AGHAZ[1]: An Expert System Based approach for the Translation of English to Urdu

Uzair Muhammad, Kashif Bilal, Atif Khan, and M. Nasir Khan

Abstract––Machine Translation (MT [3]) of English text to its Urdu equivalent is a difficult challenge. Lot of attempts has been made, but a few limited solutions are provided till now. We present a direct approach, using an expert system to translate English text into its equivalent Urdu, using The Unicode Standard, Version 4.0 (ISBN 0-321-18578-1) Range: 0600–06FF. The expert system works with a knowledge base that contains grammatical patterns of English and Urdu, as well as a tense and gender-aware dictionary of Urdu words (with their English equivalents).

Keywords—Machine Translation, Multiword Expressions, Urdu language processing, POS [2] Tagging for Urdu, Expert Systems.

I. INTRODUCTION

THE main question which comes in the mind of a human is why we translate from one human language to another. This question has many answers that reflect the importance of MT in human life. MT has important aspects socially and scientifically.

The social or political importance of MT arises in the communities where generally more than one human language is spoken.

Here the only possible alternative is to adopt a single lingua franca, which is not a good and attractive approach, since it forces to adopt a single language and dominate others one and ultimately diminishing. This loss of language becomes the cause of disappearance of a unique culture also; this is a loss that everyone should matter. Therefore translation is necessary for communication, for saving obsolescence languages and cultures and proper interaction of human in the society.

It is also a fact that one cannot provide human translator in such a great supply that can meet the translating requirement of a multilingual community. Using MT this problem can be reduced drastically and the productivity can also be increased by automating the translation task.

Uzair Muhammad is student at COMSATS Institute of Information Technology Wah Cantt, Pakistan. (phone: 0092 300 9087797; e-mail: joinuzair@yahoo.com).

Kashif Bilal is Lecturer at COMSATS Institute of Information Technology Abbottabad, NWFP, Pakistan. (phone: 0092 300 5613174; e-mail: kas_atd1@yahoo.com).

Atif Khan is student at COMSATS Institute of Information Technology Wah Cantt, Pakistan. (phone: 0092 300 5051373; e-mail: atif_ciit@yahoo.com).

M. Nasir Khan student at the COMSATS Institute of Information Technology Wah Cantt, Pakistan. (phone: 0092 943 412885; e-mail: m_nasir_khan@yahoo.com).

[1] AGHAZ is Urdu word which means "The Beginning"
[2] POS stands for Parts of Speech
[3] Machine Translation

Similarly the scientific importance of MT cannot be overlooked. It provides good testing grounds for various ideas in Artificial Intelligence, Computer Linguistics and Computer Sciences. 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 [10] and AltaVista Babelfish [9] are excellent examples of this new breed of machine translation systems that are available freely.

As the majority of Pakistani people (95% of the whole population) [6] are unaware of English, machine translation applications have an immense potential in the Pakistani market. Therefore to design a translating system from English to Urdu has a great impact on the Urdu speaking population. . Mostly all the research works and books have been written in English and a person ignorant of English language has strong sense of deprivation in this regard. MT for English to Urdu could therefore play a vital role in this regard. Urdu is a widely used language in Pakistan. Urdu/Hindi is spoken by 496 Million people in the world [7] [11].

II. PROBLEMS IN MT

Due to versatile usage of words and phrases, sometimes a well-experienced language translator encounters problems in translating a small piece of source language (SL) to any target language (TL). It is a fact that no computer in the present age has the ability to compare the knowledge of an average human being in terms of understanding the real world issues and problems.

Nowadays, the best MT systems can achieve translation accuracy of 73%. This accuracy has been achieved by machine translation system, which converts everyday Japanese to English. In the case of controlled vocabulary, the accuracy of English to German machine translation is 94.4% and that of English Spanish is 97.7% [8].

Without the ability of the comprehension of real world issues it is extremely difficult to make a computer intelligent enough to perform the task that a well-trained and experienced human translator found difficult at times. In addition to this inability, there are other problems which hinder a computer system to perform high level translation system. In the subsequent discussion we describe some of the problems faced in MT.

The following are the main problems faced by translators during the translation from Source Language to Target Language.

1. Lexical or Sense Ambiguity
2. Structural Ambiguity
3. Multiword Phrases
4. Language Differences

Out of these four main problems Lexical or Sense Ambiguity and Multiword phrases are the two most hot and unsolvable

World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:1, No:12, 2007

problems in MT. In the subsequent discussion we present the problem of multiword expressions and various solutions presented by MT specialists. Finally we present our idea to handle multiword problems that we have implemented in AGHAZ.

III. DESCRIPTION OF AGHAZ SYSTEM

AGHAZ is a Machine Translation System. It is an automatic translator from English to Urdu. AGHAZ consists of an Expert System and a knowledge base. It translates input English to Urdu language. Plus it provides an efficient solution for

- Multiword expressions.
- Proper Nouns i.e. generate an equivalent Urdu word.

AGHAZ is efficient in case of time. Normally; translators involve Reparatory Tagging, in case when they are using some off the shelf components, like QTAG is used in MUTRAJJUM [4], But in case of AGHAZ, it goes through Dictionary only once in tagging process and collects all the relevant information there. 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.

IV. AGHAZ SYSTEM MAIN COMPONENTS

There are three main components:
1. An Expert System,
2. A rich knowledge base
3. Patterns or Rules.

Both the Grammars (English and Urdu) are incompatible. English grammar doesn't consider gender difference while it affects the whole sentence in case of Urdu. For example "Dog" is masculine and "Cat" is feminine.

English: Cat is running   Dog is running

Urdu: بلی دوڑ تی ہے   کتا دوڑ تا ہے

Fig.1 Gender Difference

There is no difference at all, in case of English, except subject (Noun) but in case of Urdu there is more than one difference.

To overcome these problems we store some additional information in the knowledge base against the word. For Example English word, Meaning, Part of Speech, and some bits (Singular/Plural, Masculine/Feminine, and Multiword/Not Multiword).

As for as terminating string is concerned (used at end of Urdu sentence), our expert system does this job for us to pick the correct termination string from different patterns by applying grammar rules.

V. AGHAZ ARCHITECTURE

Figure 2 shows the architecture of AGHAZ.

A. EnglishParaSplitter: This module receives a paragraph written in English language and transforms/divides it into sentences. Paragraph may contain punctuation symbols or numbers.

B. EnglishTagger: Sentences returned from the module "A" is given as input to this module, which in turn extract tags from sentence. Each tag is as minimum as one character of length or maximum as one whole word. Special symbols i.e. @, #, &, * and _ plus apostrophe is handled here.

C. EnglishTranslator: EnglishTranslator is the composition of many sub-modules. Tags are given as input to EnglishTranslator.

C.i Extracts additional words like "Not", Helping Verbs etc., which are important only to provide some information.

C.ii *Dictionary look* up for the rest of tokens. For example "*He is not going*". The word "*is*" is helping verb, and "*not*" mark sentence as a negative one. Dictionary will be looked up for "*He*" and "*going*" only.

C.iii *Multiword Handling*: In case if a token (word) is a Multiword, then go through Multiword knowledge base. (See below the topic "AGHAZ Multiword Expression Handling")

C.iv If a word didn't find in dictionary (knowledge base) then it is passed to another module i.e. *EnglishProperNoun*, which returns its minimal solution in Urdu.

C.v *Best Search*: In case of more than one pattern are retrieved for a single word then it is responsibility of this module to select the best pattern and if there were no best solution then the most appropriate one.

C.vi Last Sub-Module performs translating and ordering. Meanings are retrieved in sub-Module ii, here we only put in order and terminating strings are concatenated to it. Both grammars are incompatible so ordering is important for semantic translation.

D. EnglishProperNoun: This module work in collaboration with EnglishTranslator (module C). When a word doesn't exist in the dictionary; EnglishProperNoun generate an equivalent Urdu solution for that word. This is normally done for proper nouns i.e. proper names and proper organizations like Ali, Zakir, Peshawar, Lahore, Spain etc.

E. MainController: MainController receives sentences returned from the above module and merge them in the form of paragraph.

World Academy of Science, Engineering and Technology
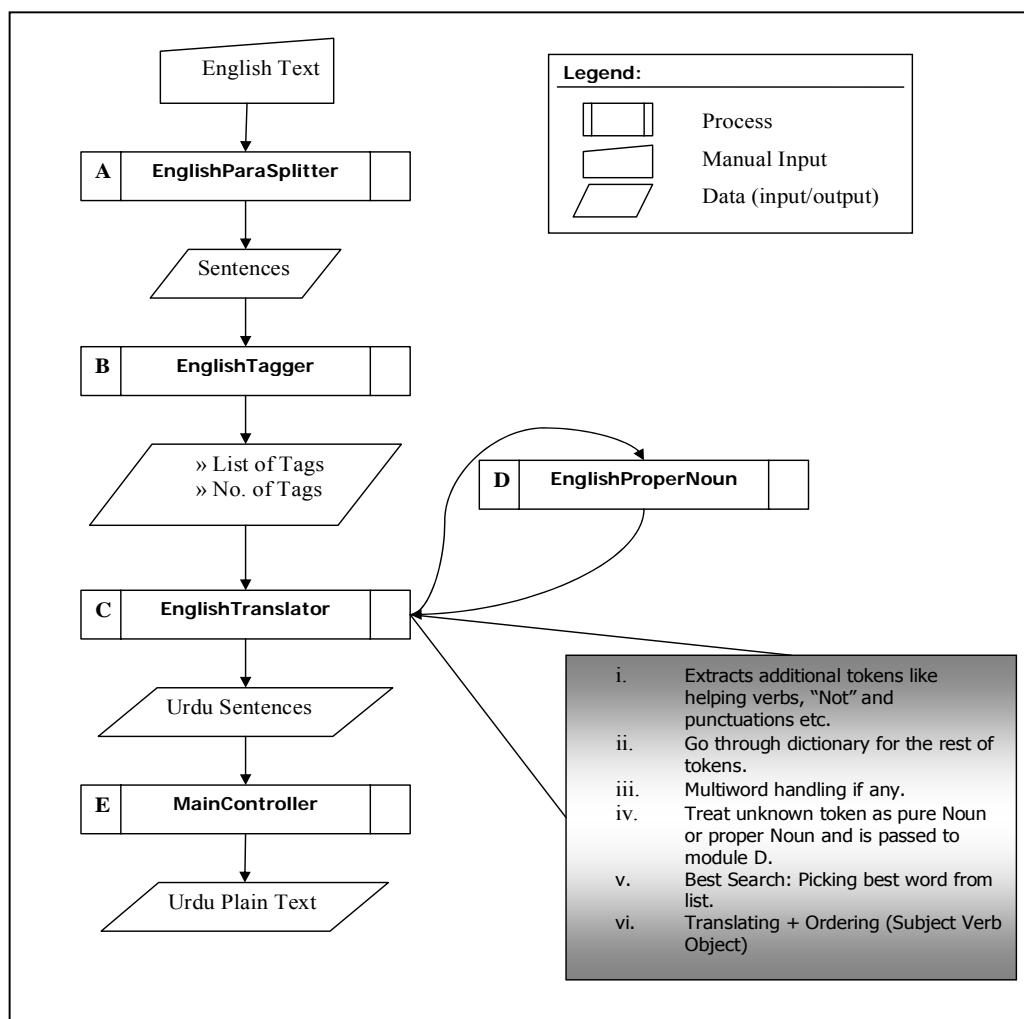International Journal of Cognitive and Language Sciences
Vol:1, No:12, 2007

Fig. 2 AGHAZ Architecture and flow

VI. HOW IT WORKS

To describe the flow of multiword resolution technique adopted in AGHAZ is consider the following sentence:
Shaukat Aziz is the prime minister of Pakistan.

[Sentence-1]

The sentence is passed to *EnglishParaSplitter* module, which will perform no action on it because this is a single sentence and this module is active only in case of paragraph.
The sentence is passed to *EnglishTagger*, which splits the sentence-1 into tokens. (Figure 3)
There are total eight tokens. Efficiency comes up here that it goes to traverse Dictionary for only three tokens i.e. token no. 5, 6 and 7. (See the following topic no. A, B for more detail).

A. Proper Noun Handling

It is superb in handling proper nouns whether they exist in the knowledge base or outside the domain of knowledge base.

No.	Token	POS Tag
1.	Shaukat	*<Prop-Noun>*
2.	Aziz	*<Prop-Noun>*
3.	Is	*<HVB>*
4.	The	*<ART>*
5.	Prime	*<Comm-Noun>*
6.	Minister	*<Comm-Noun>*
7.	of	*<Prep>*
8.	Pakistan	*<Prop-Noun>*

Fig. 3 Tagging Scheme

It considers the collection of proper noun as one whole entity like "Ali" is a proper noun and "Shaikh Zahid Bin Sultan Anahyan" is also a proper noun, so both are treated as proper noun. Unless another POS token is not encountered between proper nouns, it will be considered as one name.

World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:1, No:12, 2007

B. Helping Verb and Articles Handling

Helping verbs and articles are handled at front end and it doesn't need to look up Dictionary for them. Similarly sentence's negative and interrogative sense is also identified and handle at front end very easily.

The MainController module then merges all sentences into paragraph and shows the result. The resulted translation of sentence-1 is now:

شوکت عزیز پاکستان کا وزیراعظم ہے

Consider another example.

My name is Asad Ali. I am playing cricket.

[Sentence-2]

My name is Asad Ali. I am playing cricket.

A EnglishParaSplitter

Legend:

Process

Manual Input

Data(input/output)

My name is Asad Ali.
I am playing cricket.

B EnglishTagger

No.	Token	POS Tag
1.	My	*<Poss-Pron>*
2.	Name	*<Noun-Comn>*
3.	Is	*<HVB>*
4.	Asad	*<Prop-Noun>*
5.	Ali	*<Prop-Noun>*

No.	Token	POS Tag
1.	I	*<Pron>*
2.	am	*<HVB>*
3.	playing	*<Verb-Grnd>*
4.	cricket	*<Noun-Comn>*

Total Tokens: 5 + 4

C EnglishTranslator

Asad Ali

D EnglishProperNoun

میر ا نام اسد علی ہے.
میں کرکٹ کھیلتا ہوں.

اسد علی

E MainController

میر ا نام اسد علی ہے. میں
کرکٹ کھیلتا ہوں.

Urdu sentences are merged by the MainController into paragraph.
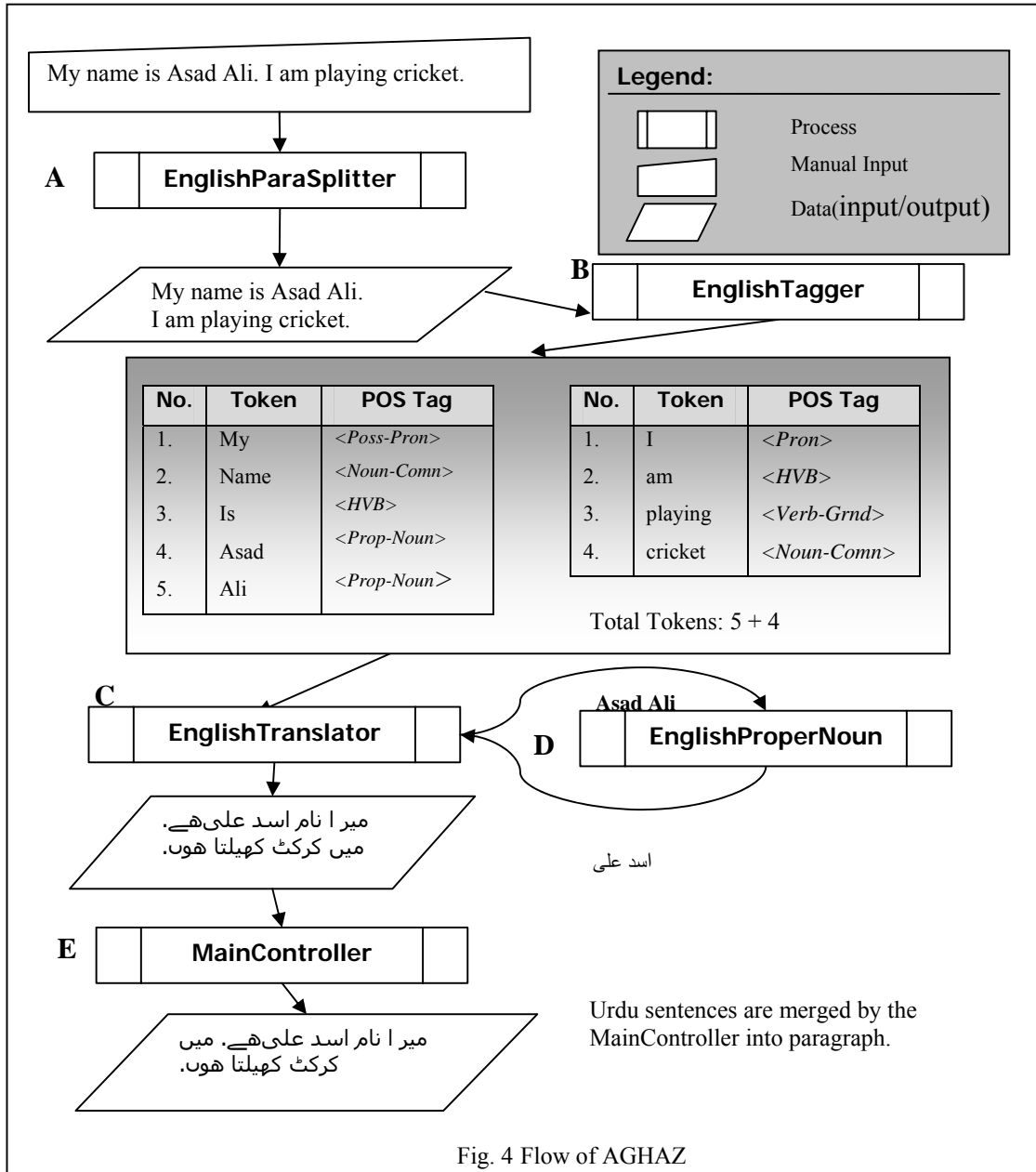
Fig. 4 Flow of AGHAZ

It can be seen from this discussion that only a slight backtracking is involved when resolving the multiword expression (complete process for handling multiword is explained in another research paper). The process has no immediate recursion and no other rules are referred from any knowledge base.

VII. CONCLUSION

The techniques used in AGHAZ are efficient and requires less lookups to dictionary, and also no recursions are involved. It does not require any lookup to dictionary for handling helping verbs, articles, not (negative sentences) etc, which make it more efficient. AGHAZ also uses efficient techniques

World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:1, No:12, 2007

to handle proper Nouns and Multiword. So overall it provides efficiency and correctness.

REFERENCES

[1] 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.

[2] Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, Tony McEnery "*Extracting Multiword Expression Using a Semantic Tagger*", Lancaster University.

[3] Ann Copestake, Fabre Lambeau, Aline Villavicecio, Francis Bond, Timothy Baldwin, Ivan A.Sag, Dan Flickinger, "*Multiword Expressions: linguistic precision and reusability*", University of Cambridge Computer Laboratory, William Gates Building, JJ Thomson Avenue, Cambridge, CB3 0FD, UK. NTT Communication Science Labortries, Hikari Dai, Seiko-cho,Soraku-gun, Kioto 619-0237, JAPAN.

[4] 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.

[5] Sarmad Hussein. "*Letter-to-Sound Conversion for Urdu Text-to-Speech System*". Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, Pakistan.

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

[7] Ethnologue, 13th Edition.

[8] T. Mitamura, E. Nyberg, E. Torrejon, D. Svoboda, A.Brunner and K. Baker, "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.

[9] AltaVista Babelfish. URL: http://babelfish.altavista.com

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

[11] 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.