Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

Analyzing Complex Predicates: Syntactic Evidences from Sampark and Google Machine Translation Systems

Md. Arfeen Zeeshan, ** Prof. Mohd. Khalid Mubashir-uz-Zafar, ***Dr. Sabahuddin Ahmad, ****Prof. Aasim Zafar, ** Dr. Sanket Pathak ****** Dr. Tauseef Qamar**

* Research Scholar, Dept. of Linguistics, A.M.U, Aligarh

** Professor, HOD, Dept. of Translation Studies, MANUU, Hyderabad

*** Associate Prof., Dept. of Linguistics, A.M.U, Aligarh

**** Professor, HOD, Dept. of Computer Science, AMU, Aligarh

***** Language Manager,R&D at Mihup Communications, Kolkata

****** Dept. of Linguistics, A.M.U., Aligarh

Abstract

The development of a machine translation (MT) system is an extremely challenging task. Because human languages are made up of both simple and complex words. Some words are differently translated depending on their contextual meaning. Therefore, this is of high importance to consider world knowledge for an MT to be able to perform the task of translation effectively. In this sense, MT also produces errors while performing the task of translation. Consequently, measuring such errors of MT output helps us to identify types of errors and subsequently enable us to improve its translation output efficacy. Hence, this study attempts to compare the errors produced by Sampark and Google translators with reference to complex predicates (CPs) while translating it from Hindi to Urdu language. Further, this paper is limited in nature in the sense that it examines Hindi CPs output in Urdu language only. The data for this study have been gathered from different secondary sources like e-dictionaries, e-newspapers, and social media platforms. Next, five hundred Hindi CPs were identified manually and translated using Sampark and Google Translator platforms and errors in their TL outputs were compared. To do so, the CPs were fed to the machine and seen how the machine maps its equivalent translational in the target language i.e. Urdu. On the basis of existing errors in these CP outputs they were divided into five groups i.e. missing word, word order, incorrect word, unknown word and wrong noun-verb agreement and illustrated with examples. Finally, source of the errors have been outlined including their possible reasons.

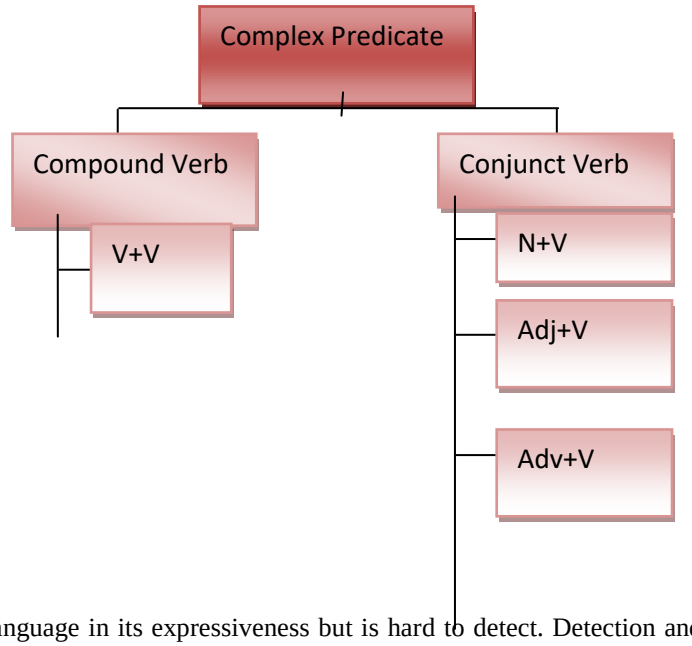
Keywords: Machine Translation**,** Sampark and Google Translator, Complex Predicates MT Output, Hindi-Urdu MT.

1. Introduction

It is believed that developing an MT is a difficult task to accomplish. Since human languages are idiosyncratic and vary from person to person within the same language (Chelbi, 2016). Therefore, it becomes of prime concern to develop an MT system by keeping world knowledge in view. However, this is impossible to capture every form of words used by its natives in day to day communication. In every human language, numerous words are translated differently based on their occurrence and context. In this context, MT also produces error while performing the task of translation. Consequently, measuring MT output error becomes necessary from two points; firstly it facilitates computational linguists with the sources and reasons of errors, secondly helps us in the improvement of MT output. In this regard, several tools have been devised and proposed that measure MT output error namely Bleu, WER, METEOR, etc. Nevertheless, they are unable to measure every type of MT output errors alone including their sources and reasons. As a result, such MT output errors can manually be measured to analyse their source and reasons. These findings can be used to improve MT system output in terms of the desired output.

A Complex Predicate in Urdu is a syntactic construction consisting of a verb, a noun, an adjective or an adverb as main predicator followed by a light verb (LV). Thus, a CP can have structures like noun+verb, adjective+verb, verb+verb or adverb+verb. CPs are commonly classified as multi-headed predicates, which are made up of multiple grammatical elements (either morphemes or words), each of which contributes a portion of the information normally associated with a head. Urdu/Hindi has been shown to exhibit various types of complex verbal constructions, including n+v, adj+v and v+vCPs (e.g., Mohanan (1994), Butt (1995), inter alia). Complex predicates (CPs) are abundantly used in Hindi and other languages of Indo-Aryan family and have been widely studied (Hook, 1974; Abbi, 1992; Verma, 1993; Mohanan, 1994; Singh, 1994; Butt, 1995; Butt and Geuder, 2001; Butt and Ramchand, 2001; Butt et al., 2003). We can easily understand the CPs structure below with the help of diagram.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.



CPs empowers the language in its expressiveness but is hard to detect. Detection and interpretation of CPs are important for several tasks of natural language processing tasks such as machine translation, information retrieval, summarization etc.

This paper aims to measure MT output errors with reference to Hindi CPs translation in Urdu language. For this we have considered two MT platforms i.e. Sampark and Google, the former one are rule-based and the latter is a neural-based MT system in order to provide a clear picture regarding existing errors. The error analysis performed in this study is manually measured. To do so, the data were gathered from Hindi e-newspapers, and social media platforms mainly Facebook and Twitter. Therefore, this work attempts to present a detailed MT output error analysis that will facilitate computational linguists in the effective handling of CPs both in rule-based and neural machine translation systems.

2. Related Works

Since measuring MT output errors using automatic measurement technique is not sufficient alone. As this only provides quantitative analysis score, that can tell us about the performance of MT system but not its strength and weaknesses. Whereas, manual measurement of MT system output errors provides us with source and reasons of MT output errors (Chelbi, 2016). Moreover, manual evaluation of MT output errors enables the tool developers to measure MT output quality and performance.

Significantly, researchers have developed a preference or hierarchy list of MT output errors using hierarchy structure 3.2 (Vilar et al., 2006). This hierarchy is classified into five broad categories i.e. *missing words, word order*, *incorrect words*, *unknown words*, and *punctuation*. We have taken into consideration only four categories as per our case study. Missing words are produced when the given equivalent word is missing in the corpora of TL. This gives birth sense that production of word wrong leads to non-sensical or ungrammatical constructions in TL that affects the overall meaning of the sentence. The second type of MT output errors is reported in 'word order' class, which is further sub-divided into two categories i.e. word level and phrase level re-ordering. In case of wrong word order generation, each word needs to be re-arranged to produce correct word order. The third type of error class is reported in the category of 'incorrect words'. Such errors are produced when MT does not find equivalent translation of a word in TL. This category is further sub-divided into five categories; first 'sense' that leads to the creation of meaningless word or sentence, second is the production of 'incorrect form' of a word, third 'extra word, in TL, fourth 'style' that refers to bad selection of translation in TL as per the compatibility of SL. Fifth type is related to 'idioms' where it is translated in literal combination of words. Last type of error category is 'unknown words' that refers to two condition i.e. 'truly unknown words' and 'unseen words'.

2.0. Methodology:

The current research has been carried out with reference to Google and Sampark translation available on the net. For our purpose we have selected Hindi as a input language or SL and Urdu as a output language or TL. As far the method we have adopted here, in the first stage, appropriate Hindi sentences were fed to the above translators as input.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

3.0. Google and Sampark MT Systems:

Google Translator: The Google is an online translation platform. It was released by the Google Inc. on Jan 12, 2010 and its revised version came in the year 2011 for translating SL text into the TL text. Presently, the platform is providing services for ninety languages in the field of MT. It is also a statistical based platform.

Samaprk MT system: Sampark is a machine translation system developed with the efforts of 11 institutions in India under the umbrella of consortium project " Indian language to India Language Machine translation" (ILMT) funded by TDIL program of Ministry of Electronics and Information Technology (MeitY), Govt. of India.

4.0. Error Analysis of Urdu CP Sampark and Google MT

The following screenshot i.e. fig. 1 (a) show existing inaccuracy in Sampark TL output of CP in Urdu language. In the first example the inaccurate output is generated due to the wrong lexical transfer into TL (Urdu). This gives rise to the sense that the system lacks equivalent lexical item of source language (SL) Hindi noun 'daan' in TL. Moreover, the following two examples have simply been transliterated which likely to happened due to non-availability of SL lexical items equivalent in TL. Taking these scenario into considerations we have proposed the desired output as per compatibility of machine and end user, as exemplified in table 1 (a) below. Please find below the attached analysis of two MT systems:

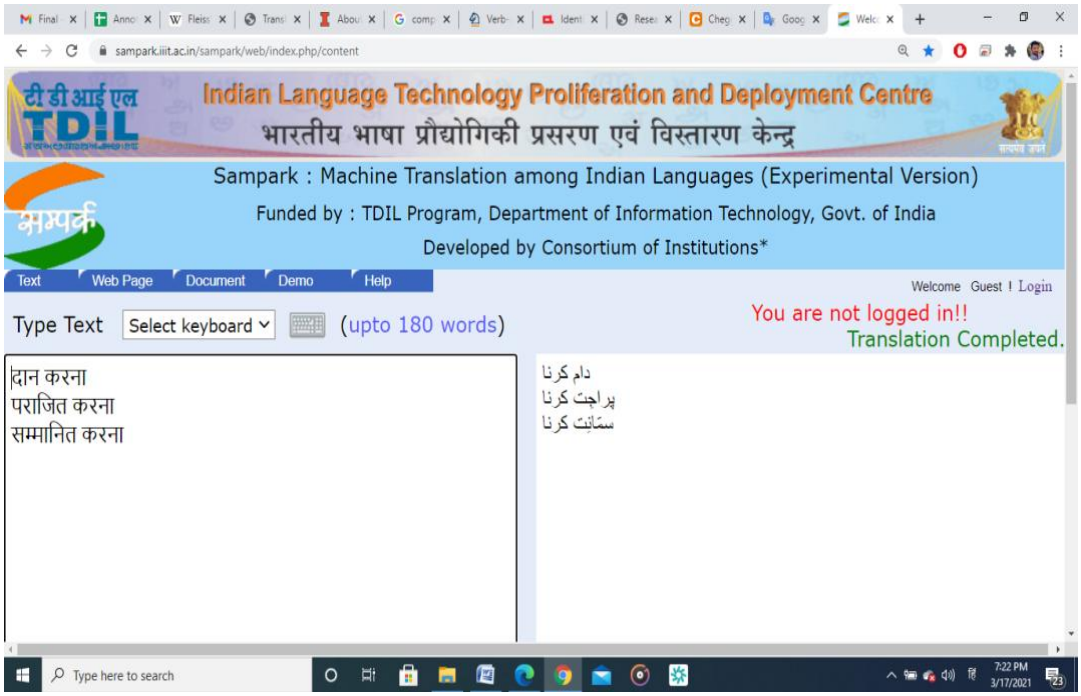


Fig1.0 : Sampark MT

Source (Hindi)	Target (Urdu)	Correct Output (Urdu)
दान करना	دام کرنا	عطیہ دینا
पराजित करना	پراجت کرنا	شکست دینا
सम्मानित करना	سمّانت کرنا	اعزاز دینا

Table 1.0. Sampark MT

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

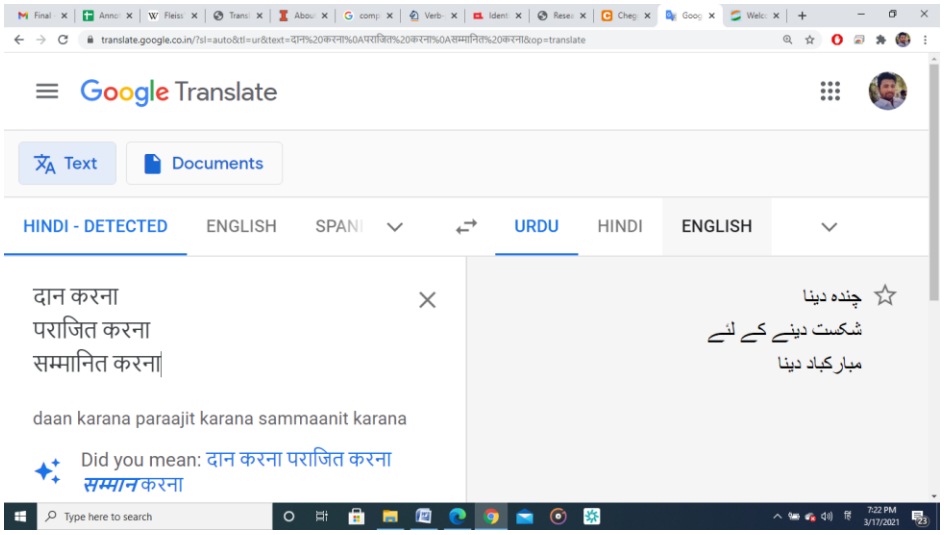


Fig. 1.1:Google Translation

Source (Hindi)	Target (Urdu)	Correct Output (Urdu)
दान करना	چندہ دینا	عطیہ دینا
पराजित करना	شکست دینے کے لئے	شکست دینا
सम्मानित करना	مبارکباد دینا	اعزاز دینا

Table1.1: Google Translator

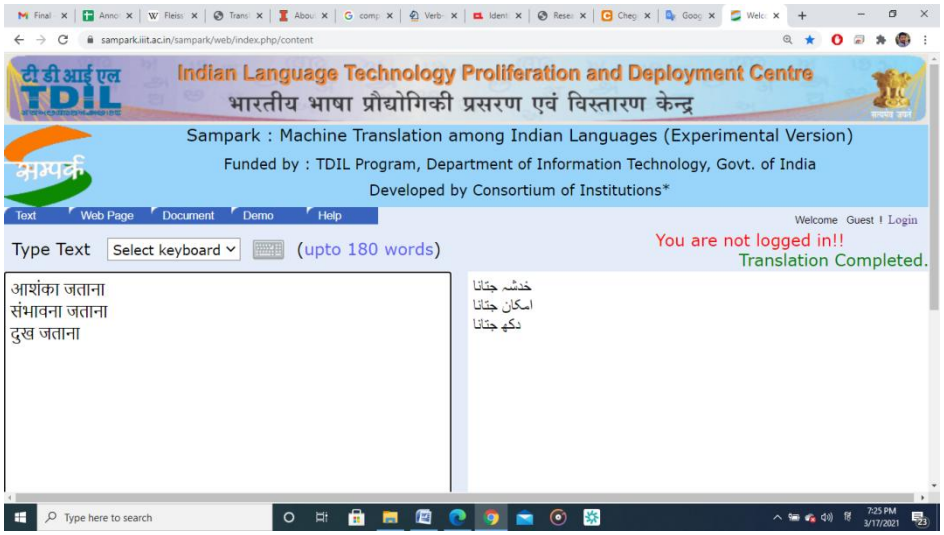


Fig 1.2: Sampark MT

Source (Hindi)	Target (Urdu)	Correct Output (Urdu)
आशंका जताना	خدشہ جتانا	خدشہ ظاہر کرنا
संभावना जताना	امکان جتانا	امکان ظاہر کرنا
दुख जताना	دکھ جتانا	غم کا اظہار کرنا

Table1.2 Sampark Translator

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

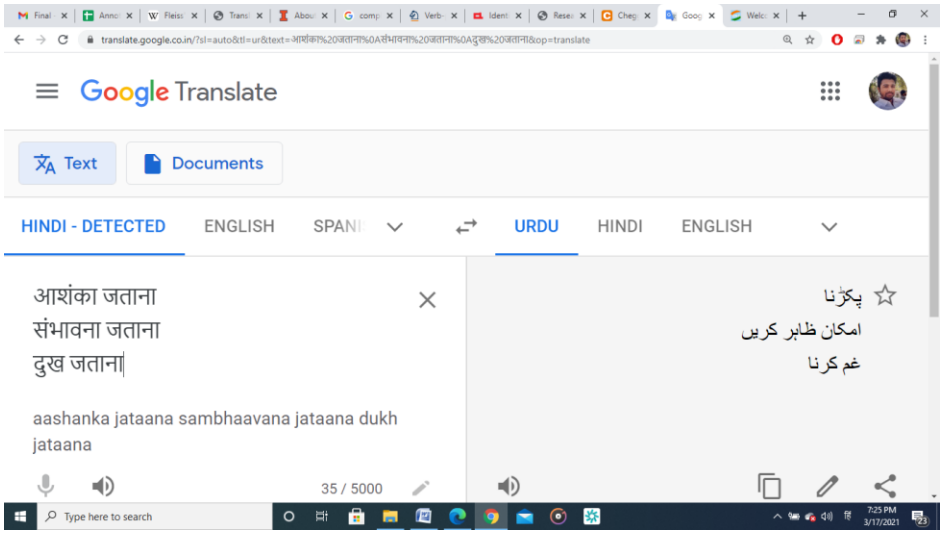


Fig 1.3 : Google Translation

Source (Hindi)	Target (Urdu)	Correct Output (Urdu)
आशंका जताना	بکڑنا	خدشہ ظاہر کرنا
संभावना जताना	امکان ظاہر کریں	امکان ظاہر کرنا
दुख जताना	غم کرنا	غم کا اظہار کرنا

Table1.3 Google Translation

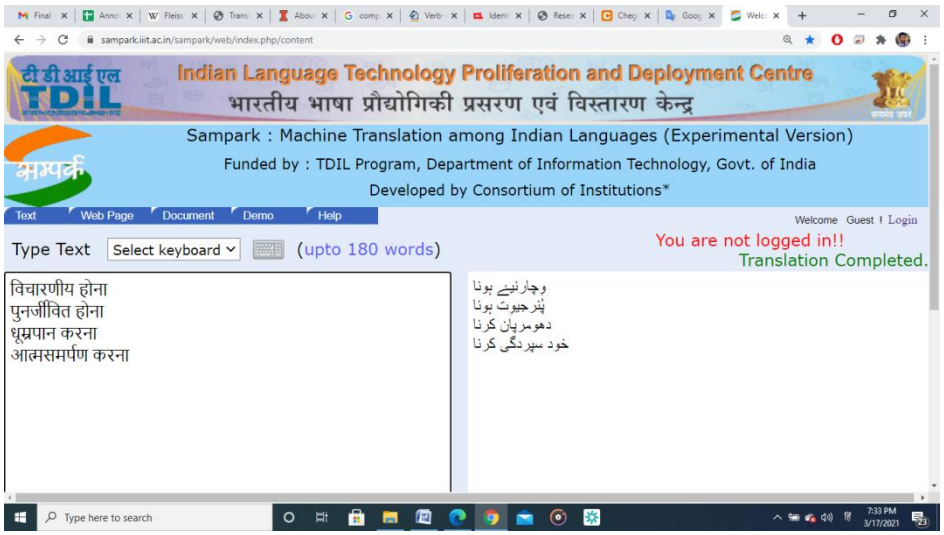


Fig 1.4: Sampark MT

Source (Hindi)	Target (Urdu)	Desired Output (Urdu)
विचारणीय होना	وچارنئے ہونا	قابل غور ہونا
पुनर्जीवित होना	پُنرجیوت ہونا	دوبارہ زندہ ہونا
धूम्रपान करना	دھومرپان کرنا	تمباکو نوشی کرنا
आत्मसमर्पण करना	خود سپردگی کرنا	خود شپردگی کرنا

Table1.4: Sampark MT

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

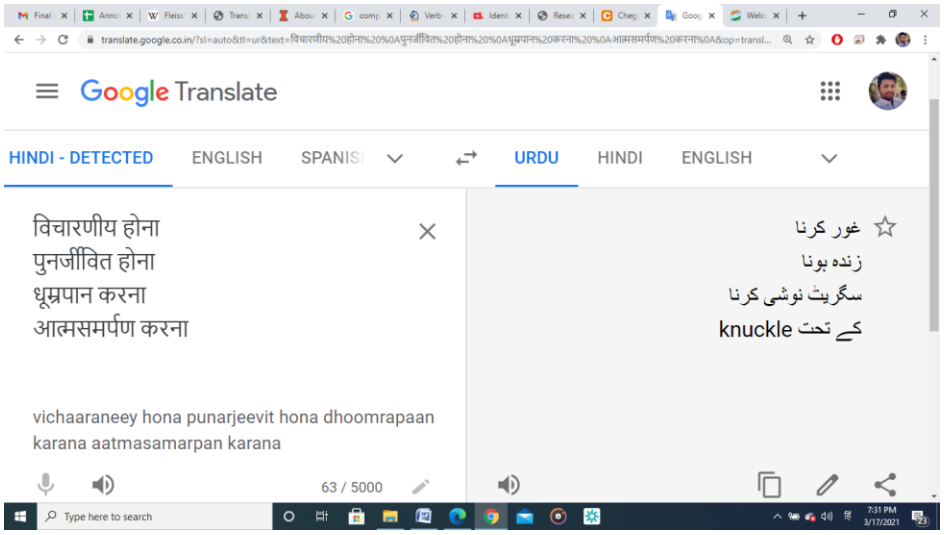


Fig 1.5: Google Translation

Source (Hindi)	Target (Urdu)	Correct Output (Urdu)
विचारणीय होना	غور کرنا	قابل غور ہونا
पुनर्जीवित होना	زندہ ہونا	دوبارہ زندہ ہونا
धूम्रपान करना	شگریٹ نوشی کرنا	تمباکو نوشی کرنا
आत्मसमर्पण करना	Knuckle کے تحت	خود شپردگی کرنا

Table1.5: Google Translation

5.0. Error Analysis of Urdu Complex Predicate Output

As we can see in fig. 1.1, the existing output of both the MT platforms i.e. Sampark and Google Translator is inaccurate. In case of Sampark only transliteration has been done however Google somehow generated better output as compared to Sampark. Here light verb construction changes from /karna/ to /dena/ which is wrong. /karna/ should remain same in such cases.

In fig. 1.2, Sampark is transferring only bi-lingual of head word. For example, /JatAnA/ should equally be transferred in Urdu. Similarly, Google also have inaccuracy in their Urdu Output. Google transfer its one of the outputs in a single unit which is totally wrong/inaccurate. Here head + light verb should equally be detected and translated. Notably, /jatAnA/ means /zAhir karna/ in Urdu but Google fails to provide the correct output of all entries of CP.

Further, in fig. 1.3, Sampark transliterated all CPs except one which is /khud supurdagi karna/. Therefore, the aim for handling such categorise is to address ambiguity. As a lexical entry may have one or more meaning. Moreover, Google also failed to detect आत्मसमर्पण करना (Atma samrpan karna) as a CP which resulted in wrong output.

6.0. Conclusion

In this paper, we have collected a few Hindi Complex Predicates belonging to categories like Noun and Adjective with light verb construction and attempted to map out their translational equivalents in Urdu using Sampark and Google Machine Translation platforms. Google produces good results in comparison to Sampark. Our focus was to find issues and analyse errors of CPs. We found three major errors that have been addressed above. It can be clearly seen that both MT fails to provide correct Urdu output. In places where the source text involves Noun, Adjective with light verbs the Sampark and Google MTs are unable to transfer the correct output.

To maintain the semantics of Complex Predicates, we need to handle it as a single unit. We need to create a separate bag/Lexicon for Urdu Complex Predicates with proper classification which includes Noun+Verb, Adjective+Verb, Verb+Verb, Adverb+Verb combinations. This may help us to overcome with one of the crucial problems of MT system.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 15853 - 15859
Received 05 March 2021; Accepted 01 April 2021.

References:

1. Chelbi, S. (2016). Error *Error analysis of a statistical machine translation system* [Bachelor's thesis, KarlsruherInstitutfürTechnologie]. Retrieved from https://isl.anthropomatik.kit.edu/pdf/Chelbi2016.pdf
2. Vilar, D., Xu, J., Luis Fernando, D. H., & Ney, H. (2006, May). Error Analysis of Statistical Machine Translation Output. In *LREC* (pp. 697-702).
3. Butt, M. (1995) The Structure of Complex Predicate in Urdu, Stanford, CSLI Publications.
4. Kachroo, Y. (1996) An Introduction to Hindi Grammar, Urban: University of Illinois.
5. Aiken, M., & Balan, Sh. (2011). An analysis of Google Translate accuracy. Translation Journal, 16(2). Retrieved June 26, 2015, from http://translationjournal.net/journal/56google.htm
6. Swati Gupta. 2004. Aligning Hindi and Urdu bilingual corpora for robust projection. Masters project dissertation, Department of Computer Science, University of Sheffield.
7. Bhat,R.N. (2001-2003) Vector in Hindi Compound Verbs, Prajna,Vol. 47-48,Part-1-2.