
# Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach

## A Dissertation

**Submitted To**

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

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

**Submitted By**
**Pravakar Ghimire**
**CDCSIT, TU**
**(September, 2011)**


# Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach

## A Dissertation

**Submitted to**

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

**In Partial Fulfillment of the Requirements for the Degree of**

**Master of Science**

**in**

**Computer Science and Information Technology**

**Submitted By**

**Pravakar Ghimire**

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

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


# Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach

Date :- ……..

## Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Pravakar Ghimire** entitled "**Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach**" be accepted as in fulfilling partial requirements for the degree of Master of Science.

**Prof. Dr. Shashidhar Ram Joshi**

Head of Department

Department of Electronics and Computer Engineering, Institute of Engineering,

Tribhuvan University

(**Supervisor**)


# Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach

Date :- ……..

## Recommendation

I hereby recommend that the dissertation prepared under my co-supervision by **Mr. Pravakar Ghimire** entitled "**Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach**" be accepted as in fulfilling partial requirements for the degree of Master of Science.

**Mr. Bikash Balami**

Instructor

Central Department of Computer Science and Information Technology,

Tribhuvan University

**(Co-Supervisor)**


**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**

| | |
|---|---|
| **Asso. Prof. Dr. Tanka Nath Dhamala** | **Prof. Dr. Shashidhar Ram Joshi** |
| **Head of Department** | **Head of Department** |
| Central Department of Computer Science and Information Technology | Department of Electronics and Computer Engineering, |
| Tribhuvan University | Institute of Engineering, |
| | Tribhuvan University |
| | (Supervisor) |
| **(External Examiner)** | **(Internal Examiner)** |

# Acknowledgement

I consider it my pleasant duty to express my sincere gratitude to all the people who supported and encouraged me to complete this thesis work entitled **"Handling Unknown Words in English to Nepali Statistical Machine Translation Using Analogical Learning Approach".**

First of all I would like to thank Tribhuvan University, Central Department of Computer Science and Information Technology for providing me this opportunity to perform this research work.

I express my sincere gratitude to my supervisor **Prof. Dr. Shashidhar Ram Joshi** for his keen supervise, able guidance and valuable suggestions. This thesis work would not be possible without his patience and support.

I'm greatly obliged to my co-supervisor **Mr. Bikash Balami** for his constant support. He was the one who was always available to deal with every obstacle that I faced during this thesis work with his insightful suggestions.

I'm also highly thankful to all the staffs and teachers of CDCSIT for providing me such a broad knowledge and enlightment in two years of study period. Their motivation and support was really appreciable.

Special thanks to my parents and family for their unconditional love, constant support and motivation. They have always been a source of inspiration for whatever I have achieved so far.

Last but not the least, I would like to acknowledge and appreciate the direct and indirect support of my friends Mr. Dinesh kumar Khadka, Mr. Roshan Silwal and Mr. Satya Bahadur Maharjan and their willing co-operation to bring this thesis work in tangible form.

I have given my best effort to make this thesis work complete and error free but still if it contains some faults, suggestions regarding those mistakes will always be welcomed.

Pravakar Ghimire

(CDCSIT, TU)

September, 2011

# Abstract

Unknown words are one of the great difficulties in the field of machine translation. In the process of translation, a system is most likely to encounter words that were not present in the available training data. While this is in part due to the segmentation issues, it is also often simply due to the lack of training data. In statistical machine translation to translate a sentence from one language to another we make use of a parallel corpus but it is not possible to a corpus to contain all the words from a whole language domain, hence the unknown word problem is obvious. In this thesis work an effective approach is used to translate those unknown words in English to Nepali statistical machine translation using word analogy. In this method the meaning of the unknown word is identified on the basis of other words presented in the corpus and the analogy between the prefixes and suffixes of those words with the unknown word.

***Dedicated to,***

***My loving dad and mom***

# Table of Contents

## CHAPTER 4



## CHAPTER 5

## List of Tables

## List of figures

## Abbreviations Used

AI – Artificial Intelligence

AT – Automatic Translation

EBMT – Example Based Machine Translation

EM – Expectation Maximization

IBM – International Business Machine

INV – In Vocabulary

LCS – Longest Common Subsequence

MT – Machine Translation

NLP – Natural Language Processing

OOV – Out of Vocabulary

SMT – Statistical Machine Translation