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-super visor 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

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

List of Tables

LIST OF ABBREVIATIONS

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

List of Figures