Journal of Information & Communication Technology
Vol. 2, No. 1, (Spring 2008) 31-39

Aghaz[1]: An Expert System Based Approach for the Translation of Urdu into English.

Uzair Muhammad*
Atif Khan*
M. Nasir Khan*
Kashif Ayyub*
Muhammad Sharif*
COMSATS Institute of Information Technology Wah Cantt, Pakistan

ABSTRACT

There is very little evidence of work in Arabian Script languages particularly in Urdu to English Translation. English and Urdu are entirely different languages in terms of their structure and writing styles. This paper represents a direct approach for using an expert system to translate a text in Urdu language into its equivalent in English Language.

INSPEC Classification : C6150C; C6170; C7820M

Keywords : Translation, Urdu, English, Parsing, Nouns, Expert Systems.

1) INTRODUCTION

In the modern world, there is an increased need for language translation. The idea of using computers to translate or help translate human languages is almost as old as the computer itself (Trujilo, 1999). Many achievements have been made in this field. The availability of "translation engines" on the Internet allows for real-time translation of arbitrary text, and even entire web sites. The Google language bar (Z.Pervez,.et.al) and AltaVista Babelfish (Google.com/language_tools) are one of the examples of the machine translation systems that are available freely on the internet.

MT for Urdu to English can play an important role in Pakistani region where a lot of population is Urdu speaking and unaware of English language (T. RAHMAN (2002), T. MITAMURA(2002), Z. PERVEZ.

* The material presented by the authors does not necessarily portray the viewpoint of the editors and the management of the Institute of Business and Technology (BIZTEK) or COMSATS Institute of Information Technology Islamabad, Pakistan.

* Uzair Muhammad : joinuzair@yahoo.com
* Atif Khan : atif_ciit@yahoo.com
* M. Nasir Khan : m_nasir_khan@yahoo.com
* Kashif Ayyub : kashifayyub@hotmail.com
* Muhammad Sharif : muhammadsharifmalik@yahoo.com

This paper is based on the AGHAZ (UZAIR MUHAMMAD, KASHIF BILAL) Machine Translation System which is a translator for Urdu to English Language. The System is efficient case of time and space (UZAIR MUHAMMAD, KASHIF BILAL). "It reserves all the retrieved information (i.e. Meaning, Part of Speech, Different Bits) in cache (temporary storage), which is used in "best search", which selects the best word's solution and drop all other solutions, and finally in Translation phase."(UZAIR MUHAMMAD, KASHIF BILAL)

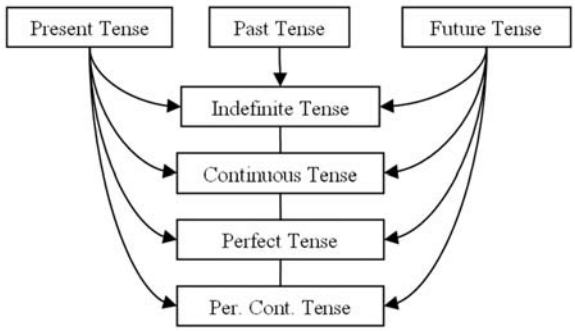
2) EXTRACTING DIFFERENT INFORMATION FROM INPUT SENTENCE

A. TENSE INFORMATION

There are three main tenses i.e. Present, Past, Future each of them has four sub tenses. One can be identified by the form of verb and helping verb etc that is used in the sentence. These parameters are very helpful for computer calculations. AGHAZ uses an Algorithm to find the tense and provide its equivalent in English.

First of all it finds the base tense by checking the last word of sentence and then goes for the sub category as shown in figure below.

Figure 1
Tense wise traversing



B. PATTERNS TO FIND THE TENSE

As we discussed earlier that last word of the sentence determine the main tense i.e. present, past or future. The second last word or some part of that word points to sub tense. English and Urdu grammars are totally incompatible. One of the differences is that of gender. In English there is no difference of gender. For example when we use pronoun "He" or "She" it has no impact on the rest of sentence only "He" is replaced with "She". Similarly names like "Asad" and "Taniya" both are treated as same in a sentence, Even in English we use "I" to point ourselves, we don't think about our gender while in Urdu it is compulsory to show our gender even in case of "I" is used as subject. Similarly "Dog" is masculine and "Cat" is feminine.

Figure 2
Example

English	**Dog is running**	**Cat is running**
Urdu	کتا دوڑ تا ھے	بلی دو ڑ تی ھے

i. Present Tense

All sentences in the present tense ends with ھو,ھوں,ھیں,ھے , See some examples in table 2 ;

Note in perfect tense sometimes یا is also used. For example; لیا ھے ,گیا ھے etc.

ii. Past Tense

Last word in the past tense sentences may be تھیں ,تھے,تھا,تھی. Please see some examples in table 3;

iii. Future Tense

Future tense uses words گا,گی,گے to terminate a sentence. Some examples are tabled in table 4;

N o t e : ھوں/ھو is used after Sub Tense in Continuous, Perfect and Perfect Continuous Tenses.

C. Patterns to find sub tense.

From the above discussion we successfully found the Main Tense token. If we see the above tables 2, 3 and 4 carefully we can find that in most cases the Sub Tense tags remain same. Some changes are encountered only in case of present perfect tense, past indefinite and future indefinite, while an additional tag is added to the rest of future tense. From Sub Tense part we get three information, first, sub type of tense and, second, gender and, third, singular/plural.

All possible sub Tense Patterns;

Table 1
Tense Sub category

Status	Sub Tense			
	Per. Cont.	Per.	Cont.	Ind.
Masc.	تا رھا	چکا	رھا	ے/ؤں/یں و/ؤ/تا
Fem.	تی رھی	چکی	رھی	تی
Plural	تے رھے	چکے	رھے	ے/ئیں/

These are the patterns that help us finding sub category of tense. These are same for all except some cases;

i. Future Tense:

ھوں/ھو is used after Sub Tense in the Future's Continuous, Perfect and Perfect Continuous Tenses.

ii. Present Perfect vs. Past Indefinite Tense

In Past Indefinite tense and present perfect tense, sometimes, a sentence can be terminated with the word that has یا in its last. This tag is concatenated in the last of verb.

Table 2

Present Tense

Main Tense	Gen.	Sub Tense				Verb	Subject
		Per. Cont.	Per.	Cont.	Ind.		
ہے	M	تا رہا	چکا	رہا	تا	جا/کھیل/پڑھ	وہ (He/She)
	F	تی رہی	چکی	رہی	تی		
ہوں	M	تا رہا	چکا	رہا	تا	جا/کھیل/پڑھ	میں (I)
	F	تی رہی	چکی	رہی	تی		
ہیں	M/F	تے رہے	چکے	رہے	تے	جا/کھیل/پڑھ	وہ/ہم (They/We)
ہو	M	تے رہے	چکے	رہے	تے	جا/کھیل/پڑھ	تم (You)
	F	تی رہی	چکی	رہی	تی		

Table 3

Past Tense

Main Tense	Gen.	Sub Tense				Verb	Subject
		Per. Cont.	Per.	Cont.	Ind.		
تھا/تھی	M	تا رہا	چکا	رہا	تا	جا/کھیل/پڑھ	وہ (He/She)
	F	تی رہی	چکی	رہی	تی		
تھا/تھی	M	تا رہا	چکا	رہا	تا	جا/کھیل/پڑھ	میں (I)
	F	تی رہی	چکی	رہی	تی		
تھے	M/F	تے رہے	چکے	رہے	تے	جا/کھیل/پڑھ	وہ/ہم (They/We)
تھیں/تھے	M	تے رہے	چکے	رہے	تے	جا/کھیل/پڑھ	تم (You)
	F	تی رہی	چکی	رہی	تی		

Table 4

Future Tense

Main Tense	Gen.	Sub Tense				Verb	Subject
		Per. Cont.	Per.	Cont.	Ind.		
گا/گی	M	تا رہا	چکا	رہا	ئے	جا/کھیل/پڑھ	وہ (He/She)
	F	تی رہی	چکی	رہی			
گا/گی	M	تا رہا	چکا	رہا	ؤں	جا/کھیل/پڑھ	میں (I)
	F	تی رہی	چکی	رہی			
گے	M/F	تے رہے	چکے	رہے	ئیں/یں	جا/کھیل/پڑھ	وہ/ہم (They/We)
گے/گی	M	تے رہے	چکے	رہے	و/ؤ/ئیں	جا/کھیل/پڑھ	تم/آپ (You)
	F	تی رہی	چکی	رہی			

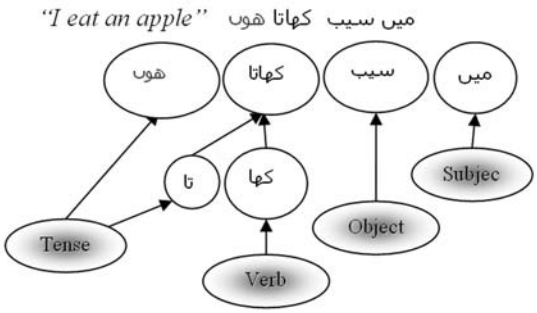
In Present perfect tense there it is used as Sub Tense and Main Tense is used to show present. When there is no Main Tense word exists and same case is encountered then we consider it as past indefinite tense.

D. VERB

In English verb is written as a single word while in Urdu, in most cases, verb is concatenated with other word.

Consider the following example;

Figure 3
verb concatenation



In this example the word that is concatenated with the verb (i.e. تا), in this case, provide us three information, first, subject is singular, second, subject is masculine and, third, tense is indefinite but the last word of sentence shows that this is present tense, so the tense is now present indefinite.

1. Sub tense how it is eliminated from verb.
2. Search from verb.

E. SUBJECT

Subject extraction is somewhat not very much difficult job. The complexity comes only when it is to be decided that whether subject is masculine or feminine, singular or plural etc. As a subject we encounter Pronouns or Nouns mostly.

i. Pronoun

In Pronouns we have some problems like for "He", "She" and "They" we have single word "وہ" in Urdu. When "وہ" is encountered in sentence one can never make a decision that which one of the three English words will replace it. In Urdu where some extra tokens are used, with verb, create problems, there, it also provides some solutions. This problem is also solved by these tokens. Consider an example in present indefinite tense;

Table 5
Pronoun Example Sentences

English	Urdu
He reads	وہ پڑھتا ھے
She reads	وہ پڑھتی ھے
They read	وہ پڑھتے ھیں

In every tense these information are available. So this problem is solved by considering the tokens just after the verb.

ii. Noun

When noun is used as subject nothing is efficient enough to handle it without a mature and complete knowledge base, because one has to extract different information like gender

and duality etc. In English "cats" is the plural of "cat" while in Urdu there are two versions used; sometime we use "بلیاں" and sometimes "بلیوں" depending of situation. So both the entries must be present in the knowledge base. If only "بلیاں" is stored in knowledge base then "بلیوں" is a new token that is unknown. In Urdu maximum plural words have two versions.

3) TRAVERSING DICTIONARY

Dictionary traversal methods are very much efficient in AGHAZ system. Also In this section we encounter some problems like in present indefinite and past indefinite when mode of sentence is changed to negative or interrogative then form of verb is also changed.

Table 6
Present Indefinite Vs Past Indefinite

Tense	Subject	Normal (Positive)	Neg./Inte.
Present Indefinite	I, We, You, They etc.	Base Form (1st form)	Base Form
	He, She, It etc.	Present Singular (s/es)	Base Form
Past Indefinite	I, We, You, They etc.	Past Tense (2nd form)	Base Form
	He, She, It etc.	Past Tense (2nd form)	Base Form

From the above table it is mandatory that there must be relationship between different forms of verbs. For example consider an example from past indefinite tense;

Table 7
Example Past Indefinite tense

Sentence Mode	English	Urdu
Positive	He **went**	وہ گیا
Negative	He did not **go**	وہ نہیں گیا
Interrogative	Did he **go**	کیا وہ گیا

There should be a mechanism in the Dictionary to pick proper form of verb. For example consider different forms of eat and go;

Table 8
Examples (eat and go)

Sr.	Form of Verb	English	Urdu	Implem-entation
1.	Base Form	eat	کھا	کھاتا
2.	Past Tense	ate	کھا	کھایا
3.	Past Participle	eaten	کھا	کھا چکا
4.	Base Form	go	جا	جاتا
5.	Past Tense	went	گیِ , گ	گیا , گئ
6.	Past Participle	gone	جا	جاچکا

In the above example single pattern is used in Urdu against all the three forms of verb. One can get any form of verb by comparing the pattern. While for "go" different patterns are used. For example in case 5 there are two patterns used for went. So in this case one can never pick go against "گیا" . It requires different forms to be interrelated.

To handle this case in AGHAZ some small dictionaries are used before the main dictionary is lookup. Changes reflected only in subject case of nouns, pronouns and some cases in different form of verbs so these are entered in the prerequisite of main knowledge base while proper nouns, adjectives and adverbs are remain same so these are laid in the main dictionary.

4)AGHAZ URDU INTO ENGLISH ARCHITECTURE AND WORKING

Figure 4 depicts the architecture and flow of AGHAZ System.

A. Sentence Parser: This module Takes the Urdu paragraph and parses it into sentences.

B. Tagger: This module put the tags to the sentences from previous module i.e the sentence parser module. The tag can only be a of one character or it can be one complete word.

C. Translator: Translator consists of many different sub-modules. The translator translates the tagged sentences into English like words. Following is the working of the sub-modules.

i. Extracts the helping words in Urdu like **ka, ke, ki** etc.,

ii. Other tokens are checked and matched from the dictionary.

iii. If the word is not matched in the dictionary values is then submitted to Urdu

Proper Noun Checker Module. This module makes the further solution in English for the translator module.

iv. The Last Sub-Module re-orders the urdu words created by Translator performs translating and ordering.

D. Urdu Proper Noun Checker: This module works with Translator. Urdu Proper Noun Checker creates equivalent word and proper nouns in English for the word which was unable to translate by the Translator Module.

E. Controller and Merger: The Controller and Merger receives sentences returned from Translator and merge these sentences to form a Translated paragraph in English.

To describe the flow and working of the algorithm, consider the following sentence:

میں سیب کھا چکا ہوں [Sentence-1]

The Sentence Parser Parse the paragraph but as the sentence is a paragraph having one sentence only, no need to parse it further.

The sentence is then passed to Tagger, which parse the sentence into tokens and tags them. (see table 9)

Figure 4
Architecture and Work Flow of AGHAZ System

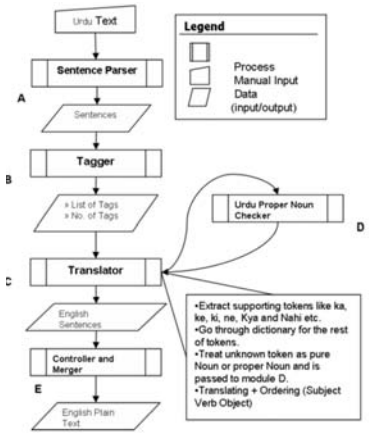


Table 9
Tokenizing

No.	Token	POS Tag
1.	میں	*<Pron>*
2.	سیب	*<Comm-Noun>*
3.	کھا	*<Verb>*
4.	چکا	*<HVB>*
5.	ھوں	*<HVB>*

Two tokens, marked as "<HVB>", shows that they are helping verbs and they help us to determine that this is present perfect tense.

5) CONCLUSION

The proposed system provides an efficient way to translate the Urdu text into English and it requires no lookup to the dictionary and for this less recursion is involved. It also do not require for handling of verbs by lookup in the dictionary. But it provides an efficient algorithm for handling proper Nouns and Multiwords.

REFERENCES

I.A. SAG, T. BALDWIN, F. BOND, A. COPESTAKE, D. FLICKINGER 2001, "Multi word Expressions: A Pain in the Neck for NLP", LinGO Working Paper No. 2001-03. Stanford University, CA.

UZAIR MUHAMMAD, KASHIF BILAL, ATIF KHAN, M. NASIR KHAN, "AGHAZ1: An Expert System Based approach for the Translation of English into Urdu", ENFORMATIKA'05 World Conferences Istanbul, Turkey. COMSATS Institute of Information Technology Wah Cantt., Pakistan.

KASHIF BILAL, UZAIR MUHAMMAD, ATIF KHAN, M. NASIR KHAN. "Extracting Multiword Expressions in Machine Translation from English into Urdu using Relational Data Approach", ENFORMATIKA'05 World Conferences Istanbul, Turkey.

Z. PERVEZ, S. KHAN, F. MUSTAFA, M. MAHMOOD, U. HASAN, "Pharasal

Consolidation Algorithm For Part Of Speech Tags In Machine Translation From English To Urdu National University of Science and Technology, Rawalpindi Pakistan.

TRUJILLO A. Translation Engines Techniques for Machine Translation, Springer (1999)

T. RAHMAN (2002). "Language Ideology and Power: Language Learning Among the Muslims of Pakistan and North India", Oxford University Press, Karachi, Pakistan.

T. MITAMURA, E. NYBERG, E. TORREJON, D. SVOBODA, A.BRUNNER AND K. BAKER,(2002) "Pronominal Anaphora Resolution in the Kantoo Multilingual Machine Translation System", Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation. Keihanna, Japan, Mar 2002.

AltaVista Babelfish.
URL: http://babelfish.altavista.com

Google Language Tool.
URL: http://www.google.com.pk/language_tools

Z. PERVEZ, S. KHAN, F. MUSTAFA, M. MAHMOOD, U. HASAN, "Pharasal Consolidation Algorithm for Part Of Speech Tags In Machine Translation from English to Urdu", NUST Institute of Information Technology, National University of Sciences and Technology.