary tools, equipment and medicines.	कहीँ अस्पताल छन् त डाक्टर छैनन् , कहीँ अस्पताल र डाक्टर छन् भने आवश्यक उपकरण र औषधी छैनन् ।
4	Doctors are not so interested to go and work in remote areas.	भौगोलिक विकटता तथा पछौटेपन ले गर्दा डाक्टरहरू दुर्गम क्षेत्र मा गएर सेवा गर्न चाहँदैनन् ।
5	They prefer to stay in cities and	नेपाल सरकार ले सबै ठाउँ मा अस्पताल

	towns, where all facilities and privileges are available.	खोल्न र डाक्टर पठाउन सकेको छैन ।
6	Contrary to it, there are no such health services in the remote areas.	भौगोलिक विकटता तथा पछौटेपन ले गर्दा डाक्टरहरू दुर्गम क्षेत्र मा गएर सेवा गर्न चाहँदैनन् ।
7	Due to the lack of education and awareness, people prefer to go to witch-doctors (dhami jhankri) instead of seeing or going to doctors.	यसको विपरीत गाउँ मा सुविधाहरू पनि छैनन् र जनचेतना को अभाव मा मानिसहरू पनि डाक्टर सँग सम्पर्क राख्न को सट्टा धामीझाँक्री को सल्लाह अनुसार काम गर्दछन् ।
8	People living in remote areas are poor and even can't pay for doctors' fee.	गरिबी को कारण ले दुर्गम क्षेत्रका मानिसहरू डाक्टर लाई माग अनुसार शुल्क दिन सक्दैनन् ।
9	The prescribed salary can't meet all the requirements of doctors.	सरकार द्वारा निर्धारण गरेको तलब ले डाक्टर का सबै किसम का आवश्यकताहरू पूरा हुन सक्दैनन् ।
10	So, the doctors like to do part time jobs for additional income in the private hospitals, clinics, etc.	त्यसैले यीनीहरू सरकारी अस्पताल मा काम सके पछि अतिरिक्त समय मा निजी अस्पताल , क्लिनिक आदि मा गई अतिरिक्त आम्दानी गर्न चाहन्छन् तर गरिबी ले गर्दा दुर्गम का सबै मानिसहरू बढी शुल्क तिरेर निजी अस्पताल मा जान सक्दैनन् ।

11	But, the poor people in rural areas can't afford to pay for expensive services of private hospitals, nursing homes and clinics.	त्यसैले यीनीहरू सरकारी अस्पताल मा काम सके पछि अतिरिक्त समय मा निजी अस्पताल , क्लिनिक आदि मा गई अतिरिक्त आम्दानी गर्न चाहन्छन् तर गरिबी ले गर्दा दुर्गम का सबै मानिसहरू बढी शुल्क तिरेर निजी अस्पताल मा जान सक्दैनन् ।
12	The status of health of the people in countries having high life-expectancy is considered to be good.	जुन् देश को औसत आयु बढी हुन्छ , त्यस देश का जनता को स्वास्थ्य को अवस्था राम्रो मानिन्छ ।
13	Here people are healthy due to the availability of nutritious food are attacked less by diseases, take regular exercise, enjoy good health care services, etc.	पर्याप्त पोषण , रोग को सङ्क्रमण को कमी , औषधी उपचार को सुविधा , व्यायाम आदि का कारण मानिस निरोगी हुन्छ र उसको औसत आयु पनि बढी हुन्छ ।
14	Life-expectancy of people in Nepal is comparatively low compared to developed countries.	हाम्रो देश को जनताहरू को औसत आयु विकसित मुलुक का मानिसहरू को तुलना मा निकै कम रहेको छ ।
15	Size of urban population in Nepal is increasing.	हाम्रो देश को जनताहरू को औसत आयु विकसित मुलुक का मानिसहरू को तुलना मा निकै कम रहेको छ ।
16	However, their health condition is deteriorating day by day in the	नियमित व्यायाम र सन्तुलित भोजन को

	absence of nutritious foods, proper exercises and health services.	अभाव ले गर्दा उनीहरू को स्वास्थ्य दिन प्रति दिन बिग्रँदै गएको छ ।
17	There is a need of balanced diet, regular exercise, and some physical labour to maintain a healthy life.	त्यस का लागि सन्तुलित खाना , नियमित व्यायाम र शारीरिक श्रम गर्नु जरुरी छ ।
18	High child mortality rate is yet another indication of poor health condition in Nepal.	स्वास्थ्य को कमजोर अवस्था ले गर्दा नेपाल मा शिशु र बाल मृत्यु दर बढी भएको देखिन्छ ।
19	Significant attempts are lately being initiated to establish and expand health posts all over the country.	नेपाल मा हाल आधारभूत स्वास्थ्य सुविधा प्रदान गर्न गाउँ गाउँ मा उपस्वास्थ्य चौकी तथा स्वास्थ्य केन्द्रहरू सञ्चालित छन् ।
20	Health workers and volunteers are mobilized throughout the country.	हरेक टोल र वडा मा स्वास्थ्य स्वयम्सेविकाहरू परिचालित छन् ।

Table A.1: Aligned Parallel Sentences in Comparable Corpus 2

Here, in this corpus following measures are found:

Total No. of Sentences in Source Corpus	20
Total No. of sentences in Target Corpus	19
No. of aligned sentences	20
No. of correctly aligned sentences	17
Correct alignments present in the Corpora	20
Precision	0.85
Recall	0.85
F-measure	0.85

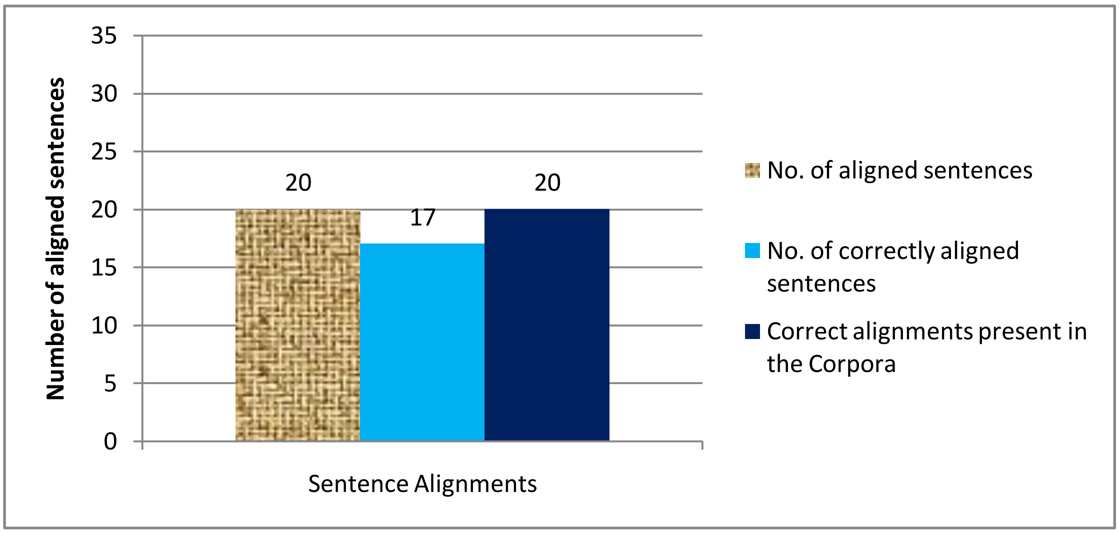


Figure A.1: - Sentence alignment in Comparable Corpus 2

Comparable Corpus 3 (English)

[Republica: English Daily Newspaper

Published on: Tuesday, August 16, 2010, Page 3]

<p> Four, including two kids, of the same family were brutally murdered at matena village of Bhimdutta Municipality - 9 in Kanchanpur district Sunday night. The wife, daughter-in-law, grandson and granddaughter of Karbir Nath, a local, were found murdered in their home located ten kilometers from Mahendranagar. Among those killed are Karbir’s 50 year-old wife Harya Nath, 25 year-old daughter-in-law Dhanadevi Nath, six year-old grandson Ravi Nath and four year-old granddaughter Renuka Nath. The bodies of all the four slain family members have been kept at Mahakali Zonal Hospital. All four have rope marks around their necks. It seems they were strangled to death by the murderer using rope. </p>

<p> Karbir’s son Dharmananda Nath is missing, prompting suspicion that he killed the four. Locals say Dharmananda has been mentally ill for more than one year, adding to the suspicion. “The family did not have enmity with anyone,” a neighbor said, adding, “It seems Dharmananda, who was mentally ill, killed them”. The suspect was with the family members until 10 pm Sunday, the neighbor said. The murder must have taken place after everybody had gone to sleep. The villagers learnt about the incident only in the morning after they forced their way into the house when nobody from inside responded to their calls. DhanaDevi and her kids were found in one room while mother-in-law Harya Nath was found in another. </p>

<p> The neighbors informed Brahma Dev Police Post about the incident. Dharmananda's father Karbir works in Mumbai, India, while elder-brother is in the Armed Police Force (APF). Post-mortem of the four bodies were done at Mahakali Zonal Hospital after Karbir's eldest son arrived from Kathmandu. The police have been searching for absconding Dharmananda while the whole village is living in terror with Baijanath Siddhanath Secondary School shut-down after the incident. </p>

Comparable Corpus 3 (Nepali)

[नागरिक: नेपाली दैनिक पत्रिका

प्रकाशित: मंगलबार, ३१ साउन २०६८, Tuesday, August 16, 2010, Page 3]

<p> कञ्चनपुर को भीमदत्त नगरपालिका - ९ मटेना मा आइतबार राति एकै घर का दुई नाबालक सहित चार को हत्या भएको छ । सदरमुकाम महेन्द्रनगर बाट १० किमि टाढा मटेना गाउँ मा स्थानीय करवीर नाथ की पत्नी , बुहारी , नाति र नातिनी को हत्या भएको हो । मारिनेहरू करवीरकी ५० बर्षीया पत्नी हरया नाथ , उनकी २५ वर्षीया बुहारी धनादेवी नाथ , ६ बर्षीय नाति रवि नाथ र चार वर्षीया नातिनी रेनुका नाथ छन् । मारिएका चारै को शव महाकाली अञ्चल अस्पताल मा राखिएको छ । चारै शव को घाँटी मा डोरी को डाम ले बनाएको घाउ छ । हत्यारा ले डोरी ले घाँटी बेरी हत्या गरेको देखिन्छ । घटना का बखत घर मा पाँच जना मात्रै थिए । </p>

<p> करवीरको पाँच जना को परिवार मा छोरो धर्मानन्द नाथ फरार रहेका ले घटना मा उनको संलग्नता आशंका गरिएको छ । राति घरपरिवार कै साथ रहेका धर्मानन्द बिहान घर मा नभेटिए पछि उनी माथि को शंका बलियो छ । केही वर्षयता धर्मानन्द को मानसिक अवस्था ठीक नभएको स्थानीय ले बताएका छन् । " कसैसित रिसइबी थिएन , झगडा पनि कसैसित भएको देखेनौं " एक छिमेकी ले भने , " मानसिक अवस्था ठीक नभएका ले धर्मानन्द ले नै चारै को हत्या गरे जस्तो देखिन्छ " । डेढ बर्ष देखि धर्मानन्द मानसिक रोगी थिए । उनको केही समय अघि भारत को बरेली मा

उपचार गरिएको थियो । “ औषधी सेवन गरिरहेका थिए , ” छिमेकी कलावती नाथ ले भनिन् , “ उनी मानसिक तनावमा देखिन्थे ” । राति दस बजे सम्म धर्मानन्द लाई परिवार सँगै देखेको ती छिमेकी ले बताए । “ राति दस बजे पछि घटना भएको छ, सबै सुतेका बेला हत्या गरेको हुन सक्छ , ” छरछिमेककाहरू भन्दै थिए । मंगलबार दाहसंस्कार गरिने पारिवारिक स्रोतले बतायो । घटना बारे गाउँले ले सोमबार बिहान मात्रै थाहा पाएका थिए । गाईबस्तु फुकाएको देखेपछि छिमेकी ले उनीहरू को खोजी गरेका थिए । बाहिर बाट बोलाउँदा कसैले केही जवाफ नदिए पछि भित्र गएर हेर्दा घटनाबारे थाहा पाएका थिए । धानादेवी र उनका दुई नाबालक एउटा कोठा मा भेटिएका थिए भने उनकी सासु हरया अर्को कोठा मा भेटिएकी थिइन् । </p>

<p> छिमेकी ले घटना बारे ब्रहमदेव प्रहरी चौकी लाई खबर गरेका थिए । हत्या पछि नाथ परिवार को घर निर्जन बनेको छ । धर्मानन्दका बाबु करवीर भारत को मुम्बई मा काम गर्छन् भने दाजु सशस्त्र प्रहरी मा कार्यरत छन् । करवीर का जेठा छोरा काठमाडौं बाट आएपछि मृतक सबै को महाकाली अञ्चल अस्पताल मा पोस्टमार्टम गरिएको थियो । घटना पछि गाउँ नै त्रासमा छ । महिला तथा बालबालिका बाहिर निस्कन डराइरहेका छन् । गाउँ को बैजनाथ सिद्धनाथ माध्यमिक विध्यालय बन्द छ । फरार धर्मानन्द को खोजी भइरहेको अनुसन्धान का लागि घटनास्थल पुगेका जिल्ला प्रहरी कार्यालय का प्रहरी निरीक्षक चक्र शाह ले बताए । ‘ विभिन्न नाकामा चेकजाँच गरिरहेका छौं , ’ उनले भने । </p>

Output of the program as a parallel corpus 3:

S.N.	English Sentence	Nepali Sentence
1	Four, including two kids, of the same family were brutally murdered at matena village of Bhimdutta	कञ्चनपुर को भीमदत्त नगरपालिका - ९ मटेना मा आइतबार राति एकै घर का दुई

	Municipality - 9 in Kanchanpur district Sunday night.	नाबालक सहित चार को हत्या भएको छ ।
2	The wife, daughter-in-law, grandson and granddaughter of Karbir Nath, a local, were found murdered in their home located ten kilometers from Mahendranagar.	सदरमुकाम महेन्द्रनगर बाट १० किमि टाढा मटेना गाउँ मा स्थानीय करवीर नाथ की पत्नी , बुहारी , नाति र नातिनी को हत्या भएको हो ।
3	Among those killed are Karbir’s 50 year-old wife Harya Nath, 25 year-old daughter-in-law Dhanadevi Nath, six year-old grandson Ravi Nath and four year-old granddaughter Renuka Nath.	मारिनेहरू करवीरकी ५० बर्षीया पत्नी हरया नाथ , उनकी २५ वर्षीया बुहारी धनादेवी नाथ , ६ बर्षीय नाति रवि नाथ र चार वर्षीया नातिनी रेनुका नाथ छन् ।
4	The bodies of all the four slain family members have been kept at Mahakali Zonal Hospital.	मारिएका चारै को शव महाकाली अञ्चल अस्पताल मा राखिएको छ ।
5	All four have rope marks around their necks.	चारै शव को घाँटी मा डोरी को डाम ले बनाएको घाउ छ ।
6	It seems they were strangled to death by the murderer using rope.	हत्यारा ले डोरी ले घाँटी बेरी हत्या गरेको देखिन्छ ।
7	Karbir’s son Dharmananda Nath is missing, prompting suspicion that he killed the four.	करवीरको पाँच जना को परिवार मा छोरो धर्मानन्द नाथ फरार रहेका ले घटना मा उनको संलग्नता आशंका गरिएको छ ।
8	Locals say Dharmananda has been mentally ill for more than one year, adding to the suspicion.	केही वर्षयता धर्मानन्द को मानसिक अवस्था ठीक नभएको स्थानीय ले बताएका छन् ।
9	“The family did not have enmity with anyone,” a neighbor said, adding, “It	“ कसैसित रिसइबी थिएन , झगडा पनि

	seems Dharmananda, who was mentally ill, killed them”.	कसैसित भएको देखेनौं ” एक छिमेकी ले भने , “ मानसिक अवस्था ठीक नभएका ले धर्मानन्द ले नै चारै को हत्या गरे जस्तो देखिन्छ ” ।
10	The suspect was with the family members until 10 pm Sunday, the neighbor said.	राति दस बजे सम्म धर्मानन्द लाई परिवार सँगै देखेको ती छिमेकी ले बताए ।
11	The murder must have taken place after everybody had gone to sleep.	“ राति दस बजे पछि घटना भएको छ, सबै सुतेका बेला हत्या गरेको हुन सक्छ , ” छरछिमेककाहरू भन्दै थिए ।
12	The villagers learnt about the incident only in the morning after they forced their way into the house when nobody from inside responded to their calls.	बाहिर बाट बोलाउँदा कसैले केही जवाफ नदिए पछि भित्र गएर हेर्दा घटनाबारे थाहा पाएका थिए ।
13	DhanaDevi and her kids were found in one room while mother-in-law Harya Nath was found in another.	धानादेवी र उनका दुई नाबालक एउटा कोठा मा भेटिएका थिए भने उनकी सासु हरया अर्को कोठा मा भेटिएकी थिइन् ।
14	The neighbors informed Brahma Dev Police Post about the incident.	छिमेकी ले घटना बारे ब्रहमदेव प्रहरी चौकी लाई खबर गरेका थिए ।
15	Dharmananda’s father Karbir works in Mumbai, India, while elder-brother is in the Armed Police Force (APF).	धर्मानन्दका बाबु करवीर भारत को मुम्बई मा काम गर्छन् भने दाजु सशस्त्र प्रहरी मा कार्यरत छन् ।
16	Post-mortem of the four bodies were done at Mahakali Zonal Hospital after Karbir’s eldest son arrived from Kathmandu.	करवीर का जेठा छोरा काठमाडौं बाट आएपछि मृतक सबै को महाकाली अञ्चल

		अस्पताल मा पोस्टमार्टम गरिएको थियो ।
17	The police have been searching for absconding Dharmananda while the whole village is living in terror with Baijanath Siddhanath Secondary School shut-down after the incident.	धर्मानन्दका बाबु करवीर भारत को मुम्बई मा काम गर्छन् भने दाजु सशस्त्र प्रहरी मा कार्यरत छन् ।

Table A.2: Aligned Parallel Sentences in Comparable Corpus 3

Here, in this corpus following measures are found:

Total No. of Sentences in Source Corpus	17
Total No. of sentences in Target Corpus	30
No. of aligned sentences	17
No. of correctly aligned sentences	16
Correct alignments present in the Corpora	17

Precision	0.941176
Recall	0.941176
F-measure	0.941176

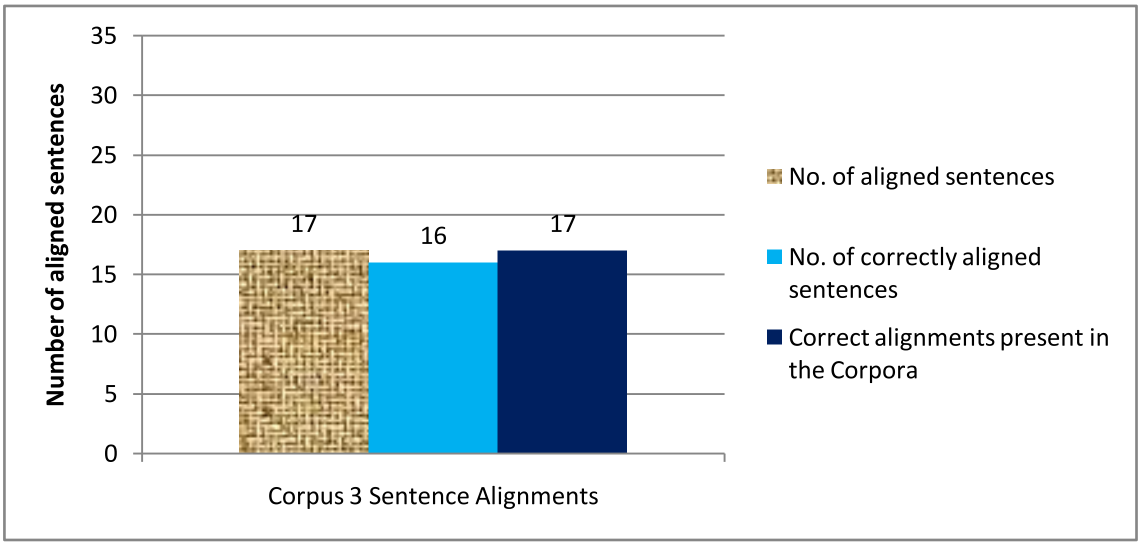


Figure A.2: - Sentence alignment in Comparable Corpus 3

Comparable Corpus 4 (English) [3; 19]

<p> This book is on the table. This is a book. I have to go to hospital. We read in class five. Boys are going to school. A boy is going to school. A boy is sitting on the table. A boy is singing a song. This is a good song. She is a good girl. I go to school. Boys are playing on the ground. A girl is sitting on the ground. She reads book. I am a teacher. A teacher teaches this lesson. This school has a good ground. This song is very popular. He teaches in the school. My country is Nepal. He has a book. He is popular. This lesson is on this book. </p>

Comparable Corpus 4 (Nepali) [3; 19]

<p> यो किताब टेबुल मा छ । म र मेरो साथी संग संगै किताब पढछौं । यो एउटा किताब हो । म काठमाण्डौं मा बस्छु । हामी कक्षा पाँच मा पढ्छौं । केटाहरू विध्यालय जांदै छन् । एउटा केटा विध्यालय जाँदै छ । एउटा केटा टेबुल मा बसिरहेको छ । एउटा केटा गीत गाउँदै छ । यो एउटा राम्रो गीत हो । ऊनी एउटा राम्रो केटी हुन् । म विध्यालय जान्छु । केटाहरू चौर मा खेल्दै छन् । एउटा केटी चौर मा बसिरहेकि छ । ऊनी किताब पढ्छिन् । म एउटा शिक्षक हुँ । मलाइ किताब पढ्न साह्रै राम्रो लाग्छ । एउटा शिक्षक यो पाठ पढाउँछ । यो विध्यालय संग राम्रो चौर छ । यो गीत धेरै प्रख्यात छ । ऊ विध्यालय मा पढाउँछ । मेरो देश नेपाल हो । ऊ संग किताब छ । ऊ प्रख्यात छ । यो पाठ यो किताब मा छ । </p>

Output of the program as a parallel corpus 4:

S.N.	English Sentence	Nepali Sentence
1	This book is on the table.	यो किताब टेबुल मा छ ।
2	This is a book.	यो किताब टेबुल मा छ ।
3	I have to go to hospital.	यो किताब टेबुल मा छ ।
4	We read in class five.	हामी कक्षा पाँच मा पढ्छौं ।

5	Boys are going to school.	केटाहरू विध्यालय जाँदै छन् ।
6	A boy is going to school.	एउटा केटा विध्यालय जाँदै छ ।
7	A boy is sitting on the table.	एउटा केटा टेबुल मा बसिरहेको छ ।
8	A boy is singing a song.	एउटा केटा गीत गाउँदै छ ।
9	This is a good song.	यो एउटा राम्रो गीत हो ।
10	She is a good girl.	ऊनी एउटा राम्रो केटी हुन् ।
11	I go to school.	म काठमाण्डौं मा बस्छु ।
12	Boys are playing on the ground.	केटाहरू चौर मा खेल्दै छन् ।
13	A girl is sitting on the ground.	एउटा केटी चौर मा बसिरहेकि छ ।
14	She reads book.	ऊनी किताब पढ्छिन् ।
15	I am a teacher.	म एउटा शिक्षक हुँ ।
16	A teacher teaches this lesson.	एउटा शिक्षक यो पाठ पढाउँछ ।
17	This school has a good ground.	यो विध्यालय संग राम्रो चौर छ ।
18	This song is very popular.	यो गीत धेरै प्रख्यात छ ।
19	He teaches in the school.	ऊ विध्यालय मा पढाउँछ ।
20	My country is Nepal.	मेरो देश नेपाल हो ।
21	He has a book.	ऊ संग किताब छ ।
22	He is popular.	ऊ संग किताब छ ।
23	This lesson is on this book.	यो पाठ यो किताब मा छ ।

Table A.3: Aligned Parallel Sentences in Comparable Corpus 4

Here, the following measures are found:

Total No. of Sentences in Source Corpus	23

Total No. of sentences in Target Corpus	25
No. of aligned sentences	23
No. of correctly aligned sentences	19
Correct alignments present in the Corpora	22

Precision	0.826087
Recall	0.863636
F-measure	0.844444

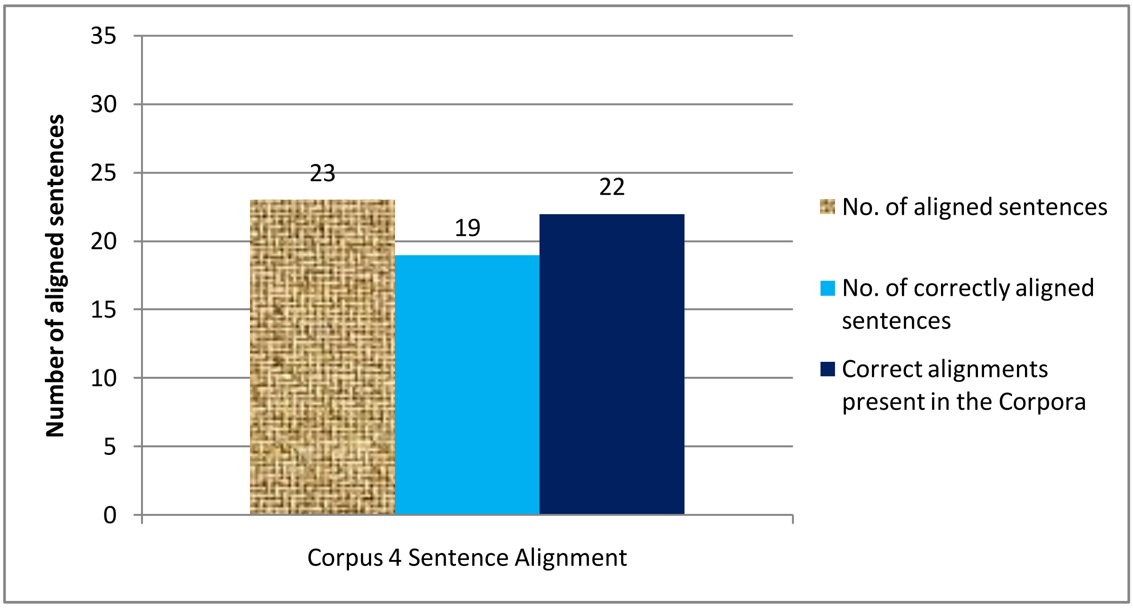


Figure A.3: - Sentence alignment in Comparable Corpus 4

Comparable Corpus 5 (English)

[*Social Studies – Grade 9*, Readmore Publishers and distributors,
Second Edition-2066 B.S., page no. 17]

<p> The basic facilities, means, resources and services available for the economic and social development are called infrastructures of development. Apart from the means and resources, a stable government and well developed infrastructures is required for the development of a country. People living in a community should be provided certain facilities and services. There is a great importance of the basic facilities and services in the economic and social development of a country. Such facilities and services are available in developed countries. Whereas, people are well educated and they unite together to bring about development in their countries. </p>

<p> Nepal is a mountainous country. Geographically, there are different kinds of landscapes ranging from high mountains (Himalayas), hills, valleys, inner plains and the Terai in Nepal. There are many rivers flowing down from the Himalayas. These rivers and rivulets are not only sources of drinking water, but they can be used to operate water-mills, small-scale hydroelectric projects, irrigation canals, etc. The undulating landscapes and varying elevations of our country are in fact nature's resources. It is important for us to utilize these means and resources properly for our economic and social development. </p>

Comparable Corpus 5 (Nepali)

[सामाजिक अध्ययन - कक्षा ९, प्रकाशक: नेपाल सरकार, शिक्षा मन्त्रालय, पाठ्यक्रम विकास केन्द्र, सानोठिमी, भक्तपुर - २०६४ पेज न.१४]

<p> मुलुक को आर्थिक तथा सामाजिक विकास का निम्ति उपलब्ध गराइने आधारभूत सेवा सुविधालाई विकास का पूर्वाधार भनिन्छ । देश विकास का लागि माथि दिइएका स्रोत र साधन का अतिरिक्त विकास का पूर्वाधार , निर्माण तथा कार्यान्वयन गर्न स्थिर सरकार को आवश्यकता पर्दछ । समुदाय मा बस्ने मानिसहरू का लागि निश्चित किसिम का सेवा र सुविधाहरू उपलब्ध गराइएको हुनु पर्दछ । मुलुक को आर्थिक तथा सामाजिक विकास का लागि आधारभूत सेवा तथा सुविधाहरू को ठूलो महत्व हुन्छ । विकसित मुलुकहरू मा त्यस्ता आधारभूत सेवा र सुविधाहरू उपलब्ध हुन्छन् , त्यस्ता मुलुकहरू मा मानिसहरू शिक्षित हुन्छन् र आफै जुटेर त्यस कार्य मा लाग्दछन् ।
</p>

<p> नेपाल एक पहाडी मुलुक हो । नेपाल को भू-धरातल हेर्ने हो भने उत्तर तिर रहेका उच्च हिमपर्वतहरू देखि लिएर दक्षिण मा समथल मैदान , पहाडी क्षेत्र , उपत्यका आदि रहेका छन् । यहाँ हिमालय बाट हिउँ पग्लेर आएका नदीहरू छन् । यी नदी , छाँगा-छहरा ले गर्दा हामी लाई पिउने पानी उपलब्ध हुनुका साथै सोही पानी को सदुपयोग गरेर ठाउँठाउँ मा पानीघट्ट , स-साना जलविध्युत् आयोजना , सिँचाइ को कुलो

पनि बनाइएका छन् । हाम्रो देश मा अग्लोहोचो भूधरातल नै प्राकृतिक वरदान बनेको छ । तिनको सदुपयोग गर्न हामीलाई तालिम प्राप्त प्राविधिक जनशक्ति को खाँचो छ । हामी ले यी प्राकृतिक स्रोत तथा साधन लाई प्रयोग मा ल्याउनु आवश्यक छ ।
</p>

Output of the program as a parallel corpus 5:

S.N.	English Sentence	Nepali Sentence
1	The basic facilities, means, resources and services available for the economic and social development are called infrastructures of development.	मुलुक को आर्थिक तथा सामाजिक विकास का निम्ति उपलब्ध गराइने आधारभूत सेवा सुविधालाई विकास का पूर्वाधार भनिन्छ ।
2	Apart from the means and resources, a stable government and well developed infrastructures is required for the development of a country.	देश विकास का लागि माथि दिइएका स्रोत र साधन का अतिरिक्त विकास का पूर्वाधार , निर्माण तथा कार्यान्वयन गर्न स्थिर सरकार को आवश्यकता पर्दछ ।
3	People living in a community should be provided certain facilities and services.	समुदाय मा बस्ने मानिसहरू का लागि निश्चित किसिम का सेवा र सुविधाहरू उपलब्ध गराइएको हुनु पर्दछ ।
4	There is a great importance of the basic facilities and services in the economic and social development of a country.	मुलुक को आर्थिक तथा सामाजिक विकास का लागि आधारभूत सेवा तथा सुविधाहरू को ठूलो महत्व हुन्छ ।
5	Such facilities and services are available in developed countries.	विकसित मुलुकहरू मा त्यस्ता आधारभूत सेवा र सुविधाहरू उपलब्ध हुन्छन् ,

		त्यस्ता मुलुकहरू मा मानिसहरू शिक्षित हुन्छन् र आफै जुटेर त्यस कार्य मा लाग्दछन् ।
6	Whereas, people are well educated and they unite together to bring about development in their countries.	विकसित मुलुकहरू मा त्यस्ता आधारभूत सेवा र सुविधाहरू उपलब्ध हुन्छन् , त्यस्ता मुलुकहरू मा मानिसहरू शिक्षित हुन्छन् र आफै जुटेर त्यस कार्य मा लाग्दछन् ।
7	Nepal is a mountainous country.	नेपाल एक पहाडी मुलुक हो ।
8	Geographically, there are different kinds of landscapes ranging from high mountains (Himalayas), hills, valleys, inner plains and the Terai in Nepal.	नेपाल को भू-धरातल हेर्ने हो भने उत्तर तिर रहेका उच्च हिमपर्वतहरू देखि लिएर दक्षिण मा समथल मैदान , पहाडी क्षेत्र , उपत्यका आदि रहेका छन् ।
9	There are many rivers flowing down from the Himalayas.	यहाँ हिमालय बाट हिउँ पग्लेर आएका नदीहरू छन् ।
10	These rivers and rivulets are not only sources of drinking water, but they can be used to operate water-mills, small-scale hydroelectric projects, irrigation canals, etc.	यी नदी , छाँगा-छहरा ले गर्दा हामी लाई पिउने पानी उपलब्ध हुनुका साथै सोही पानी को सदुपयोग गरेर ठाउँठाउँ मा पानीघट्ट , स-साना जलविध्युत् आयोजना , सिँचाइ को कुलो पनि बनाइएका छन् ।
11	The undulating landscapes and varying elevations of our country are in fact nature’s resources.	हाम्रो देश मा अग्लोहोचो भूधरातल नै प्राकृतिक वरदान बनेको छ ।

12	It is important for us to utilize these means and resources properly for our economic and social development.	यी नदी , छाँगा-छहरा ले गर्दा हामी लाई पिउने पानी उपलब्ध हुनुका साथै सोही पानी को सदुपयोग गरेर ठाउँठाउँ मा पानीघट्ट , स-साना जलविध्युत् आयोजना , सिँचाइ को कुलो पनि बनाइएका छन् ।

Table A.4: Aligned Parallel Sentences in Comparable Corpus 5

Here, in this corpus we found following measures:

Total No. of Sentences in Source Corpus	12
Total No. of sentences in Target Corpus	12
No. of aligned sentences	12
No. of correctly aligned sentences	11
Correct alignments present in the Corpora	12

Precision	0.916667
Recall	0.916667
F-measure	0.916667

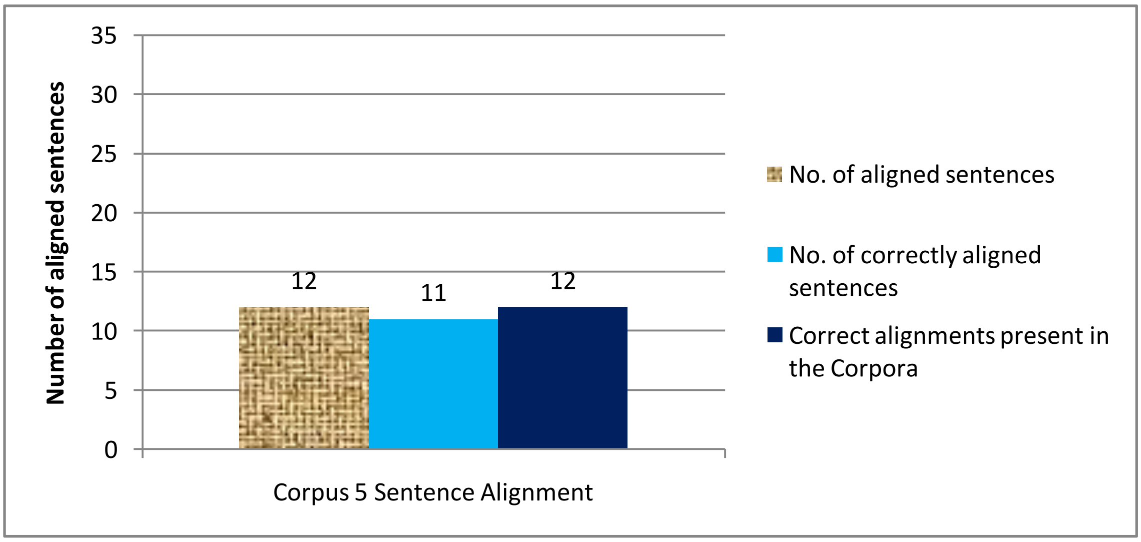


Fig: - Sentence alignment in Comparable Corpus 5

Appendix B: Code for Implementation

Source Code to make Dictionary/to read Dictionary

```
 public static String makeDictionary(String arg) throws Exception
  {
      int count = 0;
      String test[] = arg.split(" ");
      ArrayList<Integer> list = new ArrayList<Integer>();
      for(int j=0; j<test.length; j++)
      {
         count = 0;
         Scanner in = new Scanner(new File("D:\\thesis\\dictionary\\dicEng.txt"));
         while(in.hasNext())
         {
            String s = in.next();
            if(s.equalsIgnoreCase(test[j]))
            {
               list.add(count);
            }
            count++;
         }
         in.close();
      }
      String nep = "";
      for(int i=0; i<list.size(); i++)
      {
         count = 0;
         Charset ch1 = Charset.forName("UTF-8");
         FileInputStream ncorpus1 =  new
FileInputStream("D:\\thesis\\dictionary\\dicNep.txt");
         InputStreamReader irn1 = new InputStreamReader(ncorpus1,ch1);
         BufferedReader nr1 = new BufferedReader(irn1);
         while(true)
```

```
            {
                String str = nr1.readLine();
                if(count == list.get(i))
                {
                    nep = nep + str + " ";
                    break;
                }
                count++;
            }
            nr1.close();
        }
        return nep;
    }//method makeDictionary ends here
```

Source code to make paralale corpora

```
public static void makeParallelSentences() throws Exception
    {
        int countPoint = 0;
        //for wrtintg in neplai file
        Charset ch1 = Charset.forName("UTF-8");
        FileOutputStream ncorpus1 =  new
FileOutputStream("D:\\thesis\\dictionary\\paralelNep.txt");
        OutputStreamWriter irn1 = new OutputStreamWriter(ncorpus1,ch1);
        BufferedWriter nr1 = new BufferedWriter(irn1);

        //for writing in english file
        PrintStream out = new PrintStream(new
File("D:\\thesis\\dictionary\\paralelEng.txt"));

        for(int i=0; i<strEngFinalParaArr.length; i++)
        {
            int n = countLine(strEngFinalParaArr[i]);
            String engLine[] = new String[n];
            int m = countLine(strNepFinalParaArr[i]);
            String nepLine[] = new String[m];

            int index = 0;
            String s = "";
            //tokenize the english sentecnes from ith paragraph
            for(int j=0; j<strEngFinalParaArr[i].length(); j++)
            {
              s = s + String.valueOf(strEngFinalParaArr[i].charAt(j));
              if(strEngFinalParaArr[i].charAt(j) == '.')
              {
                engLine[index] = s;
                index ++;
                s = "";
              }
            }
            index  = 0;
            s = "";
            //tokenize the nepali sentecnes from ith paragraph
            for(int j=0; j<strNepFinalParaArr[i].length(); j++)
            {
              s = s + String.valueOf(strNepFinalParaArr[i].charAt(j));
              if(strNepFinalParaArr[i].charAt(j) == '।')
              {
                nepLine[index] = s;
                index ++;
                s = "";
              }
            }
FileOutputStream("D:\\thesis\\dictionary\\paralelNep.txt");
            int termFrequency = 0;
            int max = 0;
            String ansNep = "";
String hee = "";
for(int j=0; j<engLine.length; j++)
{
   max = 0;
   out.println(engLine[j]);
   out.println();
   String eq = makeDictionary(engLine[j].substring(0, engLine[j].length()-2));
   String nepArr[] = eq.split(" ");
   for(int k=0; k<nepLine.length; k++)
   {
      String nepLineArr[] = nepLine[k].split(" ");
      for(int i1=0; i1<nepArr.length; i1++)
      {
         for(int j1=0; j1<nepLineArr.length; j1++)
         {

           if(nepArr[i1].equalsIgnoreCase(nepLineArr[j1]))
                {
                   termFrequency = termFrequency + 1;
                   break;//continue;
                }
         }
      }
      if(max < termFrequency)
      {
         //System.out.println("max = "+max);
         max = termFrequency;
         ansNep = nepLine[k];
         hee = hee + ansNep + ":" + max + "\n";
            // termFrequency = 0;
      }
      termFrequency = 0;
   }
   //write the ansNep to parlalelNep file
```

```
                nr1.write(ansNep);
                nr1.newLine();
            }

            System.out.println();
      }
        nr1.close();
            out.close();
        System.out.println();
```

Source code for parsing file

```
String nepText = "";
        Charset ch1 = Charset.forName("UTF-8");
        FileInputStream ncorpus1 =  new FileInputStream("D:\\thesis\\comparable
corpora\\NepaliCorpora.txt");
        InputStreamReader irn1 = new InputStreamReader(ncorpus1,ch1);
        BufferedReader nr1 = new BufferedReader(irn1);
        while(true)
        {
            String str = nr1.readLine();
            if(str == null)
                break;
            nepText = nepText +str;
        }
        nr1.close();
        String strNepArr[] = nepText.split("<p>");

        //put all final paragraphs
        strNepFinalParaArr = new String[strNepArr.length - 1];
        int count = 0;
        for(int i=1; i<strNepArr.length; i++)
        {
            int n = strNepArr[i].length();
```

```
    strNepFinalParaArr[count] = strNepArr[i].substring(0,n-5);
    count++;
  }

  //read englsih corpus
  Scanner fileEngInput = new Scanner(new File("D:\\thesis\\comparable corpora\\
EnglishCorpora.txt"));
  String strEng = "";
  while(fileEngInput.hasNext())
  {
    strEng = strEng +  fileEngInput.nextLine();
  }
  fileEngInput.close();

  System.out.println("Englsih string is "+ strEng);
  String strEngArr[] = strEng.split("<p>");

  strEngFinalParaArr = new String[strEngArr.length-1];
  count = 0;
  for(int i=1; i<strEngArr.length; i++)
  {
    strEngFinalParaArr[count] = strEngArr[i];
    count++;
  }
```