INTERNATIONAL JOURNAL OF TRANSLATION
VOL. 23, NO. 1, JAN-JUN 2011

Automatic Translation System from Punjabi to English for Simple Sentences in Legal Domain

KAMALJEET KAUR BATRA
DAV College, Amritsar

G. S. LEHAL
Punjabi University, Patiala

ABSTRACT

The system has been developed to translate simple sentences in legal domain from Punjabi to English. Since the structure of both the languages is different, direct approach of translating word by word is not possible. So, indirect approach i.e. rule based approach of translation is used. The system has analysis, translation and synthesis component. The steps involved are preprocessing, tagging, ambiguity resolution, phrase chunking, translation and synthesis of words in target language. The accuracy is calculated for different phases of the system and the overall accuracy of the system for a particular type of sentences is about 60%.

Keywords: Tagger, Chunker, Ambiguity Resolver, Transliterator

1. INTRODUCTION

The system is a machine aided translation system as it requires certain preprocessing and post processing tasks which should be performed by human beings. The need of the system arises from the translations of the legal documents transferred from district courts of Punjab to the high court. The FIR's which are written in Punjabi language are translated to English before presenting it to the high court. The mechanization of translation has been one of humanity's oldest dreams. In the twentieth century it has become a reality, in the form of computer programs capable of translating a wide variety of texts from one natural language into another. There are no "translating machines" which, at the touch of a few buttons, can take any text in any language and produce a perfect translation in any other language without human

intervention or assistance. What has been achieved is the development of programs which can produce "raw" translations of texts in relatively well-defined subject domains, which can be revised to give good-quality translated texts which in their unedited state can be read and understood by specialists in the subject for information purposes. In some cases, with appropriate controls on the language of the input texts, translations can be produced automatically those are of higher quality needing little or no revision.

2. LITERATURE REVIEW

Machine Translation activities in India are relatively young. The earliest efforts date from the mid 80s and early 90s. The prominent among these efforts are the research and development projects at Indian Institute of Technology, Kanpur; University of Hyderabad, National Center for Software Technology, Mumbai and Center for Development of Advanced Computing (CDAC), Pune (Naskar & Bandyopadhyay 2005). Since the mid and late 90's, a few more projects have been initiated – at Indian Institute of Technology, Bombay; International Institute of Information Technology, Hyderabad; Anna University – KB Chandrasekhar Research Center, Chennai and Jadavpur University, Kolkata. There are also a couple of efforts from the private sector – from Super Infosoft Private Limited, and more recently, the IBM India Research Laboratory. Of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects. Indian Languages (TDIL) program of the Ministry of Information Technology (MIT) and also the UNDP. University Grants Commission (UGC) also started supporting minor and major research projects involving development of linguistic parsers and machine translation. Indian Institutes of Technology (IITs), Indian Institutes of Information Technology (IIITs), Centre for Development of Advanced Computing (C-DAC), Indian Institute of Science (IIS), Indian Statistical Institute (ISI), Jawaharlal Nehru University (JNU), Mahatma Gandhi International Hindi University (MGIHU), major Sanskrit universities and other institutes for significant contributions in this field. The private enterprises like Tata Institute of Fundamental Research (TIFR), Tata Consultancy Services (TCS) have also funded Indian language technology R&D.

IIT Guwahati, CDAC Kolkata, JNU New Delhi are also involved in developing the machine translation systems for different Indian languages (Naskar & Bandyopadhyay 2005). Advanced Centre for technical development of Punjabi Language, Literature and Culture,

Punjabi University Patiala has also entered into the field of Machine Translation and successfully developed Hindi-Punjabi machine translation system and vice versa. Thapar University, Patiala is also working on UNL based machine translation system.

3. APPROACH FOLLOWED

The approach followed for translation is the transfer approach. The transfer architecture not only translates at the lexical level, like the direct architecture, but syntactically and sometimes semantically. The transfer method will first parse the sentence of the source language. It then applies rules that map the grammatical segments of the source sentence to a representation in the target language. After syntactically and semantically analyzing the sentence, we can easily translate a sentence even with different structures i.e.

Subject Object Verb → Subject Verb Object
(Punjabi) → (English)

The rules, which are used for the structural transformation of sentences, for solving the ambiguity problem, all are stored in the database which we call the rule base and has been described in detail in Section 5.3. The indirect approach, first of all, divides a sentence into words, tags each word using morph database, resolves ambiguity, divide it into phrases, translates each word using bilingual dictionary, and then synthesize the translated words using rules of English language.

4. STEPS FOLLOWED FOR TRANSLATION

4.1. *Preprocessing*

Since the sentences are taken from number of legal documents, there are different types of sentences, preprocessing module change the sentences to a particular format so that it can be translated with more accuracy. Eg., system only works for simple sentences and if a sentence is either complex or compound, it is divided to two or more simple sentences. The structure of simple sentence is limited to SOV structure i.e. Subject-Object-Verb. In certain sentences, the structure contains, Object-Subject-Verb, those are not considered. The above said part of Preprocessor is manual and not automated.

It was also recognized that in a Punjabi sentence, verb phrase, which is the main component of the sentence, is further divided into different constituents i.e. main verb, conjunct verb, primary,

progressive or modal operators, even then its complexity is very high and creates problem while translating. E.g.

P: ਰਹਿਮ ਦੀ ਪਟੀਸ਼ਨ ਰੱਦ ਕਰ ਦਿੱਤੀ ਗਈ
T: *rahim dī paṭīshan radd kar dittī gaī*

P: ਆਬਕਾਰੀ ਐਕਟ ਅਧੀਨ ਮਾਮਲਾ ਦਰਜ ਕਰ ਲਿਆ ਗਿਆ ਹੈ
T: *ābkārī aikaṭ adhīn māmlā daraj kar liā giā hai*

In the above sentence, ਕਰ (*kar*) is a conjunct verb, ਦਿੱਤੀ (*dittī*) is also a conjunct verb and ਗਈ (*gaī*) is the passive operator. Both the conjunct verbs present, in the system increases complexity, such type of words are joined by using a joining database. Here ਕਰ (*kar*) and ਦਿੱਤੀ (*dittī*) are combined to ਕੀਤੀ (*kītī*) and the sentence becomes

P: ਰਹਮ ਦੀ ਪਟੀਸ਼ਨ ਰੱਦ ਕੀਤੀ ਗਈ
T: *raham dī paṭīshan radd kītī gaī*

P: ਆਬਕਾਰੀ ਐਕਟ ਅਧੀਨ ਮਾਮਲਾ ਦਰਜ ਕੀਤਾ ਗਿਆ ਹੈ
T: *ābkārī aikaṭ adhīn māmlā daraj kītā giā hai*

This part of preprocessing phase is an automated process and it combines the adjoining words from the sentence to a single word by checking them from the database created of joined words. Some of the noun phrases also contain words that can be joined and represents a single equivalent in English. E.g. ਪਿਤਾ ਜੀ (*pitā jī*), ਮਾਤਾ ਜੀ (*mātā jī*) these words have a single equivalent as father and mother.

4.2. *Tokenization*
The sentence is divided into words called tokens on the basis of spaces between them which are then passed to further phases.

4.3. *Morph analyzing and tagging*
The next step is to tag each word with the grammatical information about it. In Punjabi grammar, the parts of speech include noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, operators, auxiliary verbs etc. Tag contains the information about grammatical category of word, gender, number, person and the case in

which it can be used. The information is stored in the morph database. Tag can be arranged in the form "grammatical category-gender-person-number-case-tense-phrase-type." The fields not applicable to a particular category are left blank. E.g. Tags for the word ਭਰਾ (*bharā*) **are 'n-m- -s-d- - - -', 'n-m- -p-d- - - -', 'v-x-s-s- -f-x- -'**. The above tag for the word show that it can be used as noun with masculine gender, singular as well as plural and in direct case. It can also be used as verb with any gender, singular, second person, and future tense. The complete information for the tags is available from the morph database described in Section 5.1. In Punjabi language, a word can have number of tags as a particular word can be used in number of ways. The tagger first checks the category of each word from the database and then adds Gender, Number, Person or Tense information to it (Sharma 2008; Bharati & Sangal 1993).

For example in case of Nouns, Tagger gives information of Gender, Person, Number and Case (n-GP-NC). Tag for the word ਭਰਾ (*bharā*) – n-m-s-d (noun-masculine-singular and direct case). In case of nouns person information is not in use.

Similarly for personal pronouns, ਮੇਰੀ (*mērī*) – pp-f-f-s-d (personal pronoun – feminine, first person, singular, and in direct case)

4.4. *Ambiguity resolution*

The rules considering the tags for surrounding words are used for resolving ambiguities at different levels. Before the step of ambiguity resolution, each word is attached with number of tags. Since a particular word may have number of tags, there is need to check which tag is applicable to a particular word in a sentence. For this purpose, there is need to apply certain rules depending upon the grammatical category, number, gender or other information. These rules should be prioritized according to the information and stored in Rule base of Punjabi. (Section 5.3)

First level of ambiguity exists when a particular word can have number of tags of different grammatical category. The rules should check the grammatical category for the surrounding words so that it can conclude the tag of that particular word. E.g..

P: ਉਹਦੀ ਬੋਲੀ ਬਹੁਤ ਰੁੱਖੀ ਸੀ

T: *uhdī bōlī bahut rukkhī sī*

In the above sentence, the word **ਬੋਲੀ** (*bōlī*) has two tags, one show that it is noun and the other shows that it is verb. In the first sentence, surrounding words are demonstrative pronoun and adjective, in this case, it is used as a noun, so the tag for noun is attached with this word. But, if the sentence is

P: **ਉਹ ਸਾਡੇ ਘਰ ਆ ਕੇ ਬਹੁਤ ਉੱਚੀ ਬੋਲੀ**

T: *uh sāḍē ghar ā kē bahut uccī bōlī*

Here the word **ਬੋਲੀ** (*bōlī*) is used as verb, as it is preceded by adverb and present in verb phrase. For solving this ambiguity, a matrix is formed containing different tags for words and from that by comparing the rules for adjoining words, the tag is selected. E.g. In the above sentence, the matrix which is formed is shown below

P: **ਉਹ ਸਾਡੇ ਘਰ ਆ ਕੇ ਬਹੁਤ ਉੱਚੀ ਬੋਲੀ**

T: *uh sāḍē ghar ā kē bahut uccī bōlī*

Rows shows the word number and the column shows the category of words

	1	2	3	4	
1	d	inter	_	_	
2	pp	iaj	_	_	
3	n	_	_	_	
4	v	pri	_	_	
5	ipo	par	_	_	
6	uaj	_	_	_	
7	iaj	_	_	_	
8	n	v	_	_	

Here “d” refers to demonstrative Pronoun, “pp” – personal pronoun, “inter” – interjection, “iaj” – inflected adjective, “n” – noun, “v” – verb,

"pri" – primary operator, "ipo" – inflected postposition, "par" – particle, "uaj" – uninflected adjective.

The first word ਉਹ (*uh*) has two categories "d" and "inter" whereas the second word ਸਾਡੇ (*sāḍē*) is categorized to "pp" and "iaj." To fix the category of first word all the rules for two preceding and succeeding categories are checked. For first word, there are no preceding words so the rules

1. d pp n
2. d iaj n,
3. inter pp n,
4. inter iaj n

Above four are compared from the database and the rule with higher priority will be chosen for first word. Similarly by fixing the category of preceding words, category for succeeding words is checked.

Second level of ambiguity that has been resolved is, when there are number of tags that shows a particular word as noun, but can be used as singular or plural, as tag for the word ਬੰਦੇ (*bandē*) is, '**n-m- -s-o - - -', 'n-m- -p-d- - - -'.**

The tagged word can be noun in singular or a noun in plural. E.g., in the sentence,

P: ਬਹੁਤ ਸਾਰੇ ਬੰਦੇ ਲੜਨ ਆ ਗਏ

T: *bahut sārē bandē laṛan ā gaē*

in this case we should select the tag **'n-m- -p-d- - - - -',** and its appropriate word in English is "men," whereas in the case

P: ਉਹ ਬੰਦੇ ਨੇ ਕੁੜੀ ਨੂੰ ਛੇੜਿਆ

T: *uh bandē nē kuṛī nūṃ chēṛiā*

the tag for ਬੰਦੇ (*bandē*) should be '**n-m- -s-o- - - -**' and its appropriate meaning is "man." Such type of ambiguity can be resolved by considering the number i.e. Singular or plural of the auxiliary verb or the main verb present in the sentence. In the first sentence, ਗਏ (*gaye*) is specified as auxiliary verb with plural attribute, whereas in second sentence ਛੇੜਿਆ (*chēṛiā*) is specified as verb with singular attribute.

Similarly there are numbered tags for demonstrative pronouns.

P: ਇਹ ਮੇਰੀ ਕਿਤਾਬ ਹੈ
T: *ih mērī kitāb hai*
E: this is my book

P: ਇਹ ਮੇਰੀਆਂ ਕਿਤਾਬਾਂ ਹਨ
T: *ih mērīāṃ kitābāṃ han*
E: these are my books

ਇਹ (*ih*) has two tags, i.e. showing singular and plural and according to the number attribute of auxiliary verb, it is translated to "this" or "these." Similarly the ambiguity related with the cases (direct or oblique) and for the gender is resolved by considering the gender for surrounding words. E.g.

P: ਉਹ ਕੰਮ ਤੇ ਜਾ ਰਿਹਾ ਸੀ
T: *uh kamm tē jā rihā sī*

P : ਉਹ ਕੰਮ ਤੇ ਜਾ ਰਹੀ ਸੀ
T: *uh kamm tē jā rahī sī*

Here ਉਹ (*uh*) is translated to "he" in first sentence and "she" in second sentence depending upon the gender of the verb phrase.

Next level of ambiguity exists if a word has number of meaning but has same grammatical category, number or gender. Since the system is made domain specific, such type of ambiguity is not resolved and meaning of the word related to legal domain is considered. To consider which meaning of a word is related to legal domain, the most probable meaning of the word from the set of sentences used in training is taken.

4.5. *Phrase chunking*

Chunking involves the division of sentence into the corresponding phrases depending upon the rules. By combination of different word classes, we make phrases, such as Noun Phrases, Adjective Phrases, Prepositional Phrases and Verb phrases. A **noun phrase** consist of nouns or pronouns may preceded by its modifiers. An **adjective phrase** is a phrase with an adjective as its head. In Punjabi language, preposition is called postposition as it comes after the noun or pronoun. The preposition and its object make up a **prepositional phrase**. For example, in the sentence, ਉਹ ਮੇਰੇ ਘਰ ਵਿੱਚ ਰਹਿੰਦਾ ਹੈ (*uh mērē ghar vicc*

rahindā hai), ਮੇਰੇ ਘਰ ਵਿੱਚ (*mērē ghar vicc*) is the prepositional phrase. In the sentence ਕੁੜੀ ਕੁਰਸੀ ਤੇ ਬੈਠੀ ਹੈ (*kuṛī kurasī tē baiṭhī hai*), the prepositional phrase ਕੁਰਸੀ ਤੇ (*kurasī tē*) modifies the verb ਬੈਠੀ (*baiṭhī)*.

Verb Phrase consists of main verb, followed by operators and auxiliary verb and preceded by an adverb. Operators are of four types, Primary operator, Passive operator, Modal operator and Progressive operator. These operators help to emphasize the working of main verb. Primary operators include ਸਕ (*sak*), ਚਲ (*chal*), ਚੜ (*chad*), ਵੇਖ (*vēkh*), ਹੋ (*ho*) etc. These operators are used after the main verb as

P: ਮੈਂ ਕੰਮ ਕਰ ਸਕਦਾ ਹਾਂ
T: *maiṃ kamm kar sakdā hāṃ*
P: ਤੂੰ ਇਹ ਕੰਮ ਕਰ ਵੇਖ
T: *tūṃ ih kamm kar vēkh*

P: ਮੋਹਨ ਇੱਥੇ ਬੈਠ ਜਾ
T: *mōhan itthē baiṭh jā*

Here ਜਾ (*jā*), ਸਕਦਾ (*sakdā*), ਵੇਖ (*vēkh*) are emphasizing the verb ਕਰ (*kar*) and ਬੈਠ (*baith*)

The main component of Punjabi verb phrase is the root verb. Most of the verbs used in Punjabi language are one word, but there can be two word verbs such as double verbs and conjunct verbs. E.g. ਤੁਰ ਤੁਰ ਕੇ (*tur-tur ke*), ਖਾਧਾ ਪੀਤਾ (*khādhā pītā*) are double verbs. In English almost all the nouns can occur as verbs. But in Punjabi, verbalization of nominal is effected by combining two lexical items-noun and a simple verb. These verbs are placed in the category of conjunct verbs. These are certain verbs such as ਕਰਨਾ (*karnā)*, ਕੀਤਾ (*kītā*), ਦਿੱਤਾ (*dittā*) etc which have their translated word in English depending upon the noun which precede this word (Sinha & Thakur 2004). E.g.

P: ਮੇਰੀ ਅਰਜ਼ੀ ਮਨਜ਼ੂਰ ਕੀਤੀ ਗਈ
T: *mērī arzī manzūr kītī gaī*
E: My application was accepted

P: ਜੱਜ ਨੇ ਦੋਸ਼ੀ ਨੂੰ ਮੁਆਫ਼ ਕੀਤਾ
T: *jajj nē dōshī nūṃ muāpha kītā*

E: Judge forgive culprit

P: ਉਹ ਨੇ ਮੇਰੇ ਉਤੇ ਹਮਲਾ ਕੀਤਾ
T: *uh nē mērē utē hamlā kītā*
E: He attacked me

The preceding verb in conjunct verb can be adjective also. E.g. ਖੁਸ਼ ਕਰਨਾ (*khush karnā*)

For implementation in MT System, a different database is maintained for conjunct verbs having their English equivalent by checking the preceding word. There are certain words like ਦਿੱਤੀ (*dittī*), which can be used as main verb in a sentence or it functions as conjunct verb in another sentence, so first of all it will check it from the database of conjunct verbs and its preceding noun and if not present, then considered as main verb.

As in the sentences,

P: ਕੁੜੀ ਨੇ ਮੁੰਡੇ ਨੂੰ ਰੋਟੀ ਦਿੱਤੀ
T: *kuṛī nē muṇḍē nūṃ rōṭī dittī*

P: ਮੈਂ ਮੋਹਨ ਨੂੰ ਆਪਣੀ ਕਿਤਾਬ ਦਿੱਤੀ
T: *maiṃ mōhan nūṃ āpṇī kitāb dittī*

P: ਮੋਹਨ ਨੇ ਮੁੰਡੇ ਨੂੰ ਮਾਫੀ ਦਿੱਤੀ
T: *mōhan nē muṇḍē nūṃ māphī dittī*

In the first two sentences ਦਿੱਤੀ (*dittī*) is the main verb and its English equivalent is "gave" whereas in last sentence ਦਿੱਤੀ (*dittī*) is the conjunct verb related to the noun ਮਾਫੀ (*māphī*) and its equivalent translation is "forgive." The differences between Punjabi and English verbs give rise to language divergences in machine translating one language to another.

Phrase chunking is performed using the rules of noun phrases, adjective phrases, postpositional phrases and the verb phrases for Punjabi. The rules for division of sentence to phrases are stored in the rule base of Punjabi and the conversion rules are stored in the rule base of target language (Section 5.3).

4.6. *Addition of* karak *roles*

After chunking, the semantic information is attached with the phrases and the phrases are assigned roles according to the source sentence.

This is needed as the source language is a free order language and the position of phrase does not specify its role.

A Verbal root denotes:

- The Activity
- The result

Locus of activity is *karta* and locus of result is *karma*. The following are the relations or the *karaks*

- *Karta* – subject/agent/doer: The first noun phrase in direct case or oblique case in a sentence are referred as subject (*Karta*).
- *Karma* – Object/patient: The noun phrases with ਨੂੰ (*nūṃ*) postposition are termed as object (*Karam*). Sometimes this can also be in direct case.
- *Karana* – The nominal phrases with postposition ਨਾਲ (*nāl*), ਤੋਂ (*tōṃ*), ਰਾਹੀਂ (*rāhīṃ*) are *karan karak*, which shows the instrument with which work is to be performed.
- *Sampradaan* – The nominal phrase with postposition ਨੂੰ (*nūṃ*) or ਲਈ (*laī*) is termed as beneficiary relation or sampradaan karak if any other nominal phrase with karam karak exists in direct case.
- *Apaadaan* – The noun phrase with postposition ਤੋਂ (*tōṃ*), ਵਿੱਚੋਂ (*viccōṃ*), ਉੱਤੋਂ (*ooton*) helps to determine phrases with *apaadaan karak* which has source relationship .
- *Adhikarana* – The phrases related in location or time is *adhikarana Karak*. The postpositions ਤੇ (*tē*), ਉੱਤੇ (*uttē*), ਵਿੱਚ (*vicc*) helps to recognize these relations. Even the phrases can be without any postposition.

Examples:

P: ਉਹਨੂੰ ਨੌਕਰੀ ਤੋਂ ਬਰਖਾਸਤ ਕੀਤਾ ਜਾਵੇਗਾ

T: *uhnūṃ naukrī tōṃ barkhāsat kītā jāvēgā*

ਉਹਨੂੰ (*uhnūṃ*) – *karta karak*

ਨੌਕਰੀ ਤੋਂ (*naukrī tōṃ*) – *karan karak* (followed by ਤੋਂ (*ton*))

P: ਮੈਂ ਆਪਣੀ ਮਰਜ਼ੀ ਨਾਲ ਮਕਾਨ ਦਾ ਕਬਜ਼ਾ ਦੇ ਦਿੱਤਾ ਹੈ

T: *maiṃ āpṇī marzī nāl makān dā kabzā dē dittā hai*

ਮੈਂ (*maiṃ*)– *karta karak*

ਆਪਣੀ ਮਰਜ਼ੀ ਨਾਲ (*āpṇī marzī nāl*)– *karan karak*

ਮਕਾਨ ਦਾ ਕਬਜ਼ਾ (*makān dā kabzā*)– *karam karak*

ਦੇ ਦਿੱਤਾ ਹੈ (*dē dittā hai*)– verb phrase

After chunking, the phrases are recogonised, by recognizing the postpositions of the postpositional phrases from the database, and then those are marked with the respective relations or *karaks*. Syntactic cues help in identifying the relation types. Thus semantic information is attached with each local word group. Rules are being made to organize this information from the sentence which is stored in the rule base of Punjabi. A large number of legal sentences are considered to build the rules for that. The rules are made manually and once the rules are build using training sentences and stored in the rule base, the system use it for identifying the *karak* roles for the entire sentences. The rule base has been described in Section 5.3.

4.7. *Translation using bilingual dictionary*

Next step in translation is the use of a bilingual dictionary to translate each word in Punjabi to its English equivalent. Translation is performed by using the bilingual dictionary described in Section 5.2. There are certain words used in Punjabi language which are of English origin, as ਸਕੂਲ(*sakūl*), ਟੀਚਰ(*ṭīcar*), ਡਾਕਟਰ(*ḍākṭar*) etc. The meaning of such words stored in dictionary is the transliteration of the same.

4.8. *Transliteration of proper nouns*

After translating each word using the dictionary, there are certain words which are not present in the dictionary such as names of persons, names of cities, these all are proper nouns, and these should be transliterated. Transliteration means to write them sensing the characters in the words e.g. ਮਨਜੀਤ (*manjeet*) in Punjabi is transliterated in English as "manjeet," m for ਮ, n for ਨ, j for ਜ, ee for ੀ, t for ਤ. This transliteration process also uses a database of transliterating characters and also certain rules to insert vowels wherever needed.

4.9. *Synthesis*
After getting English equivalent of each word in Punjabi sentence, it should be synthesized first to phrases and then to the sentence using structural rules of English language. These rules of language are also stored in the rule base of English which has been described further in Section 5.3.

4.10. *Post processing*
It includes certain rules which are used to correct the sentence after translation in the target language.

4.10.1. *Omission of auxiliary verbs*
Some sentences after translating in English contains the word has, have or had as well as another auxiliary verb i.e. is, was etc (Sinha 1993).

P: ਮੇਰੇ ਕੋਲ ਇਕ ਪਿਸਤੌਲ ਸੀ
T: *mērē kōl ik pistaul sī*

Here ਕੋਲ (*kōl*) is translated as "has" whereas ਸੀ (*sī*) is being translated as "is." "Has," "have" and "had" depends upon the tense and number. In this case if has is present any other auxiliary verb should be omitted in the post processing phase.

Post processor phase also checks if adverbs and postpositions together occur in a sentence then only adverbs are taken into account not postpositions. Those should be deleted in the post processor.

4.10.2. *Addition of* **"***has,***" "***have***" and "***had***"**
Has, have and had are important constituents of English verb phrase, these are added according to the rules, i.e. if the Punjabi phrase contains the modal operator ਚੁੱਕਾ (*cukkā*), ਚੁੱਕੀ (*cukkī*) and ਚੁੱਕੇ (*cukkē*) These operators shows the perfect nature of sentence, but in some cases, perfect nature of sentence can also be shown from the vowels present at the end of main verb. If the main verb or an operator ends with ਾ (*aa*), ੇ (*ee*) or ੀ (*ieh*) followed by auxiliary verb, then "has," "have" and "had" can be added in the phrase. E.g.

P: ਉਹ ਆਪਣਾ ਕੰਮ ਖਤਮ ਕਰ ਚੁੱਕਾ ਹੈ
T: *uh āpaṇā kamm khatam kar cukkā hai*
E: He has finished his work

P: ਮੈਂ ਆਪਣਾ ਕੰਮ ਕਰ ਲਿਆ ਸੀ
T: *maiṃ āpaṇā kamm kar liā sī*
E: I had done my work

P: ਮੈਂ ਪੁਲੀਸ ਨੂੰ ਸਾਰੀ ਕਹਾਣੀ ਸੁਣਾਈ
T: *maiṃ pulīs nūṃ sārī kahāṇī suṇāī*
E: I had told whole story to police

P: ਮੈਂ ਉਹਦਾ ਘਰ ਵੇਖਿਆ ਹੈ
T: *maiṃ uhdā ghar vēkhiā hai*
E: I had seen his house

For addition of these words, sentence must be searched for the number i.e. singular or plural of the noun phrase preceding the verb phrase and also the tense of the auxiliary verb given in the phrase.

In relation to these sentences, still there are certain divergences where addition of "has" and "have" does not depend on above rules. In such type of cases, the post processing is manual.

P: ਮੇਰੇ ਦੋ ਭਰਾ ਹਨ
T: *mērē dō bharā han*
E: I have two brothers

4.10.3. *Addition of infinitive*

Since 'to' infinitive is used in place for ਨੂੰ in post positional phrases, but in verb phrases, to infinitive need to be added, before the verb if the main verb ends with ਣ*(ṇ), ਣਾ(ṇā), or ਣੇ(ṇē).*

P: ਮੈਂ ਜਾਣਾ ਚਾਹੁੰਦਾ ਹਾਂ
T: *maiṃ jāṇā cāhundā hāṃ*
E: I want to go

Both perfect word and infinitive can be added, if the verb ends with ਨਾ (*nā*), ਨੇ (*nē*) and ਨੀ (*nī*) followed by an operator or auxiliary verb.

P: ਮੈਨੂੰ ਘਰ ਛੱਡਣਾ ਪਿਆ
T: *mainūṃ ghar chaḍḍṇā piā*
E: I had to leave home

5. TOOLS USED IN TRANSLATION

5.1 *The Punjabi morphological analyzer*

Morphological analysis is the identification of a stem-form from a full word-form. For example, the analyzer must be able to interpret the root form of ਮੁੰਡੇ (*muṇḍē*) as ਮੁੰਡਾ (*muṇḍā*) and the its GNP(Gender-Number-Person) information A Punjabi morph database developed at 'Advanced centre for technical development of Punjabi language, Literature and Culture' is being used for analyzing the exact grammatical structure of the word. The morph database used in the system includes the information about every word in Punjabi, with the information about its gender, number, person, case, tense etc. Every inflected word also contains the root word from where it is derived. The database contains more than one lakh words from which 63,000 are the inflected nouns which are derived from about 18,000 root nouns. The database contains the grammatical category of each word and also the inflected words it can form. From this database, the tagger gets the information and tags each word of the sentence. Since a morph analyzer has been developed at Punjabi University, Patiala and has been in use, the morph database of that analyzer has been adapted for the system and provision has been given for its upgradation.

5.2. *The Punjabi- English dictionary*

Dictionaries are the largest components of a MT system in terms of the amount of information they hold. These are more than simple word lists. The size and quality of the dictionary limits the scope and coverage of a system, and the quality of translation that can be expected. The dictionary in use contains the English equivalent of most the Punjabi words being used. The dictionary is combined with the morph database and used for the translation of words of Punjabi sentence. There are more than one lakh words in the dictionary and it is being upgraded while translator works, if a word or its meaning is not found, it is being asked from the user and entered in the database.

5.3. *Rule base*

The rule base is a database consisting of the structural transformation rules, ambiguity rules, phrase rules etc.

The knowledge base, which contains the rules for resolving the ambiguity of number of grammatical categories of words on the basis of type of surrounding words. About 150 rules have been analyzed by training about 1000 sentences which fix the grammatical category for a

particular word. It is very easy to add new rules to the database for ambiguity resolution.

Rules are also made for chunking, i.e. division into phrases, there are rules for noun phrase, adjective phrase, postpositional phrases and verb phrases. E.g.. The noun phrase rules contain the information, i.e. the first word can be noun (ਭਰਾ – *bharā*), the first can be personal pronoun and the second is noun (ਮੇਰਾ ਭਰਾ – *mērā bharā*), it may contain the modifier also (ਮੇਰਾ ਵੱਡਾ ਭਰਾ – *mērā vaḍḍā bharā*). Similarly, there are postpositional phrase rules as "personal pronoun + noun + postposition" as ਮੇਰੀ ਕਲਾਸ ਦੇ (*mērī kalās dē*). Different tables are used to store rules for noun, postpositional, adjective or verb phrases. Since these are recursive in nature, the number of rules is not very large, there are maximum of 50 rules for a particular type of phrase and depends upon the constituents of these phrases, but in some cases, priorities are set depending upon the type of sentences for which the system is being made. While division into phrases, rules of Punjabi are used and after translation, the words are combined to English phrases according to the phrase rules of English.

Sentence rules are also present in the knowledge base, which sense the phrases and according to the rules of target language combine the different phrases into the target language sentence.

It is very easy to add rules for different purposes, if a sentence is entered and no rule is found for that, a popup window will appear giving the information for the phrases or the words and ask for entry of rules in the corresponding database and by the intervention of the user, the data can be correctly added.

Since a format for a sentence or a phrase and even the types of words is not limited in an Indian language. For the system, all the rules needed for a particular type of simple sentences have been covered in the rule base.

6. Architecture of a Machine Translation System

This section outlines the overall architecture of the Punjabi to English MT system. The system is based on the transfer approach, with three main components: an analyzer, a transfer component, and a generation component. The **analysis component** which is composed of preprocessing, morph analyzer and tagger, phrase chunker and the module which adds semantic information, assigns morph information and karak roles to the input phrases by means of Punjabi grammatical rules. The **transfer component** which has translation and transliteration modules builds target

language equivalents of the source language grammatical structures by means of a comparative grammar that relates every source language representation to some corresponding target language representation. The combination of synthesis and post processing modules forms the **generation component** which provides the target language translation (Dave, Parikh & Bhattacharya 2001; Sinha 1996)

Fig. 1 shows the block diagram for the architecture of a Punjabi to English Machine Translation System. In the figure, the rectangle shows the step followed while translation and the oval shows the databases and knowledge bases used.

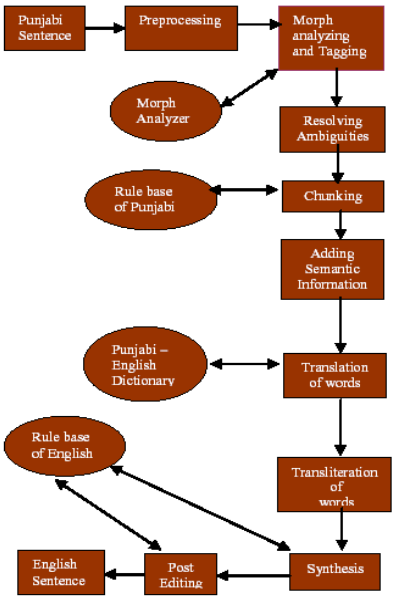


Fig 1. *Architecture of system*

7. EXAMPLE

Consider a Punjabi Sentence

P: ਮੈਂ ਫਿਰ ਅਪੀਲ ਕਰਦਾ ਹਾਂ

T: *maiṃ phir apīl karadā hāṃ*

After Tagging

ਮੈਂ (*maiṃ)* **(pp-m-f-s-d- - - -)** ਫਿਰ (*phir*) **(uad- - - - - -,v-b-s-s- -f- -x-)** ਅਪੀਲ **(**apīl) **(n-f-s- -d- - - -, n-f-s- -o- - - -)** ਕਰਦਾ (karadā) **(modal- - - - - - - -, cverb- - - - - - - -)** ਹਾਂ (hāṃ) **(av-b-f-p- -pr- - -, inter- - - - - - -)**

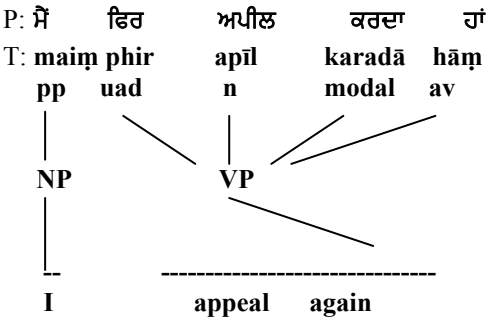
Here there are two tags for ਕਰਦਾ (*karda*) i.e. modal operator and conjunct verb, but according to the rules, it is considered as conjunct verb as the preceding word is noun. After resolving ambiguity, the tagged words are combined into phrases.

NP (ਮੈਂ (maiṃ) **(pp-m-f-s-d- - - -))**

VP(ਫਿਰ (*phir*) **(uad- - - - - -)** ਅਪੀਲ (apīl) **(n-f-s- -d- - - -)** ਕਰਦਾ (karadā) **(cverb- - - - - - - -)** ਹਾਂ (hāṃ) **(av-b-f-p- -pr- - -))**

After translating words

NP (I)
VP (appeal again)



8. TRAINING AND TESTING

After training the system with 1000 sentences, testing is performed with new 500 sentences and accuracy at different levels is calculated. The source for the legal sentences is the FIR's written in Punjabi. The other source for such sentences is the news of crime from Punjabi newspapers. While training the system, rules for ambiguity resolution, tagging, phrase chunking, sentence synthesis are manually added to the rule base after checking from the sentences, so that other same type of sentences should be translated correctly. The first phase which resolves

the ambiguity for different grammatical category and assigns tag to each word in a sentence was found to have approximately 75.54% accuracy. The accuracy is calculated by manually tagging the sentences and comparing with the results of tagged words. Similarly the accuracy of chunker is calculated and it turns about to be 86.57%, if the correct grammatical categories are feeded. For the evaluation of final translation, the intelligibility and accuracy metrics is used and evaluated for different scores and the accuracy is calculated to be 60.33%. The above said accuracy is calculated for score 2 of intelligibility test where sentences are considered generally clear and intelligible and for accuracy, it is considered that 50% of the original information passes in the translation.

9. Future Enhancements

The machine translation system being developed is limited to a particular format of a Punjabi sentence i.e. it should have a well defined subject part. It can be enhanced for other type of sentences. Moreover it can be made general purpose. The ambiguity resolution part, where a word may have more than one meaning can be solved by sensing its meaning according to the context which will increase the accuracy of the system to a great extent as there are number of words which cannot be defined for a particular domain.

References

Bhandari, V., Sinha, R. M. K. & Jain, A. 2002. Disambiguation of phrasal verb occurrence for machine translation. In proceedings of *Symposium on Translation Support Systems (STRANS2002)*, Kanpur, India, Mar 15-17.

——. & Sangal, R. 1993. Parsing free word order languages in the Paninian framework. In proceedings of *ACL 1993, New Jersey*. (pp. 105-111),

Chakrabarti, D. Rane, G. & Bhattacharyya, P. 2004. Creation of English and Hindi verb hierarchies and their application to English Hindi MT. *International Conference on Global Wordnet (GWC 04)*, Brno, Czeck Republic, Jan.

Dalal, A. Nagaraj, K. Sawant, U., Shelke, S. & Bhattacharyya, P. 2007. Building feature rich POS tagger for morphologically rich languages. *ICON 2007*, Hyderabad, India, Jan.

Dave, S. & Bhattacharya, P. 2002. Knowledge extraction from Hindi text.s *Journal of Institution of Electronic and Telecommunication Engineers*, 18/4, pages missing?

——., Parikh, J. & Bhattacharya, P. 2001. Interlingua-based English-Hindi machine translation and language divergence. *Machine Translation*, 16/4, 251-304.

Naskar, S. & Bandyopadhyay, S. 2005. Use of machine translation in India: Current status. In the proceedings of *MT SUMMIT X* (pp. 13-15), Phuket, Thailand, Sep.

Sharma, D. M. 2008. *Computational Paninian Grammar for Dependency Parsing*. Hyderabad: LTRC, IIIT. NLP Winter School 25-12-

Singh, S., Dalal, M. Vachani, V. Bhattacharya, P. & Damani, O. 2007. Hindi generation from interlingua. *Machine Translation Summit (MTS 07)*, Copenhagen, Sep.

——. 1989. A Sanskrit based word-expert model for machine translation among Indian languages. In proceedings of *Workshop on Computer Processing of Asian Languages* (pp. 82-91), Asian Institute of Technology, Bangkok, Thailand, Sep 26-28.

——. 1993. Correcting ill-formed Hindi sentences in machine translated output. In proceedings of *Natural Language Processing Pacific Rim Symposium (NLPRS'93)* (pp. 109-119), Fukuoka, Japan.

——. 1996. R & D on machine aided translation at IIT Kanpur: ANGLABHARTI and ANUBHARTI approaches. Invited paper at *Convention of Computer Society of India, (CSI'96)*, Banglore.

——. & Jain, A. 2003. AnglaHindi: An English to Hindi machine translation system. In *MT Summit IX*, New Orleans, USA, Sep 23-27.

——. & Jain, A. 2005. Divergence patterns in machine translation between Hindi and English. *10th Machine Translation Summit (MT Summit X)* (pp. 346-353), Phuket, Thailand, Sep 13-15.

——. & Thakur, A. 2004. Synthesizing verb form in English to Hindi translation: Case of mapping infinitive and gerund in English to Hindi. In proceedings of *International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS-2004)*, Tata Mc Graw Hill, New Delhi, Nov. 17-19.