Heliyon 5 (2019) e01780

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Heliyon

Transtech: development of a novel translator for Roman Urdu to English

Hafsa Masroor [a], Muhammad Saeed [a], Maryam Feroz [a], Kamran Ahsan [b], Khawar Islam [b,*]

[a] UBIT - Umaer Basha Institute of Information Technology, University of Karachi, Pakistan
[b] Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Karachi, Pakistan

ARTICLE INFO

Keywords:
Computer science
Linguistics

ABSTRACT

Advances in machine and language translation immerge new fields and research opportunities for researchers, whereas Natural Language Processing and Computational Linguistics deal with communication between natural languages and their interaction. The objective of this research is to develop and test a novel tactic to solve the issue of translation from Roman Urdu to the English language. The approach used to construct this practical model is divided into three stages; each stage works out to achieve its desired task. Self-maintained corpus alongwith its corresponding tag-set is used for tokenization. The syntactical structure is covered by writing Urdu POS tagger based on grammatical rules. We prepared the grammatical structures of different sentences for Roman Urdu to English translation. Since Roman script can be expressed in numerous ways, our grammatical structures fulfill the maximum possible needs of writing and produce the best possible English translation. We entered a sentence in Roman Urdu which gave the best possible translation in the English language. In comparison with Google Translator, Transtech worked better and gives more accurate results.

1. Introduction

Natural Language Processing is associated with natural languages and machine translation. It digs into the idea of how computers can help interpret routine sentences or phrases to produce beneficial outputs. NLP analyst plan to collect data about how people figure out and interpret language so that relevant approach and techniques can be created so that computers can manipulate and manage such languages to execute required tasks [1]. Applications of NLP cover a various perspective of study, for example, machine translation, multilingual and CLIR, speech recognition, artificial intelligence and decision support systems [2]. On another hand of machine translation is Computational Linguistics that is an integrative area of science which involves the statistical or rule-based modeling of natural language from a computational angle. It revolves around the domains of cognitive sciences, artificial intelligence, mathematics and theoretical linguistics [3]. Translation is the procedure of converting the content of one language to another, such that its significance does not change. It can be applied to written documents or in verbal communication. The primary objective of translation is to make the connotation of the source and targeted language equivalent. The importance of translation in our routine life is largely structural. Translation leads a path towards worldwide communication as well as gives access to nations to create relationships in order to lead towards technological growth, political and cultural advancements etc. [4]. All the important information for translation has been collected to translate the Roman text into the English language. Since Roman Urdu does not follow any regular standard and can be illustrated in several ways, so rule-based translation has been followed in for which dozens of grammar rules were built to implement them in a POS Tagger. Moreover, many words in Roman Urdu do not follow any specific spelling pattern and can be spelt in different ways. With a view to solve this problem, we have maintained a collection of the corpus in a knowledge base [5], in which maximum possible words are saved, and occurrence of each word in the input string is matched with all the similar words of our knowledge base.

Fig. 1 illustrates the essential steps and overview of the translator from the input source to output translation. Section 2 provides a literature review of Urdu language that we take as a sample to build Roman Urdu approach. Sect. 3, describes the method of data collection and construct a knowledge base model for a novel translator. In Sect. 4, the description of translator along with its components and how we process Roman Urdu data, normalization of text and translation from Roman Urdu English language is shown. Sects. 5 and 6 show the working of the translator with the involvement of constructed algorithm for Roman Urdu. Finally, we have discussed the results of Google translator and Transtech.

Previously, no research has been done to translate Roman Urdu language to the English language because of no attention of research

* Corresponding author.
E-mail address: khawarislam@fuuast.edu.pk (K. Islam).

https://doi.org/10.1016/j.heliyon.2019.e01780
Received 14 September 2018; Received in revised form 24 March 2019; Accepted 16 May 2019

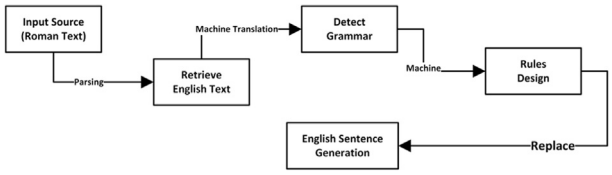

Fig. 1. A systematic overview of Roman Urdu Translator.

communities and lack of Roman Urdu resources like linguistic etc. Limited research papers were written for the translation of Urdu language into the English language which highlighted its associated problems. Most of them are focused on translation with specific wordlist one [6, 7, 8, 9]. The contribution of this paper is to develop a novel translator that converts Roman Urdu to the English language which gives the benefit to 11 million people. Key features of this translation process include:

- Spell checking with the help of a self-maintained dictionary
- Learning and inclusion of new words into Knowledge Base
- Urdu Parts of Speech tagging at runtime
- Syntax and semantic checking of grammar
- Corpus collection of Roman Urdu words
- Context Free Grammar for generation of production rules

2. Related work

We summarized all the researches and studies developed for Urdu translation. We reviewed not only Urdu translation, but also POS tagging method that provides more information on language translation. In our work, we have performed translation from Roman Urdu to the English Language, as no previous work is found to solve this problem (Section "4" refers to how we achieve language translation). Next, we studied different papers in Urdu language translation to relate papers designed for the tagging, and translation for different languages. Computational Linguistics and Data Mining tasks, like sentiment analysis, textual entailment, information extraction, topic segmentation and parts of speech tagging include a brief study of NLP. The significance of NLP in the speech processing area, such as learning phrases in machine translation, cognitive modelling, tera-scale language models, multi-task and incremental processing with neural networks and language resource extraction have critical significance in all NLP frameworks [10]. NLP frameworks for the English language are very strong and developed; however, Urdu NLP frameworks needs a lot of efforts and research to achieve a mature framework [11]. The national language of Pakistan is Urdu. According to [1], 11 million people in Pakistan and almost 300 million people from the whole world speak Urdu. As we know, English is the most common and widely spoken language of the world. Almost all the official documents are written and drafted in English [12, 13]. It has been crowned as the language of global business. After all, the English language holds such paramount importance in the global era. Therefore it is a big necessity to translate our language into English. In Asia, Urdu is the premier language for writing literature and poetry [14, 15]. Its multiple levels of politeness and meanings have been manipulated by poets for centuries to create beautiful and memorable verse. Such relevant facts depict the importance of Urdu to English Translation. The people of Pakistan prefer Urdu writing in Roman Urdu. The recent survey [1], indicates that 80% of people of Pakistan uses Roman Urdu. The effects of Roman Urdu are to decrease the capability of writing English and Urdu [16]. stated the first work on Urdu stemming and developed a new directive called *Assas-Band*. The incredible work has been done by [5], who created a dataset for Arabic Urdu script that contains two main things, one is XML format, and other is Unicode character. CLE Pakistan [17] has also developed a corpus which contains 100K Urdu words from different areas, including, education, health-related, training, etc. It contains two major categories; one is informational with 80% and second is imaginative with 20% [18]. developed a large corpus based on spoken and text Urdu. This corpus contains spoken words of about 512,000 and around 1,640,000 Urdu text words.

Table 1
Corpus collection of Roman Urdu data.

	Tum konse bazar jati thi				
English	You	which	market	go	did
Roman Urdu	Tum	konse	bazar	jati	thi

3. Materials

Data collection is always a challenging part of any research. Since Roman Urdu language is quite diverse and has got many variations, therefore it is quite difficult to cover all the grammatical aspects of Urdu language. So, we have chosen a particular domain which is going to cover the basic elementary tenses of the English language, along with their affirmative, negative and interrogative sentences. Moreover, we have also covered WH Questions and imperative sentences in our grammar. Table 1 shows an example of one sentence, of how we break it into words and achieve Roman Urdu translation.

3.1. Corpus collection

With the help of [17, 18], the target size for the corpus required for translation is around 3000 words and over 2000 different sentences. This corpus is analyzed within the research to develop the translator. This corpus is supposed to give linguists the possibility to understand different aspects of Roman Urdu language.

3.2. Knowledge base model

We have built the knowledge base model for gathering and maintaining the corpus required for the translation process. In this knowledge base, a data table for wordlist is created in which all information mandatory for syntax and semantic analysis is saved, such as the word, POS tag, its corresponding meaning and type needed for translation.

3.3. Context-free grammar

A context-free grammar contains a set of rules which determines the syntactic structure of any language. It consists of terminals (POS tags) and non-terminals, which generates a set of production rules. Several rules of CFG have been written for this translator that covers multiple tenses of Roman Urdu/English language.

4. Methods

It is a difficult task to develop an algorithm for translation of Roman Urdu to English language and work very effective in translating into another language. Expressively, the languages which have a large number of words and grammatical rules give many problems. To overcome this issue, we surveyed among people and collected words to achieve an accurate result then the translation needs more time to give the best answer. Hence, our target is to give the best answer to the user which is nearest with his typing and current context and can easily understand. We developed an algorithm which provides the translation of Roman Urdu which is not approximately accurate for complex sentences. The algorithm of language conversion is given below.

Step 1: Get Roman Urdu sentence as an input from the user.
Step 2: Split input sentence into words and determine its POS tag.
Step 3: Pass the tagged data to the machine translator as an input parameter.
Step 4: Find English words according to Roman Urdu words.
Step 5: Tokenizing each sentence
a. Check speech tagging parts.
b. Make division in chunks and generate a parse tree.
c. Find an appropriate set of grammatical rules.
d. Rearrange the English words based on rules.
Step 6: Print output sentence in English.

5. Methodology

Translation is the process of converting source language (Roman Urdu) into the target language (English). A translator consists internally of some phases; each performs its designated task to carry out the perfect translated output in the English language. It is helpful to think these phases as separate modules within the translator, and they may indeed be written as separately coded operations although in practice they are often grouped. This research is divided into three basic phases, each of which performs its own analytical and logical operations. Fig. 2 describes the internal view of language translation. It shows a conversion process of Roman Urdu into the English language.

5.1. Scanner

The scanner is the first phase of Transtech. It performs an absolute reading of source language (Roman Urdu), which is in the form of the input string. The scanner performs lexical analysis and tokenization. It converts the input string into a sequence of meaningful units called tokens which are the actual words of Roman Urdu. It does this by simply splitting the string of sentence on single space. The input of this module is a string of sentence in Roman Urdu, and the output generated is a stream of Tokens.

5.1.1. Spell checker and learning agent

The spell checker is embedded along with the scanner which performs spell checking of the tokens with the assistance of data available in the knowledge base model. For this purpose, the Levenshtein distance algorithm has been used, which calculates the degree of similarity between two strings. This distance is calculated by analyzing different number of letters among source and targeted strings. When the entered word is not available in the dictionary, it suggests the list of similar words, determined with the help of the mentioned algorithm. On arrival of a new word, the user is asked to add it along with its necessary linguistic information in the knowledge base, thus making this translator a learning agent as well.

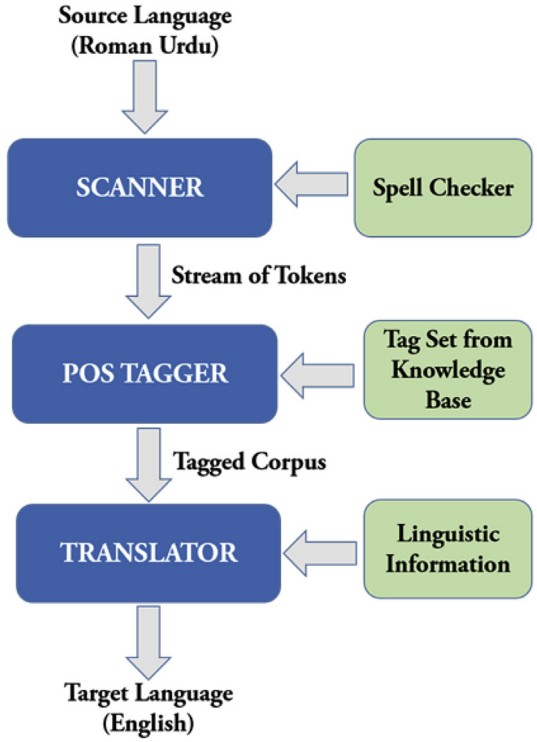

Fig. 2. Internal view of Roman Urdu Translator.

5.2. POS tagger

Parsing is the task of determining the syntax of an input sentence. The syntax of any language is usually given by the grammar rules of a context-free grammar. The basic structure used is some kind of tree, called a parse tree or syntax tree. Syntax analysis has been performed by implementing LL(1) parser along with POS Tagger. It is the procedure of allotting each word in a sentence the part of speech that it assumes to be in that sentence. The input of POS Tagger is a stream of Tokens, which are assigned its linguistic information at runtime by parsing through the syntax of grammar rules.

Consider the following sentence that has been parsed through the syntactic rules, and the tagged corpus has been generated by the POS Tagger.

Areeba/NNP *khamoshi*/RB *se*/PSP *apna*/APNA *kaam*/NN *kar*/VBF *rahi*/AUXTR *hai*/AUXTT.

5.2.1. Urdu parts of speech tag set

The following Tag Set from Center for Language Engineering [19] has been used to implement Urdu POS Tagger in Transtech (see Table 2).

5.3. Translator

It is the third phase of Transtech which performs meaningful type checking and semantic analysis. The input in this phase is a tagged corpus which with the help of linguistic information translates the sentence into the English language. Actual translation process of Transtech is carried

Table 2
The list of tag set for Urdu POS tagger.

S. No	Categories	Types	POS tag
1	Noun	Common	NN
		Proper	NNP
2	Verb	Main Verb Infinite	VBI
		Main Verb Finite	VBF
3	Auxiliary	Aspectual	AUXA
		Progressive	AUXP
		Tense	AUXT
		Modals	AUXM
		Present Tense	AUXIT
		Past Tense	AUCTP
		Future Tense	AUXTF
		Perfect Tense	AUXTC
		Continuous Tense	AUXTR
4	Pronoun	Personal	PRP
		Demonstrative	PDM
		Possessive	PRS
		Relative Demonstrative	PRD
		Relative Personal	PRR
		Reflexive	PRF
		Reflexive APNA	APNA
5	Nominal Modifier	Adjective	JJ
		Quantifier	Q
		Cardinal	CD
		Ordinal	OD
		Fraction	FR
		Multiplicative	QM
6	Adverb	Common	RB
		Negative	NEG
7	Ad Position	Preposition	PRE
		Postposition	PSP
8	Interrogative	WH Question	WH

Table 3
Comparison between Google translator & transtech.

Roman Urdu	Google 2017	Google 2019	Transtech
Tum konse bazar jati thi	You are what the market	What market did you go?	Which market did you go
Wo bohat achay kapre pehnti hai	She wears nice clothes many	Wear good clothes	She wears very good clothes
Imran waqt par ghar nahi pohanchta hai	He does not come home on time	Imran does not know home at time	Imran do not reach home on time
Ali ajkal bohat pareshan hai	Many consignment Ali today	Eli is a booming trend today	Ali is very upset now-a-days
Areeba khamoshi se apna kaam kar rahi hai	Areeba quietly doing its job	Aurabagh is doing his job quietly	Areeba is doing work silently

out in this phase. The modular approach has been followed to parse the input sentence through CFGs, which invokes different modules for semantic checking and English translation. Each module is designed to carry out the specific task, functioned with the help of grammatical rules and linguistic information. The most important module in the translation phase is the one which deals with verbs. Since in Urdu, one verb can be replaced with multiple English verbs, so it is the task of this module to determine the best possible verb according to the given sentence. It also determines the type of verb with the help of available data set and logical operations for all of its kinds. It performs the determination of pronoun as well, which is carried out with the help of leading verb in Urdu sentence. It also judges the gender and measure of a referred noun to set the best possible pronoun (he/she/it/they). Another key module of this phase examines the noun phrase. It performs quantitative analysis to determine the singular/plural information of noun, which is useful for choosing an appropriate helping verb (is/am/are/was/were) for accurate translation. Different parts of speech tags like adjectives, adverbs, pronouns and cardinal numbers are covered as well. Multiple submodules are designed which performs extraction and necessary operation required for these tags. Appropriate prepositions are also set according to the semantic information present in the sentence.

Negative, interrogative and imperative sentences are also covered, which requires the functioning of different sub-modules. WH questions are handled as well in the domain of elementary tenses. If the input sentence contains any WH tag in Urdu, it performs semantic logic to set the who, why, where, what and how accordingly.

6. Results

Table 3 describes the efficiency of Transtech as compared to Google Translator. It is clearly shown that Transtech gives much better and accurate results. It also shows the improvement of the Google machine translation system that has been made during the last two years.

7. Discussion & conclusion

We have developed a translator for Roman Urdu to the English language, which provides the best translation with maximum accuracy. Though it was challenging since Roman Urdu language does not follow any regular grammatical pattern and can be represented in different ways. Therefore we followed rule-based translation and developed various grammatical rules to carry out the process of translation in a tagger. Furthermore, several words in Roman Urdu can be spelt in various ways since there is no hard and fast rule for spellings in Roman Urdu grammar. Therefore, we managed a collection of the corpus in a knowledge base to accommodate maximum possible words and match the occurrence of each word in an input string with all the similar words of our knowledge base.

Some cases of natural language problem have been left for the future due to lack of time and unavailability of a large amount of data. Future work includes in-depth analysis of the proposed mechanism to handle complex Urdu/English sentences, different variations of same word and inclusion of more grammatical rules and vocabulary in the dataset. Translation process could also be improved by involving machine learning approach, which could train the system on the basis of its current performance.

Declarations

Author contribution statement

Hafsa Masroor, Maryam Feroz: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Muhammad Saeed: Conceived and designed the experiments.

Kamran Ahsan: Performed the experiments.

Khawar Islam: Analyzed and interpreted the data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] Daud Ali, Wahab Khan, Dunren Che, Urdu language processing: a survey, Artificial Intelligence Review, 2017, pp. 279–311.
[2] Tafseer Ahmed, Annette Hautli, Developing a basic lexical resource for Urdu using Hindi WordNet, in: Proceedings of CLT10, 2010.
[3] Qaiser Abbas, Semi-semantic part of speech annotation and evaluation, in: Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop, 2014.
[4] K. Visweswariah, V. Chenthamarakshan, N. Kambhatla, Urdu and Hindi: translation and sharing of linguistic resources, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, August, pp. 1283–1291.
[5] Dara Becker, Kashif Riaz, A study in Urdu corpus construction, in: Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization, 12, Association for Computational Linguistics, 2002, pp. 1–5.
[6] F. Adeeba, S. Hussain, Experiences in building the UrduWordNet, in: Proceedings of the 9th Workshop on Asian Language Resources, 2011, pp. 31–35.
[7] R.E.O. Roxas, S. Hussain, K.S. Choi, Proceedings of the 9th workshop on asian language resources, in: Proceedings of the 9th Workshop on Asian Language Resources, 2011.
[8] N. Durrani, S. Hussain, Urdu word segmentation, Annual Conference of the North American Chapter of the Association for Computational Linguistics, in: Human Language Technologies, 2010, pp. 528–536.
[9] A.K. Pandey, T.J. Siddiqui, Evaluating effect of stemming and stop-word removal on Hindi text retrieval, in: Proceedings of the First International Conference on Intelligent Human Computer Interaction, Springer, New Delhi, 2009, pp. 316–326.
[10] D.E. Kieras, M.A. Just, New Methods in reading Comprehension Research, Routledge, 2018.
[11] Y. Li, T. Yang, Word embedding for understanding natural language: a survey, in: Guide to Big Data Applications, Springer, Cham, 2018, pp. 83–104.
[12] Al-Shammari, Eiman Tamah, Jessica Lin, Towards an error-free Arabic stemming, in: Proceedings of the 2nd ACM Workshop on Improving Non-English Web Searching, ACM, 2008.
[13] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[14] K. Riaz, Baseline for Urdu IR evaluation, in: Proceedings of the 2nd ACM Workshop on Improving Non-English Web Searching, ACM, 2008, October, pp. 97–100.
[15] Ali Daud, et al., Knowledge discovery through directed probabilistic topic models: a survey, Front. Comput. Sci. China 4 (2) (2010) 280–301.
[16] Qurat-ul-Ain Akram, Asma Naseer, Sarmad Hussain, Assas-Band, an affix-exception-list based Urdu stemmer, in: Proceedings of the 7th Workshop on Asian Language Resources, Association for Computational Linguistics, 2009.
[17] CLE, Urdu Digest POS Tagged Corpus, 2015. http://www.cle.org.pk/software/localization.htm.
[18] A. Hardie, Developing a tagset for automated part-of-speech tagging in Urdu, in: Corpus Linguistics 2003, 2003.
[19] CLE, Urdu Parts of Speech (POS Tagset), 2013, in: http://www.cle.org.pk/Downloads/langproc/UrduPOStagger/Urdu%20POS%20Tagset%200.3.pdf.