

Algorithm of Urdu Translation Engine

Hussain Hyder Ali Khowaja, Muzammil Ahmed Khan, Kashif Shaikh
hhalik_web@gawab.com, mail4khan@yahoo.com, mshaikh@ssuet.edu.pk
 Sir Syed University of Engineering & Technology, University Road, Karachi, Pakistan

Abstract: This paper describes the procedure to translate text from English to Urdu and vice-versa using the power of dynamic dictionary, grammatical rules & NLP (Natural Language Processing) - AI (Artificial Intelligence) algorithms. It is based on standard Urdu Unicode mapped to ISO/IEC 10646 support with open-type fonts & its character controls in Operating Systems. The System uses algorithms governing linguistic rules including lexicons of words and phrases, sentences checking (semantic & syntactic) levels & parsing with grammar trees. The underlying intelligence with corrective information is funneled into a collective pool of linguistic knowledge. This engine enables us to share the knowledge & number of our people to understand business, literature, scientific researches in Urdu language.

1. Introduction

The main objective is to use smart way of automation & to reduce human efforts in translation. It can be used to translate text of web pages, books, etc. We can also use it as a component for developing other software's like browsers, voice recognition, chat messengers, SMS etc.

1.1. Urdu Unicode Numbers & Characters

The ISO Standard set of Unicode [1] provides 16-bit encoding specifications; enabling 65,535 unique characters various languages of world. The given below charts is of Arabic part also used as Urdu Unicode sets:

Symbol	Unicode	Unicode Description
٠	06F0	EASTERN ARABIC-INDIC DIGIT ZERO
١	06F1	EASTERN ARABIC-INDIC DIGIT ONE
٢	06F2	EASTERN ARABIC-INDIC DIGIT TWO
٣	06F3	EASTERN ARABIC-INDIC DIGIT THREE
٤	06F4	EASTERN ARABIC-INDIC DIGIT FOUR
٥	06F5	EASTERN ARABIC-INDIC DIGIT FIVE
٦	06F6	EASTERN ARABIC-INDIC DIGIT SIX
٧	06F7	EASTERN ARABIC-INDIC DIGIT SEVEN
٨	06F8	EASTERN ARABIC-INDIC DIGIT EIGHT
٩	06F9	EASTERN ARABIC-INDIC DIGIT NINE

Figure 1: Number Symbols

Symbol	Unicode	Unicode Description
ا	0627	ARABIC LETTER ALEF
آ	0623	ARABIC LETTER HAMZAH ON ALEF
آ	0622	ARABIC LETTER MADDAH ON ALEF
ب	0628	ARABIC LETTER BAA
پ	067E	ARABIC LETTER TAA WITH THREE DOTS BELOW = peh
ت	062A	ARABIC LETTER TAA
ٹ	0679	ARABIC LETTER TAA WITH SMALL TAH
ث	062B	ARABIC LETTER THAA
ج	062C	ARABIC LETTER JEEM
چ	0686	ARABIC LETTER HAA WITH MIDDLE THREE DOTS DOWNWARD = tchah
ح	062D	ARABIC LETTER HAA
خ	062E	ARABIC LETTER KHAA
د	062F	ARABIC LETTER DAL

Figure 2: Number Symbols

The remaining details of Unicode Charts can be available at <http://www.unicode.org/charts/>

Urdu Unicode characters [2] are mapped from Arabic range (0600-06FF) Hex, extended Presentation 'A' ranges (FB50-FDFF) Hex & Presentation 'B' ranges (FE70-FEFF) Hex along with Persian, Sindhi, Pasto, Kurdish etc. languages of Iran, Pakistan, and India. The [3] Urdu Computing Standards: Development of Urdu Zabta Takhti is used for Urdu character sets & keyboard design layout.

1.2. Open-Type Fonts:

This algorithm uses Unicode Character representation for Urdu which is available in open-type fonts or those fonts that incorporated with the ISO/IEC 10646 Unicode Support. Currently, available open-type fonts are:

- Nafeez Naskh
- Tahoma
- Times New Roman
- Arial
- Nafees Nastaleeq

4.3. Step V: Finding corresponding punctuations in array

```

1st Loop { // To get string of punctuation
  2nd loop { //To check each character with
    If this character from Unicode set then
    substitute with corresponding character
    and break if found
  } // 2nd Loop
    includes loops & various check
} // 1st Loop

```

5. Sentence Checking (Semantic & Syntactic)

When words of sentences get its appropriate meaning & form, the algorithm checks the way they written. In other words our engine tries to understand the sentences according to the language grammar or its meaning. The sentences checking are performed by set of rules, which accordingly construct the sentence into translated language. These rules are divided into two categories. (1) Semantics that involve the meaning of words. (2) Syntactic that involves don't involve the meaning of words. In this module, we encounter the sentences at both levels, however for semantic process the sentences are translated before syntactic from a separate database. This database consists of idioms, phrases & proverbs. The user can also add/subtract them from semantic database.

For example: Time flies like an arrow.

According to syntactic rules, it translate
وقت کی مکیاں پسند کرتیں ہیں تیر کو (Wrong)

But semantically, it will translate as;
وقت پانی کی رفتار سے گزر رہا ہے (Right)

5.1. Syntax Module.

The sentences governed by Context Free Grammar with [5] Active and Passive forms of Positive and Negative sentence are modeled. Along with Adverbial Phrases coming at beginning, last and middle of the sentence & Infinitive Verb Phrase (to VERB) is modeled.

- <S> → <PreNP> <NP> <VP> <PostVP>
- <NP> → <PreNP> Noun <PostNP>
- <PreNP> → <Adjectives>
- <PostNP> → <PP>
- <PP> → Prep <NP>

5.2. Sentence Rules.

- Some of rules [6] to check the syntax of sentences are:
- [1] A sentence can be a subject followed by a predicate.
 - [2] A subject can be a noun-phrase.
 - [3] A noun-phrase can be an adjective followed by a noun-phrase.

- [4] A noun-phrase can be an article followed by a noun-phrase.
- [5] A noun-phrase can be a noun.
- [6] A predicate can be a verb followed by a noun-phrase.
- [7] A noun can be: apple bear cat dog
- [8] A verb can be: eats allows gets hugs
- [9] A adjective can be: itchy jumpy
- [10] An article can be: a an the

Consider: "The itchy bear hugs the jumpy dogs"

- From syntax rules, this sentence can be generated as:
- Sentence => subject predicate Rule 1
 - => noun-phrase predicate Rule 2
 - => noun-phrase verb noun-phrase Rule 6
 - => article noun-phrase verb noun-phrase Rule 4
 - => article adjective noun-phrase verb noun-phrase Rule 3
 - => article adjective noun verb noun-phrase Rule 5
 - => article adjective noun verb article noun-phrase Rule 4
 - => article adjective noun verb article adjective noun-phrase Rule 3
 - => article adjective noun verb article adjective noun Rule 5
 - => the adjective noun verb article adjective noun Rule 10
 - => the itchy noun verb article adjective noun Rule 9
 - => the itchy bear verb article adjective noun Rule 7
 - => the itchy bear hugs article adjective noun Rule 8
 - => the itchy bear hugs the adjective noun Rule 10
 - => the itchy bear hugs the jumpy noun Rule 9
 - => the itchy bear hugs the jumpy dog Rule 7

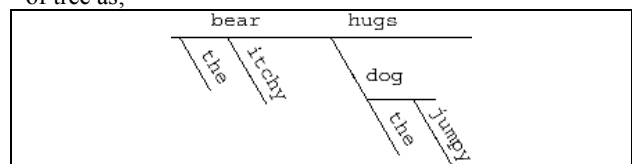
The rules allow various possibilities to translate and may also require limitation of semantics or good sense.

6. Sentence Parsing with NLP.

After the formation of grammar, there is need for processing sequences of phrases in a tree. NLP actions are embedded within the grammar to effect a translation. It takes a list of tokens as its argument. With trees consists of a node value, and one or more children; it uses bottom-up parsers on the current state of its knowledge about the constituents. A Tree consists of a node value i.e. string containing the tree's constituent type (e.g., "NP" or "VP") and one or more children that encode the hierarchical contents of tree. Each child is either a leaf or a sub-tree.

6.1. Step VI: Sentence – (Positioning & Counting) with NLP

The sentences position and counting is applied for its arrangement. They are parsed according in the form tree of tree as;



The NLP (Natural Language Processing) for sentences [5] is performed in the following phases.

6.1.1. Partial Parsing.

The system use bottom up chart parser that makes partial parsing possible. Hence it can deal sentences which have some small error (or the sentences that are not according to the grammar.)

For example: *I know him He lives here.*

6.1.2. Transformational Module.

Parse Structure from Syntactical Module is traversed and the Translation is built by re-arrangement and inflection of words and phrases. If more than one parses are generated by Syntactical Module, then it uses Heuristics for best interpretation.

- If auxiliary verb is used as main verb, it has negative weight.
- If adjective is used as noun, it has negative weight
- If verb is used as noun, it has negative weight.

6.1.3. English & Urdu Comparison.

The comparison of English & Urdu is applied as SVO (Subject Verb Object) and SOV (Subject Object Verb). It also includes order of words in phrases, many forms of adjective and prepositions, many forms of verb, object preposition and subject preposition.

English is Subject -Verb-Object Language.

Hamid writes a letter.

Urdu is Subject-Object- Verb Language.

حامد خط لکھتا ہے

6.1.4. Order of words.

Order of words for English:

<PrepPhrase> → Prep <NounPhrase>

Example: of red color.

Order of words for Urdu:

<PrepPhrase> --> <NounPhrase> Prep

Example: لال رنگ کا

6.1.5. Many Forms of Adjective and Prepositions

Blue Book, Blue Books, Blue Pen, Blue Pens

نیلی کتاب، نیلی کتابیں، نیلا قلم، نیلے قلم

Price of Book, Writer of Book

کتاب کی قیمت، کتاب کا مصنف

Blue Color

نیلا رنگ

Book of Blue Color

(Wrong) نیلا رنگ کی کتاب
(Right) نیلے رنگ کی کتاب

6.1.6. Many Forms of Verb.

It's a rule based system for verb inflection, inflection forms of verb (can) depends on tense of sentence, gender, number and person of subject or object, transitive and intransitive verb subject preposition and object preposition. For example:

Verb form depends on subject (gender, number and person) and tense

عورت کتاب خریدتی ہے

آدمی کتاب خریدتا ہے

Verb form depends on object (gender, number and person) and tense

آدمی نے کار خریدی

آدمی نے مکان خریدا

Verb form depends on verb gender and tense

آدمی نے عورت سے بات کی

عورت نے آدمی سے بات کی

6.1.7. Subject Preposition and Object Preposition.

Used in Past Indefinite Tense having Transitive Verb Commonly نے is used with Subject & کو is used as Object

اس نے تم کو بتایا

In some cases, other prepositions like سے can be used.

اس نے تم سے پوچھا

Presence and absence of object preposition depends on sense (semantic type) of verb.

He asked you

اس نے تم سے پوچھا

He asked a question

اس نے ایک سوال پوچھا

6.2. Step VII: Putting Text

From String target_text → into textarea

References

- [1] The Unicode Standard Consortium, "Urdu / Arabic Presentation Forms A & B", Version 4.0, 2004
- [2] Dr. Khaver Zia, "Towards Unicode Standard for Urdu", National Urdu Seminar 2002, KU/KHIT/Tremu, WG2 N2413-1/SC2 N35891
- [3] Dr. Sarmad Hussain and Dr. Muhammad Afzal, "Urdu Computing Standards: Development of Urdu Zabta Takhti", Proceedings of IEEE – INMIC 2001, pp: 223-228
- [4] Terence John Parr, Language Translation, MageLang Institute, Automata Publishing 1989
- [5] Tafseer Ahmed, "Urdu Translator", Mukhtadara Quami Zuban ka Tarjuman, Urdu Zaabtah Tahqeeq 2002.
- [6] Danial I.A. Cohen, Automata Theory, Hunter College, City University of New York, 1991