Towards Handling Verb Phrase Ellipsis in English-Hindi Machine Translation

Niyati Bafna
Ashoka University
niyatisanjay.bafna_ug20@ashoka.edu.in

Dipti Misra Sharma
Indian Institute of Technology, Hyderabad
dipti@iiit.ac.in

Abstract

English-Hindi machine translation systems have difficulty interpreting verb phrase ellipsis (VPE) in English, and commit errors in translating sentences with VPE. We present a solution and theoretical backing for the treatment of English VPE, with the specific scope of enabling English-Hindi MT, based on an understanding of the syntactical phenomenon of verb-stranding verb phrase ellipsis in Hindi (VVPE). We implement a rule-based system to perform the following sub-tasks: 1) Verb ellipsis identification in the English source sentence, 2) Elided verb phrase head identification 3) Identification of verb segment which needs to be induced at the site of ellipsis 4) Modify input sentence; i.e. resolving VPE and inducing the required verb segment. This system is tested in two parts. It obtains 94.83 percent precision and 83.04 percent recall on subtask (1), tested on 3900 sentences from the BNC corpus (Leech, 1992). This is competitive with state-of-the-art results. We measure accuracy of subtasks (2) and (3) together, and obtain a 91 percent accuracy on 200 sentences taken from the WSJ corpus(Paul and Baker, 1992). Finally, in order to indicate the relevance of ellipsis handling to MT, we carried out a manual analysis of the MT outputs of 100 sentences after passing it through our system. We set up a basic metric (1-5) for this evaluation, where 5 indicates drastic improvement, and obtained an average of 3.55.

1 Introduction

Verb phrase ellipsis is a particularly frequent form of ellipsis, both in speech and in text. English VPE is the elimination of a non-finite verb phrase, introduced by an auxiliary or the particle 'to'(Kenyon-Dean et al., 2016).

We observe that state-of-the-art MT systems often cannot correctly interpret and translate sentences with VPE. For example, (elliptical phrase and site of ellipsis are in bold and italics respectively):

> Ram **cooked the food** quickly, but Shyam *did not* →
>
> *Ram ne jaldi se
> Ram-ERG - quickly-ABL -
> khana banaya, lekin Shyam
> food-OBJ cook-PT, but Shyam-ERG
> ne nahi diya.
> - not give-PT
>
> (created example)

In our initial analysis, we find that Google Translate could translate only 6/50 VPE sentences, taken from the WSJ corpus, correctly from English to Hindi. This motivated us to identify and resolve VPE in English and give this modified sentence to the machine translation system to check whether it improves the quality of translation. We used our MT system Anusaaraka for this purpose.[1]

We present a rule-based approach that performs three sub-tasks in order to solve this problem: 1) ellipsis identification 2) identification of the antecedent head verb 3) addition of necessary auxiliaries and/or complements to the head verb. Finally, it modifies the input sentence by inducing the identified verb segment at the site of ellipsis. We tailor this algorithm for the purpose of English-Hindi MT, and therefore provide a special treatment to compound verbs (including phrasal verbs),

[1] https://ltrc.iiit.ac.in/Anusaaraka/anu_home.html

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 150–159
Hyderabad, India, December 2019.

serial verbs, and verb complements in English.

Our solution is based on an understanding of verb phrase ellipsis in Hindi. Hindi does not exhibit VPE in the same manner as English; however, it exhibits a phenomenon called verb-stranding verb phrase ellipsis (VVPE) (Manetta, 2018). This means that a verbal phrase, including objects and other arguments, may be completed elided, stranding the head verb, which then appears at the site of ellipsis.

We propose that English VPE can be transferred into Hindi VVPE in order to improve MT results on the modified sentence. For this, we claim that it is enough to identify the head verb antecedent in the English sentence, and perform an analysis to support the claim.

Finally, we want to see that MT systems indeed perform better on statements with ellipsis after our treatment. Since we want to test the performance of the MT on a very specific facet: i.e. the elliptical clause, we do not utilize standard evaluation metrics but set up and define our own scale, and perform a manual analysis of 100 sentences. This evaluation dataset is meant to be merely indicative of the benefits of ellipsis handling for MT. The results are explained below.

2 Background and Our Contribution

There have been several previous works addressing the problem of antecedent head resolution for VPE in English. Cheung et al adapt the Margin-Infused-Relaxed Algorithm (MIRA) for target detection and antecedent resolution and obtain an accuracy of 65 percent (Kenyon-Dean et al., 2016). Nielson, 2005, re-implements Daniel Hardt's VPE-RES algorithm on the Penn Treebank to obtain a highest Head Overlap success of 85.87 percent and a lower Exact Match success, about 78 percent on the Brown corpus (Nielsen, 2005) (Hardt, 1992). Liu et al, 2016 experiment with various joint modelling techniques, and obtain a recall of 83.46 percent for boundary identification (Liu et al., 2016). Earlier works include Daniel Hardt's linguistically motivated rule-based system, that eliminates impossible antecedents by looking at be-do conflicts, contained antecedents and assigned scores based on co-reference of the noun subjects and clausal relationships, such as 'as' constructions. (Hardt, 1992)

The necessity for tools to deal with VPE is widely recognized in literature, for the purposes of information extraction, finding event co-occurrence, etc (Kenyon-Dean et al., 2016). In the context of MT, a possible solution is to identify the antecedent, or the source verb phrase, and induce it at the site of ellipsis to gain a legitimate, simplified sentence. This solution functions by reiterating the antecedent at the site of ellipsis. By breaking the link between the ellipsis and the antecedent, we give the MT two independent clauses, which it can translate without error.

Indeed, the previous works listed above are aimed at identifying the antecedent verb phrase, which includes compulsorily a head verb, and optionally its arguments and adverbials. While it is important to pick up arguments and adverbials of the head verb for the purpose of comprehension, there are some problems that it introduces: namely, it must often disambiguate by context and therefore can be a great source of error, and in the context of MT, it might make the output sentence clumsy and unnatural.

Since we are looking at antecedent head resolution from a particular angle i.e. transforming the English sentence containing VPE in order to align with Hindi VVPE and therefore help the MT, our problem does not require us to find the antecedent boundary of the head verb, while we do provide an additional treatment of certain verb constructions in English according to the manner in which they would be translated into Hindi. As far as we are aware, this is the first attempt to tailor English VPE resolution in the context of English-Hindi MT.

3 Ellipsis in English

Verb phrase ellipsis in English is introduced by an auxiliary. We consider five classes of el-

lipsis depending upon the auxiliary at the site of ellipsis: 1) to_be, 2) to_have, 3) to_do, 4) modals, and 5) to_particle ellipsis. Cheung et al include a sixth class: the do-so anaphora, while acknowledging that modals and Do-X anaphora are not technically auxiliaries (Kenyon-Dean et al., 2016). We have chosen to identify ellipsis introduced by modals, however, because 1. the behaviour of the former is identical with ellipsis by true auxiliaries, 2. Likewise, we observe that it poses a problem to state-of-the-art MT English-Hindi systems. We do not identify Do-X anaphora in this system, however, since this is simply a pronominalization of the antecedent rather than eliding. These have a different treatment than VPE – for example, they may be directly pronominalized in Hindi as well, rather than transferred into VVPE. Indeed, MT systems are able to do this:

> Although Mr. Azoff won't **produce films** at first, it is possible that he could *do so* later, the sources said →
>
> haalaanki, shree azoph pahalee baar
> although, Mr. Azoff first time
> philmon ka nirmaan nahin
> film-PL-POSS - production not
> karenge, lekin yah sambhav hai ki
> do-F.M, but this possible is that
> vah baad mein aisa kar sake
> he later-POST - this do can
>
> (Taken from WSJ corpus)

There are certain constraints on the antecedent, depending upon the auxiliary at the site of ellipsis. For example,

1. Auxiliaries of the form to_be require antecedents with to_be auxiliaries.

2. All non-to_be forms of ellipsis do not have an antecedent with a to_be auxiliary, except for gerunds, which are permissible.

4 Antecedent Resolution in the Context of MT

We know that state-of-the-art machines cannot interpret VPE.(Voita et al., 2019). Translation to Hindi from English requires a prediction of the elided English verb. An elided VP consists of a head verb, optionally along with its object arguments and adverbials.

For example, it is clear that the adverbial is interpreted as part of the VP in the following sentence:

> I could **walk quickly**, at that age, **across traffic-filled roads**, but now I *cannot*. (created example)

One solution to eliminate errors due to VPE, is to identify the antecedent and induce it at the site of ellipsis entirely, as a preliminary step before translation, thus eliminating the ellipsis entirely. However, this naive approach has the following issues:

1. Making the decision of whether to import a particular adjunct is complicated, and might be governed by semantic context, and (when verbal), by emphasis.

2. Reiterating the entire verb phrase at the site of ellipsis may sound clumsy and unnatural, and make for a worse translation.

Identifying the boundaries of the elided verb phrase is a much harder task than identifying the antecedent head verb. We note, for example, the consistent drop in Exact Match accuracy in Neilson's study of Hardt's algorithm over different corpora (Nielsen, 2005) – sometimes dipping as much as 30 percent lower. The matter is not as simple as picking up the entire verb sub-tree - always importing all arguments and adjuncts is not permissible, since the correct interpretation often depends on surrounding knowledge. For example, in

> Ram would have liked to eat out with you on Sunday afternoon, but he can't.
> (created example)

There are more than one interpretations of what Ram can't do: eat out, eat out with you, or eat out with you on Sunday afternoon. A native speaker selects one of these in context. If we are seeking to eliminate ellipsis with our system, we must select one of these interpretations to paste at the site of ellipsis, or the whole verb phrase. We may also not adopt intermediate policies such as importing noun objects but not adjuncts, because this risks placing undue emphasis on certain arguments, and may result in an incorrect semantics.

The second problem with this strategy is that it may render sentences, after pasting, clumsy, unnatural or nonsensical, as illustrated, respectively, by the following example, taken from the WSJ corpus (Bos and Spenader, 2011).

> The Volokhs were afraid that they'd **end up like a friend of theirs who'd applied for a visa and waited for 10 years, having been demoted from his profession of theoretical mathematician to shipping clerk.** They *didn't* (end up like…shipping clerk)

5 Hindi VVPE

Verb-stranding verb phrase ellipsis is the phenomenon wherein a verb is stranded at a site of ellipsis, and its internal arguments are elided. Manetta establishes, by various diagnostics, that Hindi-Urdu do exhibit VVPE. For example,

1. Ram-ne Chomsky-ka naya lekh do baar paRha.
   Ram-ERG Chomsky-GEN new writing two time read-PFV.M.SG

   Ram read the new paper by Chomsky twice.

2. Raj-ne bhi paRha.
   Raj-ERG also read-PFV.M.SG

   Raj also read (the paper twice). (Manetta, 2018)

Here, 'paRha' provides access to the internal arguments i.e. the direct object and the adverb via VVPE. Manetta supports her analysis theoretically and by a survey across native speakers.

This provides us an intuition for a solution for our larger problem: that is, instead of disambiguating the antecedent boundaries to resolve English VPE as simplification for English-Hindi MT, we may provide the Hindi clause containing ellipsis only the head verb, which has access to the internal arguments of the antecedent. If we induce the head verb at the site of ellipsis in the English before translating, now, then we create a valid, syntactical Hindi sentence with all the interpretations of the original sentence intact. We illustrate with an example:

a (Original sentence) Ram would have liked to eat out with you on Sunday afternoon, but he *can't*.

b (With induced head verb) Ram would have liked to eat out with you on Sunday afternoon, but he can't eat.

c Ram ko aapke saath Sunday Ram-ERG - you-ERG with Sunday dopahar ko baahar khana accha afternoon-ERG - out to-eat good lagta tha, lekin vah nahi kha feel would-have, but he not eat sakta.
can-M.SG

Resolving the ellipsis in (b) we get an English sentence with 'eat' at the previous site of ellipsis, that means that Ram is incapable of eating. However, (c) in Hindi, that similarly has 'eat' at previous site of English VPE, is a perfectly acceptable translation of the original sentence, exhibiting VVPE. In (c) the stranded verb 'kha' (eat) has access to the internal arguments 'aapke saath', (with you), 'Sunday dopahar' (Sunday afternoon), etc., and therefore it carries all the interpretations of the original. This is the core idea of the system.

6 Dealing with Multi-Word Verbs in English

While we do not require to identify the boundaries of the head verb, we do need to identify all the components of the verb. This may be required in several cases, such as phrasal verbs, idioms, or serial verbs.

6.1 Compound Verbs

These can be categorized into phrasal verbs and prepositional verbs. The former is a class of verbs that consists of a head verb and a preposition, like "take over", "get around" or "sink in". Since phrasal verbs are opaque, we require to pick up its preposition to maintain the correct sense in which it appears. For example, in

She hasn‘t got over her old failures as yet, although she should.
(created example)

We want the stranded segment to be ”get over”, not simply ”get”.

Prepositional verbs are also verbs followed by a preposition e.g ”stare at”, ”care for”, but they are different from phrasal verbs in certain ways. Their meaning is derived primarily from the head verb. We can see that in the first case, ”stare at”, it would admissible to strand simply ”stare”, although in the subsequent case, the sense of ‘care’ changes without the preposition. It is necessary, therefore, to pick up the prepositional verb as a unit.

6.2 Verbal Complements

In general, when the head verb has verb complement arguments, it is not necessary to pick them up, because the stranded head verb in Hindi will have access to them. For example,

a Aditya **wants to eat** dosa, although Varun *doesn‘t*. (verb complement: to eat)

b Aditya dosa khaana chahta hai,
Aditya dosa to-eat want-PSG -,
haalanki Varun nahi chahta hai.
although Varun not wants-PSG -

(b) is a legitimate translation of (a), with the stranded verb ”wants”. However, this treatment assumes that the sequence of complements in the head verb lexically translate into a sequence of simple verbs in Hindi. This might not be the case.

a Aditya **has to go** home, but Varun *doesn‘t*.

b *Aditya ko ghar jaana padega, lekin
Aditya-ERG - home to-go has-asp, but
Varun ko nahi padega.
Varun-ERG - not have-asp.

The reason we cannot strand is because ‘padega’ (has: obligation aspect) is an auxiliary of ‘jaana’ (to go) in the Hindi sentence, not the head verb of the compound verb ”jaana padega” (has to go). The head here is ”jaana”, and therefore it must be induced in the site of ellipsis. Therefore, with the English verb ”has”, among others, we must also pick up its complements to induce at the site of ellipsis. To generalize, we require the complement cE of a head verb hE, when hE + cE results in a VV or verb-light verb complex in Hindi.

6.3 Serial Verbs

Examples of serial verbs in English are

a I’ll **go see** if she’s okay

b Why don’t you **run get** a taxi?

These can be treated as verb-complement series, as the Stanford Universal Dependency Framework treats them. Hindi also treats the verb complex as a head verb followed by complements. Taking the first example, it is acceptable to strand ”jaati” (go) from this sentence, in:

a Main jaati hu dekhne ke liye
I go-P.FSG - to-see-ERG - for
ki vah theek hai ki nahi. Vah nahi
whether she okay is or not. He not
jaayega.
will-go

7 Algorithm to Identify Ellipsis and Head Verb of Antecedent

We are using a dependency tree of the input sentence for all tasks.

7.1 Identification: Rules and Results

We assume that each word w in the input sentence that belongs to our five classes, is a site of ellipsis. Then we go through a process of elimination. We have a different set of rules for each class.

Some of the basic criteria for elimination include:

1. w is a copula (for to_be)
2. w is an auxiliary child of a verb (for to_be, to_do, to_have, modals)
3. w has a direct object noun child (for to_be, to_do, to_have, modals)
4. w is not an xcomp child (for to_particle)

We perform part 1 of our two-part testing at this stage. These rules, tested on 3900 sentences from the BNC corpus (Leech, 1992), give us a precision of **94.83 percent** and a recall of **83.04 percent**.

7.2 Antecedent Head Resolution: Algorithm

If we find an ellipsis, then we perform 2 sub-tasks:

1. Find the head verb
2. Supplement the head verb

7.2.1 Finding the Head Verb

We collect all the verbs in the input, eliminating according to the constraints on to_be/non-to_be ellipsis discussed in Section 3. We use a score-based approach, as does Hardt (Hardt, 1992).

Here are the features that we look at, for verb v as a candidate for ellipsis e:

1. Positive scores for noun subjects matching in number, negative score for noun subjects not matching in 'passivity', positive score if both noun subjects are proper, positive score for identical noun subjects.
2. Negative score if e belongs to the complement clause child of v. For example, in:
   He told me that he had passed the exam, and then he told me that John hadn't.
   "told" gets ruled out as the clause "John hadn't" is a complement child of "told".
3. Negative score if v belongs to the complement clause child of e.
4. Positive score if v has an auxiliary in the same class as e. For example:
   Sita could walk while texting, but now she can't.
5. Finally, we assign a positive score to the first verb that we obtain by backtracking up the dependency tree from e, if it is contained in our candidate verbs. This gives us the correct antecedent head several times, even when it is far away. It captures clausal relationships between antecedent verb and e, such as as...as, ...than, which Hardt also mentions in their scoring algorithm.
6. We then evaluate the scores. If there is a clash, we choose the closest verb, advantaging forward ellipsis to backward ellipsis, as the latter is widely acknowledged to be much rarer than the former.

7.2.2 Supplement Main Verb

As we said earlier, the head verb might need to be supplemented before it is 'stranded'. For example, in:

> Both banks have **been battered**, as other Arizona banks *have*, by falling real estate prices.

The above algorithm will identify "battered" as the head verb; however, we do need to supplement it with the auxiliary "been", to create a grammatical verb after inducing.

We perform three sub-tasks in supplementing the main verb v:

1. Add auxiliaries: we add any auxiliaries of v, after skipping the first if any that belongs to the same class as e. For example, we skip the auxiliary "have" in the above example.
2. Add particles: here, we check whether v is part of a phrasal verb/prepositional verb, and add the preposition(s) if so. The dependency tree marks particle dependants of the verb: however, since it doesn't always do so, we maintain and import a list of common phrasal verbs/prepositional verbs for reference.
3. Add verb complements. Similarly, we maintain and import a list of verbs of which verb complements, if any, we need to pick up, since they result in non-strandable verbs in isolation in Hindi: "let", "have". These also include verbs which may not always give a correct lexical translation in isolation, e.g. "feel", "seem", for which it is safer to also induce the complements.

The lists above can always be augmented, of course. Currently, our system does not deal with idioms, but one solution that we suggest is to maintain a list of frequently occurring idioms and induce them as a whole.

8 Results and Error Analysis for Antecedent Head Resolution

In part 2 of our testing, we tested this system on 200 sentences from the WSJ corpus and got

an accuracy of 91 percent on antecedent head resolution. Here are some errors, and their analysis:

> Mr. Dinkins also has failed to allay Jewish voters' fears about his association with the Rev. Jesse Jackson, despite the fact that few local non-Jewish politicians have **been** as vocal for Jewish causes in the past 20 years as Mr. Dinkins *has*.

Here, the algorithm identifies "failed" instead of "been". It awards "failed" for common noun subject, common auxiliary and being the first verb upon backtracking, whereas "been" is only awarded for common auxiliary.

> But Sony also says in its filing that the Warner contract "doesn't require that Guber and Peters **take** any affirmative steps to produce motion pictures; it simply rewards them when they *do* and prohibits them from producing for another entertainment company."

Here the algorithm identifies "rewards" as the source verb instead of "take". It awards "take" for noun subject number, "require" for common auxiliary, and "rewards" by backtracking. Finally, it resolves the tie by choosing "rewards" which is the closest.

> Now they **know** who you mean and you know who you mean - but no one else *does*.

The algorithm gives "mean" instead of "know". "mean" gets awarded for noun subject number, and "know" is awarded for backtracking; however, "mean" wins the tie since it is the closest.

There are errors introduced by the POS tagger; for example, in:

> A good half-hour into breakfast at the Palmer House, Mr. O'Brien looks up from his plate after Mr. Straszheim says something about people who believe interest rates are about to **nosedive** - "I'm one of them who hope they *will*, with 6 billion in debt on the books.

"nosedive" is not recognized as a verb, similar to, in another examples, "program-trade", and "move".

There are errors introduced by the parser:

> The text by Patrick O'Connor is a tough read, but the pictures make her magnetism clear and help explain why Ernest Hemmingway called Baker, "The most sensational women anybody ever **saw** – or ever *will*."

Here, the algorithm wrongly identifies "called" as the antecedent main verb, instead of "saw". The clause "or ever will" is labelled a conjugate dependent of the noun Baker, instead of a conjunct of the verb "saw". If this had been so, "saw" would have got scores for a common subject ("anybody") and being the verb obtained upon backtracking.

> "A lot of people think I will **give** away the store, but I can assure you I *will not*," he says.

The algorithm identifies "assure". This would be avoided if the dependency marked "I will not" as a complement clause of "assure" – however the dependency misses this relation. If "assure" was given a penalty on this grounds, the next highest candidate is indeed the correct one: "give".

9 Evaluation of Effect on MT outputs

We now show that inducing the head elliptical verb makes the input sentence easier for the MT system to comprehend. We perform a manual analysis of 100 sentences with ellipsis, taken from the WSJ corpus: we create a "before" and "after" translation pair for each, and compare to identify improvements. This was done by two fluent speakers of Hindi and English. We define a scale (1-5) to quantify this improvement:

5: Improvement from incoherent to perfect translation of ellipsis, and surroundings
4: The meaning is fairly clearer than it was in the original
3: The translation is as good or as bad as it

originally was
2: The meaning is fairly more obscure than it originally was
1: The sentence is rendered completely incoherent from an original good translation

We add some flags to further nuance this scale: we also mark the translations for fluidity, i.e. if the translation while rendered better is still not fluent (f) or if the translation while not making the meaning clearer is rendered more fluent (F), and for the overall meaning of the entire sentence – i.e. beyond the clauses of the antecedent and the ellipsis. These markers, however, are only for exceptional cases, since most of the translations we got, both "before" and "after" were not fluent.

The average score over 100 sentences was found to be 3.55, with 18 cases of correct non-fluent sentences, and 5 cases of incorrect sentences with especial improvement in fluency.

These are the large-scale sources of lack of improvement that we found:

1. Sentences that the MT cannot translate overall due to other complexities such as nested clauses, etc., for which its output is close to gibberish, show little to no improvement with addition of the elliptical verb. These sentences passed through a system that can handle them, sometimes Google Translate, almost always show high improvement with the addition of the verb. We had about 35 such sentences, all marked 3, sometimes marked for improvement or deterioration in fluency.

2. Most sentences of the type "as did", as in 'X **ate** apples *as did* Y' fail to show any improvement and often show a deterioration in fluency. This is because the processed sentence is rendered ungrammatical and perhaps more incomprehensible. Again, Google Translate often shows improvement on these sentences from "before" to "after".

An example of the first instance is:

(1) American Enterprise Institute scholar Norman Ornstein in the Oct. 21 TV Guide on "What TV News Doesn't **Report** About Congress – and *Should*"

The system induces the elliptical verb complex "report". Both Anusaaraka and Google output incorrect translations for the original sentence, although Google shows errors only due to the ellipsis. Therefore, it is able to improve its translation after we induce the elliptical verb:

(2) amerikee entarapraij insteetyoot ke
American Enterprise Institute-POS -
vidvaan norman orsteen ne teevee
scholar Normal Ornstin-ERG - T.V.
gaid mein "kaangres ke baare
guide in "Congress-ERG - about-in
mein kya teevee riport nahin hai -
- what T.V. report not does –
aur riport karanee chaahie":
and report does should"

(Score: 5)

Whereas Anusaarak outputs:

(3) American Enterprise sansthaan vidvaan
American Enterprise institute scholar
vastushaili Ornstein par T.V. guide
Norman Ornstein on TV guide
mein: "TV samaachaar vat congress
in "TV news what congress
about report nahi does.. aur haal
about report not does… and recently
likhta hai chahiye"
Writes-MSG - should-inf."

(Score: 3f, for deterioration in fluency)

The original gives a similarly incorrect output, though not quite as ungrammatical. This output, even after the elliptical verb has been added, conveys little to no meaning in Hindi.

Here are some micro-level sources of non-improvement:

1. When an antecedent verb is being used idiomatically, the MT system may interpret it literally when stranded in the elliptical clause, even if it catches the correct sense in the antecedent clause, possible because the former construction is more unusual.

2. In general, the VVPE construction fails if the system makes two different lexical

interpretations in the antecedent and elliptical clause. This may be for different reasons: e.g. on transitive verbs, since they appear in the elliptical verb without their objects, which is unnatural. The MT system will possibly attempt to catch an intransitive sense of the verb in such a situation. However, this is only in certain few cases of such verbs.

An example of both of the above is:

> During the takeover, Mr. Hahn said he would **put** his account **up** for review if WPP's bid were successful, but he *didn't*.

The system induces the elliptical verb complex "put up". Both Anusaaraka and Google output incorrect translations for the original sentence, although Google interprets the idiomatic meaning of "put up" correctly. Therefore, it is able to improve its translation after we induce the elliptical verb:

Anusaaraka makes error type 1 (literal interpretation of verb):

(4) Vah punarvalokan ke liye uska hisaab
He review-ERG - for his account
uthega yadi vap ke neelaam ki
lift-MSG if WPP-ERG - auction of
boli saphal the toh adhineekaran
bid successful be-PL then takeover-ERG
ke dauran, shreemaan Hahn ne
- during, Mr. Hahn-ERG -
kaha, parantu vah nahi utha tha.
said, but he not lift-PST - .

(Score: 3)

Google makes error type 2 (different lexical interpretation). *In this case*, it is minor, because the meaning of the sentence is restored from the original.

(5) adhigrahan ke dauraan, shree
takeover-ERG - during, Mr.
haahan ne kaha ki agar vah
Hahn-ERG - said that if DET
wpp kee bolee saphal rahee,
WPP-POSS - bid successful stays,
to vah sameeksha ke lie apana
then he review-ERG - for his
khaata rakh dega, lekin usane
account put give-asp, but he-ERG
nahin daala.
not put.

(Score: 4)

We note here that these figures are dependent upon how well the base translation system can translate the original. We performed the same analysis on samples from this dataset with Google Translation and got consistently better results per each batch of 10, and an average of 3.7. This indicative exercise is intended to give an idea of why targeted VPE handling for specific language pairs holds significance in bettering MT results.

10 Future Work

The concept and the system that we have introduced above, for handling ellipsis in a targeted manner to improve English-Hindi MT, are still in their nascent stages. There are three primary entry points for future work: firstly, the conceptual negotiation of the phenomenon in English and Hindi. We have decided, as we explain, only to induce the main verb. While this gives satisfactory results most of the times, it also fails in some cases: it might help, for example, to make a decision to induce object arguments in these cases. Secondly, the identification of VPE in English, and the antecedent resolution. We are already dealing with complex verbs, verbal complements etc. in a certain manner, but this treatment invites further and more rigorous work, both in terms in nuance with the treatment, and how exhaustive we are with the lists that we have drawn up, introducing, for example, treatment of idioms. Thirdly, the application of our system as over the input, before it is passed through the MT. There are certain problems that may be solved by pipelining this ellipsis handling differently into the MT system: we leave this to future investigation.

References

J. Bos and J. Spenader. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494, 2011.

D. Hardt. An algorithm for vp ellipsis. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 9–14. Association for Computational Linguistics, 1992.

K. Kenyon-Dean, J. C. K. Cheung, and D. Precup. Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. In *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing, pages 1734–1743, 2016.

G. N. Leech. 100 million words of english: the british national corpus (bnc). 1992.

Z. Liu, E. G. Pellicer, and D. Gillick. Exploring the steps of verb phrase ellipsis. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 32–40, 2016.

E. Manetta. Verb-phrase ellipsis and complex predicates in hindi-urdu. *Natural Language & Linguistic Theory*, pages 1–39, 2018.

L. A. Nielsen. *A corpus-based study of Verb Phrase Ellipsis Identification and Resolution*. PhD thesis, King's College London, 2005.

D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

E. Voita, R. Sennrich, and I. Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. *arXiv preprint arXiv:1905.05979*, 2019.