Rule Based English to Urdu Machine Translation

Naila Ata
Dept. of Computer Science
University of Karachi
Pakistan
naila_ata@yahoo.com

Bushra Jawaid
Dept. of Computer Science
University of Karachi
Pakistan
bushrajd84@hotmail.com

Amir Kamran
Dept. of Computer Science
University of Karachi
Pakistan
amirkamran@msn.com

Abstract

This paper presents an approach to machine translation for English to Urdu, primarily based on transfer approach, handling case phrases and verbs postpositions through concepts of Paninian grammar. This is done in order to get best results with optimum complexity. This paper emphasizes the architectural aspect of our translation approach and different grammatical structures which have been solved through this scheme.

1. Introduction

Machine translation (MT) implies translation from source language to target language through a computerized system. Multiple approaches are followed in MT, which can be classified as follows [8]

- Direct Architecture
- Transfer Architecture
- Interlingua Architecture

Direct Architecture is the most primitive form of translation replacing source language word with the target language word; Interlingua approach needs an in-depth semantic analysis and the target language is generated in this approach. Transfer approach lies in between the two extremes; it works on the syntactic level and involves semantics in some places. Source language's syntactic structure is analyzed to get a processed structure, mapping rules are used to transform it into target language structure, and translation is then generated with the application of target language defined rules

This MT engines falls in transfer approach, Stanford parser is used to get the parse tree of English sentence, transformation is done through algorithm and lexemes are then looked up in the dictionary for the meaning and the attributes. This way feature matrix of every lexeme is obtained and unification is done. Case markers and the post positions are then handled, through reverse Paninian approach.

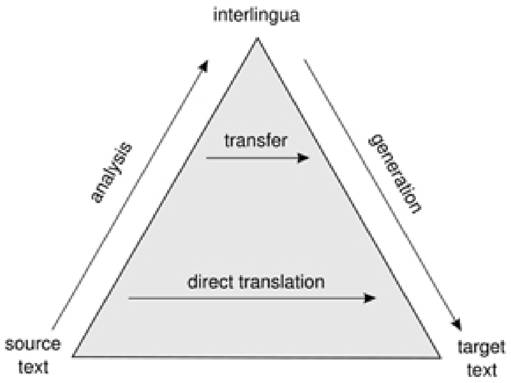

Fig 1.1: Machine Translation Paradigm

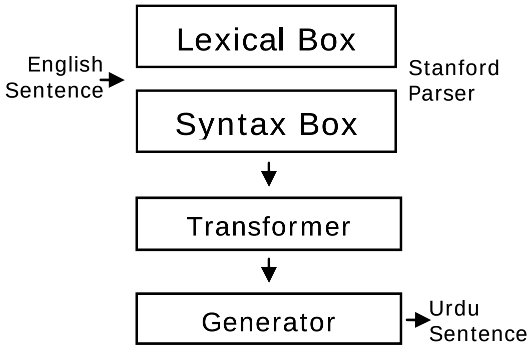

Fig 1.2: Architecture of MT Engine

2. Transformation Algorithm

Basic difference between two languages is the structure; English is fixed order SVO while Urdu is relatively free order SOV language. If each node of VP is swapped recursively, SOV structure can be obtained. Noun phrase in both languages follows the same rule; therefore swapping is not applied to it. This way we obtain a basic SOV structure.

Exceptions of this swapping rule are present and handled. Few exceptions are:

If the subject of noun phrase (NP) comprises of NP prepositional phrase (PP), transformer swap it, since the placement of PP in Urdu is before NP.

If adverb phrase (ADVP) appears before verb, swapping is not needed. ADVP in English can appears in different order depending on the type of ADVP, however, Urdu prefers ADVP before verb.

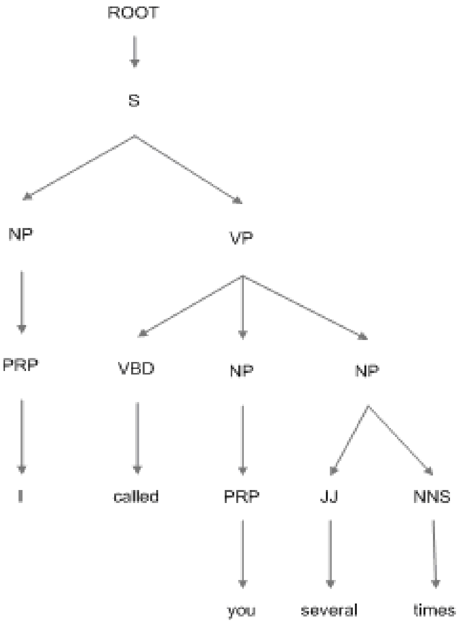

Fig: 2.1 (a) English parse tree

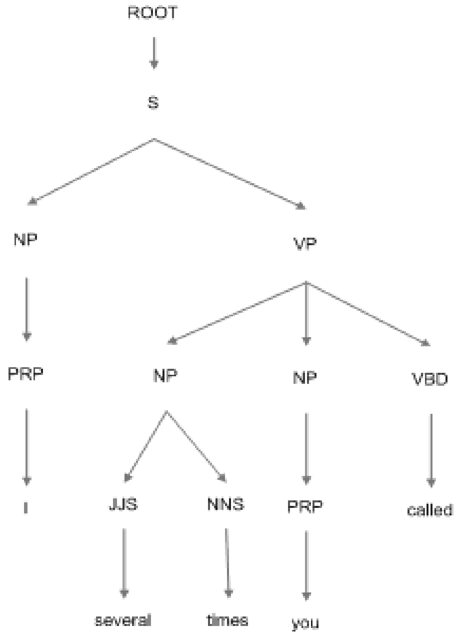

Fig: 2.1 (b) Urdu parse tree

Second phase is the agreement between NP and VP of the sentence S. Verb phrase in Urdu is inflected according to the gender, number and person (GNP) attribute of the head noun while form of the NP depends on the tense, aspect and modality of the verb phrase (VP), Urdu's adjectives are also modified by the GNP of head noun.

Context free grammar (CFG) identifies structural attributes of language, and we need annotations in the CFG for the unification between subject and object, tree transformation algorithm defined is the generalized CFG of Urdu structure, attributes for the unification is identified, lexemes are stored in dictionary with these attributes and CFG is then annotated with rules for unification.

In English sentence, if the sentence is following CFG.

$$VP_{agr,\ vagr} \rightarrow X_{agr}\ VP1_{agr,\ vagr}$$

$$*X_{agr} \rightarrow VBP \mid VBZ \mid VBG \mid VBN \mid VBD \mid VB$$

After swapping it would be

$$VP_{agr,\ vagr} \rightarrow VP1_{agr,\ vagr}\, X_{agr}$$

$$\text{Where arg} \quad = \quad \begin{bmatrix} \text{num} \\ \text{type} \\ \text{gen} \\ \text{person} \end{bmatrix}$$

$$\text{vargs} \quad = \begin{pmatrix} \text{tenseseq} \\ \text{md} \\ \text{Trans} \\ \text{PP} \\ \text{psvevoice} \end{pmatrix}$$

“num” identifies singular or plural attribute, ”type” for nominal, oblique or accusative case, “gen” for gender of Urdu noun and “person” for the 1st, 2nd or 3rd person.

“tenseseq” denotes the verb sequences (i.e. helping verb sequence for tense identification), “md” a Boolean variable denote the presence of modal auxiliary (must or should), “trans” denoting presence of transitive verb, “PP” indicating occurrence of preposition phrase.

And the unification rules are

VP1.vagr = VP.vagr
VP1.agr = VP.agr
VP1.vagr.tense += X.value
X.agr = VP.agr

Another example to make the idea clear

English CFG

$VP_{agr,\ vagr} \rightarrow X_{agr}\ NP\ PP_{in}\ NP$

After Swapping

$VP_{agr,\ vagr} \rightarrow PP_{in}\ NP\ X_{agr}$

$X_{agr} \rightarrow VBP \mid VBZ \mid VBG \mid VBN \mid VBD \mid VB$

And agreement rules are same, as described above.

* VB = verb basic form
VBP = verb, non-3sg pres
VBD = verb past tense
VBN = verb perfective form
VBG = verb gerundive form
VBZ = verb present 3-sg form

For the noun Phrase

$NP_{agr,vagr} \rightarrow DT\ \ Y_{agr}$

$^*Y_{agr} \rightarrow \quad NN \mid NNP \mid NNS \mid JJ \mid JJS$

*NN= Comon singular noun
NNP=Proper noun singular
NNS =Common plural noun
JJ=adjective
JJS =adjective superlative

NP.agr.num = Y.num
NP.agr.gen = Y.gen
NP.agr.person = Y.gen.person

If NP.vagr.tense=PI and VP.tran=true
Then NP.agr.type = Oblique

If NP.varg.mod = true or NP.vagr.psvevoice = true
Then NP.agr.type = Accusative
Else NP.agr.type = Nominal

Adjectives are also inflected according to gender and number of noun, so these attributes are passed between sister nodes of NP.

$NP_{agr,vagr} \rightarrow JJ\ \ Z_{agr}$
$Z_{agr}, \rightarrow\ NN \mid NNS$

JJ.agr = Z.agr
NP.agr = Z.agr

3. Tense Aspect Modality (TAM)

English uses auxiliaries to mark the TAM of a sentence, while Urdu verbs take post position for describing TAM; there is no one-to-one mapping between auxiliaries of English and Urdu post position. Moreover, verbs in Urdu also inflect according to the GNP of the noun phrase. Tense of a sentence is determined by the sequence and form of auxiliary verbs, before the main verb and form of the main verb. TAM table is then used to identify tense of the sentence with the help of auxiliary verb and main verb form and determine Urdu verb suffix corresponding to it. Subject NP’s attribute (gender, name and person) is required for inflection of verb. We don’t deal with morphology and auxiliary verb separately; may be it is unconventional, but it speeds up the processing and increases efficiency with seemingly no disadvantage.

We used the reverse Panini approach, since surface form of Urdu is generated from the roles and features identified.

For example, consider the following sentence:
I am going to the market
After transformation:
I to the market going am

Verb pattern (am + VBG) is looked up for the tense and TAM is then identified through Urdu TAM table.

Auxiliary and main Verb pattern	Tense	Gender	Case phrase	
			Singular	Plural
VBZ | VBP	Present Indefinite	M[1]	تا ھوں	تے ہیں
Do | does + VB		F	تی ھوں	تیں ہیں
Is | am | are + VBG	Present Continuous	M	رہا ھوں	رہے ہیں
		F	رہی ھوں	ر ھیں ہیں
Has | have + VBN	Present Perfect[3]	M	چکاھوں| یا ھے	چکے ہیں| یا ھے
		F	چکی ھوں	چکیں ہیں
Has | have + been + VBG	Present Perfect Continuous[2]	M	تا رہا ھوں	تے رہے ہیں
		F	تی رہی ھوں	تی ر ھیں ہیں
VBD	Past Indefinite[3]	M	یا | یا تھا	ئے | ئے تھے
VBD + VB		F	ئ | ئ تھی	ئیں | ئیں تھیں
Was | were + VBG	Past continuous	M	رہا تھا	رہے تھے
		F	رہی تھی	ر ھیں تھیں
Had + VBN	Past Perfect	M	چکا تھا	چکے تھے
		F	چکی تھی	چکیں تھیں

[1] For personal pronouns 'I' and 'you' gender is considered as male, since discourse analysis is not considered.
[2] If 'since' or 'for' is used in the sentence, present continuous TAM will be followed
[3] Past indefinite and present perfect follows GNP of object NP.

Table 3.1. TAM for 1st Person

Passive voice sentences are also handled from the above defined scheme.

4. Case Markers

Urdu uses case markers, which appear according to the TAM of the sentence. Grammatically, we can classify them under non-semantic preposition, as they contain negligible content. There is no one-to-one correspondence between English and Urdu words and we have to add these markers according to TAM identified.

Ergative	نے
Genitive	کا کے کی
Object	کو

Table 4.1. Urdu Case Markers

Following the Paninian approach, we have listed:

TAM	Object Type	Verb Type	Subject Marker	Object Marker
Past Indefinite / Present Perfect	Inanimate	Transitive	نے	Null
Past Indefinite / Past Perfect	Animate	Intransitive	نے	کو *

* If a pronoun is object, it will take accusative form.

Table 4.2. Urdu Subject Object Markers

Consider the sentence in fig 2.1 (a)

I called you several times.
میں نے کئ بار تمھیں بلایا

"نے" is used with the subject and object pronoun will be in accusative form.

Object marker "کو" is also present when sentence is in passive voice or the modal auxiliary "should" is present. We identify its presence through attribute.

She was called for the interview.
. اس کو انٹرویو کے لئے بلایاگیا

5. Genitive Marker

When nouns act as an adjective for the head noun, genitive case marker is used between them. Since, In Urdu adjectives also have gender attribute associated with them; these markers (کا کے کی) are used according to the number and gender of the following noun e.g.

Red brick wall
لال اینٹ کی دیوار

6. Preposition Phrase in Noun Phrase

Preposition phrase (PP) used in a noun phrase (NP) needs additional words in the translation. For Example:

Monkey on the tree is naughty.

Lexical transformation in such sentences does not produce good result.

درخت پر بندر شرارتی ھے

Therefore, we approach for approximation and translate the two prepositions (in and on) into “والا” which will be inflected according to the gender of the head noun and the translation would be:

درخت والا بندر شرارتی ھے

7. Interrogative Sentences

Interrogative sentences are broadly classified into content-type and polar questions. In English, for yes-no questions first auxiliary verb is fronted while for content-type questions sentence starts with wh-words. Urdu, being a free order language, its interrogative markers can be placed at different places, for yes-no ‘کیا’ is used to indicate interrogation. e.g.

Are they going to market?

کیا وہ بازارجارھےھیں؟

وہ کیا بازارجارھےھیں؟

وہ بازارجارھےھیں کیا؟

MT engine follows the most common placement of the markers. Parser classifies polar questions as SQ, ‘کیا’ is added in front positions of such sentences and rest of the translation follows the usual scheme.

8. Content-type Questions

Question words, wh-words, in English come at the beginning while Urdu is wh-in-situ language, question markers placement in Urdu depends on the nature of the question,

8.1. Who

“Who” is translated as “کون” and it is interrogative pronoun nominal form, therefore “کون” comes in the beginning of the sentence, at the subject place.

Who will go to Karachi?
کون کراچی جائے گا؟

However, when “be” verb is used as main verb, “کون” is not the subject and placement in Urdu changes.

Who are you?
تم کون ھو؟

With modal auxiliary “should”, accusative form of noun is used, and maps to “کس کو” .

Another usage of “کون” is

Who will come with me?
کون کون میرے ساتھ جائے گا؟

“کون کون” here depends on the context and one can not identify this sense without discourse analysis. Our MT engine in this case will translate it into “کون” .

8.2. Which

‘Which’ is interrogative determiner and as in English and Urdu determiners appear before noun, therefore no transformation is needed. However there is another issue in its translation, Urdu can map which into either ‘کون سا’ or ‘کس’ e.g.

Which machine runs fast?
کون سی مشین تیز چلتی ھے؟

Which butterfly did child chase?
کس تتلی کا بچے نے پیچھا کیا؟

When the verb is transitive, 'which' is inquiring about object noun and کس is used, while کون سی is used when verb in intransitive, in this case determiner is associated with subject NP.

8.3. Whose

"Whose" is question word for possessive adjectives, and possessive adjectives usually come before noun in both languages, Urdu question markers therefore come before the related noun.

Whose book is this?
یہ کس کی کتاب ھے؟

8.4. When, Where, Why, How

Parser identifies them as WHADVP, or WRB, and adverbial phrase comes before noun, we insert WHADVP node before verb, while transforming tree, to get the desired canonical structure.

How many years have you spent in Karachi?
تم نے کراچی میں کتنے سال گزارے؟

8.5. Whom

"Whom" is objective interrogative pronoun and generally translated in to "کس" with a case marker "کو" e.g.

Whom did you give book?
تم نے کس کو کتاب دی؟

However sometime other prepositions are used instead " کو "e.g. for sentence
To whom did you talk?
Selection of target language preposition for the source language ones is a many to many mapping, and is not dealt in the system. MT engine takes the default translation for the preposition.

8.6. What

"What" is translated as "کیا" it can be interrogative pronoun or interrogative adjective

What is happening?
کیا ھو رھا ھے؟

What did she buy from market?
اس نے بازار سے کیا خریدا؟

When used as interrogative adjective, its translation is according to the "which" translation rule.

9. Relative Clause vs. Embedded Question

Relative clause is a subordinate clause which modifies a noun. e.g.

Anyone, who thinks this will work, is crazy.

Embedded questions are questions within another statement or question. They function as noun clauses and as such should generally follow statement, not question, order

The question is who won the prize.

Parser does not differentiate between them and make sub-tree with head SBAR in both cases. In Urdu, when 'who' is used as relative clause it would be translated as "جو" while embedded question follows the rule of question sentences described above. Relative clause translation will depend on the GNP of noun preceding it while embedded question is an independent clause. Resolving this issue in translation, we mark attribute relClause + .when SBAR's appears in the tree and its immediate left node is NP and translate (who, which, that) in "جو" , and when VP precede SBAR, it is translated according to the rules of question sentences.

10. Conclusion

We have designed our MT engine with a syntactic-semantic approach, combining our novel approach with multiple existing paradigms which resulted in an efficient structure at structural transformation level. Complex structures including subordinate clauses can also be translated with this scheme.

Algorithmic structural transformation can be applied to other languages as well, we just have to analyze grammatical structure of target language and transform the tree accordingly.

It describes the usage of Paninian theory in Urdu translation, which can handle case phrases and verb post position very efficiently. Moreover, this framework can also be used for other constructs and and handle conditional and comparative sentences by designing their TAM.

11. Acknowledgement

This paper is based on our final year project of BSCS, under the supervision of Mr. Tafseer Ahmed. We are grateful to our teacher for guidance and support, without it our MT engine could not be build.

12. References

[1] Akshar Bharati, Rajeev Sangal , Parsing Free Word Order Languages in the Paninian Framework, IIT, Kanpur, 1993

[2] Akshar Bharati, Chaitanya V. and Sangal R., Natural Language Processing, A Paninian Perspective, Prentice Hall of India, New Delhi, India.

[3] Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond , A Grammar Writer's Cookbook, CSLI publications, Stanford, California.

[4] Daniel Jurafsky and James H. Martin, Speech and Language processing, University of Colorado, Boulder, Pearson.

[5] Miriam Butt, Tracy Holloway King, John T. Maxwell III, Productive Encoding of Urdu Complex Predicates in the ParGram Project,

[6] Tafseer Ahmed, Machine Translation and NLP, Department Of Computer Science, University Of Karachi.

[7] Ebba Gustavii, Target Language Preposition Selection – an Experiment with Transformation-Based Learning and Aligned Bilingual Data, Department of Linguistics and Philology, Uppsala University, Sweden.

[8] Bonnie J. Dorr, A Survey of Current Paradigms in Machine Translation, Computer Science Department, University of Maryland.

[9] http://nlp.stanford.edu/downloads/lex-parser.shtml