Handling Short Form Words for Machine Translation

A Dissertation

Submitted To

Central Department of Computer Science and Information Technology

Tribhuvan University

Kirtipur, Nepal

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

in

Computer Science and Information Technology

Submitted By

Roshan Silwal

CDCSIT, TU

(2011)

Handling Short Form Words for Machine Translation

A Dissertation

Submitted To
Central Department of Computer Science and Information Technology
Tribhuvan University
Kirtipur, Nepal

In Partial Fulfillment of the Requirements for the Degree of
Master of Science
in
Computer Science and Information Technology

Submitted By
Roshan Silwal
CDCSIT, TU

Supervisor: **Prof. Dr. Shashidhar Ram Joshi**

Co-Supervisor: **Mr. Bikash Balami**

Handling Short Form Words for Machine Translation

Date :- ……………….

Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Roshan Silwal** entitled “**Handling Short Form Words for Machine Translation”** be accepted as in fulfilling part requirements for the degree of Master of Computer Science.

Prof. Dr. Shashidhar Ram Joshi
Head of Department
Department of Electronics and Computer Engineering, Institute of Engineering Pulchowk, Lalitpur
(**Superviso**r)

Handling Short Form Words for Machine Translation

Date :- ……………….

Recommendation

I hereby recommend that the dissertation prepared under my co-supervision by **Mr. Roshan Silwal** entitled “**Handling Short Form Words for Machine Translation”** be accepted as in fulfilling part requirements for the degree of Master of Computer Science.

Mr. Bikash Balami
Lecturer
Central Department of Computer Science and Information Technology
Tribhuvan University, Kirtipur, Kathmandu
(**Co-Superviso**r)

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

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

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

Prof. Dr. Shashidhar Ram Joshi
Head of Department
Department of Electronics and
Computer Engineering,
Institute of Engineering,
Pulchowk, Nepal
(Supervisor)

(External Examiner)

(Internal Examiner)

Acknowledgement

I would like to express my gratitude to all the people who supported and accompanied me during the preparation of this dissertation for **Handling Short Form Words for Machine Translation**

This research work has been performed under Central Department of Computer Science and Information Technology (***Tribhuwan University***), Kirtipur. I also thank all team members of my department for giving me an enthusiastic support.

I would like to thank my supervisor **Professor Dr. Shashidhar Ram Joshi,** head of the Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk.. Without his help and suggestion this research could not have been done.

Thanks also to my co-supervisor **Mr. Bikash Balami** who gave me full support from the beginning to the end of the preparation of this dissertation. I would always be grateful for his insightful and guidance he provides me in order to cross the final hurdle.

Most importantly I would like to thank to respected Head of Department of Central Department of Computer Science and Information Technology, Assoc. Prof. Dr. Tanka Nath Dhamala, respected teachers Prof. Sudarsan Karanjit, Assoc. Prof. Dr. Subarna Sakya, Mr. Min Bahadur Khati, Mr. Bishnu Gautam, Mr. Dinesh Bajracharya, Mr. Navaraj Poudel, Mr. Arjun Singh Saud, Mr. Jagdish Batta, Mr. Yog Raj Joshi of CDCSIT, TU, for providing me such a broad knowledge and inspirations within the time period of two years.

Special thanks to my family, specially my dearest, lovely and respected ***Dad*** , my brother ***Rabin Silwal***, sister ***Punam Laxmi Silwal*** and ***Puspa Laxmi Silwal*** for being part of every bit of my life. Their endless motivation, constant mental support and unconditional love have been influential in whatever I have achieved so far.

I wish to thank to all my colleagues and friends especially **Mr. Dinesh Kumar Khadka,** Ms. Neelam Lama, Mr Arun Patel, Mr. Rajendra Lammichanne, Mr. Pravakar Ghimire, Mr Satya Bahadur Maharjan, for supporting me directly and indirectly in this research work.

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

Abstract

The short forms which are being originated from letter omission, word truncation and the substitution of the part of the word (which may be numeric or alphabetic) are being handled by this research. The use of such short word is being increasing day by day because people now these days are very busy; to make the communication fast and in easier manner they use such short word. So in the field of machine translation also such words may be used but the machine has to understand such word and should be capable to translate it. In this research this problem is being handled by tokenizing the input sentence to find the short word then translate the detected short word into corresponding long form that will be in the vocabulary of machine.

Objective

The objectives of this research are:

- To analyze the source sentence in order to find out the short forms
- To translate the short forms present in source sentence into their respective long forms (meaning full form)
- To improve the quality of the translation

Dedicated

To

My lovely family Dad, Mum, Brother and Sisters

Table of Contents

Details **Pages**

CHAPTER 1

Modern Communication Strategy

CHAPTER 2

Natural Language Processing

CHAPTER 3

Problem Definition

CHAPTER 4

Implementation

CHAPTER 5

Testing and Analysis

CHAPTER 6

Conclusion and Future Works

List of Tables

List of Figures

CHAPTER 1

Modern Communication Strategies

1. Introduction

Communication is the sharing of ideas or information from one person to another. Most of the people think about the communication as in oral or written form but it can be much more. A knowing look or a gentle touch can also convey a message loud and clear, as can a hard push or an angry slap. It is the process in which speech, signs or actions helps to convey or transmit information from one person to another. This definition is concise and definitive but doesn't include all the aspects of communication. In another way communication can be defined as the process that involves transmitting information from one party to another [14]. Whether or not the receiver understood the full meaning of the message that has been sent by the sender but it is called as the communication. But communication is called as better communication when both parties understand. But it can still exist even without that component. No matter the type or mechanism of communication, every instance of communication must have a message that is being transferred from sender to receiver. Sometimes while communicating, the sender and receiver may use some signs, words or signals in common with each other so the sent message can be understood. The ideal definition of communication is a two-way interaction between two parties (receiver and sender) to transmit information and mutual understanding between them self. The interchange or exchange of information from sender to receiver is best communicated when a discussion is available so the receiver can ask questions and receive answers to clarify the message. Communication requires a sender, a message, and an receiver, although the receiver need not be present or aware of the sender's intent to communicate at the time of communication; thus communication can occur across vast distances in time and space. Communication requires that the communicating parties share an area of communicative commonality. The communication process is complete once the receiver has understood the sender. The different kinds of Communication are discussed in further sections.

1.1Non Verbal Communication

Non-verbal communication can be ‘Visual’, ‘Aural’ or ‘Gestural’. Sometimes we look into some pictures, graphs, symbols, diagrams etc. and some message of that will be conveyed to us. All these are different forms of visual communication. For example, the traffic policeman showing the stop sign, a teacher showing a chart of different animals is visual communication. Bells, whistles, buzzers, horns etc. are also the instruments through which we can communicate our message. Communication through the help of these types of sounds is called 'aural' communication. For example, the bell used in schools and colleges to inform students and teachers about the beginning or end of periods, siren used in factories to inform the change of work–shift of the workers are examples of aural communication. Communication through the use of various parts of the human body, or through body language is termed as gestural communication. Saluting our national flag, motionless position during the singing of national anthem, waving of hands, nodding of head, showing anger on face, etc. are examples of gestural communication.

1.1.1 Emoticons or Smileys

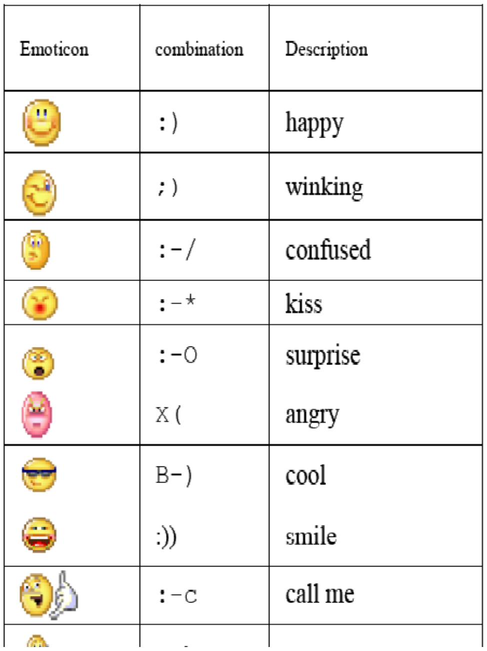
Emoticon or sometimes called as Smiley, can also be used for representation of the non-verbal communication which, is a sequence of ordinary characters found on the computer keyboard. These emoticons are generally used in e-mail while communicating chat SMS and other modes of communication especially using a computer [1]. The most popular emoticons are those ‘smiling faces generally known as smileys which are used to convey moods. In smileys the colon represents the eyes, the dash represents the nose, and the left/right parenthesis represents the mouth.

For example **:-(**

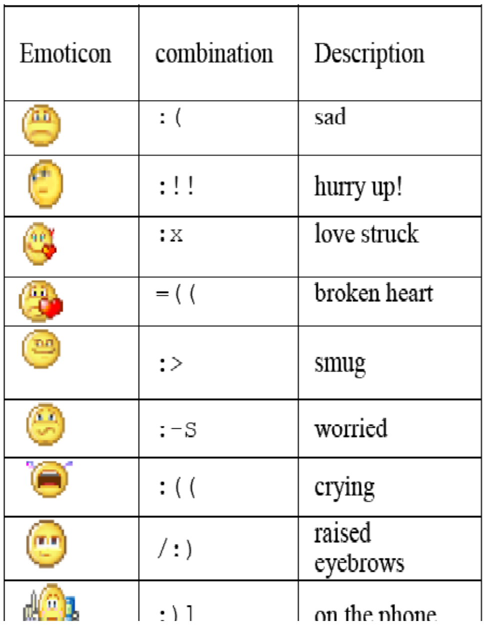
TEX	MEANING	TEXT	MEANING
(:-)	smiling with helmet	:-)=	smiling with a beard
:')	happy and crying	:-)8	smiling with bow tie
:-(	sad	:(	sad, without nose

:'-(	sad and crying	>:-(	very angry
:-~)	having a cold	>: - (	angry, yet sad
:-( )	shocked	>: -|	cross
:-*	kiss	:-\	skeptical
:-v	talking	:-#	razes
:-w	talking with two tongues	:-x	not saying a word
:-<	cheated	;)	twinkle (wink), without nose
:-<>	surprised	;-)	twinkle (wink)
:-x	small kiss	<:-|	monk / nun
:-X	big sloppy kiss	: @	shouting
:-)	smiling	:-(0)	shouting
:-9	salivating	:‘‘-D	crying with laughter
:-c	unhappy	-:-)	punk
:-||	angry	:-*	bitter
:-o	appalled	:^)	broken nose
:-O	wow	:-o zz	bored

Table 1.1 Emoticons [1]



Emoticon	combination	Description
	:)	happy
	;)	winking
	: -/	confused
	: -*	kiss
 	: -O X (	surprise angry
 	B-) :))	cool smile
	: -c	call me



Emoticon	combination	Description
	: (	sad
	: ! !	hurry up!
	: x	love struck
	= ((	broken heart
	: >	smug
	: -S	worried
	: ((	crying
	/ :)	raised eyebrows
	:)]	on the phone

Table 1.2: Smileys[1]

1.2 Verbal Communication

Verbal communication can be used by using words or sometimes either by spoken or written form. Communication through spoken words is known as oral communication, which may be in the different form sometimes it may be in the form of lectures, meetings, group discussions, conferences, telephonic conversations, radio message etc. In written communication, message is being written then it helps to transmit through written words in the form of letters, memos, circulars, notices, reports, manuals, magazines, handbooks, etc. while conveying the message by using the different communicating devices like mobile phone, computer different short words, actual word and acronyms can be used.

1.2.1 Abbreviation

An abbreviation is a shortened form of collection word or phrase. Usually, it consists of a letters or group of letters taken from the phrase. Usually abbreviations are the contraction of phrase that is used in the place of their full version, where there meaning is clear from the text. For example, we can represent the word "***WTO***" as "***World Trade Organization***".

A. Acronym

It is a technique to create of new words, in easy manner sometimes through the alphabetism and acronyms, the abbreviations usually originated. Alphabetism is an abbreviation consisting of the first letter or letters of constituent words in a phrase (for example, ***WHO*** (for World Health Organization), syllables or components of a word ***TNT*** *(*for *trinitrotoluene*), or a combination of words and syllables (***ESP*** for *extrasensory perception*) and pronounced by spelling out the letters one by one rather than as a constant word. Text message users also employ the above method in an attempt to communicate with others.

B. Alphabetism

Acronym is yet another unique method whereby each initial letter of words is taken to form a new pronounceable word such as *NATO* (North Atlantic Treaty Organization) and *UNESCO* (United Nations Educational, Scientific and Cultural Organization) or *AIDS* (Acute Immune

Deficiency Syndrome) and is pronounced as a new word. In a similar vein text message users use the initial alphabets of words with which they convey their messages economically, conveniently and speedily. These are of different types. Acronyms are often capitalized. Occasionally, acronyms form a word whose original meaning is nearly forgotten, such as "***laser***"(*l*ight *a*mplification by *s*timulated *e*mission of *r*adiation). Acronyms in SMS text messages are different from those in traditional texts, such as conference papers, scientific journals and news articles, as the users can and do coin new words whimsically.

1.2.2 Short Form Words

A short form can be defined as the shortened part of the single word. Usually such short forms are used to make the communication fast and in easy manner also. For example sometimes ***"2"*** can be used instead of ***"two"***, ***"goin"*** can be used instead of ***"going***".

.

CHAPTER 2

Natural Language Processing

2.1 Introduction

Among the various fields in AI, Natural Language Processing (NLP) is one of the researchable fields. In simpler way NLP can also be defined as the field that deals with the computer processing of natural languages, mainly evolved by people working in the field of Artificial Intelligence [9]. It is method of human-computer interaction which enables computers to extract meaning from the words and phrases that people use and respond in kind when presenting information back to them. Because of its different features it has become, very active area of research and development. In the field of natural language processing the language may be taken as the text which may be in the form of oral or written form. The goal of the Natural Language Processing (NLP) specially is to design and make software which will analyze, understand, and generate languages that humans use naturally. This goal is not easy to obtain "Understanding" language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful/ understandable way. The challenges that can be raised in this field are the sentence that is given as the input is highly ambiguous due to the nature of natural language. NLP researchers aim to gather/generate knowledge on the thing like how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks.

2.2 Machine Translation

2.2.1 Introduction

The translation of one natural language (called as the source language) into another language (i.e. target language) by the help/ use of the different computer is called the machine translation [6, 7]. As the first phase it was just like a dream (in seventeenth centaury) and has become popular in twentieth.). It may be done with or without human assistance. The source and target languages are natural languages such as English, Nepali, Chinese, French etc. here in this research the source language is contained as input as short form words and the corresponding long form of the user input short form is the target language. The importance of machine translation has increased every, especially in the field of business, economics and industrialization. Machine translation system can be further divided into two sub-systems.

2.2.2 Direct System

The first generation of the machine translation is the direct system. In such systems, translation is done word by word or phrase by phrase. Especially it is dependent on the large bilingual dictionary. For each word that is found on the source language, the dictionary specifies a set of rules for translation and then that word will be translated. The analysis that is done on this type of direct system is very less or least for each the source text. In this type of the translation, when the words are being translated, after that simple reordering rules are applied. The basic characteristic for such type of translation is that it is very simple and one needs to replace a word of source language to a word in target language using a bilingual dictionary. This kind of the machine translation has some problem like lack of any analysis of the source language causes several problems, for example

- Difficult or impossible to capture long-range reordering.
- Words are translated without disambiguation of their syntactic role

2.2.3 Indirect System

The second generation machine translation is termed out as the Indirect System. The major problem of the first generation MT i.e. Direct Translation was the lack of linguistic information about source text, researchers therefore moved on to finding ways to capture this information. This gave rise to the development of the indirect MT systems which are generally regarded as second generation MT systems. The indirect method occupies the level above direct translation in the MT pyramid and also called the Transfer Based MT System or linguistic knowledge (LK) translation. The transfer architecture not only translates at the lexical level but it also translates syntactically and sometimes semantically. The transfer method will first parse the sentence of the source language then will apply rules that map the grammatical segments of the source sentence to a representation in the target language. After syntactically and semantically analyzing the sentence, we can easily translate a sentence even with different structures. That means this approach uses two transfer rules i.e. lexical rules, Syntactic rules, in this approach word reordering is also done. Suppose in English the word order in sentence is SVO when translated into Nepali, the word order of the translated sentence will be SOV.

2.2.4 Statistical Machine Translation

Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability (but here in this research it is done with the help of frequency). The best translation, of course, is the sentence that has the highest probability [10, 13]. The key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability. A string of Source words, which is assumed here as "s", can be translated into a string of target words in many different ways. Often, knowing the broader context in which *s* occurs may serve to winnow the field of acceptable Target translations, but even so, many acceptable translations will remain; the choice among them is largely a matter of taste. In statistical translation, we take the view that every target string say *l* is a possible translation of source string say *s*. We assign to every pair of strings (*s*, *l*) a number Pr ***(l | s)***, which we interpret as the probability that a translator, when presented with *s*, will produce *l* as his translation. Given a target string l, the job of our translation system is to find the string *s* that the native speaker had in mind when he produced *l.* We minimize our chance of error by choosing that source string *s* for which Pr ***(l | s)***) is greatest. Using Bayes' theorem, we can write

$$\Pr \boldsymbol{(s\,|l)} = \frac{\boldsymbol{P}\,(\boldsymbol{s}) * \boldsymbol{P}(\boldsymbol{l}\mid\boldsymbol{s})}{\boldsymbol{P}(\boldsymbol{l})} \;\ldots\ldots\; (1)$$

Now, we arrive at the Fundamental Equation of Machine Translation:

$$\mathbf{e} \;= \text{argmax} \frac{\boldsymbol{P}\,(\boldsymbol{s}) * \boldsymbol{P}(\boldsymbol{l}\mid\boldsymbol{s})}{\boldsymbol{P}(\boldsymbol{l})} \;\ldots\ldots\ldots\ldots\; (2)$$

Here agrmax is being chosen in order to represent the maximum probability.

2.2.5 Corpus and Corpora

In simpler way corpus can be defined as the collection of large structural set of text. Depending on the types of language that are contained on the text, a corpus may be monolingual (consisting the only one language), or multilingual (consisting more than two language) or the bilingual (consisting only two language) [8]. A parallel corpus is a collection of texts in different languages where one of them is the original text and the other is their translations. Parallel corpora are very important resources for tasks in the translation field like linguistic studies, information retrieval systems development or natural language processing. The corpus here made in this research is monolingual which contains the English sentence one consisting the corpus with the short form word which has to be handle here and another consisting the corpus with long form both corpus are aligned here. So the main aim in this dissertation is to convert the short form words contained in the source sentence into respective long form by the help of the long form contained in the long form corpus.

2.3 Challenges in Machine Translation

Although the ultimate goal of MT, as in AI, may be to equal the best human efforts, the current targets are much less ambitious. MT aims not to translate literary work, but technical documents, reports, instruction manuals etc. Even here, the goal usually is not fluent translation, but only correct and understandable output. Some challenges are given below [13].

2.3.1 Ambiguities

Words and phrases in one language which is entered in a machine often map to multiple words in another language. For example, in the sentence,

I went to the bank.

It is not clear whether the "mound of sand" sense or the "financial institution" sense is being used for the word ***bank***. This will usually be clear from the context, but this kind of disambiguation is generally non-trivial. Phrasal verbs are another feature that are difficult to handle during translation. Consider the use of the phrasal verb *bring up* in the following sentence,

They brought up the child in luxury.

They brought up the table to the first floor.

They brought up the issue in the house.

Yet another kind of ambiguity that is possible is structural ambiguity:

Flying planes can be dangerous.

This can be translated as either of the following two sentences. Depending on whether it is the planes that are dangerous or the occupation of flying them that is dangerous.

2.3.2 Unknown Words

Unknown words are a major problem for every machine translation system. The word that is given by the user, if it is not included in the corpus then it will be unknown.

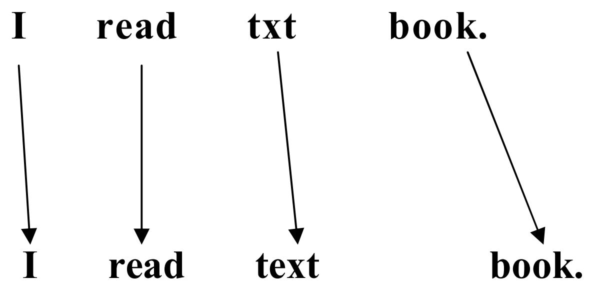
For example

Narayan Gopal is a famous singer.

If any word present in the above sentence is not present in the corpus then that word will become the unknown word.

2.3.3 Short Form Words

Short form word can be defined as the shortened part of single word. Acronyms looks like a short form word. But in real case the formation of the Acronyms is entirely different from the short form word. For example “***txt***” is a short form of ***text*** and ***www*** is a acronym of ***World Wide Web.*** For example**:**



In above example, “***txt***” is the short form words. Due to the presence of the such short forms in source sentence, often it is difficult to predict what actually the sentence is trying to say i.e. the meaning of the sentence can be unknown

CHAPTER 3

Problem Definition

3.1Background

Communication is a process by which we assign and convey meaning in order to pass on certain message/s to others. Human communication is considered the highest developed form of communication among animal communication systems. Both verbal means, as well as non-verbal methods are used to communicate with each other. This is successfully achieved in human language with the help of speech sounds or writing that conveys an agreed item of information like SMS (where short word can also be used). SMS service has developed rapidly since its introduction and is very popular throughout the world. In 2001, more than 250 billion SMS were sent, comparing to the 16 billion sent in 2000. It is particularly popular amongst young urbanites as it allows for voiceless communication, useful in noisy environments (for instance, bars) that would defeat a voice conversation, and also buffered communication since the message the sender wants to convey can be accessed by the receiver any time.

Scenario

Suppose two persona say A and B wants to communicate (remote communication like chat) with each other, but both of them were unknown to each others' language. In such case, we need translator who knew both language of A and B. instead of using human translator, we want use machine with statistical translation model that translates A's language to B's and vice versa. While communicating, A and B may use short words as usual for chatting but the translation model may not understood such short words, so in such case we need some model that can translate short word to corresponding actual words, so that the machine can understood it and can translate it without error. As a result there will be not any misunderstanding between communication between A and B.

Purpose and Advantage of Using Short Words

Authors in [11] there is limited message lengths and tiny user interface for the mobile phones, while sending the SMS from people to people, they can use abbreviation or sometimes they can use the short form words(for example while sending the message they can use the word

"GR8" in instead of **"Great"**) but such short words are generally obtained from three ways, omission of the vowel, truncation of the word and substation(numeric/alphabetic) of the some part of the word whose pronunciation will be same for both short word as well as long form but while writing only the difference can be visualized but read a similarly .Human beings can easily understands such short word, but when it is entered in the source sentence for any machine then such words can't be understood by the machine. So among the various problems for the machine translation, the translation of short word is one of the problems in the machine translation.

How Short words Make Translation difficult?

When we are talking about the translation system we should not forget its capabilities to handle the situation. There may be situation that a translation system may get the input sentence which contains short words. If the designed translation system does not have capabilities to handle such word then problem may occur that means output will be in non understandable form. So our model in this research is made for handling such short form words efficiently.

3.2 Formation of English Sentence

The words in an English sentence can be categorized into different forms. Every times it may not include the words whose meaning can be found in the lexicon. It may include some acronyms and different short forms in addition to the actual words and other. For example, Harry from **"*ktm*"** is the member of **"*UNO*".** In above example **"*ktm*"** is not the actual word but it is short form as well as **"*UNO*"** is acronym, but the ***member*** is the actual word.

3.2.1 Short Form Words

Generally short form word can be defined as the shortened part of single word. Short form word is particularly popular amongst young urbanites during the transmission of the message from one person to another. For example **"*B4*"** is the short form of the actual word great.

3.2.2 Actual word

The words that are recognized as part of the English Language and can be found in an English dictionary or lexicon are the actual word.

3.2.3 Acronym

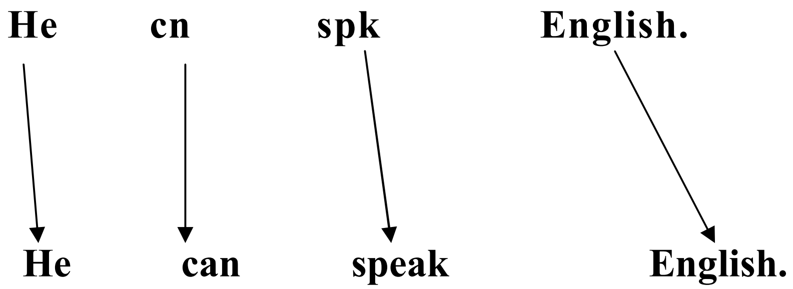
An acronym, according to the dictionary, is defined as the word formed from the first (or first few) letters of a series of words. An acronym can be made from the initial letters of words in a phrase, as in the case of the North Atlantic Treaty Organization (***NATO***), or it can be made from parts of words in a phrase, like with R*adio Detection* and *Ranging* (***RADAR***). Acronyms are arranged in such a way that they can be pronounced without needing to spell out the letters Acronyms look likes a short form word. But in real case the formation of the acronyms is entirely different from the short form word. For example *"**txt**"* is a short form and ***www*** is a acronym. As we can see, there is a great deal of overlap between short word and acronyms also. In fact, generally, every acronym is an short word. This is the case because the acronym is a shortened form of a words or phrase. However, not every short word is an acronym, This does not necessarily work the other way. Shortening the word "***Avenue***" to "***Ave***." is an short word, because it is the shortened version of the single word. However, it is not an acronym since the word AVE is not a new word comprised of the first few letters of a phrase. Acronyms can be ALL CAPS, Initial Caps or all lower case.

For Examples OPEC**, FIFA, Aids.**

3.3 Short Form Words

As we already mentioned that the English sentence may include the short form word as its constituents, at first it is necessary to understand what actually the short form is. As short forms of any word do not belong to word vocabulary, it is just used for the purpose to make communication easy, faster and in more efficient way. In machine translation model, if we enter the sentences containing such words, then for machine it will behaves as the unknown words. So for making easy for user to enter their sentence in easy manner for translation, the machine must have to understand the meaning of such short form words. So, the proposed model handles such case.

For Example:



He cn spk English.

He can speak English.

In above example, "***cn***" and "***spk***" are the short form words. Due to the presence of the such short forms in source sentence, often it is difficult to predict what actually the sentence is trying to say i.e. the meaning of the sentence can be unknown . The main target of this thesis is to convert such short form to long form in order to generate the actual meaning.

3.4 Short Form Word Identification

As in [11] occurrences of short forms in a sentence usually obtained from orthographic transformations which may be in the forms of letter omissions, word truncations and substitution of parts of words with phonetically similar letter sequences.

Examples of Short Forms

Short form	frm	b4	hv	yr
Longform	from	before	have	Year

Table 2.1: Short Forms Word Example [11]

3.4.1 Motivation for Short Form Word Identification

Distinction between short forms from acronyms is necessary because the main problem of this research is primarily concerned/ focused with short forms, with each short form expanding into one single actual word. But acronyms from the set of short forms that are formed when such short words are being expanded.

Examples:

CAN

- Computer Association of Nepal
- Cricket Association of Nepal

SMS

- Simple Mailing System
- Short Message Service

`

Short Forms, on the other hand, often can be translated into actual single words not into multiple words like acronym. Thus, it may be helpful to categorize short forms from the ways that they are being originated and the researcher may get some knowledge for the translation of such short in the field of the machine translation.

3.4.2 Orthographic Transformation and Their Types

In NLP various kinds of transformations can be existed. One of the transformations may be semiotic transformation which is the study of sign structures and sign process. But this research is focused in other transformation technique which is the translation of the short form words.

A. Letter Omission

Letter Omission is one of the ways which can be performed/production of/for the formation of the short form words in perform Orthographic transformations, where all vowels or vowel-like letters in the word are taken out. This forms a universal system as a regular system is followed by all alike as shown in following table.

Short Form	**Long Form**
frm	**from**
txt	**text**
msg	**message**

lov	love
ar	are
hv	have
shld	should
plz	please
ppl	people
thn	then
yr	year
nrg	energy
gnrl	general
cn	can
gd	good

Table 2.2 Example of Letter Omission [11]

B. **Truncation**

Truncation is yet another technique to make the short form words. Generally for the formation of the short form words by this method is to cut the sum part of the word or some times only one letter may be cut out. Because of this if the person is doing chatting or sometimes if he is sending the sms, it will help him to prevent his time because he does not has to type the whole word. For example

Short Form	Long Form
goin	going
jus	just
bn	being
doin	doing
ar	are

Table 2.3 Example of Truncation [11]

C. **Substitution**

This is another way to obtain the short form. In this method the word which can be pronounced similarly/likely is being substituted by similar letter (alphabet) or the similar digit number. In this way substation method can be categorized into two forms.

a. Alphabetic Substitution

This is another popular method to create short words. The numerals or alphabets that are phonetically similar are used to substitute during the creation of the short words. Phonetic replacements are similar in structure to the rebus, where a letter or numerical digit can replace a phonetic sound within a word.

Short form	Long form	Short form	Long form
C	SEE	YR	YEAR
U	YOU	MT	EMPTY
CU	SEE YOU	D	THE
UR	YOU ARE	Y	WHY

Table 2.4 Example of Alphabetic Substitution [1]

b. Numeric Substitution

It is the technique where each word is being substituted by numeric digits. For example

Short form	Long form	Short form	Long form
QT	Cute	SUM1	someone
D8	Date	2NITE	tonight
M8	Mate	2G4U	too good for you
H8	Hate	4GET	forget
L8	Late	W84M	wait for me
B4	before	CUL8R	see you later
W8	Wait	ACTIVE8	activate
U2	you too	IN4ML	informal

2B	to be	2MORO	tomorrow
4N		phone	
4U		for you	

Table 2.5 Example of Numeric Substitution [1]

3.5 Approach to Handle the Short Form Word

Lots of researches are done to handle the short word in an input sentence. Here some of them are discussed in short.

3.5.1 Related Approach

The author in [1] says that sentences can be taken as the combination of actual words, acronyms (alphabetism / initialism), abbreviations, and short forms. The short forms are obtained from the three different ways, in the form of letter omissions, word truncation and substitution of parts of words with phonetically similar letter sequences. Both verbal such as **“short form words”** and non-verbal methods can be used as the **“some signal”.** It is also possible to send several **“emoticons”** to convey something, as the short word also. However, orthographic transformations in the form of letter omissions, where all vowels or vowel like letters in the word are dropped to form a short word.

The authors in [2] introduced a new algorithm for extracting abbreviations and their definitions from biomedical text. Although the algorithm is extremely simple, it is highly effective, and is less specific and therefore less potentially brittle than other approaches that use carefully crafted rules. The process of extracting abbreviations and their definitions from medical text is composed of two main tasks. The first is the extraction of <short-form, long-form> pair candidates from the text. The second task is identifying the correct long form from among the candidates in the sentence that surrounds the short form. The main idea is: starting from the end of both the short form and the long form, move right to left, trying find the shortest long form that matches the short form. Every character in the short form must match a character in the long form, and the matched characters in the long form must be in the same order as the characters in the short form.

Text normalization is an important aspect of successful information retrieval from different types of documents such as clinical notes, radiology reports, discharge summaries etc. The general problem of text normalization is abbreviation and acronym disambiguation. The results of [3] suggest that using Maximum Entropy modeling for abbreviation disambiguation is a promising avenue of research as well as technical implementation for text normalization tasks. Different approaches are being popular for the machine translation, some time Rule-Based propose was popular, later on corpus based propose also became popular, for the statistical machine translation especially translation is done by the help of statistics and correct output is is taken as the result by counting the highest probability [6]. A system is presented which translates cryptic Short Messaging Service (SMS) messages with little recognized short forms into readable messages in long form. According to this paper text messages are first categorized into word, acronyms and short forms where the shortened version of the word as the short form may be obtained from process of truncation, omission of letters, or substitution of chunks of consecutive letters in a word with a shorter chunk of consecutive characters that are phonetically equivalent [11].

3.5.2 Statistical Approach to Handle the Short Form Word

For the approach we have used two corpora, one consist the sentences with short form and another corpus contains the parallel sentence with long form. First of all the short word contained on the given sentence are identified, and then it is matched to the each word of the corpus with short words and the corresponding sentences on the corpus with long form are taken. Now the frequency of each word is calculated and the word with highest frequency is chosen as the correct long form of input short form. In this way the correct output is obtained. This approach tries to overcome almost each and every deficiency faced by other approaches. The framework for this approach described in detail here [in table no 2.5].

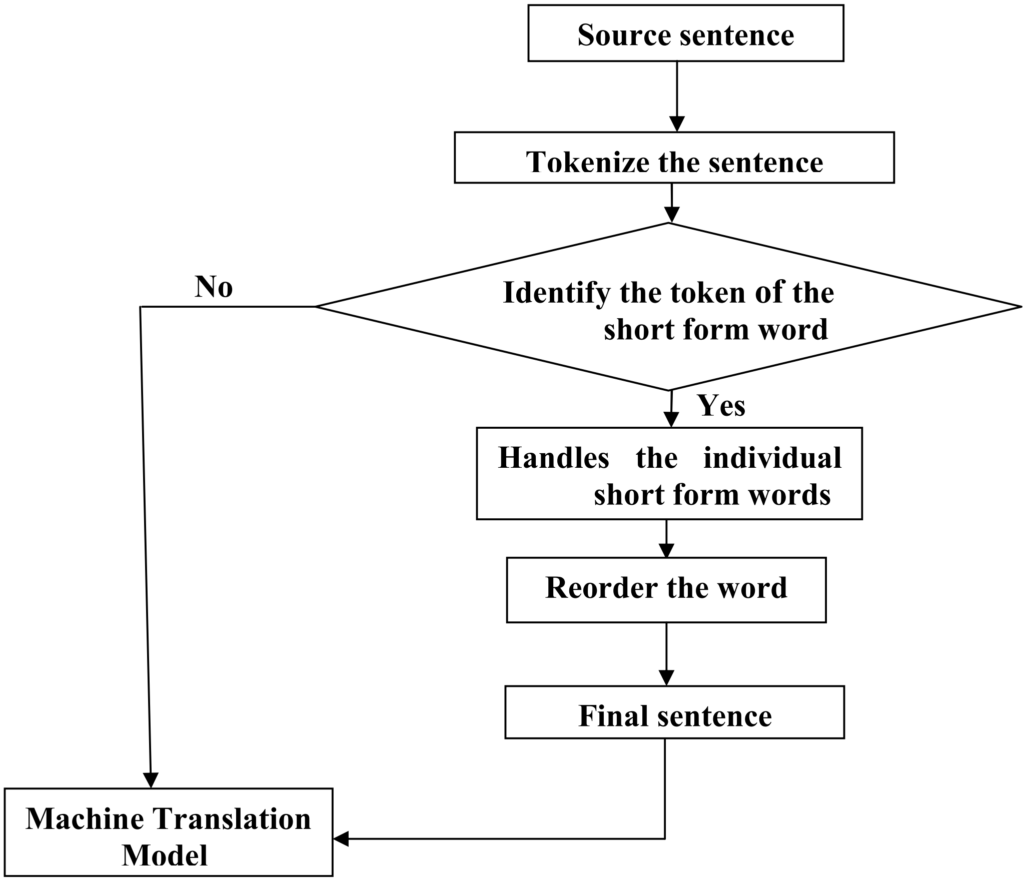

Source sentence
Tokenize the sentence
No
Identify the token of the short form word
Yes
Handles the individual short form words
Reorder the word
Final sentence
Machine Translation Model

Fig 2.6 Proposed Models for Handling Short Form Word

Description of the Proposed Model

Source sentence: The sentence that is given by the user as the input sentence is known as the source sentence.

Tokenize the sentence: In this step each word that is present in the source sentence will be separated/ tokenized in other to find out the short form word.

Identify the token of the short form: In this step each word is checked to find whether it is short form or not.

Handle the individual short form: In this step the short form will be translated into its corresponding long form by the help of the proposed model.

Reorder the word: In this step the translated words will reordered if needed/necessary.

Final sentence: Sentence obtained after handling the short words which is valid input for machine translation model.

Machine Translation Model: This model translates the so obtained English sentence as an input and translates it to another language (target language).

CHAPTER 4
Implementation

4.1 Phase of Implementation

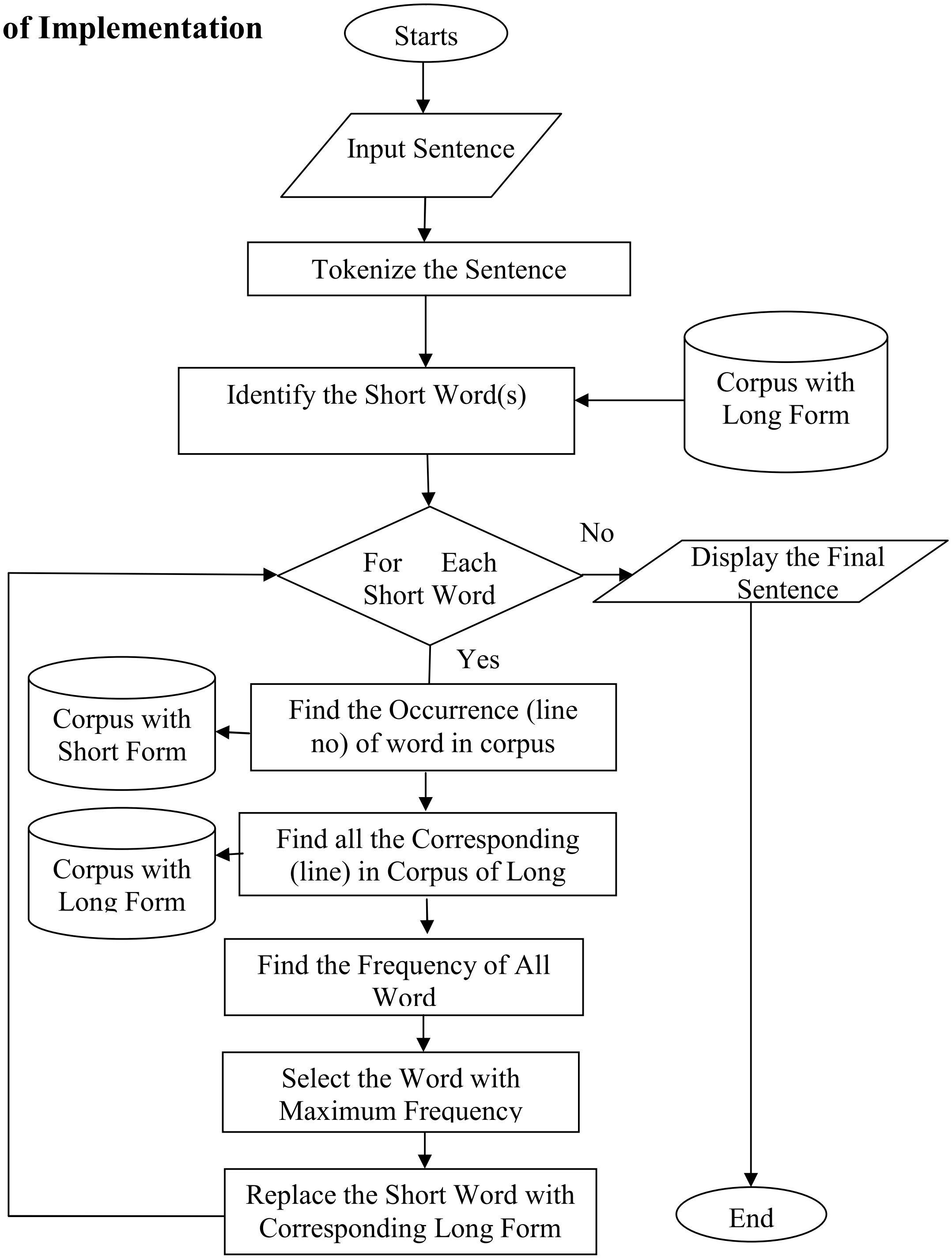

Starts
Input Sentence
Tokenize the Sentence
Identify the Short Word(s)
Corpus with Long Form
For Each Short Word
No
Display the Final Sentence
Yes
Corpus with Short Form
Find the Occurrence (line no) of word in corpus
Corpus with Long Form
Find all the Corresponding (line) in Corpus of Long
Find the Frequency of All Word
Select the Word with Maximum Frequency
Replace the Short Word with Corresponding Long Form
End

Fig. 4.1 Implementation Phases

4.2 Description of Implementation Phases

The process shown in table no 4.1 is based on statistical model for handling the short words. At first the sentence in source language to be translated is taken from the user. Then the input sentence is tokenized and individual token is analyzed whether it is the short word or not. Now after concluding the tokenized word as short word we have to handle each short word. Then, for this at first phase our model is trained with monolingual corpora, especially this consist with two , one with short form words that should be given as the input sentence and other with corresponding corpus with long forms, after tokenizing each word then the tokenized word will be matched with the long form corpora. Now the tokenized word is concluded as the short word, for each word, the machine notices the occurrence (line number) in corpus with short word. After finding the occurrence line of the short form, the parallel sentences in corpus with long forms are also noticed by same method. Hence same method is being applied to handle all the short form word that is present in the corpus with short form. Then take call the words from the corpus with long form of corresponding line number of short word in corpus with short word. Now find the word with maximum frequency, which is the solution for the taken short word. Similarly, all the others short words are handled simultaneously. After this the short words are replaced with its corresponding long words and the resulted sentence is passed to translation model for translation. Thus the long form of the tokenized sentence will be obtained

CHAPTER 5
Testing and Analysis

5.1 Architecture of Training Corpora

To test our model, at first we have to train the model and for that purpose we have built the training corpora[1]. As this thesis concerned with handling short words, so the machine should be trained with two parallel corpora. So we have used two parallel corpora as training corpora. Among them, one contains sentences with the short words and other contains the parallel sentences with all actual words of corresponding short words. These corpora are generated manually as our requirements.

5.1.1 Corpus with Long Form and Short Form Words

Short Form Word Corpus	**Long Form Word Corpus**
He is come frm Kapan with tbl	He is come from Kapan with table
Take out tbl frm cls	Take out table from class
Break this tbl	Break this table
Go out frm cls	Go out from class
We shld respect our parents	We should respect our parents
The rich man shld help poor	The rich man should help poor
This is fantastic yr	This is fantastic year
I will go aboard next yr	I will go aboard next year
Don't spk right now	Donot speak right now
Always spk the truth	Always speak the truth
I will send txt msg	I will send text message
Give me txt book	Give me text book

[1] The better result will be obtained, if we more focus on the domain based training corpora. But due to some lack of linguistic knowledge, it seems unable to complete such corpora in this thesis period. We can enhance this work on domain based training corpora.

They hv brought my book	They have brought my book
I hv to go to school 2da	I have to go to school today
I cn spk french	I can speak french
I cn hlp u	I can help you
u are a gd boy	You are a good boy
He hlp me to do my asignment	He help me to do my asignment
Plz give me your book	Please give me your book
Plz don't disturb to others	Please donot disturb to others
Hm is best place to be	Home is best place to be
I love to stay at hm during vaccation	I love to stay at home during vaccation
Some ppl are very talented	Some people are very talented
There are many ppl in market	There are many people in market
Sankalpa is gd in his study	Sankalpa is good in his study
Gd food is necessary 4 our health	Good food is necessary for our health
Cls is place where a group of students are taught together	Class is place where a group of students are taught together
Discipline shld be maintained in side Cls	Discipline should be maintained in side calss
we want bst thing in our life	we want best thing in our life
Ram is bst student in cls	Ram is best student in class
Food gives us nrg	Food gives us energy
We need nrg to do our various activities	We need energy to do our various activities
Always enjoy your wrk and be happy	Always enjoy your work and be happy
All wrk and no play makes jack a dull boy	All work and no play makes jack a dull boy
cld is high above sky	cloud is high above sky
colour of cld is white	colour of cloud is white
R you a teacher	Are you a teacher
There r many ppl in garden	There are many people in garden
Y r you so happy	why are you so happy
That's y she left so early	That's why she left so early

2da is sunny day	Today is sunny day
Is 2da your exam	Is today your exam
What's d8 today	What's date today
We will discuiss about our project at later d8	We will discuiss about our project at later date
X comes b4 z in alphabet	X comes before z in alphabet
It was some time b4 i realized truth	It was some time before i realized truth
day after 2da is 2morrow	day after today is tomorrow
2morrow is my friend's birthday	tomorrow is my friend's birthday
We should w8 for our result	We should wait for our result
Time will never W8 4 anyone	Time will never wait for any one
I am goin to market	I am going to market
He is goin to make software	He is going to make software
Jus now it is 8 o'clock	Just now it is eight o'clock
He studies in 8 cls	He studies in eight class
bus has gone just b4	bus has gone just before
Policeman is doin his duty	policeman is doing his duty
No 1 is doin work	No one is doing work
No 1 had seen ghost	No one had seen ghost
No 1 will h8 gd person	No one will hate good person
Don't h8 2 anyone	Donot hate to anyone
D da i born was monda	The day i born was monday
monda is not bd da	Monday is not Bad day
monda is bst	Monday is best
Bd lck	Bad luck
Gd lck	Good luck
Ktm is d capital of nepal	Kathmandu is the capital of nepal
C u soon	See you soon
C u 2morrow	See you tomorrow
Mt vessel does not contain anything	Empty vessel does not contain anything
We hear clr sound in Mt room	We hear clear sound in empty room

Sky is clr	Sky is clear
4n is ringing	Phone is ringing
Don't allow d kids to play with 4n set	Donot allow the kids to play with phone set
King Birendra had done D gr8 job in his life	King Birendra had done the great job in his life
we should try 2 generate gr8 idea	We should try to generate great idea
Gd msg	Good message
plz hlp me	Please help me
who cn do this	who can do this
Dharahara is located at ktm	Dharahara is located at Kathmandu
U hv 2b punctual	You have to be punctual
I love u2	I love you too
A QT woman	A cutie woman
He is in IN4ML dress	He is in Informal dress
she is my cls M8	She is my class mate
With Rgds	With regards
This is gnrl problem	This is general problem
He is taller thn me	He is taller than me

Table 5.1: Corpus with Short Form and Long Form Words

5.2 Testing and Analysis

Test Case1: (Paragraph with only one short word.)

Input: Break this tbl. These boys aren’t bad. This is fantastic yr. I will go aboard next yr. Always spk the truth. Give me txt book. They hv brought my book. Plz give me your book. They haven’t knowledge. We want bst thing in our life. Food gives us nrg. cld is high above sky. Colour of cld is white. R you a teacher.

Output: Break this table. This is fantastic year. I will go aboard next year. Always speak the truth. Give me text book. They have brought my book. Please give me your book. We want best thing in our life. Food gives us energy. Cloud is high above sky. Colour of cloud is white. Are you a teacher

Analysis:

Number of input sentence: 14

Number of correct output: 12

Accuracy: (12/14)*100 = 85.71 %

Test Case 2: (Paragraph with comparably more short words)

Input: Go out frm cls. Don't spk right now. I hv to go to school 2da. This book is nt gd. U are a gd boy. Gd food is necessary 4 our health. No bd cn do this. Discipline shld be maintained inside the cls. King Birendra had done d gr8 job in his life. This is gd msg. Plz hlp me.

Output: Go out from class. Donot speak right now. I have to go to school today. This book is not good. You are a good boy. Good food is necessary for our health. Discipline should be maintained inside the class. King Birendra had done the great job in his life. This is good message. Please help me.

Analysis:

Number of input sentence: 11

Number of correct output: 9

Accuracy: (9/11)*100 = 81.8 %

Test Case 3: (Paragraph with large number of short words)

Input: We hear clr sound in the mt cls room. d da i born was Monda. No 1 will h8 gd person. Don't h8 2 anyone. d day after 2da is 2morrow. r u a gd teacher. There r many ppl in the cls. Y r u so happy 2da. u hv 2 be punctual. Ram n shyam r celfis. Plz hlp me to send this msg. don't allow d kids to play with 4n set.

Output: We hear clear sound in the empty class room. The day I born was Monday. Donot hate to anyone. The day after today is tomorrow. Are u a good teacher. There are many people in the class. Why are you so happy today. You have to be punctual. Please help me to send this message. Don't allow the kids to play with phone set.

Analysis:

Number of input sentence: 12

Number of correct output: 10

Accuracy: (10/12)*100 = 83.3 %

More Testing example is performed in Appendix 1.

The Precision and recall and F-measure are mostly preferred evaluation techniques [5, 12], can be calculated as

$$\text{Precession} = \frac{\text{The number of short words correctly translated}}{\text{The output of the translation system}} = \frac{46}{56} = 0.82$$

$$\text{Recall} = \frac{\text{The number of short words correctly translated}}{\text{The translation by human expert manually}} = \frac{46}{50} = 0.92$$

$$\text{F-measure} = \frac{(2*\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})} = \frac{(2*0.82*0.92)}{(0.82+0.92)} = 0.86$$

CHAPTER 6
Conclusion and Future Works

6.1 Conclusion

Nowadays due to the availability of bilingual text, -readable texts has stimulated interest in methods for extracting linguistically valuable information from such texts. Along with the rapid progress in technologies the trends and techniques have also being changing. Let's take an example of sending message. In earlier days the message send by the sender was contained only actual words, by the time this trend has changed. Now most of the word in the message of sender is contained Short Words. But the problem is whether the receiver interprets these short words correctly or not. For example if sender write "lv" in his message to express "love" but the receiver may interpret it as "live".

These days it can be said that the short words are being the necessary part of communication of each and every human beings. So the increasing trends of using short words make us feel that there should be the efficient way to handle such short word or we can say slang. Different researchers use different approaches which are discussed in literature review section. The statistical approach that I suggest here is another approach to handle the short words. It statistically finds out the short words in the user given sentence and translates it into its appropriate long form.

6.2 Future Works and Limitations

Especially the approach which is suggested here in this research is focused to handle the short words. Because the acronyms have their own meaning, they do not need to expand in most of the case. This research has some limitations. It is mostly focused to handle the short words only but still there are emoticons and other different sign languages remains to handle. The new researcher may get some ideas knowledge to move on. Moreover, while doing the same task, the training corpora can be made on domain based to get the better result and to make the model more practical and usable.

Appendices

Appendix 1(Testing)

Input Sentence	I will send txt msg
Short Forms	txt, msg
Expected Long Forms	text, message
Output	I will send text message
Input Sentence	Cld is high above sky
Short Forms	Cld
Expected Long Forms	Cloud
Output	Cloud is high above sky
Input Sentence	Colour of cld is white
Short Forms	cld
Expected Long Forms	could
Output	Colour of cloud is white
Input Sentence	R you a teacher
Short Forms	R
Expected Long Forms	Are
Output	Are you a teacher
Input Sentence	There r many ppl in garden
Short Forms	r, ppl
Expected Long Forms	Are, people
Output	There Are many people in garden
Input Sentence	Y r you so happy
Short Forms	Y, r
Expected Long Forms	why, are

Output	why are you so happy
Input Sentence	That's y she left so early
Short Forms	y
Expected Long Forms	why
Output	That's why she left so early
Input Sentence	2da is sunny day
Short Forms	2da
Expected Long Forms	Today
Output	Today is sunny day
Input Sentence	Is 2da your exam
Short Forms	2da
Expected Long Forms	today
Output	Is today your exam
Input Sentence	What's d8 today
Short Forms	d8
Expected Long Forms	date
Output	What's date today
Input Sentence	We will discuss about our project at later d8
Short Forms	d8
Expected Long Forms	date
Output	We will discuss about our project at later date
Input Sentence	X comes b4 z in alphabet
Short Forms	b4
Expected Long Forms	before
Output	X comes before z in alphabet

Input Sentence	It was some time b4 i realized truth
Short Forms	b4
Expected Long Forms	before
Output	It was some time before i realized truth
Input Sentence	Day after 2da is 2morrow
Short Forms	2da, 2morrow
Expected Long Forms	today, tomorrow
Output	Day after today is tomorrow
Input Sentence	2morrow is my friend's birthday
Short Forms	2morrow
Expected Long Forms	Tomorrow
Output	Tomorrow is my friend's birthday
Input Sentence	We should w8 for our result
Short Forms	w8
Expected Long Forms	wait
Output	We should wait for our result
Input Sentence	Time will never w8 4 anyone
Short Forms	w8, 4
Expected Long Forms	wait, for
Output	Time will never wait for anyone
Input Sentence	I am goin to market
Short Forms	goin
Expected Long Forms	going
Output	I am going to market

Input Sentence	He is goin to make software
Short Forms	goin
Expected Long Forms	going
Output	He is going to make software
Input Sentence	Jus now it is 8 o'clock
Short Forms	jus, 8
Expected Long Forms	just, eight
Output	Just now it is eight o'clock
Input Sentence	He studies in 8 cls
Short Forms	8, cls
Expected Long Forms	eight, class
Output	He studies in eight class
Input Sentence	Bus has gone just b4
Short Forms	b4
Expected Long Forms	before
Output	Bus has gone just before
Input Sentence	Policeman is doin his duty
Short Forms	doin
Expected Long Forms	doing
Output	Policeman is is his duty
Input Sentence	No 1 is doin work
Short Forms	No 1
Expected Long Forms	No one
Output	No No is is work
Input Sentence	No 1 had seen ghost

Short Forms	No 1
Expected Long Forms	No one
Output	No No had seen ghost
Input Sentence	No 1 will h8 gd person
Short Forms	No 1, h8, gd
Expected Long Forms	No one, hate, good
Output	No No will hate good person
Input Sentence	Don't h8 2 anyone
Short Forms	Don’t, h8, 2
Expected Long Forms	Donot, hate, to
Output	Donot hate to anyone
Input Sentence	D da i born was monda
Short Forms	D, da, monda
Expected Long Forms	The, day, monday
Output	The day i born was monday
Input Sentence	Monda is not bd da
Short Forms	Monda, bd,da
Expected Long Forms	Monday, bd, da
Output	Monday is not Bad day
Input Sentence	Monda is bst
Short Forms	Monda, bst
Expected Long Forms	Monday, best
Output	Monday is best
Input Sentence	Bd lck
Short Forms	Bd, lck

Expected Long Forms	Bad luck
Output	Bad luck
Input Sentence	Gd lck
Short Forms	Gd lck
Expected Long Forms	Good luck
Output	Good luck
Input Sentence	Ktm is d capital of Nepal
Short Forms	Ktm, d
Expected Long Forms	Kathmandu, the
Output	Kathmandu is the capital of Nepal
Input Sentence	C u soon
Short Forms	C, u
Expected Long Forms	See, you
Output	See you soon
Input Sentence	C u 2morrow
Short Forms	C, u, 2morow
Expected Long Forms	See you tomorrow
Output	See you tomorrow
Input Sentence	Mt vessel does not contain anything
Short Forms	Mt
Expected Long Forms	Empty
Output	Empty vessel does not contain anything
Input Sentence	We hear clr sound in Mt room
Short Forms	clr, Mt
Expected Long Forms	clear, empty

Output	We hear clear sound in empty room
Input Sentence	Sky is clr
Short Forms	clr
Expected Long Forms	clear
Output	Sky is clear
Input Sentence	4n is ringing
Short Forms	4n
Expected Long Forms	Phone
Output	Phone is ringing

Input Sentence	He is come frm Kapan with tbl
Short Forms	frm, tbl
Expected Long Forms	from, table
Output	He is come from Kapan with table
Input Sentence	Take out tbl frm cls
Short Forms	tbl, frm, cls
Expected Long Forms	table, from, class
Output	Take out table from class
Input Sentence	Break this tbl
Short Forms	tbl
Expected Long Forms	table
Output	Break this table
Input Sentence	Go out frm cls
Short Forms	frm, cls
Expected Long Forms	from, class
Output	Go out from class

Input Sentence	We shld respect our parents
Short Forms	shld
Expected Long Forms	should
Output	We should respect our parents
Input Sentence	The rich man shld help poor
Short Forms	shld
Expected Long Forms	should
Output	The rich man should help poor
Input Sentence	This is fantastic yr
Short Forms	yr
Expected Long Forms	year
Output	This is fantastic year
Input Sentence	I will go aboard next yr
Short Forms	yr
Expected Long Forms	year
Output	I will go aboard next year
Input Sentence	Don't spk right now
Short Forms	Don’t, spk
Expected Long Forms	Donot, speak
Output	Donot speak right now
Input Sentence	Always spk the truth
Short Forms	spk
Expected Long Forms	speak
Output	Always speak the truth

Input Sentence	Give me txt book
Short Forms	txt
Expected Long Forms	text
Output	Give me text book
Input Sentence	They hv brought my book
Short Forms	hv
Expected Long Forms	have
Output	They have brought my book
Input Sentence	I hv to go to school 2da
Short Forms	hv, 2da
Expected Long Forms	have, today
Output	I have to go to school today
Input Sentence	I cn spk french
Short Forms	cn, spk
Expected Long Forms	can, speak
Output	I can speak french
Input Sentence	I cn hlp u
Short Forms	cn, hlp, u
Expected Long Forms	can, help, you
Output	I can help you
Input Sentence	u are a gd boy
Short Forms	u, gd
Expected Long Forms	you, good
Output	you are a good boy
Input Sentence	He hlp me to do my asignment

Short Forms	hlp
Expected Long Forms	help
Output	He help me to do my asignment
Input Sentence	Plz give me your book
Short Forms	plz
Expected Long Forms	Please
Output	Please give me your book
Input Sentence	Plz don't disturb to others
Short Forms	plz, don’t
Expected Long Forms	please, donot
Output	Please donot disturb to others
Input Sentence	Hm is best place to be
Short Forms	Hm
Expected Long Forms	Home
Output	Home is best place to be
Input Sentence	I love to stay at hm during vaccation
Short Forms	hm
Expected Long Forms	home
Output	I love to stay at home during vaccation
Input Sentence	Some ppl are very talented
Short Forms	ppl
Expected Long Forms	people
Output	Some people are very talented
Input Sentence	There are many ppl in market
Short Forms	ppl

Expected Long Forms	people
Output	There are many people in market
Input Sentence	Sankalpa is gd in his study
Short Forms	gd
Expected Long Forms	good
Output	Sankalpa is good in his study
Input Sentence	Gd food is necessary 4 our health
Short Forms	Gd, 4
Expected Long Forms	Good, for
Output	Good food is necessary for our health
Input Sentence	Cls is place where a group of students are taught together
Short Forms	Cls
Expected Long Forms	Class
Output	Class is place where a group of students are taught together
Input Sentence	Discipline shld be maintained in side cls
Short Forms	shld, cls
Expected Long Forms	should, class
Output	Discipline should be maintained in side class
Input Sentence	We want bst thing in our life
Short Forms	bst
Expected Long Forms	best
Output	We want best thing in our life
Input Sentence	Ram is bst student in cls
Short Forms	bst, cls
Expected Long Forms	best, class

Output	Ram is best student in class
Input Sentence	Food gives us nrg
Short Forms	nrg
Expected Long Forms	energy
Output	Food gives us energy
Input Sentence	We need nrg to do our various activities
Short Forms	nrg
Expected Long Forms	energy
Output	We need energy to do our various activities
Input Sentence	Always enjoy your wrk and be happy
Short Forms	wrk
Expected Long Forms	work
Output	Always enjoy your work and be happy
Input Sentence	All wrk and no play makes jack a dull boy
Short Forms	wrk
Expected Long Forms	work
Output	All work and no play makes jack a dull boy
Input Sentence	Don't allow d kids to play with 4n set
Short Forms	Don’t, d,4n
Expected Long Forms	Donot, the, Phone
Output	Donot allow the kids to play with Phone set
Input Sentence	King Birendra had done D gr8 job in his life
Short Forms	D, gr8
Expected Long Forms	the, great
Output	King Birendra had done the great job in his life

Input Sentence	we should try 2 generate gr8 idea
Short Forms	2, gr8
Expected Long Forms	to, great
Output	we should try to generate great idea
Input Sentence	Gd msg
Short Forms	Gd, msg
Expected Long Forms	Good, message
Output	Good message
Input Sentence	Plz hlp me
Short Forms	plz
Expected Long Forms	Please
Output	Please help me
Input Sentence	who cn do this
Short Forms	cn
Expected Long Forms	can
Output	who can do this
Input Sentence	Dharahara is located at Ktm
Short Forms	Ktm
Expected Long Forms	Kathmandu
Output	Dharahara is located at Kathmandu
Input Sentence	U hv 2b punctual
Short Forms	U, hv, 2b
Expected Long Forms	You, have, to be
Output	you have You punctual

Input Sentence	I love u2
Short Forms	u2
Expected Long Forms	you, too
Output	I love I
Input Sentence	A QT woman
Short Forms	QT
Expected Long Forms	Cutie
Output	A A woman
Input Sentence	He is in IN4ML dress
Short Forms	IN4ML
Expected Long Forms	Informal
Output	He is in He dress
Input Sentence	She is my cls M8
Short Forms	cls, M8
Expected Long Forms	class, Mate
Output	She is my class She
Input Sentence	With regds
Short Forms	regds
Expected Long Forms	regards
Output	With With
Input Sentence	This is gnrl problem
Short Forms	gnrl
Expected Long Forms	general
Output	This is This problem
Input Sentence	He is taller thn me

Short Forms	thn
Expected Long Forms	than
Output	He is taller He me

Appendix 2(Code to Handle the Short Form word)

Appendix 2.1 (Source Code for short word detection)

```
//find the short words
String strInput = fldInput.getText();
String strInputArr[] = strInput.split(" ");
String strSlangWords = "";
int flag = 0;
ArrayList<String> listEqWords = new ArrayList<String>();
try
{
for(int i=0; i<strInputArr.length; i++)
{
flag = 0;
Scanner in = new Scanner(new File("F:\\Thesis\\Implementation\\Corpus\\longForm.txt"));
while(in.hasNext())
{
String str = in.next();
if(str.equalsIgnoreCase(strInputArr[i]))
{
flag = 1;
break;
}
}
in.close();
if(flag == 0)
{
strSlangWords = strSlangWords + strInputArr[i] + " ";
}
}
}
catch(Exception e)
{
//print the exception
}
```

Appendix 2.2 (Source code to handle the short words)

```
//now finds the equivalent meaning for short words
String strMulSlangWordsArr[] = strSlangWords.split(" ");

for(int i = 0; i<strMulSlangWordsArr.length; i++)
{
try
{
String lineNumber = getLineNumber(strMulSlangWordsArr[i]);
//System.out.println("adad" +lineNumber);

String strPossible = findPossibleString(lineNumber);
//System.out.println("adad" +strPossible);

//now choose the best one
String best = "";
int max = 0;
String strPos[] = strPossible.split(" ");
for(int j=0; j<strPos.length; j++)
{
int count = countFrequencyOfWords(strPos[j], strPossible);
if(max < count)
{
best = strPos[j];
max = count;
}
}
System.out.println(best);
listEqWords.add(strMulSlangWordsArr[i] + " " + best);
}
catch(Exception e)
{

}
}
//find the line number of slang words in the file
public String getLineNumber(String str) throws Exception
{
int line = 0;
String strLine = "";
Scanner in = new Scanner(new File("F:\\Thesis\\Implementation\\Corpus\\shortForm.txt"));
while(in.hasNext())
{
String s1 = in.nextLine();
String s[] = s1.split(" ");
for(int j=0; j<s.length; j++)
{
if(s[j].equalsIgnoreCase(str))
{
strLine = strLine + String.valueOf(line) + " ";
}
}
```

```
line++;
}
in.close();
return strLine;
}//function getLineNumber ends here

//find the corresspomding possible all words of slang words
public String findPossibleString(String line) throws Exception
{
String strEq = "";
int l = 0;
Scanner in = new Scanner(new File("F:\\Thesis\\Implementation\\Corpus\\longForm.txt"));
String str[] = line.split(" ");
while(in.hasNext())
{
String s = in.nextLine();
for(int k=0; k<str.length; k++)
{
if(String.valueOf(l).equalsIgnoreCase(str[k]))
strEq = strEq + s + " ";
}
l ++;
}
in.close();
return strEq;
}//function findPossibleString ends here


//find the frequency of words on given sentence
public int countFrequencyOfWords(String word, String sentence)
{
```

```
int c = 0;
String test[] = sentence.split(" ");
for(int j=0; j<test.length; j++)
{
if(word.equalsIgnoreCase(test[j]))
c++;
}
return c;
}//function countFrequencyOfWords ends here
```

Appendix 2.3 (Source code for generating final sentence)

```
String strFinal = "";
String strStart[] = strInput.split(" ");
for(int j=0; j<strStart.length; j++)
{
flag1 = 0;
for(int i=0; i<listEqWords.size(); i++)
{
String t = listEqWords.get(i);
String t1[] = t.split(" ");
if(strStart[j].equalsIgnoreCase(t1[0]))
{
strFinal = strFinal + t1[1] + " ";
flag1 = 1;
break;
}
}
if(flag1 == 0)
{
strFinal = strFinal + strStart[j] + " ";}
}
```

References

[1] Jayantha Wannisinghe, Semiotic Analysis of Short Message Service (SMS), Sri Palee Campus, University of Colombo, Sri Lanka, 2011

[2] Ariels S. Schwartz, Marti A. Hearst, A Simple Algorithm for Identifying Abbrevation Definitions in Biomedical Text, Computer Science Division University of California, berkeley, 2003

[3] Serguei Pakhomov, Ph.D, Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts, Mayo Foundation, Rochester, MN, 2002

[4] Wen-Hsiang Lu, Jiun-Hung Lin, and Yao - Sheng Chang, Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names, The Association for Computational Linguistics and Chinese Language Processing, 2008

[5] Stuart Yeates, David Bainbridge and Ian H. Witten, Using compression to identify acronyms in text, Department of Computer Science, University of Waikato Hamilton, New Zealand,2001

[6] W. J. Hutchins, Machine Translation: History and General Principles, 1994

[7] Adam Lopez, Statistical Machine Translation, University of Edinburgh., 2008

[8] Bikash Balami, A Chunk Level Statistical Machine Translation, Tribhuvan University, 2010

[9] David B. Leake, Artificial Intelligence, Indiana University, 2002

[10] Peter E Brown, Vincent J. Della Pietra, The Mathematics of Statistical Machine Translation:, IBM T.J. Watson Research Center, Volume 19, Number 2

[11] Lee Ming Fung, Short Messaging Service Short Form Identification and Codec, National University of Singapore, 2004/2005

[12] Dana Dennells, Recognizing Swedish Acronym and Their Definition in Biomedical Lecture, Department of Swedish Language, Goteberg, University, 2005

[13] Ananthakrishnan Ramanathan, Statistical Machine Translation, Ph.D. Seminar Report,Department of Computer Science and Engineering Indian Institute of Technology, Bombay Mumbai,2002

[14] Ted Slater, A Definition and Model for Communication. International encyclopedia of communications, vol. 3 (pp. 36-44). New York: Oxford University