Rule Based Machine Translation of Noun Phrases from Punjabi to English

Kamaljeet Kaur Batra[1] and G S Lehal[2]

[1]Dept. of Comp Sc. & IT, DAV College,
Amritsar, Punjab, India
kamaljit_batra@yahoo.com

[2]Dept of Comp Sc & Engg., Punjabi University,
Patiala, Punjab, India
gslehal@gmail.com

Abstract

The paper presents automatic translation of noun phrases from Punjabi to English using transfer approach. The system has analysis, translation and synthesis component. The steps involved are pre processing, tagging, ambiguity resolution, translation and synthesis of words in target language. The accuracy is calculated for each step and the overall accuracy of the system is calculated to be about 85% for a particular type of noun phrases.
Keywords: *Tagger, Ambiguity resolver, Transliteration*

1 Introduction

Machine Translation (MT), also known as "automatic translation" or "mechanical translation", is the name for computerized methods that automate all or part of the process of translating from one human language to another.[2] Machine Translation is the need of the hour. It helps in bridging the digital divide and is an important technology for globalization. The mechanization of translation has been one of humanity's oldest dreams. The work is done to convert a noun phrase from Punjabi to English.

2 Approach followed

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. The rules, which are used for the structural transformation of phrase, for solving the ambiguity problem, all are stored in the database. The indirect approach, first of all, divides a phrase into words, tags each word using morph database, resolves ambiguity, translates each word using bilingual dictionary, and then synthesize the translated words using rules of English language.

3 Steps followed for translation

3.1 Pre processing

Since the phrases are taken from number of sentences, there are different types of phrases, Pre processing module change the phrase to a particular format so that it can be translated with more accuracy. Eg System only works for simple noun phrases and if a phrase is either complex or compound, it is divided into two or more simple phrases. The structure of simple phrase is limited to a particular format. The above said part of Pre processor is manual and not automated.
The automated part of pre-processor performs the following tasks.

3.1.1 Identifying Collocations

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. Eg ਪਿਤਾ ਜੀ (pita ji), ਮਾਤਾ ਜੀ (mata ji), these words have a single equivalent as father and mother.

3.1.2 Identifying Named Entities

In certain cases named entities can be recognized by their preceeding words which can be ਸ੍ਰੀ, ਸਰਦਾਰ, ਸਰਦਾਰਨੀ, ਸ੍ਰੀਮਤੀ, ਕੁਮਾਰੀ in the input phrase.

ਸ਼੍ਰੀ ਰਮੇਸ਼ ਚਾਵਲਾ (shri ramesh chawla), ਸਰਦਾਰ ਹਰਪ੍ਰੀਤ ਸਿੰਘ (sardar harpreet singh)These named entities will then be send to transliteration module.

3.2 Tokenization

The output of pre processor is then send to the tokenizer which divides the given phrase on the basis of spaces between them into constituents called tokens which are then passed to further phases.

3.3 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 for noun phrase include noun, pronoun, adjective, preposition, conjunction 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. The fields not applicable to a particular category are left blank. E.g. Tags for the word 'ਭਰਾ'(Bhra) are 'n-m- -s-d', 'n-m- -p-d'. The above tag for the word shows that it can be used as noun with masculine gender, singular as well as plural and in direct case. The complete information for the tags is available from the morph database. In Punjabi, 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 Case information to it. [6,7] For example, in case of nouns person information is not in use whereas for personal pronouns person information is used.

3.4 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 example a word present in a noun phrase of Punjabi can be tagged with a noun as well as an adjective tag. For this purpose, there is need to apply certain rules depending upon the grammatical category of preceding or succeeding words. These rules should be prioritized.
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.
Eg. Consider the two noun phrases ਜਵਾਨ ਮੁੰਡਾ (javan munda) and the phrase ਸਾਰੇ ਜਵਾਨ (sarey javan). In the first phrase, 'ਜਵਾਨ' is an adjective followed by a noun and its English equivalent is 'young' whereas in the second phrase, it is a noun preceded by an adjective which should be translated as 'soldier'.

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 tags for the word ਬੰਦੇ(bandey) are 'n-m- -s-o' and 'n-m- -p-d'. The tagged word can be noun in singular or a noun in plural. Eg. In the phrase, ਬਹੁਤ ਸਾਰੇ ਬੰਦੇ (bahut sarey bandey). In this case we should select the tag 'n-m- -p-d' and its appropriate word in English is men,whereas in the phrase ਮੋਟੇ ਬੰਦੇ ਨੇ (mote bandey ne), the tag for ਬੰਦੇ(bandey)should be 'n-m- -s-o' and its appropriate meaning is man. Such type of ambiguity can be resolved by considering the number ie. Singular or plural of the sentence in which the phrase should be used. Similarly the ambiguity related with the number and gender for demonstrative pronouns is resolved by considering the gender and number for the sentence.

3.5 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. There are certain words used in Punjabi language which are of English origin,as 'ਸਕੂਲ', 'ਟੀਚਰ', 'ਡਾਕਟਰ' etc. Such words should be written as it is.

3.6 Transliteration of Proper nouns

While translating each word using the dictionary, there are certain out of vocabulary words such as names of persons, names of cities etc., these all are proper nouns, and these should be passed to the transliteration module. Also there are certain words which are recognised at the preprocessing phase as names of persons, those should also be transliterated. Transliteration means to write them sensing the characters in the words e.g. 'ਮਨਜੀਤ' 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 uses a database of transliterating characters and also certain rules to insert vowels wherever needed.

3.7 Synthesis

After getting English equivalent of each word in Punjabi sentence, it should be synthesized to the phrase in English. Since the order of occurrence of words is different in target language than the source language, the approach used while synthesis is indirect approach, so certain rules have been build to synthesize the phrases to target language. These rules of language are also stored in the rule base of English.

4 Tools used in Translation

4.1 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" as "ਮੁੰਡਾ" and the its GNP(Gender-Number-Person) information A Punjabi morph analyzer developed at 'Advanced centre for technical development of Punjabi language' 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 tag each word of the phrase.

4.2 The Punjabi- English Dictionary

Dictionaries are the largest components of a MT system in terms of the amount of information they hold. If they 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 contains the English equivalent of all the Punjabi words. The dictionary is combined with the morph database and used for the translation of words of Punjabi Phrase. There are more than one lac words in the dictionary and it is being upgraded.

4.3 Rule Base

The rule base is a database consisting of the structural transformation rules, ambiguity rules, phrase rules etc. The knowledge base contains the rules for resolving the ambiguity of number of grammatical categories of words on the basis of type of surrounding words. Rules, not only check the grammatical category, but also number, gender or person in some cases. Rule base also contains the information about its synthesis, that while it is of same order or different. All the rules in the database are arranged according to priority. Phrase Rules are represented as context free grammar. Since these are recursive in nature, the number of rules is not very large, but in some cases, priorities are set depending upon the type of phrases for which the system is being made.

5 Architecture of a Machine Translation System

This section outlines the overall architecture of the Punjabi to English MT system for noun phrases. 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 assigns tags to the input phrases by means of Punjabi grammatical rules. The transfer component 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 generation component which provides the target language translation.[2,13]

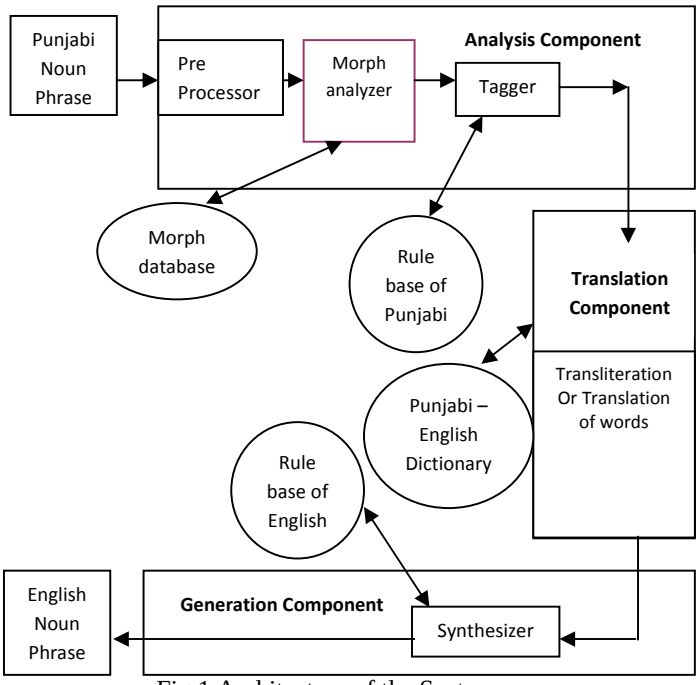


Fig 1 Architecture of the System

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.

6 Example

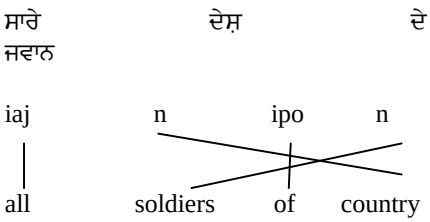
Consider a Punjabi Noun Phrase

ਸਾਰੇ ਦੇਸ਼ ਦੇ ਜਵਾਨ

After Tagging

ਸਾਰੇ (iaj-m- - -) ਦੇਸ਼(n-m-s- -d,n-m-p-d) ਦੇ(ipo- - - -) ਜਵਾਨ(n-m-s- d-, n-m-p- -d,iaj-b- - -)

Here there are two tags for ਜਵਾਨ ie inflected adjective and noun, but according to the rules, it is considered as noun with plural as there is no succeeding noun and the adjective signifies the plural. After resolving ambiguity, the tagged words are the translated and combined into target phrase.



7 Training and Testing

After training the system with about 2000 phrases, testing is performed with new 500 sentences and and accuracy at different levels are calculated. 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. Overall accuracy of translation is 85.33%. In case of translation, the output phrase is considered correct, even if the translated equivalent may not be grammatically very correct, but signifies the true meaning of the Punjabi phrase.

References

[1] R.M.K. Sinha and Ajay Jain, AnglaHindi:An English to Hindi Machine Translation System, MT Summit IX, New Orleans, USA, Sept.23-27, 2003.

[2] S. Dave, J. Parikh and P. Bhattacharyaa. Interlingua-based English-Hindi Machine Translation and Language Divergence. Machine Translation 16(4) (2001) 251-304.

[3] R.M.K. Sinha and Anil Thakur, Divergence Patterns in Machine Translation between Hindi and English, 10th Machine Translation summit (MT Summit X), Phuket, Thailand, September 13-15, (2005), 346-353.

[4] Aniket Dalal, Kumara Nagaraj, Uma Sawant,Sandeep Shelke and Pushpak Bhattacharyya, Building Feature Rich POS Tagger for Morphologically Rich Languages, ICON 2007, Hyderabad, India, Jan, 2007.

[5]Akshar Bharati, Vineet Chaitanya, Amba P. Kulkarni, Rajeev Sangal Anusaaraka: Overcoming the Language Barrier in India. (informal publication) Electronic Edition (link) BibTeX [cs.CL/0308018]

[6] Computational Paninian Grammar for Dependency Parsing Dipti Misra Sharma,LTRC, IIIT,Hyderabad, NLP Winter School 25-12-2008

[7] Akshar Bharati, Rajeev Sangal: Parsing Free Word Order Languages in the Paninian Framework. ACL 1993: 105-111

[8] Akshar Bharati, Rajeev Sangal: A Karaka Based Approach to Parsing of Indian Languages. COLING 1990: 25-29

[9] R M K Sinha, Some thoughts on computer processing of natural Hindi.. Annual convention of Computer Society of India, 1978, pp 151-165.

[10] Shachi Dave and P Bhattacharya – Knowledge Extraction from Hindi Text, Journal of institution of Electronic and Telecommunication Engineers Vol.18, No.4 July 2002.

[11] Vartika Bhandari, R M K Sinha and Ajai Jain, Disambiguation of Phrasal Verb Occurrence for Machine Translation, Proc. Symposium on Translation Support Systems (STRANS2002), Kanpur, India, March 15-17, 2002.

[12] R M K Sinha, ‘A Sanskrit based Word-expert model for machine translation among Indian languages., Proc of workshop on Computer Processing of Asian Languages, Asian Institute of Technology, Bangkok, Thailand, Sept.26-28, 1989, pp 82-91.

[13] R M K Sinha, 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, 1996.

[14] R M K Sinha, Correcting ill-formed Hindi sentences in machine translated output. Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS.93), Fukuoka, Japan, 1993, pp 109-119.

[15] R Jain, R M K Sinha, A Jain, Translation between English and Indian Languages, Journal of Computer Science and Informatics, March 1997, pp 19-25.

[16] R M K Sinha and others, ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi., IEEE International Conference on systems, Man and Cybernetics,Vancouver, Canada, 1995, pp 1609-1614.

[17] R.M.K. Sinha and Anil Thakur, Synthesizing Verb Form in English to Hindi Translation: Case of mapping Infinitive and Gerund in English to Hindi, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System(iSTRANS-2004), November 17- 19,2004, Tata Mc Graw Hill, New Delhi, pp:52-55

[18] Smriti Singh, Mrugank Dalal, Vishal Vachani, Pushpak Bhattacharya and Om Damani, Hindi Generation from Interlingua, Machine Translation Summit (MTS 07), Copenhagen, September, 2007.

[19] Debasri Chakrabarti, Gajanan Rane and Pushpak Bhattacharyya, Creation of English and Hindi Verb Hierarchies and their Application to English Hindi MT, International Conference on Global Wordnet (GWC 04), Brno, Czeck Republic, January, 2004.