Chapter 1
Introduction

Homo sapiens have a great tendency of getting their work done by machines and as a result of that they have invented a number of machines. The invention of computers in previous century has resulted into opening of a number of research areas. Machine Translation (MT), also known as automatic translation or mechanical translation, is an area of research that has attracted many researchers during last couple of decades. MT includes the computerized methods that automate all or part of the translation process from one natural language to another. 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, like direct MT, rule-based MT (transfer-based and interlingua-based), corpus-based MT and knowledge-based MT.

Universal Networking Language (UNL) based MT (developed on interlingua-based approach) 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 and it is currently supported by Universal Networking Digital Language (UNDL) foundation, an autonomous organization. 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. UNL system has the potential to bridge the language barriers across the world with the developments of $2n$ components, while traditional approaches requires the $n^*{(n-1)}$ components, where $n$ is the number of languages. UNL represents the information at sentence level in the form of Universal Words (UWs), UNL relations and UNL attributes. The concepts are represented by UWs and UNL relations are used to specify the role of each word in a sentence. The subjective
meaning of the sentence is expressed through UNL attributes. UNL system makes use of word dictionary that stores that information in the form of root words of the language and the corresponding UWs. In the work carried out in this PhD thesis, two components, namely, EnConverter and DeConverter have been developed for Punjabi language. Punjabi language can be written using two scripts, namely, Gurmukhi and Shahmukhi script. Gurmukhi script is used in eastern Punjab (India), and Shahmukhi script in western Punjab (Pakistan). The proposed work handles the Punjabi sentences written in Gurmukhi script. The examples given in this thesis work are in Gurmukhi script along with their transliteration and gloss in Roman script. For inline examples, transliteration is provided in italics and gloss is provided in italics with in single quotes, e.g., ਮਾਤਾ mātā ‘mother’. The transliteration provided in this thesis is based on Gurmukhi to Roman transliteration software ‘GTrans’, developed by Punjabi University, Patiala, India. The UNL graphs presented in this thesis are developed by using UNL graphs web verifier and visualizer tool developed by Spanish Language Center (2004).

In this chapter, the need and challenges of machine translation, approaches of machine translation, objectives of this research, methodology adopted to achieve these objectives and the features of Punjabi language have been presented.

1.1 Need of Machine Translation

MT has a great social and commercial importance. The following points highlight the need of MT in today’s world.

1.1.1 Socio-Political importance of MT

There are communities in the world where more than one language is generally spoken. The dominance of one language may prove disadvantageous to the speakers of other language and may even cause the disappearance of a language. The loss of a language often results into the disappearance of a distinctive culture, and a way of thinking. Thus, MT becomes a social and political necessity for modern societies, who do not wish to impose a common language on their members. It is also a necessity of organizations like the United Nations and the European Community, for whom multilingualism is both a basic principle and a fact of everyday life (Arnold et al., 1994).
1.1.2 Commercial importance of MT
MT has become a necessity in the scenario of global economy. Translation of user guides and user interfaces in multiple languages is the key of success for multi-national companies. Manual translation is slow and proves to be costly owing to the requirement of highly skilled manpower (Arnold et al., 1994).

1.1.3 Coverage of literary works by MT
MT can break the language barriers by making the availability of rich sources of literature to people across the world in their native language. It is really amazing to think of an MT system that can translate literary works from any language into our native language (Siddiqui and Tiwary, 2008).

1.1.4 MT to bridge the digital divide
MT can help to overcome technological barriers. Internet has a great influence on the working of human beings these days. One can find vast amount of information on Internet. The usage of this information is beyond the reach of a significant portion of society as most of this information is in English. MT can help in bridging the digital divide, by translating web pages and electronic mail messages into the native language of a user (Siddiqui and Tiwary, 2008).

1.1.5 MT to assist human translator
Online versions of electronic dictionaries and translation systems can assist human translators in the process of translation. These systems combine multilingual word processing, OCR facilities, terminology management software and translation memories to facilitate human translators. The translation memories enable translators to store original texts and their translated versions side by side. The translator can search here for phrases or even full sentences in one language and can get corresponding phrases in the other language (Hutchins, 2003).

1.1.6 MT for cross language information retrieval
In multilingual environment, there is a need of information retrieval systems capable of searching text in many languages. Cross Language Information Retrieval (CLIR) deals with retrieving information written in a language that is different from the language of the user’s query. For example, a user may pose his query in Punjabi and can retrieve relevant documents written in English. CLIR makes use of MT in two ways: it uses MT system to
translate foreign language documents into the language of the user’s query and it translates the user’s query into target language. The target language query is then used to retrieve target language documents using classical information retrieval techniques (Siddiqui and Tiwary, 2008).

Thus, MT has a potential to overcome language barriers and to make communication between users of different languages much easier.

1.2 Challenges of Machine Translation

There are number of challenging issues in MT that make it a difficult problem to solve. Some of the important issues are structural differences between languages, ambiguity, multiword units etc. In this section, some of the issues are presented in brief.

1.2.1 Word order

The arrangement of words in a sentence varies across languages. For example, English language is Subject-Verb-Object (SVO) type language whereas most Indian languages, including Punjabi, are Subject-Object-Verb (SOV) type language. This makes the approach of word by word translation impractical.

1.2.2 Lexical differences

Sometimes, a word used in one language has no single-word equivalent in another language which results into lexical differences between languages. The example sentence given in (1.1) illustrates lexical differences between English and Punjabi.

English sentence: Ram hugged Rahat. ...(1.1)

Equivalent Punjabi sentence: ਰਾਮ ਨੇ ਰਾਹਤ ਨੂੰ ਗਦਲਾਈਆ। ...(1.2)

Transliterated Punjabi sentence: rām nē rāhat nūm gē lagāiā.

Here, ‘hug’ does not have a single-word equivalent in Punjabi; it has ‘ਗੜਲਾਈਆ’ as multiple-words equivalent in Punjabi.

1.2.3 Lexical ambiguity

If a word has more than one meaning, then it is classified as lexically ambiguous. The example Punjabi sentences given in (1.3), (1.4) and (1.5) illustrates concept of lexical ambiguity.

Punjabi sentence: ਕੱਪੜੇ ਵਿਚ ਜੱਟ ਹਨ। ...(1.3)

Transliterated sentence: kappṛē vic vaṭṭhan.
Equivalent English sentence: There are wrinkles in the cloth.

Punjabi sentence: ਇਹ ਿੱਟ ਸਾਡੇ ਖੇਤ ਵਿਚ ਹੈ। 
Transliterated sentence: ih vaṭṭ sāḍē khēt vic hai.
Equivalent English sentence: This mud path is in our field.

Punjabi sentence: ਮੈਂ ਰੱਸੀ ਿੱਟ ਰਹੀ ਹਾਾਂ। 
Transliterated sentence: maiṃ rassī vaṭṭ rahī hāṃ.

Equivalent English sentence: I am winding the rope.

Here, ਿੱਟ vaṭṭ can be replaced by ‘wrinkle’, ‘mud path’, or ‘wind’ depending on the sense implied in the structure. The correct sense must first be identified for each of the words before selecting the appropriate replacement (Bharati et al., 1994).

1.2.4 Structural ambiguity

A phrase or sentence can be interpreted in more than one way. This kind of ambiguity is known as structural ambiguity. This ambiguity is especially challenging because it requires a deep understanding of the speaker’s intention, and we can often not be certain of what exactly the speaker meant. The sentence structure must be interpreted correctly for translation. The example English sentence given in (1.6) illustrates this concept.

I saw Ram on the hill with the telescope. 

Here, the ‘telescope’ could have been the instrument of seeing, or ‘Ram’ could have been carrying the ‘telescope’, or it is the ‘hill’ with ‘telescope’. Hence, it would be important to identify the relationship of the telescope correctly (Bharati et al., 1994).

1.2.5 Pronoun resolution

The unresolved references of pronoun may lead to incorrect translation. The example English sentence given in (1.7) illustrates this concept.

A dog saw a cow on the road. It started barking on seeing it. 

The first ‘it’ in the example sentence given above can refer to a ‘dog’, ‘cow’ or ‘road’. For its translation into Punjabi, the gender information about the referent will be important because the gender of the verb depends on it. If ‘it’ was referring to the ‘dog’ then the gender will be masculine but for the ‘road’ and ‘cow’ it will be feminine (Bharati et al., 1994). So, it is very important to resolve ‘dog’ or ‘cow’ for ‘it’ in the
example sentence given in (1.7).

1.2.6 Idioms and phrases

It is difficult to translate a sentence with idiomatic expressions, because idioms are composed of words that do not directly contribute to their meaning (Siddiqui and Tiwary, 2008). The example English sentence given in (1.8) illustrates this concept.

The old man finally kicked the bucket. ...(1.8)

If the system does not recognize the idiom ‘kicked the bucket’ which means ‘to die’ then, its translation in Punjabi will end up as given in (1.9), which is a nonsense translation.

буੜਦ ਆਦਮੀ ਨੇ ਆਖਰ ਬਾਲ਼ਟੀ ਨੂੰ ਲਾਟ ਮਾਰੀ। ...(1.9)

These are some of the important challenges that one faces in the development and implementation of a machine translation system.

1.3 Translation Architecture

The various approaches of machine translation can be described with the help of Vauquois triangle (Jurafsky and Martin, 2000) as shown in Figure 1.1. In this triangle, vertical direction (height of the triangle) indicates the increasing depth of analysis and the horizontal direction (width of the triangle) indicates the amount of effort required for transfer. It is evident from Figure 1.1 that the base of the triangle needs the most transfer and the least analysis and generation, while the top of the triangle needs least transfer and most analysis and generation.

Figure 1.1: Vauquois triangle
In the next section, a brief detail about different machine translation approaches has been provided.

1.4 Machine Translation approaches

Machine Translation approaches can be classified into four categories, namely, direct MT, rule-based MT, corpus-based MT and knowledge-based MT, as depicted in Figure 1.2. The rule-based approach can further be classified into transfer-based approach and interlingua approach. The corpus-based translation approach can also further be classified as statistical machine translation and example-based machine translation. In this section, these approaches are briefly described.

![Machine Translation Approaches](image)

**Figure 1.2: Machine Translation approaches**

1.4.1 Direct MT

Direct MT technique was developed during 1950s to make use of newly invented computers for MT. It is based on a straightforward and easily implementable technique, keeping in view less processing power of computers available at that time. This method of translation is depicted in Figure 1.3.

A direct translation system carries out word-by-word translation with the help of bilingual dictionary. As such, it is also known as dictionary driven machine translation approach. It involves a parser, which performs preliminary analysis of the source language sentence to produce its parts of speech information. This information is processed by a rule base to transform the source language sentence into a target language.
sentence. These rules include bilingual dictionary rules and rules to re-order the words. The direct machine translation system with parser and rule-base is also known as Transformer. The transformer architecture is shown in Figure 1.4 (Arnold et al., 1994).

The example English sentence given in (1.10) illustrates the concept of direct machine translation approach.

Amar slept in the garden. ...(1.10)

To translate this sentence into Punjabi, a direct translation system will first look up a dictionary to get target language words for each word appearing in the source language sentence. Then, the words are re-ordered to match the default sentence structure of the target language, i.e., SOV in case of Punjabi. The output of these steps is given in (1.11) and (1.12).

Word-by-word translation:

ਅਮਰ ਸੁੱਟਾ ਵਿਚ ਬਾਗ

amar suttā vic bāģ

Syntactic rearrangement:

ਅਮਰ ਬਾਗ ਵਿਚ ਸੁੱਟਾ

amar bāģ vic suttā

Translation among Indian languages is not very complex owing to the fact that most Indian languages have similar sentence format. Hence, a simple word-by-word
translation, with small set of rules may lead to an acceptable translation in many cases as shown in the example of Hindi-Punjabi translation given in (1.13) and (1.14).

Example Hindi source sentence:
राम ने राहत को प्यार से गले लगाया।
…(1.13)
\( rām nē rāhat kō payār sē galē lagāyā. \)

Equivalent English sentence:
Ram hugged Rahat with love.

Equivalent Punjabi translated sentence:
ਰਾਮ ਨੇ ਰਾਹਤ ਨੂੰ ਵਪਾਰ ਨਾਲ ਗਲ਼ੇ ਲਗਾਇਆ।
…(1.14)
\( rām nē rāhat nū ਪਿਆਰ ਨਾਲ ਗਲ਼ੇ ਲਗਾਇਆ। \)

The unidirectional nature of direct machine translation system is its major disadvantage. In a multilingual scenario, this may be quite expensive. For example, in order to provide translation capability for \( n \) number of languages, we need to develop \( n^*(n-1) \) MT systems (Siddiqui and Tiwary, 2008). Direct machine translation is not suitable for translation of complex sentences, because it requires complex grammatical analysis and word ordering rules. This approach is also not suitable for ambiguous sentences (Seasly, 2003).

1.4.2 Rule-Based MT system

The rule-based MT is used to remove major shortcomings of direct machine translation system. It parses the source text and produces an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation. These systems rely on specification of rules for morphology, syntax, lexical selection, semantic analysis, transfer and generation process. Due to the extensive use of rule-base, these systems are known as Rule-based systems. Depending on the intermediate representation used, these systems are further categorized as transfer-based machine translation and interlingua machine translation.

1.4.2.1 Transfer-Based MT

Transfer approach occupies the level above direct translation in the Vauquois triangle as shown in Figure 1.1. It is also known as Indirect or Linguistic Knowledge (LK) translation. It is used to capture the meaning of the original sentence to produce an intermediate representation. It has three intermediate stages in translation process that
include an analysis stage: to analyze the source text to produce source structure, a transfer stage: to transfer source structure to target structure and synthesis stage: to generate target language text from target structure as shown in Figure 1.5 (Arnold et al., 1994).

An advantage of this approach is its modular structure. Analysis of source language text is independent of language generator. All language-pair specific differences are captured in the transfer stage. To provide translation capability among multiple languages, we need an analyzer and a generator component for each language and a transfer component for each pair of such languages. For example, to provide translation capability for five languages, we need five analyzers, five generators, and twenty transfer components, while in direct translation approach 20 complete translation systems are required (Siddiqui and Tiwary, 2008).

Transfer approach requires a large number of rules for its implementation in multilingual machine translation systems. This is a bottleneck of such a system (Arnold et al., 1994).

1.4.2.2 Interlingua-Based MT

The interlingua approach appears at the apex of the Vauquois triangle as shown in Figure 1.1. It is inspired by Chomsky's findings that regardless of varying surface syntactic structures, languages share a common deep structure. In interlingua-based MT approach, the source language text is converted into a language independent meaning representation...
called Interlingua. As stated by Jurafsky and Martin (2000) ‘An interlingua represents all sentences that mean the same thing in the same way regardless of the source language they happen to be in’.

Interlingua-based MT system, involves two stages in the translation process, including analysis stage: to deeply analyze the source sentence for producing a language independent representation (interlingua); and synthesis stage: the target language is generated from the interlingua. This concept is shown in Figure 1.6. The system involves extensive use of source and target grammar rules during this process.

Interlingua approach is best suited for multilingual machine translation, as it requires only two components for each language: one for conversion from source language to interlingua and other for interlingua to target language. As such, in this approach we require $2n$ components as shown in Figure 1.7 (Subramanian and Narayanan, 1990). The meaning based representation known as interlingua makes this approach suitable for information retrieval applications as well.

There are few problems with interlingua approach. The analysis of source text requires a deep semantic analysis, which sometime results into the loss of information. Another problem is the creation of an adequate interlingua, because it should be both abstract and independent of the source and target languages, which make this task difficult in case of
large multilingual machine translation systems involving the languages with wider differences.

1.4.3 Corpus-Based MT

Corpus-based MT systems have become popular in recent years. These are fully automatic systems that require significantly less human labor than traditional rule-based approaches. However, they require sentence aligned parallel text for each language pair and cannot be used for language pairs for which such corpora do not exist. The corpus-based approach is further classified into statistical and example-based machine translation approaches (Siddiqui and Tiwary, 2008).

1.4.3.1 Statistical MT

Statistical Machine Translation (SMT) uses statistical models for translation whose parameters are derived from the analysis of bilingual text corpora. It does not make use of linguistic rules. SMT was introduced by Warren Weaver in 1949. SMT was re-introduced in 1991 by researchers at IBM. The essence of this method is first to align phrases, word groups and individual words of the parallel texts, and then calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in other language. SMT has given more acceptable results by picking the word(s) that has the highest probability of occupying its current position, given the surrounding words (Seasly, 2003).

SMT considers the MT as a noisy channel metaphor process. It means that, if a user wants to translate a given sentence ‘f’ in the source language ‘F’ to a sentence ‘e’ in the target language ‘E’, the noisy channel model handles this situation in the following way:

Suppose that the sentence ‘f’ to be translated was initially conceived in language ‘E’ as some sentence ‘e’. During communication, ‘e’ was corrupted by the channel to ‘f’. Now, a user will assume that each sentence in ‘E’ is a translation of ‘f’ with some probability,
and the sentence that a user chooses as the translation (‘b’) is the one that has the highest probability as shown in mathematical expression given in (1.15) (Brown et al., 1990).

\[ b = \operatorname{arg\, max} Q_e \, f \]  

...(1.15)

Here, \( Q_e \, f \) depends on two factors, (i) the kind of sentences that are likely in the language ‘E’. This is known as the Language Model (LM) and is represented mathematically as \( Q_e \). (ii) the way sentences in ‘E’ get converted to sentences in ‘F’.

This is called the Translation Model (TM) and represented mathematically as \( Q_f \, e \).

Thus, by applying Bayes’ theorem in the expression given in (1.15), will result the equivalent mathematical expression as given in (1.16).

\[ b = \operatorname{arg\, max} Q_e \, f \, e \]  

...(1.16)

Since, ‘f’ is fixed, \( Q_f \, e \) can be omitted from the expression given in (1.16) to obtain its equivalent mathematical expression given in (1.17) (Ramanathan, 2008).

\[ b = \operatorname{arg\, max} Q_e \, e \, f \]  

...(1.17)

The concept of SMT is illustrated with an example English sentence given in (1.18).

He is walking.  

...(1.18)

The possible Punjabi translations of this example sentence are given in (1.19), (1.20) and (1.21).

ਉਹ ਚੱਲ ਵਰਹਾ ਹੈ।  

...(1.19)

\[ uh \, call \, rihā \, hai. \]  

...(1.20)

\[ call \, rihā \, hai \, uh. \]  

...(1.21)

\[ uh \, tur \, rihā \, hai. \]  

The expression given in (1.17) helps to select the appropriate translation of the source sentence. Because, though \( Q_f \, e \) would be the same for the three sentences, the language model would rule out the last two sentences given in (1.20) and (1.21), \( i.e. \), the first translation given in (1.19) would receive a much higher value of \( Q_e \) than the other two sentences. This leads to another perspective on the statistical MT model, \( i.e.. \), the best translation is the sentence that is both faithful to the original sentence and fluent in the
target language (Jurafsky and Martin, 2000). Thus, in the expression given in (1.17), $\mathbb{Q}_e$ represents fluency and $\mathbb{Q}_f |_e$ represents faithfulness.

The open source tools like CMU-Statistical language modeling toolkit and ‘SRILM’ are used for the creation of LM; ‘GIZA++’ and ‘MGIZA’ are used for the creation of TM; ‘Moses’ and ‘ISI ReWrite’ are used for creation of decoder in the process of SMT.

1.4.3.2 Example-Based MT

The Example-Based Machine Translation (EBMT) approach was suggested by Makoto Nagao in 1984. The EBMT approach requires a bilingual corpus with parallel texts. This approach works on the principle of translation by analogy. This principle is encoded in EBMT through example translations. This concept has been illustrated with an example of sample bilingual corpus of English-Punjabi as shown in Table 1.1.

Table 1.1: Example bilingual English-Punjabi corpus

<table>
<thead>
<tr>
<th>English corpus</th>
<th>Punjabi corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>For how much is that red umbrella?</td>
<td>ਲਾਲ ਛਤਰੀ ਵਕੂੰਨੇ ਦੀ ਹੈ?</td>
</tr>
<tr>
<td>For how much is that book?</td>
<td>ਵਕਤਾਬ ਵਕੂੰਨੇ ਦੀ ਹੈ?</td>
</tr>
</tbody>
</table>

The sample corpus given in Table 1.1 makes it simpler to learn translations of sub-sentential units. After getting trained on sample corpus, the system learns the correspondences as given in Table 1.2.

Table 1.2: Correspondences learned by system

<table>
<thead>
<tr>
<th>English sub-sentential units</th>
<th>Correspondence to Punjabi</th>
</tr>
</thead>
<tbody>
<tr>
<td>For how much is that ‘X’?</td>
<td>X ਵਕੂੰਨੇ ਦੀ ਹੈ?</td>
</tr>
<tr>
<td></td>
<td>$X$ kinnē dī hai?</td>
</tr>
<tr>
<td>Red umbrella</td>
<td>ਲਾਲ ਛਤਰੀ</td>
</tr>
<tr>
<td></td>
<td>lāl chtarī</td>
</tr>
<tr>
<td>Book</td>
<td>ਵਕਤਾਬ</td>
</tr>
<tr>
<td></td>
<td>kitāb</td>
</tr>
</tbody>
</table>

By composing these units, the system can produce novel translations in the future. The architecture of EBMT system is shown in Figure 1.8. An EBMT system has two main
modules, namely, retrieval and adaptation. The retrieval module is used to retrieve translation examples from example-base or Translation Memory (TMEM) for a given input and adaptation is used to carry out necessary modifications in the retrieved example pair to generate translation of target language sentence. The modification may involve addition, deletion, and replacement of morphological words.

EBMT system has certain advantages over rule-based and SMT systems. In the rule-based system, knowledge about the syntax and semantics of source and target, need to be represented as rules, which are difficult to write. The problem with SMT systems is that they need a huge, aligned, parallel corpus. The availability of huge corpora is scarce. Therefore, SMT approach cannot be widely used. EBMT systems require neither a large set of rules, nor a huge parallel corpus. It only requires an example-base, which can be easily created from resources like multi-lingual official notices and reports (Siddiqui and Tiwary, 2008).

1.4.4 Knowledge-Based MT

The important process in knowledge-based translation is to capture as much linguistic knowledge as possible from the source language sentences and store this into the translation system’s knowledge base. For this, the system makes the use of source and target language dictionaries; source and target language structures and rules; word meanings in different contexts and language constructs; domain specific terminology; previously translated words, phrases, sentences, paragraphs; ontological and lexical knowledge; language style and cultural differences etc. By capturing all these knowledge sources, the system produces a high quality output. It is implemented on the interlingua architecture, but differs from interlingual technique by the depth with which it analyzes the source language and its reliance on explicit knowledge of the world. It makes use of augmenter as shown in architecture of KBMT given in Figure 1.9. The only problem of
KBMT is that it is quite expensive to produce such a system because it requires a large amount of knowledge (Seasly, 2003).

1.4.5 Hybrid approaches

The experiences of researchers in last couple of decades have shown that it is not possible to produce highly accurate system by relying on a single machine translation approach. Thus, nowadays, researchers are using hybrid approaches to improve the quality of existing systems. Most commonly, it involves the use of a linguistic method to parse the source text, and a non-linguistic method, such as statistical-based or example-based, to assist with finding the proper interpretation (Siddiqui and Tiwary, 2008).

1.5 Need of UNL based machine translation system for Punjabi language

There is an immense need to develop a machine translation system for Punjabi language. Some machine translation systems for Punjabi Language are at developing stages but they are limited only to a pair of languages, i.e., Hindi-Punjabi. In multi-lingual environment, it is very important to have machine translation systems for large number of languages. In order to have a multi-lingual machine translation system for Punjabi language, it would require a separate system for every language, i.e., \( n^*(n-1) \) language systems will be required for \( n \) number of languages. As such, there is a need to develop an internationally standardized system in which fewer components are needed for a multi-lingual machine translation. An interligua approach based machine translation system is one solution for this problem. UNL is an internationally standardized interligua that has the potential to bridge the language barriers across all the languages of the world.
Such a system for Punjabi language will certainly be very helpful for more than 91 million Punjabi language users (Lewis and Paul, 2009).

1.6 Objectives of this research

The main objective of this study is to design and develop a multi-lingual machine translation system for Punjabi language. In order to perform this task, following objectives were proposed to be carried out.

i) An electronic Punjabi-UW dictionary of about 40,000 root words was to be created in the proposed work.

ii) A DeConverter for Punjabi language was to be developed which can convert a UNL document to corresponding Punjabi language text.

iii) An EnConverter for Punjabi language was to be developed which can convert input Punjabi text to corresponding UNL output.

iv) A web interface was to be designed with the provisions for input of Punjabi sentences for corresponding UNL output and for input of a UNL document for corresponding Punjabi text output.

v) The scope of work was proposed to be limited to simple sentences with an approximate accuracy of 70%.

1.7 Major contributions and achievements

A major outcome of this research work is the development of Punjabi EnConverter and Punjabi DeConverter. A Punjabi EnConverter has been developed that converts input Punjabi sentence to corresponding UNL expression. A web interface has been designed for online EnConversion of input Punjabi sentence to UNL expression. It enables the conversion of input Punjabi sentence to other languages having their own UNL DeConverter. A Punjabi DeConverter has also been developed that converts a UNL expression to corresponding Punjabi language text. A web interface has also been designed for online DeConversion of UNL expression to corresponding Punjabi sentence. It enables the Punjabi readers to read the sentences in their local language that are originally written in different languages having their equivalent UNL expression present on the web.

This system will also provide an opportunity to research scholars working on MT to explore UNL as an interlingua.
### 1.8 Features of Punjabi language

Punjabi language is an Indo-Aryan language and is one of the constitutionally recognized languages of India. Indo-Aryan languages form a subgroup of the Indo-Iranian group of languages, which in turn belong to Indo-European family of languages. Punjabi is widely spoken in north-west India, Pakistan, United States, Australia, United Kingdom and Canada. There are more than 91 million native speakers of Punjabi language, which makes it approximately the 12th most widely spoken language in the world (Lewis and Paul, 2009). Gill (2008) has explained the features of Punjabi language. The following paragraphs give a comprehensive review of these features.

Punjabi has word classes in the form of noun, pronoun, adjective, cardinal, ordinal, main verb, auxiliary verb, adverb, postposition, conjunction, interjection and particle. Punjabi nouns change forms for number (singular or plural) and case in sentences. Punjabi nouns have assigned gender (masculine or feminine). For example, ਕੂੰਧ kandh ‘fence’, ਕ ਰਸੀ kurasī ‘chair’, ਸੜਕ saṛak ‘road’ etc. are used in feminine gender, and ਮੇਜ਼ mēz ‘table’, ਟਰੱਕ ṭarakk ‘truck’, ਦਿਨ din ‘day’ etc. are used in masculine gender.

Punjabi has six types of pronouns. These are: personal pronouns, *e.g.*, ਮੈਂ maiṁ ‘i’, ਤੂੰ tūṁ ‘you’; reflexive pronouns, *e.g.*, ਅਪ āp (some what equivalent to honorific form of English second person ‘you’); demonstrative pronouns, *e.g.*, ਉਹ uh ‘that’ and ਇਹ ih ‘this’; indefinite pronouns, *e.g.*, ਕੋਈ kōī, ਕ ਝ kujh, ਸਾਰੇ sārē etc.; relative pronouns (to join two clauses in a complex sentence), *e.g.*, ਜੋ jō and ਜ਼ਿਹ ਤਾ jihṛā and interrogative pronouns, *e.g.*, ਕੌਣ kauṇ ‘who’, ਕੀ kī ‘what’ etc.

In Punjabi language, adjectives usually precede the nouns but follow the pronouns. For example, ਸੋਹਣਾ sōhṇā ‘handsome’, ਕਾਲਾ kālā ‘black’ are functioning as adjectives and they precede nouns in ਸੋਹਣਾ ਮਨੰਡ sōhṇā muṇḍā ‘handsome boy’, ਕਾਲਾ ਗਾਂਡ kālā ghōṛā ‘black horse’, respectively. The examples of adjectives following pronouns are, ਮੈ ਸੋਹਣਾ hāṁ ‘I am handsome’, ਤੂ ਸੋਹਣਾ haiṁ ‘you are handsome’. Punjabi adjectives can be classified into two categories, inflected and uninflected. For example, ਸੋਹਣਾ sōhṇā ‘handsome’ inflections are ਸੋਹਣਾ sōhṇā ‘handsome’ (direct singular), ਸੋਹਣੇ sōhṇē ‘handsome’ (oblique singular and direct plural),
and ਸੋਹਵਣਆਂ sōhṇāṁ ‘handsome’ (oblique plural). The uninflected adjectives are ਲਾਲ lāl ‘red’, ਮਿਹਨਤੀ mihnatī ‘hardworking’, ਮਸ਼ਹਰ mashhūr ‘famous’ etc.

Cardinals and Ordinals can be used for both the genders and change forms for case (direct and oblique). Except, ਇੱਕ ikk ‘one’, which is in singular number, all the remaining cardinals (ਦੋ dō ‘two’, ਪੰਜ pañj ‘five’, ਅਠਾਰਾ aṭhārāṁ ‘eighteen’ etc.) are in plural number. Generally, all the ordinals are used in singular number. For example, ਪੰਜਵਾਂ pañjvāṁ ‘fifth’, ਚੇਵਾਂ chēvāṁ ‘sixth’ etc. are all ordinals.

In a Punjabi sentence, verbs must agree with the subject or object of the sentence in terms of gender, number, and person. Punjabi verbs change forms for gender, number, person, and tense. The verbs have assigned transitivity and causality. In Punjabi, there are two auxiliary verbs – ਹੈ hai for present tense (e.g., ਰਾਮ ਅੂੰਬ ਖਾਣਦਾ ਹੈ। rām amb khāndā hai. ‘Ram eats mango’) and ਸੀ sī for past tense (e.g., ਰਾਮ ਨੇ ਅੂੰਬ ਖਾਧਾ ਸੀ। rām nē amb khādhā sī. ‘Ram had eaten mango’). All the forms of these two auxiliary verbs can equally be used for both the genders. For future tense in sentences, ‘EGA’ form of main verb is used and in those sentences auxiliary verb is thus not used (e.g., ਰਾਮ ਅੂੰਬ ਖਾਿੇਗਾ। rām amb khāvēgā. ‘Ram will eat mango’).

Adverbs can indicate manner, time, place, condition etc. For example, ਉੱਪਰ uppar ‘upon’, ਉੱਤੇ uttē ‘over’, ਹੇਠ hēṭhā ‘below’ etc. are some Punjabi adverbs. Postpositions are similar to prepositions in English. These link noun, pronoun, and phrases to other parts of the sentence. Some Punjabi postpositions are ਨੇ nē, ਨੂੰ nūṁ, ਉੱਤੇ uttē ‘over’, ਦਾ dā ‘of’ etc. In Punjabi, postpositions follow the noun or pronoun unlike English, where these precede the noun or pronoun, and thus termed prepositions. In Punjabi, the postpositions can be classified into two types, namely, inflected postpositions and uninflected postposition. For example, ਦਾ dā ‘of’ (marker of possessive case) postposition is an inflected postposition because it changes forms for gender, number and case. There are a large number of postpositions, which do not change forms at all, e.g., ਨੇ nē (instrumental or agentive case marker), ਨੂੰ nūṁ (generally used with objects), and ਤੋਂ tōm ‘from’ are known as uninflected postpositions.
Conjunctions are used to join words, phrases, or independent clauses in compound sentences. For example, ਅਤੇ/ਤੇ atē/tē ‘and’, ਜਾਂ jāṃ ‘or’ are acting as co-ordinate conjunctions. Sub-ordinate conjunctions are typically used to introduce dependent clauses in complex sentences. For example, ਜੇ jē, ਉੱਂ uññ, ਤਾਂ tāṃ etc.

Interjection is used to express emotions in sentences. For example, ਹਾਏ hāē, ਆਹਾ āhā, ਸਦਕੇ sadkē, ‘ਸੀ ਆਹਾ ਮੇਂ jī āiāṃ nūṃ’ ‘welcome’ all are interjections. There are various particles that are used in the Punjabi sentences for emphasis, negation etc. These particles do not change forms for any of the grammatical categories. These include, emphatic particles that are used to emphasize or put stress on some part of the sentences, e.g., ਦੀ ਦੀ, ਦੀ hī, ਦੀ ਦੀ ਦੀ tāṃ, ਦੀ ਦੀ ਤੇ tē, ਦੀ ਦੀ ਖਾ ਖਾ khā āīē etc.; negative particles that are normally used for negation effect in sentences, e.g., ਦੀ ਦੀ ਦੀ ਤੇ nā, ਦੀ ਦੀ ਤੇ ਨਹੀਂ nāhī; honorific particles that are used for giving respect, e.g., ਦੀ ਦੀ ਦੀ ਤੇ jīō, ਦੀ ਦੀ ਕਲਾ bāhī ahī etc. and vocative particles that are commonly used in vocative case, i.e., to call someone, e.g., ਕਰੀ ਕਰੀ ਦੀ ਕਲਾ aਫ਼ਾ ਦੀ ਕਲਾ aਫ਼ਾ (masculine singular), ਕਵਿੰਦੀ aਫ਼ਾ (masculine plural), ਕਵਿੰਦੀ aਫ਼ਾ (feminine singular) etc.

Punjabi phrases can be broadly classified into two types, namely, nominal phrases (built using the words of various word classes like noun, pronoun, adjective etc.) and verb phrases (built using primarily the words of main verb and auxiliary verb word classes) (Gill, 2008).

1.9 Organization of thesis

This PhD thesis is divided into seven chapters. The details about the literature review carried out for this work is presented in Chapter 2. Introduction to UNL framework and creation of Punjabi-Universal Word lexicon has been discussed in Chapter 3. The details on the design and development of Punjabi-UNL EnConverter have been given in Chapter 4. Chapter 5 covers the design and development of UNL-Punjabi DeConverter. The results and discussions on the implementation of proposed work have been provided in Chapter 6. Chapter 7 of thesis discusses the conclusion and future scope of the work.
Chapter Summary

In this chapter, the need and challenges of machine translation, approaches of machine translation, objectives of this research, methodology adopted to achieve these objectives and the features of Punjabi language have been presented. The machine translation has socio-political and commercial importance in the modern world. It helps to provide wide readership to literary works and used to bridge the digital divide. It is also used for cross language information retrieval. Automated Machine Translation is a complex problem involving the challenges of word order, lexical differences, lexical ambiguity and structural ambiguity among the languages. Machine Translation approaches can be classified into four categories, namely, direct MT, rule-based MT, corpus-based MT and knowledge-based MT. A direct translation system carries out word-by-word translation with the help of bilingual dictionary. Direct machine translation is not suitable for translation of complex sentences and for multilingual translations. The rule-based MT is further categorized as transfer-based machine translation and interlingua machine translation. A transfer-based machine translation system has three intermediate stages in translation process that include an analysis stage, a transfer stage and synthesis stage. Interlingua-based MT system, involves two stages in the translation process, including analysis stage and synthesis stage. The corpus-based approach is classified into statistical and example-based machine translation. Statistical machine translation requires three key components, namely, Language Model (LM), Translation Model (TM) and decoder. Example-Based Machine Translation (EBMT) approach requires a bilingual corpus with parallel texts. An EBMT system has two main modules, namely, retrieval and adaptation. The main process in knowledge-based translation is to capture as much linguistic knowledge as possible from the source language sentences and store this into the translation system’s knowledge-base. Nowadays, researchers are using hybrid approaches to improve the quality of existing systems.

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.