

Syntactical Translation System for English to Sindhi Translation

Saeed Ahmed, Shah Zaman and Qurat-ul-Ain

saeedtalpur@gmail.com, shahzamannizamani@yahoo.com and agro@usindh.edu.pk

Institute of Mathematics & Computer Science, University of Sindh, Jamshoro, Pakistan

Abstract: Machine translation is a field concerned with translation of one Natural language to another. The task involves Natural Language Processing techniques, a potential area of Artificial Intelligence. The research done in this field up till now was targeted for European languages only. There is very little evidence of such work in Arabian Script languages particularly in English to Sindhi Translation. The aim of present research is to present a system for English to Sindhi translation. English and Sindhi are entirely different languages in terms of their structure, set of alphabets, and writing styles. A number of problems were addressed successfully in this research to generate a basic English to Sindhi translator.

1. INTRODUCTION

1.1 Introduction to Language Translation

Translation can be traced to conversations and correspondence between Andrew D. Booth, a British crystallographer, and Warren Weaver of the Rockefeller Foundation in 1947, and more specifically to a memorandum written by Weaver in 1949 to the Rockefeller Foundation [1].

At the time of writing, the use of Language Translation or indeed, any sort of computerized tool for translation support is completely unknown to the vast majority of individuals and organizations in the world, even those involved in the so called 'language industries', like translators, terminologists, technical writers, etc. Human Translators actually deploy at least five distinct kinds of knowledge:

- Knowledge of the source language.
- Knowledge of the target language. This allows them to produce texts that are acceptable in the target language.
- Knowledge of various correspondences between source language and target language (at the simplest level, this is knowledge of how individual words can be translated).
- Knowledge of the subject matter, including ordinary general knowledge and 'common sense'.
- This, along with knowledge of the source language, allows them to understand what the text to be translated means.
- Knowledge of the culture, social conventions, customs, and expectations, etc. of the speakers of the source and target languages.

1.2 Dissimilarities Between English and Sindhi

Sindhi language is a vast language, so it is difficult to describe all characteristics of this language. Given below are few important characteristics which are important to make the translation..

1. Sindhi language is written from right hand side while English from left hand side.

Example

This is English. هي سنڌي آهي.

2. In Sindhi language the auxiliaries such as ٿو، ٿا، آهي، ٿن appear at the end of the sentence, while in English auxiliaries appear in middle of the sentence.

Example

This is my book. هي مونهنجو ڪتاب آهي..

3. In Sindhi language verbs come after the object of the sentence, while in English language verbs come before the object of the sentence.

Example

I am eating. آئون کائي رهيو آهيان.

4. In Sindhi language preposition comes after the object, while in the English language object comes after preposition .

Example

I am going to School. آئون اسڪول ڏانهن وڃي رهيو آهيان.

5. If an English sentence contains a verb, preposition and adverb then the Sindhi sentence will have adverb first followed by preposition and then verb.

2. DESIGN & SPECIFICATION

2.1 Translation of verbs in Sindhi

In the English language there are five possible forms of the verb. For instance the verb GO has five forms i-e GO, WENT, GONE, GOES, GOING, in same way there

are more than 59 possible forms of verb in the Sindhi language [2], [3]. In order to make the translation of sentence from English to Sindhi or vice versa, all possible meanings of a verb (for Sindhi) in the database had to be added. In order to reduce the size of database a new methodology is developed. For each verb some letters are put in the database, and by adding some letters at the end of these letters, the required meaning of the verb (in Sindhi) can be achieved. Different letters have to be added according to different tenses and different forms of verbs, refer table1, table2 and table 3.

Table 1: Characters required for the present simple, continuous and perfect continuous tenses.

Word	I	We	He	She	It
Add	ان	ون	ي	ي	ي
Word	You 1	You 2	They	F Name	M Name
Add	ين	و	ن	ي	ي

Table 2. Characters required for the past simple tense.

Word	I	We	He	She	It
Add	س	اسين	و	ي	و
Word	You 1	You 2	They	F Name	M Name
Add	ن	ا	ا	ي	و

Table 3. Characters required for the present perfect, past perfect, past continuous and past perfect continuous.

Word	I	We	He	She	It
Add	و	ا	و	ي	و
Word	You 1	You 2	They	F Name	M Name
Add	و	ا	ا	ي	و

Secondly the meaning of pronoun is also affected in Sindhi according to various situations but they remain same in English.

e.g.

I went to School آئون اسڪول ويس
I ate mango مون انب کاڌو

It can be seen that the meaning of I is آئون in first sentence and مون in the second sentence.

In the remaining tenses it is required to observe whether the subject is singular or plural and the second consideration is to check the gender of the subject.

2.2 Rules for Sindhi Sentences

One of the important tasks for translation was to define structural rules for both languages. English language rules were defined by James H. Martin and Davis Jurasky [1], but there were no rules available for Sindhi language. A successful attempt was made in this regard and rules were defined for Sindhi language. The current research also aims to define the structure of the Sindhi language. In order to make translation, it is required to change English sentence according to the sequence of Sindhi

sentence. Table 4 contains rules for English and Sindhi sentences.

Table 4. English and Sindhi sentence rules

English Sentence Rules	Sindhi Sentence Rules
NP VP	NP VP
NP AUX VP	NP VP AUX
WhNP NP AUX VP	NP WhNP VP AUX
WhNP NP VP	WhNP NP VP
WhNP VP	WhNP VP
WhNP NP AUX	NP WhNP AUX
WhNP VP AUX	WhNP VP AUX
WhNP AUX VP	WhNP AUX VP
WhNP NP AUX PP	PP NP WhNP AUX
Verb NP	NP Verb
Verb NP PP	NP PP Verb
Verb NP PP1 PP2	PP2 PP1 NP Verb
Verb PP	PP Verb
NP PP	PP NP
Verb Adverb	Adverb Verb
Verb NP Adverb	Adverb NP Verb
English Sentence Rules	Sindhi Sentence Rules
Verb PP Adverb	Adverb PP Verb
Verb NP PP Adverb	Adverb PP NP Verb
Verb NP1 NP2	NP2 NP1 Verb
Verb NP1 NP2 NP3	NP3 NP2 NP1 Verb
Verb NP1 NP2 PP	PP NP2 NP1 Verb
Verb PP1 PP2	PP2 PP1 Verb
Verb PP1 PP2 PP3	PP3 PP2 PP1 Verb
Verb PP NP	NP PP Verb
Verb NP1 PP NP2	NP2 PP NP1 Verb
Adverb Adjective	Adjective Adverb

2.3 Degree of Adjectives for Sindhi

The meaning of adjective in Sindhi language changes according to the gender of the object. The meaning further changes according to the singularity or plurality of the object. In order to put all the possible meanings a large database was required. The problem is addressed by adding some letters at the end of each word to complete the meaning of an adjective according to the need of sentence. These letters are given in table 5.

Table 5. Characters for adjectives

	Singular	Plural
Feminine	ي	يون
Masculine	و	ا

When the object is singular feminine then character “ي” is added and “و” when it is singular masculine. When the object is plural feminine the characters “يون” are added, and character “ا” is added when it is plural masculine.

There are three degrees of an adjective base, comparative and superlative. For instance hard, harder, hardest. For getting the meaning of all forms of adjective, only the

meaning of base form of adjective is required. For the meaning of comparative adjective only “وڌيڪ” is added before the meaning of base form of adjective, and in the same way by adding word “تمام”, meaning of superlative adjective is achieved.

Example	Hard سخت	Harder وڌيڪ سخت	Hardest تمام سخت
	Big وڏو	Bigger وڌيڪ وڏو	Biggest تمام وڏو

2.4 Translation of Numbers Written in Words

The numbers in English, which are written in words, is a tedious task to examine and convert to Sindhi. An algorithm is designed for this conversion, which nicely converts even very long numbers into numerical format, and then with the help of database, the numerical value is converted to words in Sindhi.

Usually, the largest value in English is billion. The values larger than Billion are important for scientists and not for common man. So billion is supposed to be the largest word used in daily transactions.

The algorithm follows the following steps in order to convert a number made up of words into the corresponding numerical value:

1. Recognize a combination of continues words to be a number from the input collection of words representing a sentence.
2. Take the first word from the stream. If the word is one of the following:
 - “one”, “two”, ..., “ninety”, then
 - i) If the current word is the last word then assign the word’s value to the “temp” variable.
 - ii) If the current word is not the last one, add the word’s value to the “temp” variable.
3. If the word is “hundred” then
 - i) If it is the last word in stream, assign the word’s value to the temp variable
 - ii) Else, multiply the word’s value with the “temp” variable
4. if the word is “thousand”, “million”, “billion”, ..., then multiply “temp”’s value to the word, and put the value in appropriate variable (“thousand”, “million”, etc.). Assign the value Zero to the “temp” variable
5. Take the next word out from the array of words of the number, and repeat steps 2 to 4 for it.
6. Reaching the end, the final numerical value of the number can be calculated as:

$$\text{Number} = \text{Billion} + \text{Million} + \text{Thousand} + \text{temp}$$

In general the algorithm says that, if any word has a value larger than the previous word, multiply both. Otherwise if any word has a value smaller than the previous word, add the word to the total.

After getting the value of the number in words, it is translated into words in Sindhi. This is done by first dividing the number into its component parts. Sindhi has a unit system different from that of English. It uses multiples of hundred for describing the units, in contrast with English that uses multiples of thousand for describing the units. Sindhi’s units are given in table 6:

Table 6. Unit system for Sindhi

English	Value	Sindhi
Hundred	100	Sow
Thousand	1,000	Hazar
Hundred Thousand	1,00,000	Lac
Ten Million	1,00,00,000	Crone
Billion	1,00,00,00,000	Arb

First, the input numerical value is converted into its Sindhi component parts. Then with the help of database, the proper names of units are found from database and values and then they are put in proper order, to get the number in words of Sindhi.

3. CONCLUSION AND FUTURE WORK

A language translation system has been developed on a very basic level that will help the developers to get in the world of language translation. Researchers have tried to cover the three levels of grammar tenses (Present, Continuous & Past and their four sub tenses (Indefinite, Continuous, Perfect & Perfect Continuous).

The present research successfully defines rules for Sindhi language through which structure of Sindhi sentence can be identified.

The translation of sentences that contain the model auxiliaries, for example can, and would and so on is also achieved.

This project handles the problem of proper noun. It takes proper nouns and automatically translates them in Sindhi word, without the help of database.

The present work in no way can be termed as complete and fully operational. A number of enhancements and refinements are possible. During the research more stress was on functionality therefore the database is limited and should be enhanced in future.

In the process of translation there were many ambiguities, few of them have been resolved, others still have to be addressed, e.g Semantic rules are not applied; these rules should be implemented in future in order to increase the accuracy of translation.

The project is not targeted for complex sentences and the sentences that contain idioms. It is recommended for the future researchers to work on the complex sentences and idioms.

As this project has limited features, so the error detection from the sentence is not achieved. It is also recommended for the future researchers to work over this.

ACKNOWLEDGEMENT

All praises and gratitude to Almighty Allah for His Benevolence. After this, we wish to express our sincere appreciation to **Mr. Aslam Pervaiz Memon** for his help and support in this research. The completion of this research was made possible only by **Miss. Qurat-ul-Ain Agro**, and her dedicated and committed supervision. She was available to us at all times solving our problems and guiding to us.

The amount of support and affection from our families is uncountable. Last but not the least; we would like to thank all faculty members of Institute of Mathematics and

Computer Science, University of Sindh for their Cooperation.

REFERENCES

- [1] John Hutchins. In proceedings to MT Summit VI: past, present, future, pp. 14-23. San Diego, California, USA, 29 October - 1 November 1997.
- [2] Moti Lal, Naoun Sindhi Grammar, Edition 2nd, National Printing Press Sukkur Sindh, 1979.
- [3] John E. Warriner, English Composition and Grammar, Edition 1st, Holt Rinehart and Winston.
- [4] James H. Martin, Speech & language processing, First edition, Prentice Hall, 2000.