

**Tribhuvan University**  
**Institute of Science and Technology**

**Automatic Construction Of Dictionary On English–Nepali  
Parallel Corpus.**

**Dissertation**

Submitted to

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

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

By

**Lokendra Bahadur Saud**

November, 2010



**Tribhuvan University**

**Institute of Science and Technology**

**Automatic Construction Of Dictionary On English–Nepali Parallel  
Corpus.**

**Dissertation**

Submitted to

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

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

By

**Lokendra Bahadur Saud**

**Nov, 2010**

Supervisor

**Prof. Dr. Shashidhar Ram Joshi**

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk, Nepal

(Head)



**Tribhuvan University**  
**Institute of Science and Technology**  
**Central Department of Computer Science and Information Technology**

**Student's Declaration**

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

.....  
Lokendra Bahadur Saud  
**Date: Nov, 2010**

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

## LETTER OF CERTIFICATE

This is to certify that the dissertation work entitled “**Automatic Construction Of Dictionary On English –Nepali Parallel Corpus.**”, submitted by Mr. Lokendra Bahadur Saud has carried out under my supervision and guidance. In my best knowledge this is an original work in computer science and no part of this dissertation has been published or submitted for the award of any degree else where in the past.

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

## LETTER OF APPROVAL

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

### Evaluation Committee

-----  
**Dr. Jeevan Joyti Nakarmi**  
Head, Central Department of Computer  
Science and Information Technology  
Tribhuvan University, Nepal

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

-----  
(External Examiner)

-----  
(Internal Examiner)

Date:

## **ACKNOWLEDGEMENT**

It is a great pleasure for me to acknowledge the contributions of a large number of personal to this work. I deeply extend my heartily acknowledgement to my respected teacher and dissertation supervisor Prof. Dr. Shashidhar Ram Joshi, Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk, for giving me an opportunity to work under his supervision and for providing me guidance and support throughout this work. With this regard, I deeply extend my acknowledgement to my dissertation co-supervisor Mr. Yoga Raj Joshi, (center department of Computer science and Information Technology) for his whole time help in this research work, I also wish to extend my sincere appreciation to respected Head of the Central Department of Computer Science and Information Technology, Prof. Dr. Jeevan joyti Nakarmi for his kind help, encouragement and suggestions. I am very much grateful and thankful to the Madan Puraskar Pustakalaya for providing the corpus and other supports.

I would like to express my gratitude to the respected teachers Prof. Dr. Onkar P. Sharma (Marist College, USA), Dr. Subarna Shakya, Prof. Sudarshan Karanjeet, Asst. Prof. Min Bahadur Khati, Mr. Samujjwal Bhandari, Mr. Bishnu Gautam, Mr. Hemant B.G.C, Mr. Dinesh Bajracharaya, Mr. Kedarjung Thapa ,Mr Arjun Singh Saud, Mr Jagdish Bhatt and others for granting me broad knowledge and inspirations within the time period of two years.

I can not remain without admiring the efforts put by my friend Mr. Tej Bahadur Shahi, Mr. Bikash Balami for their exceptional participation on this work. Last but not least, I would like to thank my family members for their constant support and encouragement.

**Lokendra Bahadur Saud**  
**November, 2010**

## **Abstract**

This dissertation describes an approach based on word alignment on parallel corpora, which aims at facilitating the lexicographic work of dictionary building. The proposed model of dictionary construction first perform the tokenization of input text and then TnT tagger is used for tagging the tokenized text, after this the word alignment is done to find out the word pair form source (English)language and target (Nepali) language. Finally our dictionary generation algorithm generate the sample dictionary formed from the word that made the given input text. Our model does rely on the information from tagging as well. Hence the model accuracy not only depends on the alignment algorithms and the training corpus but also depends on the accuracy of tagger. This corpus-driven technique, in particular the exploitation of parallel corpora, proved to be helpful in the creation of bilingual dictionaries for several reasons. Most importantly, a parallel corpus of appropriate size guarantees that the most relevant translations are included in the dictionary.

**To my Parents**



## TABLE OF CONTENTS

### DETAILS

### PAGE NUMBER

### CHAPTER I

<b>1. INTRODUCTION</b>	<b>1-12</b>
1.1 Dictionary	1
1.2 Corpus	2
1.2.1 Parallel Corpus	2
1.2.2 Bilingual Parallel Corpus	2
1.2.3 Application of parallel corpora	3
1.3 Identifying words and sentences	3
1.4 Text Alignment	4
1.4.1 Document Alignment	4
1.4.2 Paragraph Alignment	4
1.4.3 Sentence Alignment	4
1.4.4 Word Alignment	6
1.5 Stage in automatic dictionary construction	8
1.5.1 Construction of parallel corpus	8
1.5.2 Sentence alignment in Parallel Corpus	8
1.5.3 Tokenization	8
1.5.4 Annotation and Categorization	9
1.5.5 POS Tagging and its Approaches	9
1.5.5.1 Rule Based Approach	9
1.5.5.2 Probabilistic Approach	10
1.6 Alignment of words	11
1.7 Application overview of bilingual dictionary	11
1.7.1 Word sense disambiguation	11
1.7.2 Cross information retrieval	12
1.7.3 Multilingual Query Translation	12

## CHAPTER II

<b>2.BACKGROUND AND PROBLEM DEFINITION</b>	<b>14-24</b>
2.1 Background	14
2.2 Problem Definition	15
2.3 Approaches on dictionary creation	16
2.3.1 Association approaches	16
2.3.1.1 Co-occurrence Measures	17
2.3.1.2 String Similarity Measures	17
2.3.2 Estimation approach	18
2.4 Statistical Word Alignment models	18
2.4.1 IBM Models	20
2.5 Problems in Word Alignment	23

## CHAPTER III

<b>3. IMPLEMENTATION</b>	<b>25-38</b>
3.1 Specification of the model	25
3.2 Description of the model	26
3.2.1 Parallel Bilingual corpus	26
3.2.2. Tokenization	26
3.2.3 Tagging	27
3.2.4. TnT POS Tagger	27
3.2.5 Specification and description of the tagset used for Nepali and English language	28
3.3 Alignment	31
3.3.1 Expectation Maximization: The Intuition	31
3.3.2 Expectation Maximization (EM) Algorithm for Training	32
3.3.3 Alignment Algorithm	34
3.4 Word class and category	37
3.5 Dictionary generation algorithm	38

## **CHAPTER IV**

<b>4. TESTING AND ANALYSIS</b>	<b>39-45</b>
4.1 Gold standards	39
4.2 Training and Test corpus	40
4.3 Input /Output of program	41
4.4 Analysis	45
4.5 Limitation of proposed model	45

## **CHAPTER V**

<b>5. CONCLUSION AND FUTURE WORK</b>	<b>46-47</b>
5.1 Conclusions	46
5.2 Further Recommendation	47
<b>References</b>	<b>48</b>
<b>Appendix A</b> <b>[Training Corpus]</b>	<b>51</b>
<b>Appendix B</b> <b>[Test Corpus]</b>	<b>55</b>
<b>Appendix C</b> <b>[Code for Implementation]</b>	<b>58</b>

## List of Tables

<b>Details</b>	<b>Page number</b>
<b>Table 3.1: List of Part-of-Speech tags for Nepali Language</b>	<b>28</b>
<b>Table 3.2: List of Part-of-Speech tags for English Language</b>	<b>30</b>
<b>Table 3.3: List of word categories used for both language</b>	<b>37</b>

## LIST OF ABBREVIATIONS

<b>AI</b>	Artificial Intelligence
<b>EBMT</b>	Example Based Machine Translation
<b>EM</b>	Expectation Maximization
<b>HMM</b>	Hidden Markov Model
<b>ITG</b>	Inversion Transduction Grammar
<b>KB</b>	Knowledge Base
<b>LU</b>	Lexical Unit
<b>MAP</b>	Maximum A Posterior
<b>MT</b>	Machine Translation
<b>NLP</b>	Natural Language Processing
<b>POS</b>	Part of speech
<b>RBMT</b>	Rule Based Machine Translation
<b>SL</b>	Source Language
<b>SMT</b>	Statistical Machine Translation
<b>TAM</b>	Tense Aspect and Modality
<b>TL</b>	Target Language
<b>WSJ</b>	Wall Street Journal

## List of Figures

<b>Details</b>	<b>Page number</b>
Fig 1.1: Stage in automatic dictionary construction	8
Fig 2.1: An example of word alignment on English-Nepali parallel sentence	23
Fig 2.2: Problems in word alignment which the first three IBM models try to solve	24
Fig 3.1: Work flow model of the dictionary construction	25