Issues in Urdu-Hindi NER Output of Google and Bing Translator: An Orthographic Perspective

Md. Tauseef Qamar, Juhi Yasmeen

***Abstract:** Named Entity Recognition (NER) is a sub-task of information extraction in which names are extracted both from the text and linguistic corpora which is still a tough nut to crack for NLP researchers in existing Machine Translation (MT) system due to its long tail. Since decades, NER has been an area of great interest both in MT and computational linguistics, thus, several tools have been designed for their handling in different languages. Therefore, this paper aims to compare the end user output of both Google and Bing translator with special reference to Urdu-Hindi NER. This will provide more insights in the development of intelligent language tools. Thus, on the one hand, the paper deals with orthographic challenges pertaining to Urdu-Hindi NER in general, while on the other hand, the paper also sheds light on the transliteration issues in particular. Further, we have also investigated the personal names, and named entity of Urdu, especially ezafat constructions. Consequently, the paper also proposes to handle NER from the language engineering point of view based on the existing end user output quality. Furthermore, the MT output of both Google and Bing has been ranked on the scale of 0 to 1, where 0 assigned to the correct output while 1 given to the wrong or inaccurate output.*

***Keywords:** Named Entity Recognition, Urdu Orthographic Challenges, Ezafat, Googl and Bing NER Urdu-Hindi Output.*

I. INTRODUCTION

Humans speak a number of different languages around the globe for the effective communication, and notably they varies greatly from each other in a number of ways. Apart from significant differences, there are certain linguistic components shared by all languages, and one among them is *NER* or *Naming Entity*. Thus, researchers from different fields like; linguistics, literature, and computer science, etc. study and research human language to unfold its linguistic properties and their underlying specialties. Similarly, human language has also drawn the attention of technology tycoons like; Google and Microsoft, in addition to computer science engineers. As a result, a number of intelligent language tools have been designed to process the human language under the hood of natural language processing (NLP), some of them are; Machine Translation (MT), Text to Speech (TTS), and Voice Assistants, etc. Significantly, in NLP the NER is one of the vast and active area of research for the last 25 years.

NER is a sub-task of information extraction whereby names are extracted and classified in a text because it plays an extremely crucial role in NLP, especially in MT. Similarly, MT is a sub-field of computational linguistics whereby the meaning of source language (SL) gets converted into an equivalent meaning of target language (TL). Therefore, it is imperative to deal with the crucial role of NER while generating the end user output. Evidently, several tools have designed for the processing of NER in resource rich language which produces high accuracy, for example, the English NER extraction tool. This tool significantly produces high accuracy in terms of end user output. But, still, there are languages whose corpora is not sufficient enough to process NER with high accuracy. For example, Urdu and Hindi especially from the orthographic perspective (mainly the glottalic/vocalic sound loaned from Arabic to Urdu) and chiefly those names made of ezafat. Therefore, the primary objective of this paper is to deal with the need for diacritics and ezafat adoption in Hindi as per the systems of Urdu language. As a result, we also propose careful attention to the transliteration which may pave a path for better transliteration output in generating the desired end user output in terms of ezafat into the TL, i.e. Hindi. Further, transliteration is a process where the phonological characters of SL gets transferred into the equivalent phonological character of TL.

Significantly, this paper proposes to identify the need and consequent challenges of NER transliteration in Urdu-Hindi scenario. These needs and challenges can be adopted to improve the inaccuracies in existing Google and Bing translator’s end user output. Since NER is a vast area, so existing paper aims to cover only Urdu ezafat names. Further, in order to test our collected names, we have translated them on both the translation platforms (Google and Bing) which shows noticeable inaccuracies in existing end user output, i.e. Hindi. The existing end user output is not satisfactory both from orthographic (vocalic/glottalic sound and ezafat) and translation purposes. Furthermore, the existing output also presents a clear picture that the inaccuracies in end user output resulted due to the homographic (especially from diacritics ‘ehrab’ point of view) challenges of Urdu in general and ezafat in particular.

This paper is organized in the following ways: section one (I) deals with the introduction about the NER from orthographic perspective in general, while section two (II) sheds light on the related work to NER in Urdu and Hindi in particular in addition to English. The subsequent section three (III) and four (IV) unfolds the orthographic challenge and ezafat (including its types) in the source language (Urdu), respectively. Further, the section five (V) presents the overall picture about the existing inaccuracies in end user output of Google and Bing translator with special reference to ezafat transliteration, while section six (VI) focuses on the handling of NER from a linguistics point of view. Furthermore, sections seven (VII) and eight (VIII) demonstrates the existing output of Google and Bing followed by the discussion, sequentially. Finally, section nine (IX) outlines the concluding remarks about this paper.

Revised Manuscript Received on November 22, 2019

Md. Tauseef Qamar, Ph.D. Scholar, D/O Linguistics, AMU, Aligarh tauseefqamar007@gmail.com

Dr. Juhi Yasmeen, Ph.D. in Linguistics, AMU, Aligarh juhi1421@gmail.com

Published By:
Blue Eyes Intelligence Engineering
&Sciences Publication

II. RELATED WORK

Significantly, NER has drawn a great amount of research interests by NLP research especially in the last decade (Chinchor, 1995, 1998). Among the few notable tasks in NER extractions are English-Spanish NER by Bikel et al. (1999), Urdu NER by Kashif (2010), Bengali NER by Ekbal et. al. (2008), Hindi NER by Saha et al. (2008), and so on. Further, existing literature also reveals that where there is so many works have been done on other aspects of Urdu NER but a little amount of attention has been given to the ezafat names. Consequently, Urdu NER end user output is lagging in terms of ezafat in Urdu-Hindi MT. Furthermore, already published work also reveals that there are two classical approaches used for NER extraction mentioned below:

i) Linguistic Approach: This approach purely works on the rule-based technique, whereby rules of the languages are hand written by the linguists and grammarians.

ii) Machine Learning Approach: This approach works on the technique in which a large amount of annotated data for the acquisition of high level language specific knowledge is used which results in obtaining high accuracy of end user output. There are number of significant machine learning tools designed for the acquisition of NER: Hidden Markov Model (HMM), Maximum Entropy Model (MaxEnt), Decision Tree, Support Vector Machines, and Conditional Random Field (CRFs).

III. ORTHOGRAPHIC AND TRANSLITERATION ISSUES IN URDU-HINDI MT

Notably, the sound system of Urdu differs slightly from Hindi in both ways: place and manner of articulation, especially in terms of vocalic and glottalic sounds (loaned from Arabic). This invites the sound system asymmetry between the two language. Consequently, Urdu retains vocalic/glottalic sounds loaned from Arabic (ع غ ق ح خ). Therefore, to present their sound effect diacritic /./ is used in Hindi also known as Halanta. Further, there are characters in Urdu which has multiple representation while Hindi has only one. For example, instead of Urdu /ا/ and /ع/ Hindi has only /अ/, similarly, for /غ/ and /گ/ Hindi has only /ग/ and so on. Therefore, such sounds are frequently used in Urdu names belonging to the persons, places, and organizations, which exerts influence on Hindi sound systems to retain both vocalic and glottalic sounds, especially from translation and transliteration point of view. Additionally, Urdu also uses ezafat loaned mainly from two languages Arabic and Persian. Evidently, 'ال' and 'و' are of Arabic origin, for example, قمر الدین 'qamaruddin' and فکر و نظر 'fikr-o-nazar', while /ِ/ or /e/ (usually not written but spoken) taken from Persian, e.g., وزیر اعظم (wazeer azam) 'prime minister'. Consequently, due to the need for maintaining sound effect and reader's effect Hindi also uses ezafat and presented by 'ए' (मुगल ए आज़म). Therefore, the handling of NER in Urdu-Hindi MT system becomes important mainly from two following perspective:

i) Writing Style/System

There are two significant points which are important to mention from writing point of view: a) firstly, Urdu is written from right to left and there are letters called joiners (total number of such characters is 27) which conjoins with either preceding or the following character, and b) secondly, there are words in Urdu which are written differently and pronounced differently due to its writing style/script i.e., Perso-Arabic, which is one of the key challenge in Urdu to be handled in target language (Hindi). For example, written as خواه مخواه (khawaah makhwaah) but spoken as 'khaah makhaah'. There are plenty of such examples available in Urdu.

ii) Multiple Characters/Alphabets

This appears to be another crucial issue in Urdu-Hindi which needs a significant amount of attention. Notably, the issue of multiple characters of Urdu needs to be solved mainly from translation point of view. Further, there are eight (8) characters of Hindi which has multiple representations in Urdu, presented in the following table:

Table 1: Showing the Multiple Characters of Urdu

Hindi Characters	Urdu Characters
अ	ا ع
क	ک ق
त	ط ت
स	س ص ث
ह	ه ح
ज	ز ض ظ ذ
ं	ن ں
ँ	ن ں

IV. HURF-E-IZAFAT (EZAFAT) IN URDU

The concept of ezafat has become more popular in 20th century among the languages like; English and Hindi due to the emergence of translation. Chiefly, the concept of ezafat in Urdu has been borrowed from Persian and used at the end of the first word which is presented by zer /ِ/ but not necessarily written. The native speaker realizes its articulation at the time of reading (examples are presented in table 2). While in Arabic, it stands for genitive formation, whereas in Urdu, it refers to the 'addition' or 'extension'. In Urdu, Ezafat acts as an enclitic short vowel 'e' which conjoins either two nouns or an adjective and a noun (Schmidt, 2004). This also acts like a possessive marker in Urdu. Generally, there are two types of ezafat structure found in Urdu according to their functions, i.e., noun+e+noun and adjective+e+noun. Further, noun+e+noun mainly deals with possessive affinity where first noun is dependent to the second or following noun in terms of their meaning. In the following examples both the combination of ezafat has been presented, respectively:

Noun+Noun Ezafat

Table 2: Showing the Examples of Noun+Noun Ezafat in Urdu

Ezafat	Transliteration	Gloss
جشن ریختہ	Jash-e-rekhta	Celebration of Rekhta
حکومت ہند	Hukumt-e-hindustan	Government of India
دعوت ولیمہ	Dawat-e-waleema	Invitation of reception

Adjective+Noun Ezafat

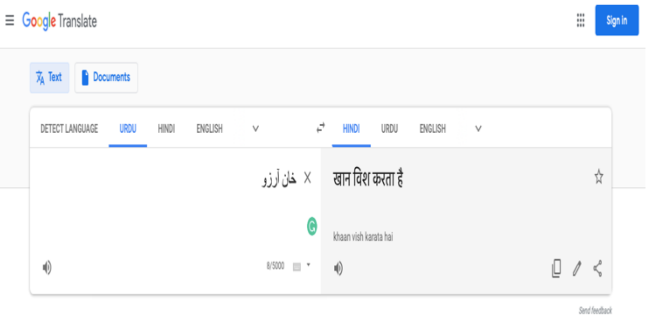
Table 3: Showing the Examples of Adjective+Noun Ezafat in Urdu

Ezafat	Transliteration	Gloss
وزیر اعظم	wazeer-e-azam	Prime minister
اہل زبان	Ahl-e-zaban	Native speakers

The present examples in table 2 above clearly shows that in Urdu ezafat is presented without diacritic 'zer' but its native speaker realizes their articulation in the speech from the surrounding words that it is an ezafat construction. However, it is also important to note that sometimes the diacritic zer is written but not necessary in general. Such an instance can be observed from the example mentioned in table 3. Notably, Hindi also uses ezafat especially in the names borrowed from Urdu, Persian, and Arabic languages, for example, a famous movie name مغل اعظم 'mughl-e-azam'. Therefore, it becomes central to the discussion of ezafat in Urdu-Hindi MT system end user output, especially from an orthographic perspective. Consequently, it is also important to transfer and generate the ezafat accurately in the target language i.e., Hindi, while going from the SL (Urdu) to the TL (Hindi) scenario, in addition to the vocalic and glottalic sounds.

V. ISSUES IN GOOGLE AND BING TRANSLATE EXISTING OUTPUT

Existing end user output of Google and Bing translator, still pose problems/inaccuracies, especially in terms of ezafat constructions of Urdu into Hindi which seems to happen due to the Urdu orthographic challenges, inadequate corpus in general and improper implementation of ehrab in particular. Following are the two screenshots from both the translator (Google and Bing) which shows the actual scenario about glottalic/vocalic and missing ezafat in the target language.

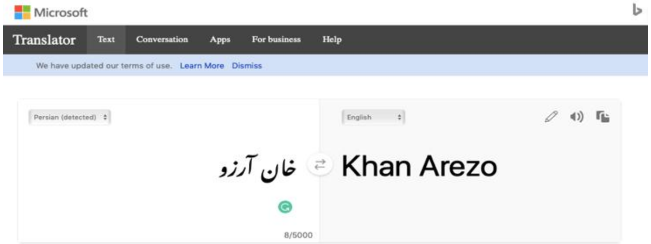


خان آرزو

खान विश करता है

khaan vish karata hai

Figure 1: Showing the Ezafat Construction Output of Google



We have updated our terms of use. Learn More Dismiss

خان آرزو

Khan Arezo

Figure 2: Showing the Ezafat Construction Output of Bing

From the above screenshots, it can be clearly said that the end user output in existing Google and Bing MT is inaccurate. As fig. 1, showing the example of Google in which the selected pair for the translation is Urdu to Hindi where the source text (Urdu) is not translated in the target language (Hindi) as per the source orthography. Further, in fig. 2 which presents the output of Bing, where the end user output of source text (Urdu) has been generated in English despite choosing the Hindi as the target language (which is also inaccurate), and this resulted due to the separate lexical transfer and insufficient information in their linguistic corpora.

Notably, to get the desired output all the NER should be treated and transferred as a single unit, and bilingual corpora needs to be developed as per the requirements. However, the end user out of Google is better as compared to the Bing both from orthographic and ezafat point of view. Furthermore, the inaccuracies have been noticed with both the translator's output in general. Therefore, the author believes that inaccuracies in end user output are taking place due to the missing word, word order issues (in terms of ezafat), missing diacritics in the source language corpus and the same is resulting in the target language (Hindi).

VI. HANDLING/PROCESSING NER

Both the languages belong to the same language family, i.e., Indo-Aryan. Though, Urdu belongs to Indo-Aryan but uses Perso-Arabic script in addition to the rich amount of lexical items borrowed from both Arabic and Person. Therefore, they differ noticeably with each-other especially in terms of naming convention and lexical choices. Therefore, Urdu includes the name of both Arabic and Persian origin. Contrary, Hindi is a descendant of Sanskrit, therefore, it includes the name of Sanskrit origin. As a result, there are certain challenge occurs (mentioned in section 4 and 5) in NER transliteration which needs to be handled in existing Google and Bing Translators.

Therefore, the classification of NER needs careful attention due to their complex and long tail, which is usually transferred separately in existing Google and Bing translator which results in inadequate end user output. To handle NER following steps are important to deal:

i) Detection
ii) Classification
iii) Handling

To handle NER, the first step is to identify them both in text and speech. Once the identification is performed, then their categorization needs to be done into several categories followed by some sub-categories like; person, place, and organization, etc. For example, 'bab e rahmat' in Urdu will not be classified as the personal name of a person, rather it refers to a particular gate which is located in 'Holy places like; Mecca and Medina'. Further, it should also be treated as single unit (despite the fact that it contains multiple individual names in a complex name) but not as a separate unit, because if transferred separately then 'baab' may be treated as a noun (adhayaya), and 'rahmat' either as an adjective (blessings) or as a personal noun (name of a person) which will result in inaccurate end user output like 'dayaa adhayaya or daya daya'.

VII. END USER OUTPUT OF GOOGLE AND BING

In the following figure 3, the blue line in the circle showing the google output accuracy, while the yellow line shows the output accuracy of Bing translator. The Urdu source names (mainly names made of ezafat) translated into Hindi by using both the translators Google and Bing and the end user output is compared based on a 0 to 1 scale. Where zero (0) stands for correct output and one (1) stands for an incorrect one. The ranking has been done manually as the sample size is small (only 30 ezafat names has been translated) but the author surmises that if these basic ezafat names are not adequately translated into the target language Hindi, then how the more complex ezafat names will be translated accurately (especially those names which consists of two or more ezafat).

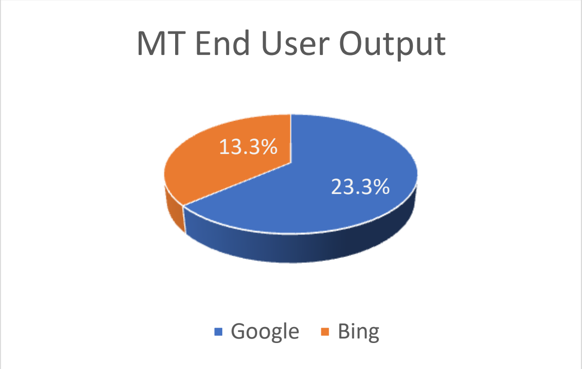


Figure 3: Showing the Actual Picture of Existing Inaccuracies in Google and Bing MT output

The above figure 3 also reveals that the accuracy of Google end user output is better as compared to the Bing translator. Therefore, the author opines that the existing inaccuracy in end user output is taking place due to the inadequate corpora which lacks the word order in terms of ezafat and orthographic information related to vocalic and glottalic sounds. This leads to the inaccurate generation of end user output. Further, based on the sample size the accuracy of Google is 23.3%, while Bing's output produces 13.3% accuracy. From this, it can clearly be said that the accuracy of end user output, especially in terms of ezafat construction is not up to mark.

Notably, there is certain output which is generated in a combination of both English and Hindi (as a translation output in Hindi). From this, we mean that the output is not as per the desire. Furthermore, the other observation can be noted that at least Google generates the end user output in Hindi text only either in form of translation or transliteration (no English word is found) but Bing has generated the mixture of both Hindi and English as an output.

VIII. RESULTS AND DISCUSSION

This paper is an attempt to outline the issues pertaining to inaccuracies in NER output of both Google and Bing translator, with special reference to ezafat. In figure 1, a small sample of ezfata NER data has been presented, but gives a clear idea about the seriousness of inaccurate output, in both the translators (Google and Bing). For the present study, the examples taken from Urdu are basic ezafat construction and translated from Urdu to Hindi using both Google and Bing translate. Further, the existing output also signifies that, if the size of ezafat construction (more than two word combinations) will be longer then the result would be more affected. The generated output of ezafat is merely not only unacceptable, rather it indicates the seriousness of meaning and missing characters of Urdu in Hindi. Therefore, it is a question which needs to be solved from linguistic point of view. Existing Google and Bing MT systems need such improvements in order to fix such issues or inaccuracies while generating end user outputs. Furthermore, the generated output raises the question from two points of view: primarily from translation/transliteration point of view, and secondly, from a linguistic point of view. These wrong outputs could be misleading for both native and non-native speakers and it also raises the question on the reliability of such translator's end user output qualities.

IX. CONCLUSIONS

This work has simply outlined the inaccuracies in existing Google and Bing MT output with special reference to ezafat names. Further, the end user output of both the translators has also been compared and found that the accuracy of Google (23.3%) is better than Bing's (13.3%) end user output. Based on the existing inaccuracies in end user output, it can be surmised that this is happening due to the insufficient linguistic corpora and their poor linguistic classification. Therefore, depending on the highlighted errors in end user output, we can strategize for better handling of ezafat names in Urdu-Hindi MT systems which will further help to obtain the desired TL output in Google and Bing MT. Notably, one of the main challenges in Urdu-Hindi scenario is about retaining vocalic sounds in the target language (Hindi) according to Urdu in order to improve both end user output quality and readability. Therefore, the author also recommends that to avoid the confusion in terms of ezafat construction, the diacritic 'zer' /◌ِ/ should necessarily be used in the source language (Urdu) and consequently needs to be generated in the target language (Hindi) 'ए' as per the requirements.

Furthermore, the adoption of diacritics 'zer' and 'halanata' will also help the speakers of both the languages to identify and read the words in actual pronunciation easily. This will also ensure the quality of the translation, which is one of the main objective of doing translation.

REFERENCES

1. Ahmed, T. (2009). Roman to Urdu transliteration using wordlist. In Proceedings of the Conference on Language and Technology (Vol. 305, p. 309).
2. Anwar, W., Wang, X., & Wang, X. L. (2006, August). A survey of automatic Urdu language processing. In 2006 International Conference on Machine Learning and Cybernetics (pp. 4489-4494). IEEE.
3. Bharati, A., Chaitanya, V., Sangal, R., & Ramakrishnamacharyulu, K. V. (1995). Natural language processing: a Paninian perspective (pp. 65-106). New Delhi: Prentice-Hall of India.
4. Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). An algorithm that learns what's in Name. Machine Learning, 34(1-3), 211-231.
5. Chandra, P., & Kumar, R. (2013). Urdu possession: An instance of Ezafe. Indian Linguistics, 74(3-4), 101-109.
6. Ekbal, A., Haque, R., Das, A., Poka, V., & Bandyopadhyay, S. (2008). Language independent named entity recognition in indian languages. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.
7. Jawaid, B., & Ahmed, T. (2009). Hindi to Urdu conversion: beyond simple transliteration. In Conference on Language and Technology.
8. Khan, S. A., Anwar, W., & Bajwa, U. I. (2011, November). Challenges in developing a Rule-based Urdu stemmer. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP) (pp. 46-51).
9. Kudo, T., & Matsumoto, Y. (2000). Use of support vector learning for chunk identification. In Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop.
10. Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In Proceeding of 18th International Conference on Machine Learning, 282-289.
11. Li, W., & McCallum, A. (2003). Rapid development of Hindi named entity recognition using conditional random fields and feature induction. ACM Transactions on Asian Language Information Processing (TALIP), 2(3), 290-294.
12. Moldovan, D. I., Sanda M. H., Roxana, G., Morarescu, P., Lacatusu, V. F., Novischi, A., Badulescu, A., & Bolohan, O. (2002). LCC tools for question answering. In Proceedings of the TREC, 1-10.
13. Rehman, Z., Anwar, W., & Bajwa, U. I. (2011, November). Challenges in Urdu text tokenization and sentence boundary disambiguation. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP)(pp. 40-45).
14. Riaz, K. (2010, July). Rule-based named entity recognition in Urdu. In Proceedings of the 2010 named entities workshop (pp. 126-135). Association for Computational Linguistics.
15. Sekine, S. (1998). Description of the Japanese NE system used for MET-2. In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.
16. Shah, D. N., & Bhadka, H. (2017). A survey on various approach used in named entity recognition for indian languages. International Journal of Computer Applications, 167(1).
17. Sharma, P., Sharma, U., & Kalita, J. (2010). The first steps towards Assamese named entity recognition. In Brisbane Convention Center (Vol. 1, pp. 1-11).
18. Takeuchi, K., & Collier, N. (2002, August). Use of support vector machines in extended named entity recognition. In proceedings of the 6th conference on Natural language learning-Volume 20 (pp. 1-7). Association for Computational Linguistics.
19. Vapnik, V. N. (1995). The nature of statistical learning. Theory. Springer.
20. Yamada, H., Taku K., & Yuji, M. (2002). Japanese named entity extraction using support vector machine. J. Inf. Process. Soc. Jpn., 43(1), 44-53.

AUTHOR'S PROFILE

Md. Tauseef Qamar, is a research scholar in the D/O Linguistics at Aligarh Muslim University, India. He is a young linguist and researcher who have published and presented several papers in national and international conferences mainly in the area of machine translation and other fields of linguistics like; phonology, morphology, and syntax. His research interests are: phonology, morpho-syntax, NLP and MT, etc.

Juhi Yasmeen, has recently received her PhD in Applied Linguistics from Aligarh Muslim University at Aligarh, India. She is an active linguist who accepts challenges to explore the emerging areas in the field of academics. She presented and published in several national and international seminar & conferences. Her research interest lie in the field of translation studies and advertising media.