Journal of Information & Communication Technology
Vol. 1, No. 2, (Fall 2007) 56-64

Handling Proper Nouns in Machine Translation from English into Urdu

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

ABSTRACT

Handling proper nouns using machine translation (MT) is really a harder job that has not been tackled with much success to date. In this paper a solution for handling the proper nouns phrases in machine translation is proposed which uses an expert system. For this purpose, the Unicode Standard, Version 4.0 (ISBN 0-321-18578-1) Range: 0600–06FF for Urdu language characters has been used.

Inspec Classification: C6180N, C7820M, C123O, D2010O

Keywords : Machine Translation, Urdu language processing, Natural Language Processing, Expert Systems.

1) INTRODUCTION

During the translation process of English or Urdu words, their meanings are extracted from dictionary. This process was necessary since we are interested in the exact and actual meaning of the words of the source languages to translate them in the target language (Sarmad Hussein). Proper nouns have the exception in this case. To store proper nouns and their equivalent Unicode in the dictionary is a solution but an unattractive one. If this solution follows for every proper noun then every noun should have been stored in the dictionary, which is a tedious and time-consuming task. This solution is also inefficient from algorithmic point of view, since the dictionary look up process for every proper noun slows down the process of translation and if the noun does not exist in the dictionary then time has also wasted and also the word goes un-translated. Another reason not to store proper nouns in the dictionary is that proper nouns are unlimited in number and have many kinds and they vary from country to country and culture to culture. Therefore the idea is proposed for storing proper nouns in the dictionary and the problem is analyzed thoroughly and finally came up with a solution, which was able to handle the translation of proper nouns on run time.

* 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

2) BACKGROUND

"AGHAZ is a Machine Translation System. It is an automatic bilingual translator, i.e. from English into Urdu and vice versa, that has been successfully implemented. AGHAZ consists of an Expert System and a knowledge base" (Uzair Muhammad, 2005), (Kashif Bilal, 2005). It translates input language into target language and hence it is a compiler like activity. Two major achievements have been made in translation from English into Urdu i.e. First solution for handling multiword expressions and secondly the current one that is handling proper nouns in MT (Uzair Muhammad, 2005).

3) DESCRIPTION OF THE ANALYSIS

Everyone knows that the number of alphabetic characters is twenty-six and that of Urdu are thirty-six. To represent a single character of English in Urdu there are one or many characters available in Urdu and vice versa (Uzair Muhammad, 2005). This situation also creates a problem in selecting exact equivalent character on run time. For example for the English alphabet "I" there are four possible equivalent Urdu characters, i.e., Ain(ع) Yee(ي)Zair(ے)and Alif(ا). This shows that in practice it is not possible to select the exact equivalent target character and thus 100% accuracy could never be achieved. After a lot of wearisome and meticulous analysis and study from previous research (Sarmad Hussein), (I.A. Sag, 2001), (Azza Abd El-Mohammad, 2000), (T. Mitamura, 2002), (Z. Pervez), (Uzair Muhammad, 2005), (Kashif Bilal, 2005), we designed a list of English and their equivalent character given as follow:

Figure 1:
English and their equivalent chart in Urdu

Serial	English Alphabets	Equivalent Urdu Alphabets
1	A	ع,آ,ا
2	B	ب
3	C	ک,س
4	D	ڈ,د
5	E	ے, ي
6	F	ف
7	G	ج,گ
8	H	ھ,ہ,ح
9	I	ا,ي,ع
10	J	ج
11	K	ک
12	L	ل
13	M	م
14	N	ن
15	O	ع,و,او
16	P	پ
17	Q	ق
18	R	ر, ڑ
19	S	ژ,ص,س
20	T	ط , ث,ت
21	U	ع ,ي,و,, ا
22	V	و,
23	W	و,,
24	X	Combination of ک,س
25	Y	ي,ی
26	Z	ظ,ض,ذ,ز

Figure 1 shows that one or more than one Urdu characters can be replaced for every English character. This creates an ambiguous situation where the appropriate selection becomes difficult for replacement of characters from source to target language. In Urdu the phonetics f o r ز, ذ, ظ, ضare same and are pronounced almost same and are represented in English with only one character that is Z. For example to translate the proper noun "Zia" from English to Urdu, we have four alternatives in Urdu i.e.ظیا ,ضیا ,زیا ,ذیا.More examples depicting this situation are given in the following table. This table shows only two possible outcome of a noun in English in Urdu.

Figure 2:
Two possible outcomes of English word in Urdu

S#	English Nouns	Urdu Equivalent 1	Urdu Equivalent 2
1	*A*sif	اسف	اصف
2	*S*amrina	سمرینا	ثمرینا
3	Maim*o*na	میمونا	میمانا
4	Kash*i*f	کاشف	کاشعف
5	*Z*aheer	ضہیر	ظہیر
6	*H*anifa	حنیفہ	ہنیفہ
7	*T*akia	تاکیا	ڈاکیا
8	*T*ahir	طاہر	تاہر
9	*U*mair	عمیر	امیر
10	*G*ul	گل	جل
11	Danish	دانش	ڈانش

Figure 2 shows the difficulty in choosing correct replaceable character from source to target language. This also shows that 100% intelligence in this regard cannot be achieved. This is not the end of this problem. In English, combination of more than one character has single phonetic effect in Urdu (I.A. Sag, 2001). The table given below shows some of them.

Figure 3:
More than one English alphabets equivalent to one Urdu character

Serial	Combination of English Characters	Equivalent Urdu
1	Bh	بھ
2	Ch	چ
3	Dh	ڈھ, دھ
4	Gh	غ,گھ
5	Jh	جھ
6	Kh	خ,کھ
7	Ph	پھ
8	Rh	رھ, ڑھ
9	Sh	ش

Closely analyzing Figure 3 shows that some of the combination of English alphabet characters has single or combination of two alphabet characters equivalent in Urdu. Their occurrence in the word cannot be restricted; therefore it was also one of the hard issues for handling nouns during translation.

Various examples can be presented depicting this dilemma. To write the name شازیا there is no one to one relationship between English and Urdu alphabet characters, instead we have to use the combination of S and h i.e. Sh in English for the character ش. Similarly for the name چراغ we do not have any single character for the character چ in English therefore we must use the combination of C and h i.e. Chiragh.

Furthermore the alphabets of Urdu use some extra characters to pronounce a noun or word correctly like ◌َ (Zabar), ◌ِ (Zair), ◌ُ (Pesh) etc. These characters affect the pronunciation like in Urdu ا and س may make Es and Us depending on their use.

Now to better clarify the idea we give the following table which shows our entire analysis for every character of English and its mapping in Urdu alphabets with certain examples.

Figure 4:
English and Urdu characters mapping with example

Serial	English Alphabets	Equivalent Urdu Alphabets	Examples			
1	A	◌َ,ع,ا	**A**kmal	اکمل	**A**leem	علیم
			Akram	اکرم	**JA**mil	جمیل
2	B	ب	**B**aqir	باقر	Ak**B**ar	اکبر
			i**B**rar	ابرار	Muhee**B**	مهیب
3	C	ک,س	Cola	کولا	office	آفس
4	D	ڈ,د	Danyal	دانیال	Dafli	ڈفلی
5	E	ي,	Ehsaan	احسان	Maheen	مهین
6	F	ف	Faraz	فراز	Kashif	کاشف
7	G	گ,ج	Gandhi	گاندھی		
8	H	ه,ہ,ح	Hamid	حامد	Hazara	هزارا
9	I	ا,ي,ع	Irfan	عرفان	Arif	عارف
			Ishaq	اسحاق		
10	J	ج	Javed	جاوید	Amjad	امجد
11	K	ک	Karachi	کراچی	Akram	اکرم
12	L	ل	Lahore	لاهور	Laila	لیلی
13	M	م	Maham	ماهم	Majid	ماجد
14	N	ن	Nusrat	نصرت	Anila	انیلا
15	O	ع,و,او	Omar	عمر	Osama	اسامه
			Zahoor	ظهور		
16	P	پ	Pervez	پرویز	Pasha	پاشا
17	Q	ق	Qamar	قمر	Aqeel	عقیل
18	R	ر, ڑ	Rubi	روبی	Khiora	کھیوڑا
19	S	ث,ص,س	Sarwat	ثروت	Sabir	صابر
			Sarmad	سرمد	Wasif	واصف
20	T	ت,ط,ٹ	Tahir	طاهر	Tausif	توصیف
			Tommy	ٹامی	Tabasum	تبسم
21	U	ع,ُ,ُو,ا	Zubair	زبیر	Umar	عمر
			Qurban	قربان	Suraj	سورج
22	V	و,	Vinash	وناش	Vitr	وتر
23	W	و,ُ	Warid	وريد	Jawad	جواد
24	X	Combination of ک,س	Dixit	ڈکشٹ	Zerox	زیراکس
25	Y	ي,ئ	Yamin	یمین	Faryal	فریال
26	Z	ظ,ض,ذ,ز	Zahid	زاهد	Zia	ضیا
			Zaheer	ظهیر	Zarab	ذاراب

Similarly Examples for the combination of characters are as follows:

Figure 5:
Combination of English characters and Urdu equivalent mapping with example

S #	Combination of English Characters	Equivalent Urdu	Examples	
1	Bh	بھ	Bharat	بھارت
2	Ch	چ	Chawla	چاولہ
3	Dh	ڈھ, دھ	Dhoom	دھوم
4	Gh	غ,گھ	Ghoda	گھوڑا
			Ghunwa	غنوا
5	Jh	جھ	Jhang	جھنگ
6	Kh	خ,کھ	Khurshid	خورشید
			Khokhar	کھوکھر
7	Ph	پھ,ف	Phool	پھول
			Philadelphia	فلادلفیا
8	Rh	رھ, ڑھ	Rheel	رھیل
9	Sh	ش	Shabeer	شبیر

The examples given in the above table are not enough to get inside and deduce rules. To deduce implemental rules a detailed and thorough analysis was done. At the end of the analysis we came up with some satisfactory solution. After deducing the solution it became also clear that one couldn't get 100% accurate results. The reason has already been discussed.

4) DESCRIPTION OF THE SOLUTION

In the consequent discussion, the solution for the noun problem is presented. We do not mention here the entire complex detail of the solution. Only a bird eye of the solution is given.

Figure 6:
English alphabets and their replacement rules

S#	English Alphabets	Replacement Rules for Urdu Alphabets
1	A	ا,أ,ع,
		If A is at start and is followed by "Li,T,Q,Z,D", it will be ع
		If A is at start and is followed by the characters other than above, it will be ا
		If comes at the end it will be ا
		If AA is at start it will be آ otherwise ا

Continue

2	B	ب
		In every case it would be ب
3	C	ک,س
		If C is followed by "A,E,Y, it will be س
		If C is followed by the characters other than above, it will be ک
4	D	ڈ,د
		د has maximum usage frequency for D, so we preferred instead of ڈ
5	E	ي, ۔
		If E is at start it will be ا
		If E comes at the end it will be ي
		If E is followed by "C, E, I, O, P, Q, U , I, X, Y" it will be . otherwise it will be ي
		EE will be replaced by ي
6	F	ف
		F will be replaced by ف in all cases
7	G	ج,گ
		If G is followed by "A,E,Y" it will be ج
		Other than above it will be گ
8	H	ھ,ہ,ح
		If H is at start it will be ح
		Other than starting position it will be ھ
9	I	ا,ي,ع
		If I is at start and is followed by "M,R,L,F", it will be ع otherwise it will be ا
		If I comes at the end it will be ي
		If I comes other than starting and end and is followed by "A, B, C, D, G, L, M, N, O, P" it will be ي otherwise it will be .
10	J	J is simply replaced by ج in all cases.
11	K	K is simply replaced by ک in all cases
12	L	L is simply replaced by ل

Continue

13	N	N is simply replaced by ن
14	M	M is simply replaced by م
15	O	ع,و,او
		If O is at start it will be او except it follows M
		If O follows M it will be ع
		If OO comes it will be و
16	P	P is simply replaced by پ
17	Q	Q is simply replaced by ق
18	R	ر, ڑ
		ر Has more usage frequency therefore we prefer ر instead of ڑ
19	S	ث,ص,س
		س Has the most usage frequency for S in compare to ص and ث. So we prefer س
20	T	ط , ث,ت
		If T follows "I,AL,,AR,AY" comes it will be ط otherwise it will be ت
		T creates an ambiguous situation.
21	U	ع ,ُو,ا
		If U is at start and is followed by "M,S,Q" it will be ع otherwise it will be ا
		If U comes at end it will be و
		Other than starting and end it would be ُ
22	V	V is simply replaced by و
23	W	V is simply replaced by و
24	X	Combination of ک,س
25	Y	ي Will be used except where it comes at end
		If Y comes at the end it will be ی
26	Z	ظ,ض,ذ,ز
		Study

Figure 5 displays the mapping rule for single character. Now we present the rules for two English characters having single or more than one phonetic effect.

Figure 7:
Replacement rule for two English characters have single or more than one phonetic effect

S#	Combination of English Characters	Rules
1	Bh	بھ
		Bh will be replaced by بھ
2	Ch	چ
		Ch will be replaced by چ
3	Dh	ڈھ, دھ
		Dh will be replaced by دھ
4	Gh	غ,گھ
		غ Will be used for Gh
5	Jh	جھ
		Jh will be replaced by جھ
6	Kh	خ,کھ
		Kh will be replaced by خ
7	Ph	پھ
		Ph will be replaced simply by پھ
8	Rh	رھ, ڑھ
		Rh will be replaced رھ
9	Sh	ش
		Sh will be replaced by ش

5) LIMITATION OF THE ALGORITHM

The so far discussion shows an approach to resolve the proper noun problems. The replacement technique that has been adapted for noun resolution in AGHAZ chooses only those replacement characters, which have more common occurrences. However this solution fails where there is no one to one or one to two correspondence between the alphabetic characters of the source (English) and target (Urdu) language. In other words the algorithm is unable to translate the exact meaning of those proper nouns, which have totally different replacement for every character in the source language. For example the name of the planets and zodiac sign names cannot exactly be generated by this solution.

Consider the situation where we want to generate the meaning of the noun "Mars" in Urdu. By applying our solution the meaning would be مارس which is not, however, the correct and exact meaning in Urdu. Rather the exact meaning in Urdu is مریخ. Similarly the meaning for the zodiac sign Leo, according to our replacement rules, would be لیو a n d the correct meaning is اسد. Similarly the plants and animals vast nomenclature also depict this kind of problem. After a meticulous analysis we reached to the result to cope with this problem we have only one way i.e. to hard-code these name and their meanings in the dictionary.

REFERENCES

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

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.

AZZA ABD EL-MOHAMMAD.2000."Machine Translation of Noun Phrases: From English to Arabic", Cairo University, GIZA Egypt.

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.

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.

"UZAIR MUHAMMAD, KASHIF BILAL, ATIF KHAN, AND M. NASIR KHAN, 2005" AGHAZ: An Expert System Based approach for the Translation of English to Urdu, Proceedings Of World Academy Of Science, Engineering And Technology Volume 6 June 2005 ISSN 1307-6884

KASHIF BILAL, UZAIR MUHAMMAD, ATIF KHAN, AND M. NASIR KHAN, 2005 "Extracting Multiword Expressions in Machine Translation from English to Urdu using Relational Data Approach", Proceedings Of World Academy Of Science, Engineering And Technology Volume 6 June cISSN 1307-6884