

An Empirical Analysis of Pronoun Translation Evaluation Metrics from English to Hindi Machine Translation Perspective

Chetan Agarwal, Kamlesh Dutta, Pardeep Singh

Computer Science and Engineering Department,

National Institute of Technology, Hamirpur (H.P.), INDIA.

(e-mail: chetanswadeshi@gmail.com, kd@nith.ac.in, pardeep@nith.ac.in)

HOW TO CITE:

Chetan Agarwal, Kamlesh Dutta, Pardeep Singh (2026). An Empirical Analysis of Pronoun Translation Evaluation Metrics from English to Hindi Machine Translation Perspective. International Journal of Special Education, 41(4s), 511-529.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Background: Recognizing the incompatibility of traditional or general-purpose automatic evaluation metrics in measuring the pronoun translation quality of a *machine translation output* (MTO), researchers evolved some pronoun specific metrics to measure the pronoun translation quality.

From this point of view, this research paper is aimed: (i) to have a systematic study about such pronoun specific evaluation metrics, (ii) to check their suitability for an *English to Hindi machine translation* (EHMT) system and (iii) to perform an empirical analysis of these metrics with respect to an EHMT system.

Method: Accounting to three major selection criteria viz. ease of implementation, language independent methodology and scope of performance enhancement, related literature was explored and surveyed. Based on this survey and exploration, out of eight evaluation scripts, two evaluation scripts namely as *Precision, Recall and F1-measure* (PRF) and *Accuracy of Pronoun Translation* (APT) metric were adopted for the purpose of this study and analysis. Since both evaluation metrics originally evolved for some other source and target language pair, both metrics were modified appropriately making it suitable for an EHMT system, then conducted an experiment to analyze the performance.

Result: Results show that APT shows very close correlation with human judgements. F-measure attained by PRF and APT was 53% and 68% respectively.

Conclusion: Study and analysis revealed that PRF and APT both metrics are enough suitable for an EHMT system and can be adapted with certain language specific modifications. Experiment results are quite encouraging and show that with accurate word alignment, appropriate configuration setting and case wise weight assignment, APT metric outperform over PRF.

Keywords: Natural Language Processing, Pronoun Translation Evaluation, Anaphora Resolution, Evaluation Metrics, Machine Translation, Precision and Recall

1. INTRODUCTION

Since manual evaluation methods used for *machine translation* (MT) quality assessment are costlier and complicated, as well as non-reusable and time consuming also (Chatzikoumi, 2019; Bentivogli et al, 2018), MT researchers evolved some automatic evaluation metrics to guide their development efforts. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), WER (Vidal, 1997), mWER (Niessen et al., 2000) etc. are some of the renowned general-purpose MT evaluation metrics. Since all these evaluation metrics evolved to measure the overall translation quality of a MTO, their evaluation methodology, underlying framework, and logical building blocks, were not found suitable to assess the pronoun translation quality effectively (Jwalapuram et al., 2019).

Advocating the need of pronoun translation metric, Jwalapuram et al., (2019) stated that, “Though, going beyond the sentence-level may assist in resolving various discourse-level ambiguities (e.g. pronominal anaphora) hence may yield better translations. But typically, there are very few changes in the translation which may appear substantial for humans but typical MT evaluation measures like BLEU would not be able to notice these changes.

While discussing pronominal divergence for “it”, Dutta et. al. (2009) gave examples (Table 1) to show how translation of “it” depends upon the type of divergence. These changes are reflected in translation but general-purpose evaluation metrics cannot capture these small changes.

Table 1: Pronominal divergence of “it” in Hindi Translation (Dutta et. al. 2009).

Anaphora	Translation of “it” in Hindi	Divergence
Nominal Anaphora	us-ko/use	Case-based
Clausal Anaphora	yeh	Case-based
Pro-action	us-ko/use	Case-based
Cataphoric	veh	Case-based
Discourse Topic	-	Pronominal

Pleonastic	-	Pronominal
Idiomatic	-	Pronominal

Following example illustrates in detail how general-purpose evaluation metrics are not appropriate to measure pronoun translation quality of an MTO.

Ex 1

S1: *His death has caused great damage to the political world.*

RT: उनके निधन से राजनीतिक जगत को गहरा नुकसान हुआ है।

Roman Gloss: (*Unke nidhan se raajneetik jagat ko gabraa nuksaan hua hai* |) (Manual translation)

CT: उनकी मौत से राजनीतिक दुनिया को काफी नुकसान हुआ है।

Roman Gloss: (*Unki mritu se raajneetik duniyaa ko kaafee nuksaan hua hai* |)

(Bingo Translation, performed on 20 Nov 2023)

NT: उनके मौत से राजनीतिक दुनिया को काफी नुकसान हुआ है।

Roman Gloss: (*Unka mritu se raajneetik duniyaa ko kaafee nuksaan hua hai* |)

(Noisy Translation, manipulated for comparison)

S2: Last week the WHO allowed the Covid-19 vaccine BBIBP-CorV made in China to use emergency. **This** means that **this** vaccine can now be used all over the world.

RT: पिछले हफ्ते WHO ने चीन में बने कोविड-19 वैक्सीन BBIBP-CorV को आपातकाल में उपयोग करने की अनुमति दी। इसका मतलब है कि यह टीका अब दुनिया भर में इस्तेमाल किया जा सकता है।

(Manual translation)

Roman Gloss: (*Pichle haphte WHO ne cheen mein bane kovid-19 vaicseen BBIBP-CorV ko aapaatkaal mein upyog karne ki anumati dee* | *Iskaa matlab hai ki yah teekaa ab duniyaa bhar mein istemaal kiya jaa saktaa hai* |)

CT: पिछले हफ्ते WHO ने चीन में बनी कोविड-19 वैक्सीन BBIBP-CorV को इमरजेंसी इस्तेमाल की इजाजत दे दी। अर्थ इस टीके का इस्तेमाल अब पूरी दुनिया में हो सकेगा।

(Bingo Translation, performed on 20 Nov 2023)

Roman Gloss: (Pichle haphthe WHO ne cheen mein bane sinofarm ki kovid-19 vaicseen BBIBP-CorV ko imergency istemaal kee izaatat de dee | Atharth is teeke ka istemaal ab pooree duniyaa mein ho sakega |)

NT: पिछले हफ्ते WHO ने चीन में बनी कोविड-19 वैक्सीन BBIBP-CorV को इमरजेंसी इस्तेमाल की इजाजत दे दी | अथार्थ यह टीके का इस्तेमाल अब पूरी दुनिया में हो सकेगा |

(Bingo Translation, performed on 20 Nov 2023)

Roman Gloss: (Pichle haphthe WHO ne cheen mein bane sinofarm ki kovid-19 vaicseen BBIBP-CorV ko imergency istemaal kee izaatat de dee | Atharth yah teeke ka istemaal ab pooree duniyaa mein ho sakega |)

(Noisy Translation, manipulated for comparison)

In example 1, both source sentences S1 and S2 are part of some news articles, while RT, CT and NT are their corresponding reference translation (i.e. actual translation accomplished by human translator) and candidate translations (i.e. machine translation under study) and noisy translations (i.e. manipulated knowingly for comparison of Evaluation)

Here, it can be observed that for pronoun “his” in S1, “उनके” and “उनकी” are corresponding translation in RT and CT. Similarly, in sentence S2, for both instances of pronoun “this”, there are different translations in RT and CT. On one hand, there is no candidate translation in CT for first instance of “this”, while in RT, its corresponding translation is “इसका”. Similarly, second instance of “this” also have different translations, “यह” in RT and “इस” in CT.

Though both translations are correct, however for general purpose automatic MT evaluation metrics like BLUE, TER etc., this accuracy not only remains unnoticeable even may account as word mismatch penalty. Translation score obtained by different metrics, shown in Table 2, illustrates this phenomenon more explicitly.

Here, for both sentences, it can be observed that though CT is more accurate compared to NT, but still BLUE score and METEOR score of NT is little bit higher than CT, while TER score of NT is little bit lagging than CT.

Table 2: Evaluation score measured by different metrics for both sentences.

Score	Evaluation Metric					
	BLEU		TER		METEOR	
	CT	NT	CT	NT	CT	NT
S1	.28	.29	.36	.35	.61	.62
S2	.23	.24	.57	.55	.47	.48

To overcome such limitation of general-purpose evaluation metrics, researchers attempted to evolve some pronoun specific metrics.

According to the literature review, Hardmeier and Federico (2010) attempted first in this direction and proposed a precision and recall based F-measure metric. Later, Guillou and Hardmeier (2016) proposed a test suite named PROTEST, Werlen and Belis (2017) proposed an APT metric. Jwalapuram et al. (2019) proposed another evaluation measure to differentiate good and bad pronoun translation.

Since all these evaluation measures were originally proposed for a group of European languages, an evaluation measure particularly from English to Hindi translation perspective was required. So being aligned in line of this goal, there was a need either to evolve some new evaluation approach or to adapt and modify existing evaluation approaches and then have a performance analysis of the same.

Towards this aim, evaluation metrics included for the purpose of study and analysis are as; (i) Precision, Recall and F1-measure (PRF) metric and (ii) Accuracy of Pronoun Translation (APT) metric. Though, both evaluation metrics originally evolved for some other source and target language pair, however because of their generic nature i.e., language independent, both evaluation scripts have been adapted and modified appropriately, making it suitable for an EHMT system.

In the light of the above discussion about the background of problem, this study aims to address following research questions (RQs).

RQ1: Is there any specific pronoun category which is imposing challenge in pronoun translation?

RQ1 aims to identify those pronouns whose incorrect or inappropriate translations frequency is comparatively high and hence imposing challenges for pronoun translation. This static may help to

identify the reason why certain pronouns only are becoming problematic and may suggest the remedy as well.

RQ2: Which one of these two metrics is more suitable for EHMT systems?

RQ2 is concerned with the study and analysis of underlying structure, performance measure, pros and cons, implementation complexity of both these metrics. This study might help to decide which one of these should be adopted for EHMT system.

RQ3: What changes are required in these metrics to make them suitable for an EHMT system?

RQ3 aims to go into the implementation detail of both metrics and to identify if any code section needs to be changed to make them compatible for EHMT system.

RQ4: Can both of metrics be combined in one?

RQ4 aims to identify possibilities of integrating the features of both metrics in one metric.

Table 3 maps research questions with corresponding sections of this article which deals with these research questions.

After introducing about the objectives of the paper in section 1, remaining part of the paper has been structured under the following sections: Section 2 starts with a discussion about important issues in pronoun translation quality measurement up to a significant extent. Section 3 provides a review of related literature. Section 4 highlights evaluation methodology and basic working principles of evaluation metrics considered for study and analysis in this paper. Section 5 describes the establishment of experimental setup and data set required for performance evaluation of metrics under study. Section 6 discusses the results of experiments conducted for PRF and APT. At the end, section 7 depicts the findings of this study along with future research direction.

To incorporate the PRF metric feature in APT metric, a module to APT metric has also been proposed in section 7.3. This additional module would enable the APT metric to measure the accuracy of pronoun translation at individual level as well as a comprehensive score for all pronouns.

2. ISSUES IN PRONOUN TRANSLATION QUALITY EVALUATION

After a thorough analysis and examination, researchers identified some major issues related to pronoun translation evaluation. Diversity in pronouns functionality, evaluation metric, multiple possible translation, omitted translation, pronoun alignment and differences in pronoun system of source-target language pair are among important issues of pronoun translation related problem. This section describes above mentioned pronoun translation evaluation related issues from English to Hindi translation point of view.

2.1 Evaluation Metric

Evaluation poses one of the major problems for researchers interested in MT, specifically in pronoun generation. Ex 2 illustrates how evaluation metric poses the challenge in pronoun translation evaluation.

Ex 2

S1: *BJP's veteran leader Prof Chaman Lal Gupta has a long political journey of five decades. His death has caused great damage to the political world.*

RT: भाजपा के दिग्गज नेता प्रो चमन लाल गुप्ता का पांच दशक का लंबा राजनीतिक सफर है | उनके निधन से राजनीतिक जगत को भारी नुकसान पहुंचा है | (Manual translation)

Roman Gloss: (BJP ke diggaj neta Prof. Chaman Lal Gupta ka paanch dashak lambaa raajneetik safar hai | Unke nidhan se rajneetik jagat ko bhaaree nuksaan pahunchaa hai |)

CT: भाजपा के दिग्गज नेता प्रो चमन लाल गुप्ता का पांच दशक का लंबा राजनीतिक सफर है | उनकी मौत से राजनीतिक दुनिया को काफी नुकसान हुआ है | (Bingo Translation, performed on 20 Nov 2023)

Roman Gloss: (Bhajapaa ke diggaj neta Prof. Chaman Lal Gupta ka paanch dashak lambaa raajneetik safar hai | Unki maot se rajneetik duniya ko kaafi nuksaan hua hai |)

Although, in Sentence S1 of Ex 2, for English pronoun 'his', both translation "उनके" and

“उनकी” are correct in respective translation T1 and T2. Nevertheless, if one of these translations is considered as reference translation and another translation is considered as candidate translation, then during pronoun translation evaluation this correct translation too would be considered as mismatch and hence would account for low pronoun translation accuracy.

2.2 Pronoun Alignment

For the accuracy of any reference-based pronoun translation evaluation metric, a precise alignment of source and target pronouns for both translations, reference as well as candidate, is one of the essential requirements. Therefore, given the list of source pronouns considered for evaluation, the first step is to obtain their corresponding alignments in the target language texts. This requirement of precise alignment between source and target pronoun also poses a major challenge in pronoun translation evaluation. On one hand, because of the high cost and associated difficulties with manual word alignment, researchers must rely on automatic alignment tools. On the other hand, poor performance of automatic word alignment tools makes evaluation metrics performance unreliable. Because of this improper word alignment, there might be situations when a matched pronoun translation pair may in fact be wrong and vice versa also. In the case of the candidate translation, the alignment can be directly obtained from the MT system if such an option is available. However, in the case of reference, it is necessary to perform automatic word alignment. Ex 3 illustrates the impact of pronoun alignment on evaluation measure.

Ex 3

SS: *At (1) present (2) she (3) lives (4) in (5) the (6) Afghan (7) Basti (8) of (9) Delhi (10).*

RT: *फिलहाल (1) वह (2) दिल्ली (3) की (4) अफगान (5) बस्ती (6) में (7) रहती (8) है (9) | (Manual translation)*

Roman Gloss: *(Filhaal vah Delhi kee Afgaan Basti mein rahtee hai |)*

SS-RT Alignment: {1-1}, {2-1}, {3-2}, {4-8}, {4-8}, {5-7}, {6-0}, {7-4}, {8-5}, {9-4}, {10-3} *(Manual word alignment)*

CT: *वर्तमान (1) में (2) वह (3) दिल्ली (4) की (5) अफगान (6) बस्ती (7) में (8) रहती (9) है (10) | (Bingo Translation, performed on 20 Nov 2023)*

Roman Gloss: *(Vartmaan mein vah Delhi kee Afgaan Basti mein rahtee hai |)*

SS-CT Alignment: {1-1}, {2-1}, {3-0}, {4-8}, {5-8}, {6-0}, {7-5}, {8-6}, {9-4}, {10-3} *(Giza++ word alignment, (Och and Ney, 2000))*

English pronoun “*she*” appeared at 3rd index in source sentence SS has been translated as “*वह*” in both translations, reference as well as candidate translation. In reference translation “*वह*” has appeared at 2nd index while in candidate translation “*वह*” has appeared at 3rd index. Now it can be observed that, though RT and CT both has same translation, however in SS-RT alignment (correct manual alignment) it has been mapped correctly as {3-2}, but in SS-CT alignment (automatic word alignment tool GIZA++) it has been mapped incorrectly as {3-0}. Because of such improper word alignment, correct translation too would be considered as mismatch, hence would produce false result and accounting poor pronoun translation accuracy.

Table 3: RQs mapping with article sections.

RQs	RQ title	Objective	Process	Sections addressing the RQ
RQ1	Is there any specific pronoun category imposing challenge in pronoun translation?	Identify the reason of the problem and suggest the possible remedy.	Experimental result analysis.	6.1 and 6.2
RQ2	Are metrics compatible for an EHMT system?	Determine which metric is more suitable for an EHMT system.	Study of the basic working principles.	7.4
RQ3	How evaluation process is being implemented?	Recognize required changes to make metrics suitable for an EHMT system.	Understand the mapping of functional block with code block	4.1 and 4.2
RQ4	Can both metrics be combined in one?	Explore the possibility of combining the features of both metrics into one.	Add or modify the required code modules.	7.3

Multiple and Omitted Translation

In some cases, there might be more than one correct translation for a given pronoun. Similarly, in some cases there might be situations when source pronoun is translated in reference translation but omitted in candidate translation or vice versa. Ex 4 illustrates the issue arising out of the multiple and omitted pronoun translation.

Ex 4

S1: *Mrs. Sushma Swaraj had a long political career. She had a chance to represent in UN also.*

RT: श्रीमती सुषमा स्वराज का लंबा राजनीतिक जीवन रहा। उन्हें यूएन में भी प्रतिनिधित्व करने का अवसर मिला।

(Manual translation)

Roman Gloss: (*Shrimati Sushmaa swaraaj ka lamba raajneetik jeevan rahaa hai | Unhen UN mein bhi pratinidhitva karne kaa avsar milaa |*)

CT: श्रीमती सुषमा स्वराज का लंबा राजनीतिक जीवन रहा। **उनको** यूएन में भी प्रतिनिधित्व करने का मौका मिला।

Roman Gloss: (*Shrimati Sushmaa swaraaj ka lamba raajneetik jeevan rahaa hai | Unkon UN mein bhi pratinidhitva karne kaa mauka milaa |*)

Although, in Sentence S1 of Ex 4, for English pronoun 'she', both translation “**उन्हें**” and “**उनको**” are correct in respective translation RT and CT and can be exchanged without affecting the translation. However, just because of the lexical mapping consideration, both translations would be counted as mismatched translation and hence would account for low pronoun translation accuracy.

Table 4: Feature based Comparison between PRF, APT, PROTEST and GBPT

Feature	Evaluation Metrics			
	PRF	APT	PROTEST	GBPT
Proposed by	Hardmeier and Federico (2010)	Werlen and Popescu-Belis (2017)	Guillou and Hardmeier (2016)	Jwalapuram et al. (2019)
Evaluation	Automatic	Automatic	Semi-automatic	Automatic
Context	Without context	Without context	With and without context	With and without context
Initial language pair	English to French	Spanish to English	English to German	French to English
Performance measurement level	Single comprehensive score as well as at level of individual pronoun	Single comprehensive score over all pronoun	At level of individual pronoun	Single comprehensive score over all pronoun
Complexity	Simple	Simple	Moderate	High
Language dependent	Independent	Independent	Dependent	Independent
Methodology	Matching of reference and candidate translation	Matching of reference and candidate translation with case wise weight selection	Matching of reference and candidate translation with antecedent agreement	Distinguish between good and bad translation
Multiple system output	Single candidate translation	Single candidate translation	Multiple candidate translations	Single candidate translation
Source language input	Single	Single	Single	Multiple
Performance measure	Number of translation pairs matched	Number of translation pairs matched	Number of translation pairs matched and	Ranking Based

3. RELATED WORK

Review of related literature reveals that a limited work only is available for the pronoun translation measure and maximum of relative work has come out in last one decade only. The emergence of the need of pronoun translation evaluation has its root in MT evaluation and role of anaphora resolution in MT. According to the available literature, though anaphora resolution was a topic of interest for NLP and MT researchers from the very early days of MT evolution itself, but MT evaluation metrics which are widely known and used, came into existence in the beginning of 21st century only.

3.1 General Purpose Evaluation Metrics

“BLEU” (Papineni et al., 2002) is considered as one of such and most widely used automatic evaluation metric. The BLEU was designed by IBM and its evaluation was based on n-grams matching from multiple references. A close variant of BLEU metric was proposed by Doddington (2002), which was referred to as “NIST”. Banerjee and Lavie (2005) presented another evaluation measure, referred to as “METEOR”. This evaluation metric uses unigram matches on the words and their stems, then it counts the number

of exact words matches between the system output and reference. A new intuitive measure of translation named as “TER”, was presented by Snover et al. (2006). Basic idea behind TER was to compute the amount required for editing that a human would have to perform to change a system out-put, so it exactly matches a reference translation.

Besides these most widely known and used MT evaluation metrics, several other metrics also have been proposed. WER (Vidal (1997); Tillmann et al. (1997)), mWER (Niessen et al., 2000), GTM (Melamed et al., 2003), CDER (Leusch et al., 2006), METEOR-Hindi (Gupta et al., 2010), METEOR-Universal (Denkowski et al., 2014) etc. are among the other widely known MT evaluation metrics. Segment level MTE (Shimanaka et al., 2018) and AdaBLEU (Chauhan et al., 2021) have been noticed as the most recent MT evaluation metrics.

3.2 Pronoun Translation Evaluation Metrics

Since all these evaluation metrics discussed so far evolved as traditional and general-purpose evaluation metrics, therefore were not suitable for the evaluation of various discourse-level aspects such as discourse connectives translations, pronominal anaphors translation etc.

- (a) PRF metric: According to study and observation of literature, first noticeable effort attempted for pronoun translation evaluation was made by Hardmeier and Federico (2010). With a basic idea of counting the number of correctly translated pronouns, Hardmeier and Federico proposed a direct method to measure the precision and recall of pronoun translations.
- (b) ACT metric: To measure the accuracy of discourse connectives translation, Hajlaoui and Popescu-Belis (2013) presented an automatic metric “ACT”. ACT automatically compares the translation of discourse connectives in MT output with those in the reference translation.
- (c) Manual Evaluation: Pronoun translation has been the target of a shared task at DiscoMT and WMT workshops in 2015-2017 (Hardmeier et al., 2015; Guillou et al., 2016; Loaiciga et al., 2017). Some other researchers proposed manual evaluation methods also, like

pronoun selection task used in the DiscoMT 2015 shared task evaluation (Hardmeier et al., 2015), methods based on manual counting (Le Nagard and Koehn, 2010; Guillou, 2012; Novak et al., 2013).

- (d) PROTEST: Guillou and Hardmeier (2016) proposed another test suite, “PROTEST” to be used in semi-automatic evaluation measure. PROTEST was comprised of 250 hand selected pronoun tokens and an automatic evaluation script. The evaluation script was used to compare the pronoun in output of a MT system against pronoun in a reference translation. If in case pronoun translation of MT output does not match with given reference, evaluation process was aided with manual observation.
- (e) APT metric: To evaluate the accuracy of pronoun translation, another reference-based metric, “APT” was presented by Werlen and Popescu-Belis (2017). APT also relied on a reference translation and on a comparison of the candidate translation with the reference translation to compute the evaluation scores.
- (f) Probability based metric: Based on probability conversion over antecedents to the features of translated pronouns, Luong et al. (2017) proposed a probabilistic model and improved the translation of personal and possessive pronouns for Spanish-English translation system. They combined the anaphora resolution approach (used for Spanish) with the phrase-based MT system. Another semi-automatic system is described in Guillou et al. (2018). It focused on just two pronouns, “it” and “they”, and was applied to a single language pair.
- (g) GBPT metric: Jwalapuram et al. (2019) proposed a novel ranking based and source language independent pronoun evaluation measure to differentiate good and bad pronoun translation. Cai and Xiong (2019) proposed a test suite to evaluate three common discourse phenomena: pronoun, discourse connective and ellipsis, in Document-level NMT.

Table 4 highlights the key differences between four pronoun translation evaluation metrics: PRF, APT, PROTEST and GBPT metrics.

4. EVALUATION METHODOLOGY OF PRF, APT AND PROTEST METRICES

This section first presents basic working principles and concepts of three specific pronoun translation evaluation metrics and then presents their comparative study.

4.1 PRF Metric

According to the study of available literature, PRF metric proposed by Hardmeier and Federico (2010) was the first attempt in the direction of measuring the pronoun translation quality of an MTO. Resembling the BLEU metric (Papineni et al., 2002), this metric also gives an indication about similarity in candidate and reference translation of an individual pronoun. In other words, this metric is a measure of appropriateness of a MT system in translating an individual pronoun in compared to the reference translation of the same pronoun. *Precision* and *Recall* are computed based on three calculations; (i) total number of translations appeared for an individual pronoun in candidate translation, represented as $|c|$, (ii) total number of translations appeared for the same pronoun in reference translation, represented as $|r|$ and (iii) total number of translations matched in candidate and reference translation for that pronoun, represented as $|r \cap c|$. From these three calculations, at first, precision and recall are computed and then for the purpose of normalization, harmonic mean of *Precision* and *Recall*, termed as *F-measure* is computed. The “*Precision*” represents the ratio of total number of translations matched for a particular pronoun in candidate and reference translation, i.e., $|r \cap c|$, and total number of translations appeared for that pronoun in candidate translation, i.e., $|c|$.

Similarly, “*Recall*” is the ratio of total number of translations matched for a particular pronoun in candidate and reference translation, i.e., $|r \cap c|$, and total number of translations appeared for that pronoun in reference translation, i.e., $|r|$. Equations 1, 2 and 3 (Hardmeier and Federico, 2010) represents the formula used for precision, recall and F-measure computation, while figure 1 represents the flow chart of precision and recall computation procedure.

$$\text{Precision}(P) = \frac{|r \cap c|}{|c|}$$

Eq. 1

$$\text{Recall}(R) = \frac{|r \cap c|}{|r|}$$

Eq. 2

$$F - \text{Measure}(F) = \frac{2 * P * R}{P + R}$$

Eq. 3

4.2 APT

In recent, Werlen and Popescu-Belis (2017) had another important attempt to evaluate the accuracy of pronoun translation in an MTO and proposed another reference-based metric “*APT*”. Unlike precision/recall and PROTEST, which were designed to support analysis of system performance at the level of individual pronoun groups, APT metric provides a single aggregate measure over all pronouns.

Basic idea behind the APT score computation is to compute the ratio of total number of correctly translated pronouns and total number of pronouns taken into consideration for translation. For this purpose, at first APT metric matches the candidate translation with reference translation for every appearance of source pronoun taken into consideration and then counts the total number of identical, equivalent, and different translation pairs. If for a given pronoun, candidate and reference translation is identical, then translation is considered as correct, and a weight associated to this case type is 1. If in case, candidate translation and reference translation does not match, then two situations might be there; either candidate translation is incorrect, or it might have been translated by another equivalent target pronoun. Two target pronouns are considered as “*equivalent pronouns*” if these pronouns can be exchanged in most contexts without affecting the meaning of the sentence. If candidate translation is incorrect, then the weight associated with this case type is 0 and if candidate translation is equivalent then weight associated with this case type is 1. In this way, after assigning a weight to each case type, the total number of identical, equivalent, and different translation pairs are counted for each case type and then this count is multiplied by their associated weight. Finally, these multiplications of associated

weight and number of translation pairs of a particular case type are summed up and then divided by total counts of translation pairs.

Equation 4 (Werlen and Popescu-Belis, 2017) represents the formula used for APT score computation and figure 2 represents the flow chart of APT score computation.

$$APT = \frac{\left(\sum_{i=1, c_i \in c_d}^m w_i * n_{c_i} \right)}{\left(\sum_{i=1, c_i \in c_d}^m n_{c_i} \right)} \quad Eq.4$$

5. EXPERIMENTAL SETUP

This section describes composition and establishment of experimental setup along with data base used for pronoun translation quality evaluation metrics.

5.1 Composition of Test Data Set

Since both metrics are reference based metrics, both metrics require a candidate translation and a reference translation for a given source text. For this purpose, a parallel corpus of 3000 sentences was constituted from various sources like Newspapers, TED talks, social-media, and websites content etc. For obtaining the candidate translation, source sentences were translated by MS Word's online translation service, Bingo translator. Rather to be too specific for two or three pronouns like PRF and to generalize the experimental setup, under this experimental set up evaluation metrics were tested for a set of 30 different English pronouns. Hence both metrics compute the pronoun translation quality for only these 30 pronouns.

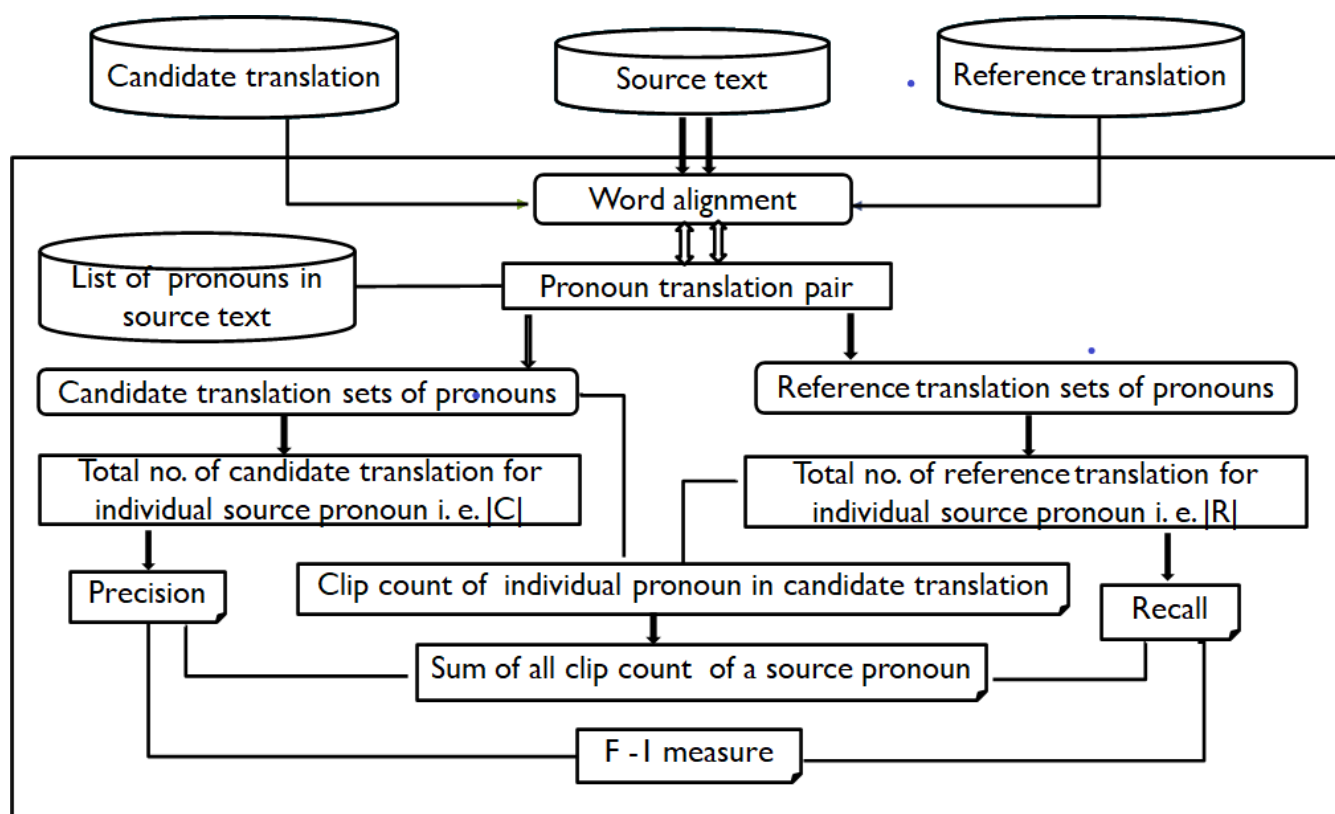


Figure 1: Flow chart for precision and recall computation.

Table 5: Strength and weakness (Comparison) of PRF and of APT metric

PRF	APT
Weakness	Strength
High frequency of pleonastic pronouns would cause reduced efficacy.	Such type of situation can be handled by assigning appropriate weight case and increasing entries in dictionary.
Matching of reference and candidate translation	Matching of reference and candidate translation with case wise weight selection
Strength	Weakness
support analysis of system performance at the level of individual pronoun groups.	APT metric provides a single aggregate measure over all pronouns

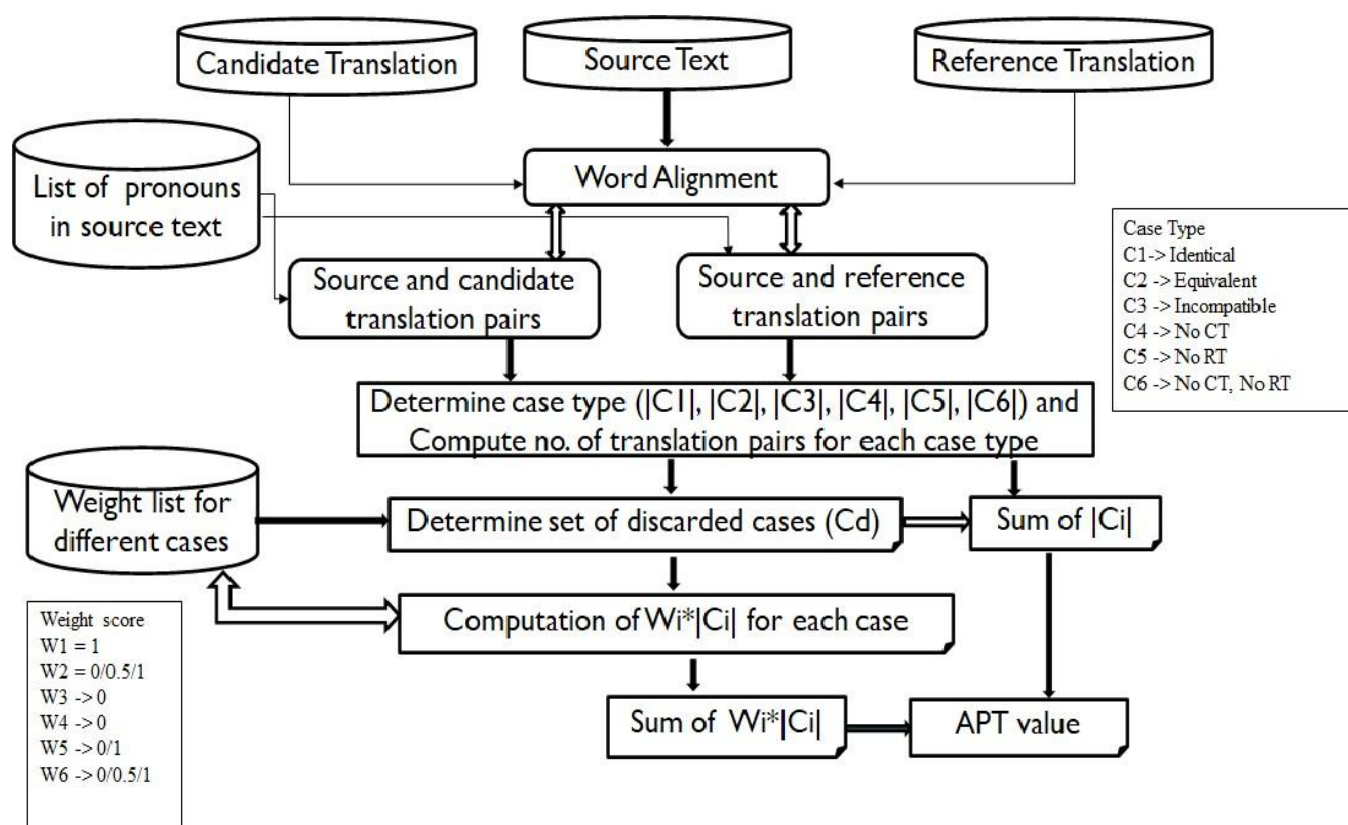


Figure 2: Flow chart for APT score computation

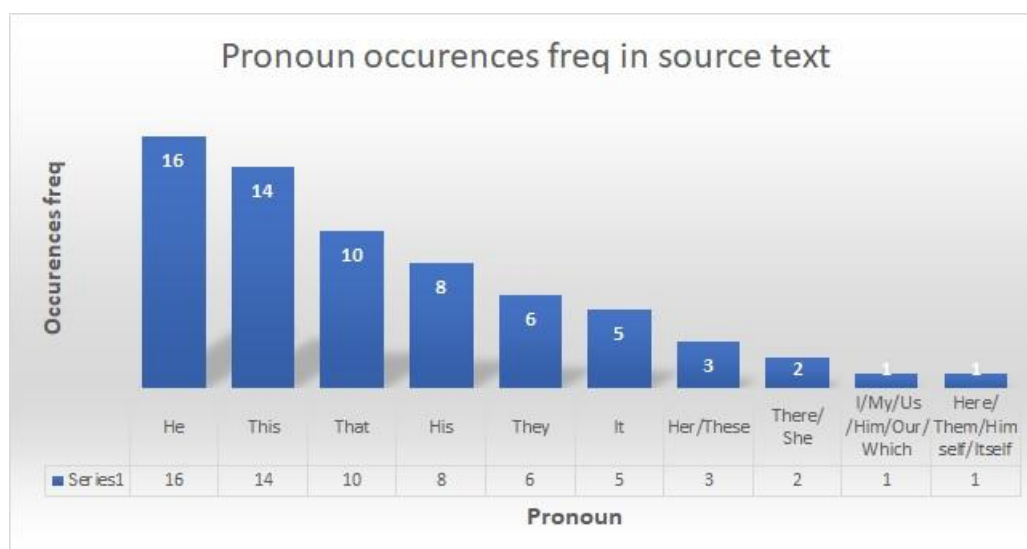


Figure 3: Pronoun occurrence Frequency Chart

Figure 3 provides a view of occurrences frequency for all 30 pronouns considered for evaluation in the given English source text. Among the almost 120 pronouns of English language (www.thefreedictionary.com/List-of-pronouns.htm), all these 30 pronouns belong to different pronoun categories and occurs very frequently in English text. Out of these 30 pronouns, 7 pronouns (I, he, she, we, they, it, you) are personal, 7 pronouns (his, her, my, your, its, our, their) are possessive or relative, 5 pronouns (itself, myself, himself, herself, ourself) are

reflexive or intensive pronoun, 4 pronouns (me, him, us, them) are dative or objective, 4 pronouns (this, that, these, those) are demonstrative and 3 pronouns (which, here, there) are place pronouns.

PRF facilitates the efficient evaluation and inspection of the translation of individual pronouns, while APT produces a comprehensive score for all pronouns. The samples examined in this case study are drawn from the English-Hindi parallel corpus developed as a data set for this pronoun translation evaluation task.

Table 6: Four different configurations of APT metric

Parameters	config_1	config_2	config_3	config_4
list_source_pronoun	Compulsory	Compulsory	Compulsory	Compulsory
list_target_pronoun	No	No	Yes	Yes
dictionary of equal pronouns	No	No	No	No
dictionary of similar pronouns	No	Yes	Yes	Yes
source pronouns	NA	NA	NA	NA
target pronouns	NA	NA	NA	NA
Weight Case	1, 0, 0, 0, 0, 0	1, 0.5, 0, 0, 0, 0.5	1, 0.5, 0, 0, 0, 0.5	1, 0.5, 0, 0, 0, 0.5
count_OTHER_as equal	False	False	False	True

Table 7: Precision, Recall and F-score computed by PRF metric for all individual pronouns

Pronoun	M/CO	Precision	M/RO	Recall	F1
He	6/16	0.3750	6/16	0.3750	0.3750
Her	1/2	0.5000	1/3	0.3333	0.4000
Here	1/1	1.0000	1/1	1.0000	1.0000
Him	1/1	1.0000	1/1	1.0000	1.0000
Himself	1/1	1.0000	1/1	1.0000	1.0000
His	6/8	0.7500	6/8	0.7500	0.7500
I	1/1	1.0000	1/1	1.0000	1.0000
It	0/2	0.0000	0/2	0.0000	NaN
Itself	1/1	1.0000	1/1	1.0000	1.0000
My	1/1	1.0000	1/1	1.0000	1.0000
Our	0/1	0.0000	0/1	0.0000	NaN
She	0/2	0.0000	0/2	0.0000	NaN
That	5/9	0.5556	5/9	0.5556	0.5556
Them	0/1	0.0000	0/1	0.0000	NaN
There	2/2	1.0000	2/2	1.0000	1.0000
These	0/2	0.0000	0/2	0.0000	NaN
They	4/6	0.6667	4/6	0.3333	0.3333
This	8/14	0.5714	8/14	0.5714	0.5714
Us	0/1	0.0000	0/1	0.0000	NaN
Which	1/1	1.0000	1/1	1.0000	1.0000

To check the maximum achievable efficiency or accuracy of the above metrics, rather than obtaining the word alignment with the help of any word alignment tool, like mGIZA or GIZA++, word alignment was done manually.

5.2 Weight and Parameter Selection for APT Metric

To improve the APT score as well as to reflect an improvement in the human perception of pronoun translation quality, i.e., to increase the human correlation, APT metric has

a provision of tuning its parameters under different configuration settings.

For example, depending upon the language pair involved in translation process and set of pronouns under consideration, employing “list of target pronoun”, “dictionary of equal pronouns”, “dictionary of similar pronouns”, “count_OTHER_as_equal” and associating appropriate weights with different cases, etc., would enable APT metric to achieve a high a score. Table 6 depicts the summary of four such configuration setting of APT metric used for

pronoun translation evaluation for English-Hindi language pair.

6. RESULT ANALYSIS

This section presents the result of experiments conducted for evaluation of pronoun translation quality in an MTO.

6.1 PRF Metric

Though, in the original source text, total occurrences of pronouns considered for evaluation is 79, but total candidate translations are only 73 while total reference translation are only 74. It means that 4 pronouns (three instances of ‘it’ and one instance of ‘her’, ‘these’ and ‘that’) did not have candidate translation, while 5 pronouns (one instance of ‘there’ and ‘that’ and three instances of ‘it’) did not have reference translation. Observation reveals the fact that among these 6 pronouns, 4 pronouns are those that have been used as pleonastic pronoun and hence both translation (candidate as well as reference) were missing. In other words, high frequency of pleonastic pronouns would cause reduced efficacy. The

overall precision, recall and F-score computed by PRF metric are respectively 0.5342, 0.5270 and 0.5306. Table 7 provides the precision, recall and F1-measure for all 30 pronouns considered for evaluation in given English source text.

6.2 APT Metric

Although, there is conceptual difference in working of APT and PRF metric, however with simplest configuration (i.e., with only source pronoun list), result of APT is almost like PRF metric. Reflection of this similarity can be seen in closeness of final F1-measure of PRF and comprehensive accuracy score of APT metrics, as well as translation accuracy of individual pronoun as shown in table 3 and table 6 (config_1 column). With simplest configuration, F1-measure of APT is 0.5316 while F1-measure of PRF is 0.5306. The

reason for this little gap in performance, i.e., lagging of PRF is due to those pronouns whose candidate or reference translation are missing, hence have not been considered by PRF metric.

Table 9 depicts the overall performance of APT metric under different configurations as mentioned in table 4, while table 8 provides performance of individual pronoun for different configurations. Analysis of results also revealed functional diversity and pleonastic property of pronouns play major role in pronoun translation. For example, from English to Hindi translation point view, for pronouns like 'he', 'she', 'you' etc., which show higher functional divergence, translation accuracy is low while for pronouns like 'here', there, etc., translation accuracy is fine.

Table 8: Performance analysis of APT metric with different configurations

Parameters	config_1	config_2	config_3	config_4
Finding-per case	42, 0, 29, 3, 2, 3	42, 16, 13, 3, 2,3	41, 18, 12, 3, 2, 3	43, 18, 10, 3, 2, 3
Accuracy measured	0.5316	0.6329	0.6456	0.6772

Table 9: Accuracy measure of individual pronoun in APT metric

Pronoun	config_1	config_2	config_3	config_4
He	6/16	9/16	10.5/16	10.5/16
Her	1/3	1.5/3	1.5/3	1.5/3
Here	1/1	1/1	1/1	1/1
Him	1/1	1/1	1/1	1/1
Himself	1/1	1/1	1/1	1/1
His	6/8	6.5/8	6.5/8	7/8
I	1/1	1/1	1/1	1/1
It	0/5	1/5	1/5	1/5
Itself	1/1	1/1	1/1	1/1
My	1/1	1/1	1/1	1/1
Our	0/1	0/1	0/1	0/1
She	0/2	1/2	1/2	1/2
That	6/10	6/10	6/10	6/10
Them	0/1	0.5/1	0.5/1	0.5/1
There	2/3	2.5/3	2.5/3	2.5/3
These	0/2	1/2	1/2	1/2
They	4/6	4.5/6	4.5/6	4.5/6
This	10/14	10.5/14	11/14	11.5/14
Us	0/1	0/1	0/1	0/1

Which	1/1	1/1	1/1	1/1
-------	-----	-----	-----	-----

ALGORITHM 1: Individual Pronoun Precision and Recall Computation Algorithm

```

compute_individual_pronoun_precision_recall (output_detail_file)
open output_detail_file in read mode
pronoun_list = []
read output_detail_file line by line
    for every line of output_detail_file
        src_pronoun, case_type, case_weight = extract (source_pronoun, case_type, case_weight)
        if src_pronoun not in pronoun_list
            add src_pronoun in pronoun_list
        depending upon case type and case weight
        update (no_of_matches, no_ref_trns, no_cnd_trns)
    end
compute (pronoun_list, precision, recall)
print_deatil (pronoun_list, precision, recall)

```

7. FUTURE RESEARCH DIRECTION

7.1 Need of accurate word alignment

Since word alignment accuracy is too crucial for all reference based or editing cost-based evaluation metrics, this is a challenge for PRF and APT metric also. In fact, the result of any reference-based evaluation metric used with tool-based word alignments are not reliable and hence may not be reflecting the correct measures. Moreover, because of this poor word alignment, sometime a good metric also might show poor performance for a good translation while better performance for a bad translation. It can be observed very easily that, compared to manually obtained word alignments, if any of these evaluation metrics are employed with word alignments obtained by any word alignment tool, there is high degradation in performance of these metrics. But on the other hand, it is very tedious and cumbersome to perform manual word alignment for a large corpus. A separate study needs to be conducted to measure the impact of word alignment quality on accuracy of translation quality evaluation metric.

7.2 Addressing pronoun translation issues

Some pronoun translation related issues have been discussed in section 2 and examples illustrated the impact of these issues on pronoun translation and pronoun translation quality measurement. Analysis of experiment result also endorsed this point. For example, in case of multiple translations, (like “तुम” and “आप”, which are exchangeable translation for pronoun “you” most of the time), evaluation metric would count it as different translation, i.e., as a mismatch translation and in turn would account for reduced accuracy. Similarly, divergence and pleonastic issues increases the ambiguity in pronoun translation and hence in translation quality measurement also.

Experimental result endorsed that resolving pronoun translation related issues would affect pronoun translation as well overall translation quality.

7.3 Improvement in APT metric by blending PRF features.

Study of underlying evaluation methodology revealed that PRF provides precision and recall of individual pronouns along with an aggregate score for all pronouns, while APT provides a single

comprehensive score for all pronouns. But addition of a small module in APT metric may enable APT metric also to provide precision and recall of individual pronoun.

Algorithm 1 shows the algorithm for addition of that module.

7.4 Suitability of APT for EHMT system

Though the performance of PRF metric and APT metric with simplest configuration is almost same, but other configuration of APT have shown a high jump in performance.

Results shown that use of target pronoun list, dictionary of equivalent pronoun and dictionary of similar pronouns and assigning appropriate case weight in different configuration improved the translation accuracy measurement.

underlying framework and evaluation methodology of two pronoun translation evaluation metric, PRF and APT. After making them conform for an EHMT systems and developing accordingly a test data set, experiments were conducted to see their suitability for EHMT systems.

Results show that with a few modifications both metrics can be used for EHMT systems also.

This paper was also aimed to find the answer of four RQs and fortunately all four RQs were addressed. RQ1 has been addressed in section 6. RQ2, RQ3 and RQ4 have been addressed in section 7. As discussed in section 7.4, APT metric is more suitable for EHMT system hence RQ2 addressed. As section 7.3 indicates that PRF feature can be embedded in APT through addition of a small module, hence RQ4 addressed.

8. CONCLUSION

This paper was aimed to study and analyze

References

- Eirini Chatzikoumi. 2019. How to Evaluate Machine Translation: A Review of Automated and Human Metrics, Natural Language Engineering. Sept 2019 Pages 1-25
- Bentivogli L., Cettolo M., Federico M. and Federmann C. (2018). Machine translation human evaluation: An investigation of evaluation based on post-editing and its relationship with Direct Assessment. In Proceedings of the International Workshop on Spoken Language Translation, Bruges, Belgium, pp. 62–69
- Rachel Bawden, Rico Sennrich, Alexandra Birch and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation
- Alon Lavie and Michael J. Denkowski. 2009. The Meteor: Metric for Automatic Evaluation of Machine Translation, In Proceedings of Machine Translations, Vol 23, pages 105-115
- Ankush Gupta, Sriram Venkatapathy and Rajeev Sangal. 2010. METEOR-Hindi: Automatic MT Evaluation Metric for Hindi as a Target Language, In Proceedings of 8th International Conference on Natural Language Processing (ICON-2010)
- Bonnie J. Dorr. (1994). Classification of Machine Translation Divergences and a Proposed Solution. Computational Linguistics, 20, 4 (1994), pages 597-633
- C. Tillmann, S. Vogel, H. Ney, H. Sawaf and A. Zubiaga. 1997. Accelerated DP Based Search for Statistical Translation. In Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece
- Chetan Agarwal and Kamlesh Dutta. 2021. मशीन आधारित भाषा अनुवाद में संदर्भ निराकरण का महत्व. Bhartiya Vaidyanik Evam Audhyogik Anusandhan Patrika, 29, 1 (June 2021), pages 49-54
- Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Paris, France, pages 283–289
- Christian Hardmeier, Joakim Nivre, and Jorg Tiede-mann. 2012. Document Wide Decoding for Phrase Based Statistical Machine Translation. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12), Jeju Island, Korea, pages 1179–1190

- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jorg Tiedemann, Yannick Versley and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In Proceedings of the 2nd Workshop on Discourse in Machine Translation (DiscoMT), Lisbon, Portugal, pages 1– 16
- Christian Hardmeier. 2014. Discourse in Statistical Machine Translation. PhD thesis, Uppsala University, Sweden.
- David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005), Ann Arbor, Michigan, pages 263– 270
- Deepa Gupta and N. Chatterjee. 2003. Identification of Divergence for English to Hindi EBMT. In Proceedings of 3rd MT Summit, pages 141-148
- E. Vidal 1997. Finite-State Speech-to-Speech Translation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, Germany
- Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics
- George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrences Statistics. In Human Language Technology: Notebook Proceedings, pages 128–132
- Gregor Leusch, Nicola Ueffing and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics
- Hiroki Shimanaka, Tomoyuki Kajiwar and Mamoru Komachi. 2018. Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2018, June 2-4), New Orleans, Louisiana, pages 106–111
- <https://www.thefreedictionary.com/List-of-pronouns.htm>
- Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003), Vol 2, (May 2003), Edmonton, Alberta, pages 61-63
- Jesus Peral and Antonio Ferrandez. 2003. Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation, Journal of Artificial Intelligence Research 18 (2003) pages 117-147
- Jochen Stefan Weiner. 2014. Pronominal Anaphora in Machine Translation, Master Thesis, Institute for Anthropometrics and Robotics Interactive Systems Lab (ISL), Karlsruhe Institute of Technology – University of the State of Baden-Wuerttemberg and National Laboratory of the Helmholtz Association
- Kamlesh Dutta, Nupur Prakash and Saroj Kaushik. 2009. Application of Pronominal Divergence and Anaphora Resolution in English-Hindi Machine Translation. Research journal "POLIBITS" Computer Science and Computer Engineering with Applications, Issue 39, pages 55-58
- Kishore Papineni, Salim Roukos, Todd Ward and WeiJing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Philadelphia, PA, pages 311–318
- Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). In Proceedings of the 3rd Workshop on Discourse in Machine Translation: Association for Computational Linguistics, pages 17–25.
- Liane Guillou 2012. Improving pronoun translation for statistical machine translation. In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon (France), pages 1–10
- Liane Guillou and Christian Hardmeier. 2016. Protest: A Test Suite for Evaluating Pronouns in Machine Translation. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Paris, France

- Liane Guillou, Christian Hardmeier, A. Smith, Jorg Tiedemann and B. J. Webber. 2014. ParCor 1.0: A Parallel Pronoun Co-Reference Corpus to Support Statistical MT. In Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2014), Reykjavík (Iceland), pages 3191–3198
- Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski and Sharid Loaiciga. 2018. A Pronoun Test Suite Evaluation of the English German MT Systems at WMT 2018. In Proceedings of the 13th Conference on Machine Translation: Shared Task Papers, pages 570–577
- Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jorg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In Proceedings of the First Conference on Machine Translation. Association for Computational Linguistics, Berlin, Germany, pages 525–542
- Liane Guillou. 2011. Improving Pronoun Translation for Statistical Machine Translation (SMT). Master's thesis, University of Edinburgh, School of Informatics. www.inf.ed.ac.uk/publications/thesis/online/IM110943.pdf.
- Liane Guillou. 2016. Incorporating Pronoun Function into Statistical Machine Translation. PhD thesis, University of Edinburgh, UK
- Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation Using Word Sense Disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-2006), Cambridge, MA, August, pages 223-231
- Michael Denkowski and Alon Lavie. 2014. METEOR Universal: Language Specific Translation Evaluation for Any Target Language, In Proceedings of the 9th Workshop on Statistical Machine Translation: The Association for Computer Linguistics (WMT@ACL 2014, June 26-27), Baltimore, Maryland, USA. 2014, ISBN 978-1-941643-17-4 pages 376-380
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2004), Boston, Massachusetts, USA, pages 273–280
- Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing), Springer-Verlag, LNCS 7817, Samos, Greece, pages 236–247
- Ngoc Quang Luong, Andrei Popescu-Belis, Annette Rios Gonzales, and Don Tuggener. 2017. Machine Translation of Spanish Personal and Possessive Pronouns Using Anaphora Probabilities, In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (ACL 2017- April 3-7), Valencia, Spain, Vol. 2, pages 631–636
- Peter E. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19, 2
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase Based Translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003), Edmonton, Canada, 1, pages 48–54
- Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova and Preslav Nakov. 2019. Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pages 2964–2975
- R. Le Nagard and Philip Koehn, 2010. Aiding Pronoun Translation with Co-Reference Resolution. In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics, MATR, Uppsala (Sweden), pages 252–261
- R. Mahesh K. Sinha and Anil Thakur. 2005. Divergence Patterns in Machine Translation between Hindi and English

- S. Niessen, F. J. Och, G. Leusch and H. Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005, pages 65-72
- Sharid Loaigiga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jorg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In Proceedings of the 3rd Workshop on Discourse in Machine Translation (DiscoMT). Copenhagen, Denmark
- Shweta Chauhan, Philemon Daniel, Archita Mishra, and Abhay Kumar. 2021. AdaBLEU: A Modified BLEU Score for Morphologically Rich Languages, IETE Journal of Research, Vol. 67(5) (Aug. 2021)
- Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007) Prague, Czech Republic, pages 33–40