Acknowledgements

With the grace of God, it is very special moment to acknowledge the contributions of those who supported me during the journey to achieve this task. This research work is the outcome of very valuable guidance from my supervisor Dr. R.K. Sharma. He has always been there to guide me towards the right direction whenever I got stuck with the ideas. I want to thank him for his valuable evenings and week-ends that he spent with me in finalizing the writings of this thesis work. I also want to thank Mrs. Sharma to provide a conducive environment in her home during discussions on this proposed work.

The motivation of this research work is provided by Dr. Pushpak Bhattacharyya, Prof. Department of Computer Science and Engineering, IIT Bombay. He introduced me to the concept of UNL. He is very kind to provide me the required NLP resources and guidance. I will always remain indebted to him for showing me the path of development of Punjabi EnConverter and DeConverter.

I express my gratitude to the Doctoral Committee for monitoring the progress and providing valuable suggestions for improvement of my research work. I feel very proud and thankful to the Department of Computer Science and Engineering, Thapar University, Patiala for providing all the resources for a good research work. I am very thankful to the administration of Thapar University for granting me a teaching off for one semester. This period helped me immensely to compile my research work. I am also very thankful to Dr. Seema Bawa, Dr. Maninder Singh and all my colleagues for their constant support and cooperation during my research work.

I take this opportunity to say thanks to my family members, who are always there to support me morally and emotionally. With the grace of God, this work provides me the opportunity to make my parents; Mr. Ved Kumar, Mrs. Jagdish Bhatia, S. Dalip Singh and Mrs. Joginder Kaur feel proud. I want to thank my wife Mrs. Sanmeet Kaur, kids Rahat and Rishan for providing a very loving, cool and calm environment in our home.

(Parteek Kumar)
Abstract

Machine Translation (MT) has been an area of immense interest among researchers during last couple of decades. This area has witnessed a few lows and highs during its life span and has also witnessed integration of research works from different fields including linguistics, computer science, artificial intelligence, statistics, mathematics, philosophy and others. Researchers have proposed different paradigms for machine translation across natural languages with reasonable success. Universal Networking Language (UNL) based MT is also an effort in this direction. The UNL programme was launched in 1996 in Institute of Advanced Studies (IAS) of United Nations University (UNU), Tokyo, Japan. The approach in UNL revolves around the development of an EnConverter and a DeConverter for a natural language. The EnConverter is used to convert a given sentence in natural language to an equivalent UNL expression; and the DeConverter is used to convert a given UNL expression to an equivalent natural language sentence. In the work carried out in this PhD project, these two components, namely, EnConverter and DeConverter have been developed for Punjabi language.

The PhD thesis on the work carried out in this project is divided into seven chapters. These chapters are: Introduction, Review of Literature; UNL Framework and Creation of Punjabi-Universal Word Lexicon; Punjabi-UNL EnConverter; UNL-Punjabi DeConverter; Results and Discussions; Conclusion and Future Scope of the work.

First chapter contains introduction to Machine Translation and its need in the age of Information Technology. In this chapter, the challenges of MT, approaches of MT, objectives of this research, methodology adopted to achieve these objectives and the features of Punjabi language have been presented. Machine Translation approaches that can be classified into four categories, namely, direct MT, rule-based MT, corpus-based MT and knowledge-based MT, have also been discussed in this chapter. The main objective of this study is to design and develop a multi-lingual machine translation system for Punjabi language. A major outcome of this research work is the development of Punjabi EnConverter, Punjabi DeConverter and a web interface for online EnConversion and DeConversion task.
Second chapter of the thesis provides the details of the findings of research in the area of MT. The review of literature on MT has been performed by tracking the historic developments in this field. The literature reported in this work has been so organised as to include the state of MT before the invention of computers, beginning of automated MT (1946-1954), decade of high expectations and disillusion (1955-1966), post ALPAC decade (1967-1976), the revival of MT research (1977 to 1989), decade of 1990 to 2000, research since 2000 and MT for Indian languages and research activities in UNL.

Computerized translation was first performed by Georgetown Automatic Translation (GAT) system at Georgetown University, USA. During the period of 1955-1966, RAND Corporation’s statistical analysis of a large corpus of Russian Physics texts, ‘CETA’ a hybrid system, Mechanical Translation and Analysis of Languages (METAL) with the use of Chomsky's transformational paradigm, and ‘Logos’ machine translation system had been some of the important developments in this area. During the post ALPAC period (1967-1976), ‘Ariane’ system was developed and ‘TAUM’ project was undertaken by using syntactic transfer for English-French translation. The revival of research on MT took place during 1977-89 with ‘Ariane’ system that was developed on linguistics based transfer approach. Other important MT projects such as ‘ATLAS’ system, ‘UNITRAN’ system, IBM’s ‘Candide’ system, Universal Networking Language (UNL) based MT system, ‘KANTOO’ system, ‘ALEPH’ a pure example-based machine translation system, ‘SisHiTra’ a hybrid Machine Translation system, Google Translation and ‘OpenLogos’ system have also been discussed in this chapter.

The review of literature on MT for Indian languages has also been presented in this chapter. The important machine translation systems for Indian languages like ‘AnglaBharti’, ‘AnuBharati’, ‘Anusaaraka’, ‘MANTRA’ (MAchiNe assisted TRAnslation tool) system, English-Bangla-ANUBAD system, a translation system for bi-lingual Hindi-English (Hinglish) text, ‘Shakti’ system, ‘MaTra’ system, a Punjabi to Hindi MT system, English to Urdu translation system, and ‘Sampark’ a hybrid system for translation among Indian languages have been discussed in this chapter.

The research activities in UNL have been presented in three distinct sections in this chapter. These sections are: development of EnConversion and DeConversion modules; applications of UNL in other contexts; and use of external lexical and ontological
resources to enhance some of the processes of UNL. The work on conversion of Brazilian Portuguese into UNL and vice-versa, EnConversion and DeConversion tools for Tamil language, ‘HERMETO’ system, French EnConverter and French DeConverter, UNL DeConverter for Chinese language, ‘Manati’ DeConversion model, UNL-Nepali DeConverter, UNL-Hindi DeConverter, Arabic MT system based on UNL and a Bangla EnConversion system have also been reviewed in this chapter.

Third chapter discusses UNL format for information representation that includes the details on UWs and their four types (Basic UWs, Restricted UWs, Extra UWs and Temporary UWs), UNL relations, UNL attributes and formats to write UNL sentence. Compound UWs in UNL are used to denote compound concepts that are to be interpreted as a whole so that one can use their parts at the same time. This chapter also provides the details on UNL system that consists of EnConverter, DeConverter, Dictionary Builder, Grammar Rules, UNL Key Concept in Context (KCIC), UW Gate Universal Parser, UNL Verifier, Language Server and Word Dictionary (Language-UW Dictionary). UNL system makes use of word dictionaries in the form of Language-UW lexicon of respective languages for its processing. An entry of the word dictionary contains three parts, namely, a headword, a UW and a set of morphological, syntactic and semantic attributes. This chapter discusses the grammatical attributes of Punjabi-UW lexicon, important issues in creation of Language-UW dictionary and creation of Punjabi-UW dictionary. A Punjabi-UW dictionary having 1,15,000 entries has been developed in this work by taking Hindi-UW dictionary as a reference.

Fourth chapter of this thesis provides the details of EnConversion system for conversion of input Punjabi sentences into UNL. This chapter also discusses the framework for designing the EnConverter for Punjabi language with a special focus on generation of UNL attributes and relations from Punjabi source text. The architecture of Punjabi EnConverter has been divided into seven phases. These phases are: (i) Parser phase (to parse the input sentence with Punjabi shallow parser), (ii) Linked list creation phase, (iii) Universal Word lookup phase, (iv) Case marker lookup phase, (v) Unknown word handling phase, (vi) User interaction phase (this phase is optional) and (vii) UNL generation phase. These phases have been implemented using Java for the development of proposed system. All these phases are illustrated by giving example sentences.
Fifth chapter discusses UNL to Punjabi DeConverter that generates natural language Punjabi sentence from a given input UNL expression. The architecture of Punjabi DeConverter has been divided into five phases, namely, (i) UNL parser phase, (ii) Lexeme selection phase, (iii) Morphology generation phase, (iv) Function word insertion phase and (v) Syntax planning phase. The first stage of a DeConverter is UNL parser which parses the input UNL expression to build a node-net from the input UNL expression. During lexeme selection phase, Punjabi root words and their dictionary attributes are selected for the given UWs in the input UNL expression from the Punjabi-UW dictionary. After that, the nodes are ready for generation of morphology according to the target language in the morphology phase. The proposed system makes use of morphology rule base for Punjabi language to handle attribute label resolution morphology; relation label resolution morphology; and noun, adjective, pronoun and verb morphology. In function word insertion phase, the function words are inserted to the morphed words. These function words are inserted in the generated sentence based on nine column rule base. Finally, the syntax planning phase is used to define the word order in the generated sentence so that output matches with a natural language sentence.

The pseudocodes and algorithms for building data structures in UNL parser; processing of noun morphology and adjective morphology rule base; processing of function word insertion rule base; controlling the syntax planning of nodes of simple UNL graph; syntax planning of UNL graph with a scope node; handling of untraversed parent node and nodes with multiple parents nodes during syntax planning, handling of some special cases of syntax planning and syntax planning of noun, adjective and adverb clause sentences have been presented in this chapter. All these pseudocodes and algorithms have been implemented in Java to develop the proposed UNL-Punjabi DeConverter.

Sixth chapter of this thesis contains the results and discussion on the work done in this project. The evaluation of proposed system, consisting of Punjabi EnConverter and DeConverter, has been performed with the help of one thousand Punjabi sentences. These sentences have been selected in such a way that generation of all possible UNL relations and attributes can be tested. For testing purpose, Spanish UNL Language Server and agricultural domain threads developed by IIT Bombay, India are considered as gold-standards.
Spanish Language Server contains English sentences with their corresponding UNL expressions generated by the system (Spanish Language Center, 2004) while agricultural domain threads developed by IIT Bombay have Hindi language sentences with their equivalent UNL expressions. These sentences were translated manually into equivalent Punjabi sentences and then inputted to the proposed Punjabi EnConverter system for their EnConversion to UNL. The UNL expression generated by proposed system is compared with the UNL expression given by the gold-standard EnConverter. These two UNL expressions match with each other if the UNL relations, including associated UWs and UNL attributes present in the expressions are same. It has been seen that proposed system handles the resolution of UNL relations and generation of attributes for these sentences with a very reasonable accuracy. Proposed Punjabi DeConverter is also evaluated by inputting UNL expressions generated by Punjabi EnConverter to it and the output of Punjabi DeConverter is compared with input Punjabi sentence given to Punjabi EnConverter. The two Punjabi sentences (original sentence inputted to Punjabi EnConverter and the sentence generated by DeConverter) are compared and the system is evaluated based on adequacy test and BLEU score.

Subjective tests like adequacy and fluency tests have been performed on the proposed system. BLEU score has been calculated to evaluate the quality of output. The quantitative test, namely, error analysis has also been performed by calculating Sentence Error Rate (SER) and Word Error Rate (WER). From this analysis, it has been concluded that proposed system generates 89.0% intelligible sentences and generates 92.0% sentences that are faithful to the original sentences. The system could achieve a fluency score of 3.61 (on a 4-point scale) and adequacy score of 3.70 (on a 4-point scale). The proposed system is able to achieve a BLEU score of 0.72. The proposed system has a word error rate of 5.43% and sentence error rate of 20.8%. These scores of the proposed system can be improved further by improving the rule base and lexicon.

Chapter seven presents the conclusion, limitations and future scope of work in order to refine the proposed development of EnConverter and DeConverter for Punjabi language.
List of Figures

Figure 1.1: Vauquois triangle 6
Figure 1.2: Machine Translation approaches 7
Figure 1.3: Direct translation method 8
Figure 1.4: Transformer architecture 8
Figure 1.5: Transfer-based architecture 10
Figure 1.6: Architecture of interlingua system 11
Figure 1.7: Interlingua-based translation among $n$ languages 12
Figure 1.8: Architecture of EBMT 15
Figure 1.9: Architecture of KBMT 16
Figure 3.1: A UNL system 48
Figure 3.2: UNL graph for example sentence given in (3.2) 64
Figure 3.3: UNL graph with compound UW 65
Figure 3.4: Conversion process in a UNL system 68
Figure 3.5: Structure of the UNL system and applications 68
Figure 3.6: Structure of EnConverter 70
Figure 3.7: Structure of DeConverter 71
Figure 3.8: Structure of ‘DicBld’ 72
Figure 3.9: Structure of Universal Parser 73
Figure 3.10: Flowchart of UNL verifier 73
Figure 3.11: Working of LSs over Internet 74
Figure 4.1: Analysis and condition windows of Punjabi EnConverter 83
Figure 4.2: Flowchart for working of Punjabi EnConverter 93
Figure 4.3: Structure of a node 94
Figure 4.4: UNL graph for the UNL expression in (4.35) 104
Figure 4.5: UNL graph for UNL expression given in (4.57) 108
Figure 4.6: UNL Graph for UNL expression given in (4.59) 117
Figure 4.7: UNL graph for UNL expression given in (4.61) 124
Figure 4.8: UNL graph for UNL expression given in (4.63) 129
Figure 4.9: UNL graph for UNL expression given in (4.65) 134
Figure 4.10: UNL graph for UNL expression given in (4.67) 140
Figure 5.1: Architecture of UNL-Punjabi DeConverter 145
Figure 5.2: UNL graph generated by UNL parser for UNL expression given in (5.2) 146
Figure 5.3: A UNL graph representation for UNL binary relation \( \text{rel}(\text{UW1}, \text{UW2}) \) 148
Figure 5.4: UNL graph with a node having two parents 148
Figure 5.5: Back edge in UNL graph for \( \text{rel}(\text{UW1}, \text{UW2}) \) 149
Figure 5.6: UNL graph with back edges to handle a node having two parents 149
Figure 5.7: UNL graph with scope node for UNL expression given in (5.9) 150
Figure 5.8: UNL graph of two nodes with same parent 184
Figure 5.9: Matrix representation for \((N_1 L N_2)\) structure 184
Figure 5.10: UNL graph for UNL expression given in (5.52) 185
Figure 5.11: Priority matrix for syntax plan of UNL expression given in (5.52) 186
Figure 5.12: UNL graph for UNL expression (5.62) 188
Figure 5.13: Priority matrix for ‘agt’ and ‘obj’ relations 189
Figure 5.14: UNL graph with a scope node for UNL expression (5.67) 191
Figure 5.15: Priority matrix for ‘agt’, ‘plc’ and ‘tim’ relations 191
Figure 5.16: UNL graph of UNL expression (5.73) 192
Figure 5.17: Modified UNL graph for a node having untraversed parent 193
Figure 5.18: UNL graph having multiple parents for UNL expression (5.77) 195
Figure 5.19: Modified UNL graph for a node having multiple parents 196
Figure 5.20: UNL graph for UNL expression (5.81) 198
Figure 5.21: UNL graph for UNL expression (5.83) 198
Figure 5.22: UNL graph having relation ‘and’ for UNL expression (5.85) 199
Figure 5.23: UNL graph having relation ‘or’ for UNL expression (5.91) 201
Figure 5.24: UNL graph having relation ‘fmt’ for UNL expression (5.95) 202
Figure 5.25: UNL graph having ‘cnt’ relation for UNL expression (5.100) 204
Figure 5.26: UNL graph having ‘seq’ relation and parent with 206
‘@reference’ attribute

Figure 5.27: UNL graph having ‘seq’ relation and child node with ‘@reference’ attribute  

Figure 5.28: UNL graph for noun clause sentence given in (5.112)  
Figure 5.29: UNL graph for adjective clause sentence given in (5.120)  
Figure 5.30: UNL graph with adverb clause for time for example sentence (5.131)  

Figure 6.1: Results of evaluation of Punjabi EnConverter for UNL relations  
Figure 6.2: Approach for evaluating UNL based MT system for Punjabi language  
Figure 6.3: Distribution of adequacy scores for various fluency scores  
Figure 6.4: Word Error Analysis of the proposed system
List of Tables

Table 1.1: Example bilingual English-Punjabi corpus 14
Table 1.2: Correspondences learned by system 14
Table 3.1: Description of UNL relations 50
Table 3.2: Description of UNL attributes 61
Table 3.3: Syntax of UNL sentence in table format 62
Table 3.4: Syntax of UNL sentence in list format 63
Table 3.5: Tags used in a UNL document 66
Table 3.6: Description of Word dictionary entries 74
Table 3.7: Some of morphological and semantic attributes used in Punjabi-UW lexicon 76
Table 4.1: Types of adverb clause 109
Table 4.2: EnConversion process of adverb clause for time for example sentence given in (4.58) 110
Table 4.3: EnConversion process of adverb clause of condition sentence for example sentence given in (4.60) 118
Table 4.4: EnConversion process of adverb clause of manner for example sentence given in (4.62) 125
Table 4.5: EnConversion process of adverb clause for place for example sentence given in (4.64) 130
Table 4.6: EnConversion process of adjective clause example sentence given in (4.66) 135
Table 5.1: Noun morphology for Punjabi DeConverter 155
Table 5.2: Generation of noun morphology rules 160
Table 5.3: Adjective morphology 162
Table 5.4: Generation of adjective morphology rules 164
Table 5.5: Generation of pronoun morphology rules 169
Table 5.6: Some examples of conjunct verb morphology for ਕਰੋ kar 173
‘do’ type of verbs

Table 5.7: Some examples of conjunct verb morphology for ī hō 174

‘be’ type of verbs

Table 5.8: ‘kaarak’ system with respect to UNL relations and function words 176
Table 6.1: 4-point scale of fluency score 224
Table 6.2: 4-point scale of adequacy score 225
## List of Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>UNL relation resolution and generation of attributes</td>
<td>97</td>
</tr>
<tr>
<td>5.1</td>
<td>Processing of noun morphology rule base</td>
<td>161</td>
</tr>
<tr>
<td>5.2</td>
<td>Processing of adjective morphology rule base</td>
<td>165</td>
</tr>
<tr>
<td>5.3</td>
<td>Processing of function word insertion rule base</td>
<td>179</td>
</tr>
<tr>
<td>5.4</td>
<td>Processing the nodes of UNL graph with a scope node</td>
<td>190</td>
</tr>
<tr>
<td>5.5</td>
<td>Processing of noun clause sentences</td>
<td>211</td>
</tr>
<tr>
<td>5.6</td>
<td>Handling of noun and adjective clause sentences</td>
<td>213</td>
</tr>
<tr>
<td>5.7</td>
<td>Handling of noun, adjective and adverb clause sentences</td>
<td>217</td>
</tr>
</tbody>
</table>
## List of PseudoCodes

<table>
<thead>
<tr>
<th>PseudoCode</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Building data structures in UNL parser</td>
<td>150</td>
</tr>
<tr>
<td>5.2</td>
<td>To control the processing sequence of nodes of simple UNL graph</td>
<td>186</td>
</tr>
<tr>
<td>5.3</td>
<td>Handling of untraversed parent node</td>
<td>193</td>
</tr>
<tr>
<td>5.4</td>
<td>Handling nodes with multiple parents nodes</td>
<td>195</td>
</tr>
<tr>
<td>5.5</td>
<td>Handling of ‘and’ and/or ‘or’ relations</td>
<td>202</td>
</tr>
<tr>
<td>5.6</td>
<td>Handling of ‘fmt’ relation</td>
<td>203</td>
</tr>
<tr>
<td>5.7</td>
<td>Handling of ‘cnt’ relation</td>
<td>205</td>
</tr>
<tr>
<td>5.8</td>
<td>Implementation of step (i) of strategy for ‘seq’ relation</td>
<td>208</td>
</tr>
<tr>
<td>5.9</td>
<td>Implementation of step (ii) of strategy of ‘seq’ relation</td>
<td>209</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certificate</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Algorithms</td>
<td>xiii</td>
</tr>
<tr>
<td>List of PseudoCodes</td>
<td>xiv</td>
</tr>
<tr>
<td><strong>Chapter 1: Introduction</strong></td>
<td>1-21</td>
</tr>
<tr>
<td>1.1 Need of Machine Translation</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Socio-Political importance of MT</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Commercial importance of MT</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3 Coverage of literary works by MT</td>
<td>3</td>
</tr>
<tr>
<td>1.1.4 MT to bridge the digital divide</td>
<td>3</td>
</tr>
<tr>
<td>1.1.5 MT to assist human translator</td>
<td>3</td>
</tr>
<tr>
<td>1.1.6 MT for cross language information retrieval</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Challenges of Machine Translation</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Word order</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Lexical differences</td>
<td>4</td>
</tr>
<tr>
<td>1.2.3 Lexical ambiguity</td>
<td>4</td>
</tr>
<tr>
<td>1.2.4 Structural ambiguity</td>
<td>5</td>
</tr>
<tr>
<td>1.2.5 Pronoun resolution</td>
<td>5</td>
</tr>
<tr>
<td>1.2.6 Idioms and phrases</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Translation Architecture</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Machine Translation approaches</td>
<td>7</td>
</tr>
<tr>
<td>1.4.1 Direct MT</td>
<td>7</td>
</tr>
<tr>
<td>1.4.2 Rule-Based MT system</td>
<td>9</td>
</tr>
<tr>
<td>1.4.2.1 Transfer-Based MT</td>
<td>9</td>
</tr>
</tbody>
</table>
1.4.2.2 Interlingua-Based MT
1.4.3 Corpus-Based MT
  1.4.3.1 Statistical MT
  1.4.3.2 Example-Based MT
1.4.4 Knowledge-Based MT
1.4.5 Hybrid approaches

1.5 Need of UNL based machine translation system for Punjabi language
1.6 Objectives of this research
1.7 Major contributions and achievements
1.8 Features of Punjabi language
1.9 Organization of thesis

Chapter 2: Review of Literature

2.1 Introduction
2.2 State of Machine Translation before the invention of computers
2.3 Beginning of computerized MT (1946-1954)
2.4 Decade of high expectations and disillusion (1955-1966)
2.5 Post ALPAC decade (1967-1976)
2.6 The revival of research (1977 to 1989)
2.7 Decade of 1990 to 2000
2.8 Research since 2000
2.9 MT for Indian languages
  2.9.1 Indian languages MT systems
2.10 Research activities in Universal Networking Language
  2.10.1 Development of EnConversion and DeConversion modules
  2.10.2 Application of UNL in other contexts
  2.10.3 Use of external resources to enhance UNL processes

Chapter 3: UNL Framework and Creation of Punjabi-Universal Word Lexicon

3.1 Introduction
3.2 UNL format for information representation
  3.2.1 Universal Words

xvi
3.2.2 UNL relations 50
3.2.3 UNL attributes 60
3.3 UNL sentence 62
3.4 Compound UWs in UNL 64
3.5 UNL document 65
3.6 UNL knowledge-base 66
3.7 UNL system 67
  3.7.1 EnConverter 69
  3.7.2 DeConverter 70
  3.7.3 Dictionary Builder 71
  3.7.4 Grammar rules 72
  3.7.5 UNL Key Concept in Context (KCIC) 72
  3.7.6 UW Gate 72
  3.7.7 Universal Parser 72
  3.7.8 UNL verifier 73
  3.7.9 Language Server 74
  3.7.10 Word dictionary (Language-UW dictionary) 74
3.8 Grammatical attributes of Punjabi-UW lexicon 75
3.9 Important issues in creation of L-UW dictionary 79
3.10 Creation of Punjabi-UW dictionary 80

Chapter 4: Punjabi-UNL EnConverter 83-141

4.1 Introduction 83
4.2 Working of Punjabi EnConverter 83
4.3 Format of EnConversion analysis rules 84
  4.3.1 Left composition rule (+) 84
  4.3.2 Right composition rule (-) 85
  4.3.3 Left modification rule (<) 85
  4.3.4 Right modification rule (>) 85
  4.3.5 Attribute changing rule (:) 85
4.4 Punjabi shallow parser 86
  4.4.1 Tokenizer 87
4.4.2 Morph analyzer 87
4.4.3 Part-of-speech tagger 88
4.4.4 Chunker 89
4.4.5 Pruning 89
4.4.6 ‘head’ computation 90
4.4.7 ‘vibhakti’ computation 91

4.5 Punjabi EnConverter architecture 92
   4.5.1 Parser phase 92
   4.5.2 Linked list creation phase 92
   4.5.3 Universal Word lookup phase 94
   4.5.4 Case marker lookup phase 95
   4.5.5 Unknown word handling phase 96
   4.5.6 User interaction phase 97
   4.5.7 UNL generation phase 97

4.6 EnConversion of a simple Punjabi sentence to UNL expression 98

4.7 EnConversion of complex sentences 104
   4.7.1 EnConversion process of noun clause sentence 105
   4.7.2 EnConversion process of adverb clause sentences 109
      4.7.2.1 EnConversion process of sentences containing adverb clause
            for time 109
      4.7.2.2 EnConversion process of sentences containing adverb clause
            for condition 117
      4.7.2.3 EnConversion process of sentences containing adverb clause
            for manner 124
      4.7.2.4 EnConversion process of sentences containing adverb clause
            for place 129
   4.7.3 EnConversion of sentences containing adjective clause 134

Chapter 5: UNL-Punjabi DeConverter 143-219
5.1 Introduction 143
5.2 Natural Language Generation system 143
5.3 Punjabi DeConverter as an NLG system 143
5.3.1 Lexeme selection
5.3.2 Morphology generation of lexical words
5.3.3 Function word insertion or case marking
5.3.4 Syntax planning
5.4 Architecture of UNL-Punjabi DeConverter
5.5 UNL parser
5.6 Lexeme selection
5.7 Morphology generation
5.7.1 Attribute label resolution morphology
5.7.2 Relation label resolution morphology
5.7.3 Noun, adjective, pronoun and verb morphology
  5.7.3.1 Noun morphology
  5.7.3.2 Adjective morphology
  5.7.3.3 Pronoun morphology
  5.7.3.4 Verb morphology
5.8 Function word insertion phase
5.8.1 Issues in direct mapping of function words with UNL relations
5.8.2 Rule base for function word insertion
5.8.3 Implementation of the function word insertion rule base
5.9 Syntax planning phase
5.9.1 Major issues in syntax planning
5.9.2 Matrix based priority of relations
5.9.3 Syntax planning for simple sentences
5.9.4 Syntax planning of UNL graph with a scope node
5.9.5 Untraversed parent handling
5.9.6 Handling of multiple parents
5.9.7 Special cases in syntax planning
  5.9.7.1 Strategy for ‘and’ and/or ‘or’ relation(s)
  5.9.7.2 Strategy for ‘fmt’ relation
  5.9.7.3 Strategy for ‘cnt’ relation
  5.9.7.4 Strategy for ‘seq’ relation
5.9.8 Syntax planning for clausal sentences

5.9.8.1 Syntax planning of sentences with noun clause

5.9.8.2 Syntax planning of sentences with adjective clause

5.9.8.3 Syntax planning of sentences with adverb clause

Chapter 6: Results and Discussion

6.1 Introduction

6.2 Selection of sentences for test cases

6.3 Evaluation of Punjabi EnConverter

6.4 Evaluation of Punjabi DeConverter

6.5 Tests for the evaluation

6.5.1 Fluency test

6.5.2 Adequacy test

6.5.3 BLEU score

6.5.4 Tests based on quantitative metrics

6.6 Experiments

6.7 Results for fluency score

6.8 Results for adequacy score

6.9 Results for BLEU score

6.10 Error Analysis

Chapter 7: Conclusion and Future Scope

7.1 Conclusion

7.2 Limitations of the proposed system

7.3 Future scope of the proposed work

Appendix-A: Some important EnConversion rules

Appendix-B: Attribute label morphology for masculine gender

Appendix-C: Some important verb paradigms for first person singular masculine gender

Appendix D: Rule base for function word insertion

List of Publications by the Author

References