Rule Based Machine Translation of Noun Phrases from Punjabi to English

Kamaljeet Kaur Batra and G S Lehal

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

2Dept 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 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.

3.1.2 Identifying Named Entities

In certain cases named entities can be recognized by their preceeding words which can be माता, भाई, बहू, पिता in the input phrase.
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 'जवन' (javan) are 'n-m- -s-o' and 'n-m- -p-d'. The tagged word can be noun in singular or a noun in plural. In the phrase, जवन मने बने (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 मोटे बने (mote bandey ne), the tag for बने (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.

3.4 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 जवन मने बने (javan munda) and the phrase जवन मने सरे (sarey javan). In the first phrase, ‘जवन’ 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 बने (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, बहुत मने बने (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 मोटे बने (mote bandey ne), the tag for बने (bandey) 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 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 ‘स्कूल’, ‘टीचर’, ‘फ्यूटर’ 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. ‘मन्जीत’ in Punjabi is transliterated in English as ‘manjeet’, m for म, n for न, j for ज, ‘ए’ for ए, ‘ट’ for ट 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 “ਸਫ਼ਲਾਂਤ” as “ਸਫ਼ਲਾ” 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 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

मरे देस में सैनिक

After Tagging

मरे (iaj-m- - -) देस (n-m-s- d-, n-m-p-d) सैनिक (n-m-s- d-, n-m-p- -d, iaj-b- - -)

Here there are two tags for सैनिक 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 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


