MACHINE TRANSLATION EVALUATION STRATEGY

Table of Contents

ACKNOWLEDEGEMENT

Revision History

Date	Version	Pages affected	Reason	Author(s)/Changes By
25 April 2014	1.0	Draft MT evaluation strategy	Internal Review	Neha Aphale, R Doctor, Babita Shinde
30 April 2014	1.1		Internal Review	Neha Aphale
06 May 2014	1.2	2,8,10,14,15,21,24	Internal Review	Neha Aphale
30 January 2015	1.3	7,9,12,17,19,20,26,33	Revision after feedback received from Mr. Manoj Jain, Ms.Lele, Mr. PushpakB	Anand Kulkarni, Neha Aphale

CHAPTER 1

INTRODUCTION

1. Introduction

India is a multilingual country. There is a great demand for translation of documents from one language to another. This will ensure larger flow of information across different languages. Machine translation is the process of converting the text from one natural language into another natural language. To build machine translation systems different institutes & organizations have been working for several years to overcome the language barriers and have generated a large repository of linguistic tools & resources, pertinent to machine translation. Nowadays, MT systems are in great demand as it can help to reduce the language barrier and enable easier communication.

Currently various machine translation projects are running in the consortium mode and TDIL - DeitY is playing an instrumental role in funding these projects. These projects are either from English to various Indian languages or within Indian languages. Though huge efforts are being taken, Machine translation is still an open problem.

As machine translation is becoming more popular, methodology for estimating the quality of MT system is becoming very important, so not only development of MT but evaluation of MT also involves research component.

1.1 Approaches Used For Machine Translation

There exist a number of approaches used for development of machine translation system. Following are the few of approaches used in Indian languages MT systems:

- Rule based MT
- Example based MT
- Statistical MT
- Tree Adjoining Grammar based MT
- Analyze-Transfer-Generate based MT
- Hybrid MT

And the choice of approach depends upon the available resources and the kind of languages involved. However, quality of MT system is important irrespective of approaches being used for MT development.

1.2 Purpose

Purpose of this document is to evolve the strategy for evaluating output of different MT engines and provide a methodology at the national level for MT evaluation, which will focus on linguistic analysis as well as provide quantitative

measure to provide end to end system performance irrespective of approaches being used for MT development; which can be referred by testing agency for determining the quality of a MT system.

This is not necessary that all the mentioned evaluation parameters should be deemed fit to the ~~MT~~ application under test, the evaluation parameters will be decided by the development agency and only that will be checked by the testing agency.

1.3 Scope

This document is intended to serve as an evaluation guide for MT system. It can also be used by the C-DAC, GIST members, to educate their users about the evaluation of the MT system.

The document is segregated in four parts as follows:

Part I: Sets things in perspective,

Part II: of the paper presents aims of the machine translation evaluation.

Part III: deals with the various methods for MT evaluation, all the more so, since the evaluation methodology is common for all the machine translation systems.

Part IV: gives the details of evaluation procedure, which includes importance of test data, evaluator's selection criteria.

1.4 Abbreviation, Acronyms & Definitions

TDIL-DeitY	Technology Development for Indian Languages- Department of Electronics & Information Technology
MT	Machine Translation
ARPA	Advanced Research Projects Agency
ALPAC	Automatic Language Processing Advisory Committee
BLEU	Bi Lingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit Ordering
NIST	National Institute of Standards and Technology
SVO	Subject Verb Object
SOV	Subject Object Verb
NER	Named Entity Recognition
POS	Part of Speech

Terminology	Definitions
Fluency	Fluency refers to the degree to which the target is well formed according to the rules of the target language. The objective of fluency evaluation is to determine how much like "good fluent" a translation appears to be without taking into account the correctness of information.
Adequacy	Adequacy refers to the degree to which information present in the source text is communicated in target translation. The objective of the adequacy is to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the language in the candidate translation.
Comprehensibility	Comprehensibility is a measure of how easy is a text to understand.
Intelligibility	Intelligibility is a measure of how "understandable" the sentence is. Intelligibility is measured without reference to the original source sentence.
Fidelity	Fidelity is measurement of correctness of the information transferred from source language to the target language. It is a subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation.

CHAPTER 2

AIMS OF MACHINE TRANSLATION EVALUATION

2 Aims of Machine Translation Evaluation

In machine translation development, evaluation is a very important activity but difficult task. The difficulty arises due to some inherent characteristics of the language pairs, like simple word-level discrepancies to more difficult structural variations for English to Indian languages, such as reduplication of words (चलते- चलते), free word order and many more.

Several researchers have worked on evaluation techniques of Machine Translation systems and many measures and methods have been developed for this purpose. Many attempts have been made to produce best suited evaluation schemes. It is found that evaluation is carried out for a purpose which varies from case to case.

A. Andereewsky & G. Van Slype considers that evaluation of a translation system can vary according to the standpoint from which it is viewed and depends on the types of persons concerned and their motivation. From an analysis of aims of evaluation, the evaluation criteria can be deduced. Hence considering all the stakeholders of machine translation system, following are the broad aims of evaluation of MT system.

Sr. No.	stakeholders Involved	Aims of Evaluation
1.	Funding agency	1. To evolve the strategy for evaluating output of different MT engines and provide a methodology at the national level for MT evaluation, which will focus on linguistic analysis as well as provide quantitative measure to provide end to end system performance. 2. To assess the potential Market based on performance figures 3. System optimization 4. Qualitative machine translation output
2.	MT developers	1. Qualitative machine translation output 2. Potential Market 3. System optimization 4. To evolve the strategy for evaluating output of different MT engines and provide a methodology at the national level for MT evaluation, which will focus on linguistic analysis as well as provide quantitative measure to provide end to end system performance.

3.	Testing & evaluation team	1. To evolve the strategy for evaluating output of different MT engines and provide a methodology at the national level for MT evaluation, which will focus on linguistic analysis as well as provide quantitative measure to provide end to end system performance. 2. Qualitative machine translation output 3. To detect the problems in translation. 4. Translation quality is not as absolute concept hence to assess the MT development. 5. To provide constructive feedback to the developers, this will help them to work towards quality improvement. 6. To check how close and how far MT output is from end user's expectations.
4.	End users of MT system (Language Translators + Linguists)	1. End user satisfaction 2. Effective translation of source language to target language 3. Services in the form of user friendly applications

CHAPTER 3

EVALUATION METHODOLOGIES

3 Evaluation Methodologies

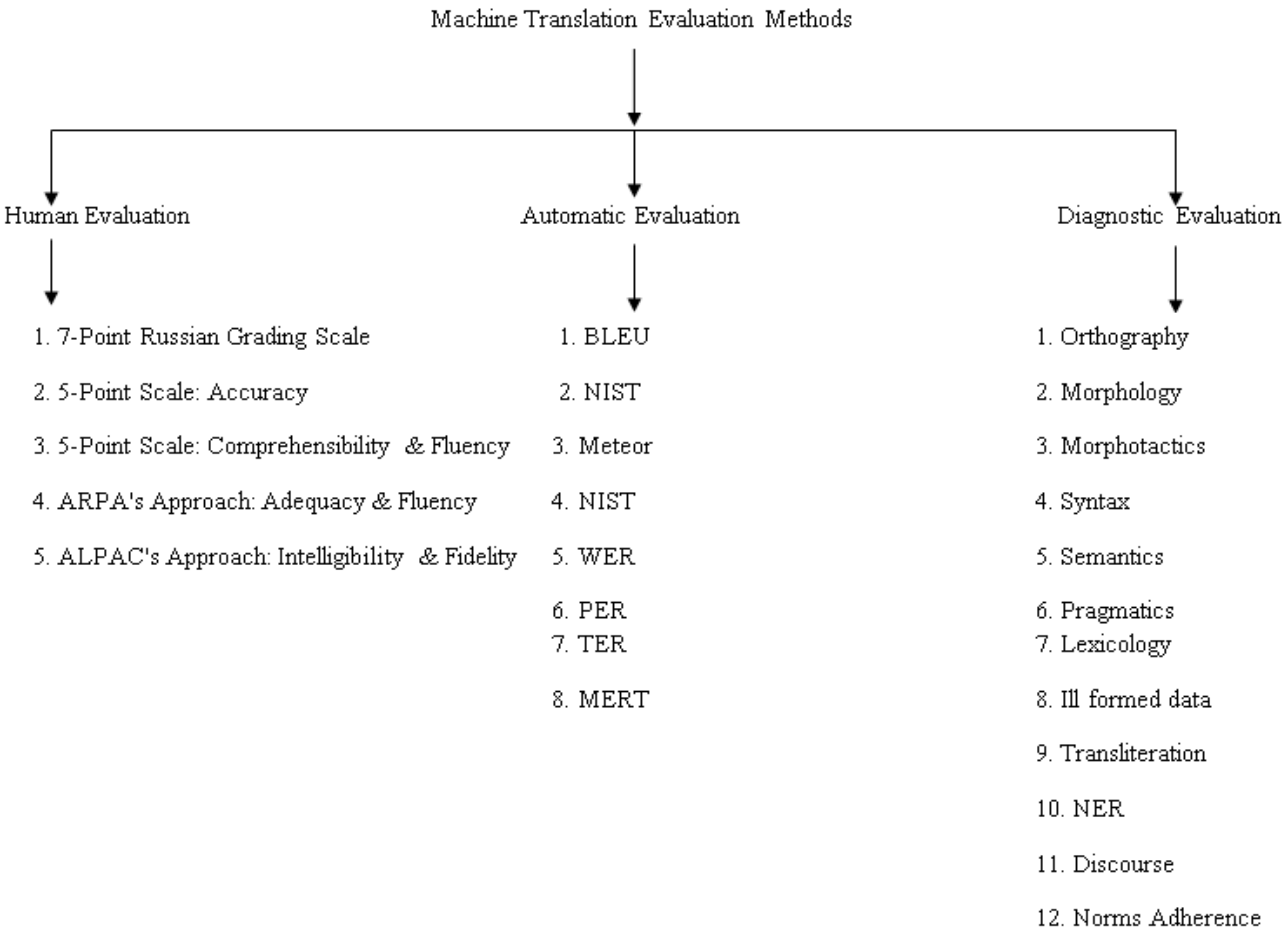
Once the objectives are set in place, evaluations in conformity with the objectives need to be defined. Evaluation is necessarily a two pronged process.

1. Creating a strategy: - A strategy for evaluation needs to be defined in terms of what is being targeted.

2. Implementation of the strategy: - Once the strategy is formulated and finalized, a method for implementing the strategy needs to be developed.

Translation quality is not an absolute concept, it needs to be assessed. Considering this, evaluation methodologies are divided in following three major groups

1. Human Evaluation methodologies
2. Automatic Evaluation methodologies
3. Diagnostic Evaluation methodologies



3.1 Previously Used Human Evaluation Methods

In case of Indian languages there is no single correct translation, but multiple good translation options can exist. So instead of comparing the translation with single reference, subjective evaluation is more useful. Human evaluation allows measuring the quality of MT system over a set of end users. Translations are produced for end users hence end users are the right measure of quality of translation and end users can recognize and weigh errors in translation correctly. Because of these strong arguments, subjective human evaluation is important.

In this section, we give an overview on the most well known evaluation methods which focuses on the users and their needs. The main focus is on different types of subjective evaluation methods. Following are the grading systems used for evaluation of MT system. Initially, 7-Point Russian Grading System was proposed then 5-point scale for Accuracy was proposed and used. Afterwards, 5-Point Scale: Comprehensibility & Fluency was used. Details are provided below.

3.1.1 7-Point Russian Grading System

This grading system is in conformity to standards laid down by Russian Experts and which closely parallels BLEU and Meteor.

Usability and transmission of information are the prime a criterion on which grading systems is based. The 7-point Russian grading system deals with translation as a process of visibility of text.

Sr. No.	Grade	Description
1.	**Opacity**	The rendering is absolutely useless for any purpose. Such a rendering shall be deemed as of **UNACCEPTABLE** quality.
2.	**Semi-Opaque**	Some parts are comprehensible, but on the whole the picture still remains difficult to get and the text evades the target user. Such a rendering shall be deemed as **POOR** quality.
3.	**Part Visibility**	The user can get a grasp of the over-all intention of the text, but on the whole, the user has to work hard to get at the meaning of the text and large fragments are practically opaque and incomprehensible. Such a rendering shall be deemed as **LOW** quality.
4.	**Half Visibility**	The rendered text is quite comprehensible to the target reader and can be used by him/her as can be used as a rough draft for improvement. Such a rendering shall be deemed as **DRAFT** quality.

5.	**Near Visibility**	Text is clear enough and all pertinent information can be drawn from it. However, the text is hard to read due to language errors and require further filtration. Such a rendering shall be deemed as of **ACCEPTABLE** quality.
6.	**Near-total visibility**	The rendering has stylistic errors and also some difficult grammatical, syntactic, lexical issues are not clarified. However, it transmits the information needed to the target user. Such a rendering shall be deemed as of **SATISFACTORY** quality.
7.	**Total visibility**	The rendering passes muster though not stylistically perfect. Such a rendering shall be deemed as of **HIGH** quality.

Problems:

- More the number of scales, the more will be errors in human judgment
- More training is required for human evaluators.
- As it is based on linguistic testing, these parameters cannot be applied on basic MT systems.

3.1.2 5-Point Scale: Accuracy (Proposed Scale for Indian Languages)

Above mentioned 7-point grading scale is based on linguistic testing which may not be for basic MT systems. Hence, under the TDIL-DeitY funded MT projects, following 5-point scale was evolved. A new approach is formulated with stress on Sprachgefühl i.e. focus on usability and the native speaker's expectations and the translation quality is provided in terms of comprehensibility of output.

Grade 0	No output provided by the engine concerned.
Grade 1	The translated output is not comprehensible.
Grade 2	Comprehensible after accessing the source text.
Grade 3	Comprehensible with difficulty.
Grade 4	Acceptable since the text is comprehensible.

Using this 5-point scale, two types of testing need to undertake, Blind testing and Open testing. In case of Blind Testing, evaluator does not have access to Source sentence. And in Open Testing s/he is provided machine translated output along with Source sentence. The graded output through both blind and open testing, provided by each evaluator for a given language was processed using following formulae.

$$\text{Final Evaluation Mean For Each Sentence (Within 0 - 4)} = \frac{\sum_{T=1}^{T=3} \text{Grade Given By Tester T}}{3}$$

$$\text{Total Average} = \sum_{S=1}^{S=100} \text{Final Evaluation Mean Of Sentence S}$$

$$\text{Accuracy Of Engine (In \%)} = \frac{\text{Total Average}}{\text{Number Of Sentences}} * 25$$

Problems:

- "Grade 2- comprehensible after accessing the source text" is useful only in open testing.
- Grade-2 has no significance in blind testing; this makes a difference in result of MT systems.

3.1.3 5-Point Scale: Comprehensibility & Fluency

5-Point scale for accuracy mentioned above has Grade 0 for "no output provided by system". This may affect largely on the performance of the system. So Dr. Sangal suggested a modification of the above 5-point scale. The 5-point scale was thus

modified in consultation with Dr. Sangal, Dr. RMK Sinha, Dr. Darbari, Dr. Ramanan, Mr. V.N.Shukla & DeitY officials; in which NO OUTPUT by system and buffer clearance issues are graded as -1.

This Modified 5-point scale is provided below, where performance of the MT system is given on two parameters (1) Comprehensibility & (2) Fluency

This led to the following rating system when the systems were not stable.

Grade -1	No Output OR buffer clearance issue
Grade 0	Nonsense (If the sentence doesn't make any sense at all – it is like someone speaking to you in a language you don't know)
Grade 1	Some parts make sense but is not comprehensible over all (e.g., listening to a language which has lots of borrowed words from your language – you understand those words but nothing more)
Grade 2	Comprehensible but has quite a few errors (e.g., someone who can speak your language but would make lots of errors. However, you can make sense out of what is being said)
Grade 3	Comprehensible, occasional errors (e.g., someone speaking Hindi getting all its genders wrong)
Grade 4	Perfect (e.g., someone who knows the language)

The results are thus calculated on two parameters, (a) comprehensibility and (b) fluency. Both (a) and (b) are to be calculated by considering the average of the score given by all the evaluators for every sentence.

(a) Comprehensibility is calculated by taking out the percentage of the number of sentences getting an average score between 2 to 4 out of the total number of sentences in the set.

Specifically, let Si be the number of sentences with a grade of i (i=0, 1, 2, 3, 4) Where N is total number of sentences excluding sentences having grade as "-1".

Comprehensibility C = sum_over{i=2, 3,4}(Si)/ N.

(b) For fluency, the average scores is measured against [4 * total number of sentences in the set]. Specifically, let Si be the number of sentences with a grade of i (i=0, 1, 2, 3, 4). Then

Fluency F = sum_over_all_i (i *Si)/N.

Problems:

- From the feedback of evaluators, Grade 2 & Grade 3 is confusing; as well as Grade 0 & Grade 1 are also confusing.
- By using this evaluation technique there will be some cases
  - Where comprehensibility can be 100% but fluency can be considerably low (case where all sentences have grade between 2 to 4)
  - Where fluency is higher and comprehensibility is 0%, which is incorrect (this is case when all the sentences have grade 0 & 1)

3.2 Challenges Faced In Human Evaluation

Though human evaluation is very useful, informative; it has several inherent challenges & limitations. These are listed below.

1. Setting up a quality panel is very difficult since, it require having a number of evaluators together in one place for several days.
2. To keep the evaluator's pool ready as and when required is very difficult. It is not possible to have all the evaluators at one place at one time. Availability of the evaluators is a major concerned.
3. In case of evaluation of Indian language machine translation (ILMT) it requires evaluators with the knowledge of two different languages. Evaluators knowing English and one Indian language or Hindi and any other Indian language are easy to find. But to find the evaluators having knowledge of languages like Tamil - Malayalam and Tamil - Telugu is a challenging job.
4. Training to evaluators: - Once evaluators has been identified, trainings/ & guidelines to them are also important. To clarify each scale in detail to the evaluators is difficult, if definition wise there is a not much difference in scales. It may include the risk of wrong grading by evaluators.
5. Expensive & slow: - Human evaluation is a time consuming activity and also labor-intensive, as evaluating each sentence evaluator needs to consider several quality criteria.
6. From the feedback of evaluators, quality of MT system improves and therefore human assessments are not reusable as system improves with timely feedback.
7. Inter-annotator agreement: - Generally machine translation output is given to multiple evaluators for evaluation. No two evaluators will give the same score to the translation, though the same evaluation guidelines are given. Because

along with evaluation guidelines evaluators always consider their perception about language, choice of words, which may be different. So, in human evaluation inter-annotator agreement issue will always be present because human evaluation is subjective.

3.3 Proposed Evaluation Methods

To overcome the issues faced in existing evaluation techniques, following evaluation methods are proposed.

- ARPA's Approach: Fluency & Adequacy
- ALPAC's Approach: Intelligibility & Fidelity
- 4-Point Scale By Google Translation

Along with above human evaluation techniques, diagnostic evaluation method is proposed which covers

- Evaluation of major Linguistic areas
- Error analysis

3.3.1 Human Evaluation Methods

3.3.1.1 ARPA's Approach: Fluency & Adequacy

As a part of Human Language Technologies program, ARPA has created an evaluation methodology for MT. This evaluation program was instigated in 1991, which involves different methods. The methods were- comprehension evaluation, quality panel evaluation & evaluation based on adequacy & fluency.

Out of these 3 methods, the most popular and widely used method is to ask the evaluators to provide ratings for translation, for fluency and adequacy.

Fluency:-

Definitions:

1. Fluency is a rating of how good the language of translation is. (http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Advanced_Research_Projects_Agency_.28ARPA.29)

2. Fluency refers to the degree to which the target is well formed according to the rules of the target language. Fluency is intended to capture translation intelligibility. (http://nlp.lsi.upc.edu/papers/gimenez08-thesis.pdf)

3. The objective of fluency evaluation is to determine how much like "good fluent" a translation appears to be without taking into account the correctness of information. (http://ufal.mff.cuni.cz/pbml/100/art-chatzitheodorou-chatzistamatis.pdf)

Grade	Description
Grade 1	Incomprehensible
Grade 2	Disfluent language
Grade 3	Non-native language
Grade 4	Good language
Grade 5	Flawless language

For fluency assessment, only translations being evaluated should be given to the evaluators and not the source sentences or its reference translations. Then the evaluator is asked to provide the rating from above 5-point scale. Fluency ratings can be given by the evaluators knowing only target language.

Adequacy:-

Definitions:

1. Adequacy is a rating of how much information is transferred between the original and translation. (http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Advanced_Research_Projects_Agency_.28ARPA.29)

2. Adequacy refers to the degree to which information present in the source text is communicated in target translation. Adequacy is intended to capture translation fidelity. (http://nlp.lsi.upc.edu/papers/gimenez08-thesis.pdf)

3. The objective of the adequacy is to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the language in the candidate translation. (http://ufal.mff.cuni.cz/pbml/100/art-chatzitheodorou-chatzistamatis.pdf)

Grade	Description
Grade 1	None
Grade 2	Little Meaning
Grade 3	Much Meaning

Grade 4	Most Meaning
Grade 5	All Meaning

For adequacy assessment, along with the translations, evaluators should have access to original source sentences and optionally reference translation. Then the evaluator is asked to provide the rating from above 5-point scale based on the meaning in translations as compared with source text. For this evaluator should have knowledge of both the source as well as target languages.

Advantages:

- Easy to deploy
- It does not require expert judgment.
- Separate scales for adequacy and fluency were developed assuming that a translation might be disfluent but contain all the information from the source.

Along with the informativeness, evaluation based on adequacy and fluency is these days the standard methodology for ARPA evaluation program.

Problems:

- Practically it is very difficult to separate these two aspects of translation i.e. fluency & adequacy
- It is difficult to provide the guidelines in terms of how to quantify the meaning and how many and what type of grammatical errors separates the different levels of fluency

3.3.1.2 ALPAC's Approach: Intelligibility & Fidelity

One of the constituent parts of ALPAC report was a study comparing different levels of human translation with machine translation output, using human subjects as judges. For this two variables were considered

1. Intelligibility

2. Fidelity

Intelligibility:

Definitions:-

1. Intelligibility is a measure of how "understandable" the sentence is. It is measured without reference to the original source sentence. Only translated sentence is given to evaluator and asked to rate. Initially it was measured on a scale of 1 - 9. But more the number of scales, the more will be errors in human judgment. Hence modified 4-point scale is used most widely.
(http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Automatic_Language_Processing_Advisory_Committee_.28ALPAC.29)

2. Intelligibility tells the degree of comprehensibility and clarity of the translation. It is affected by grammatical errors, mistranslation, untranslated words. *(G. Van slype*)

Following 4-point scale has been defined by *G. Van Slype* for Intelligibility, in which only translated text is given to evaluator.

Grade	Scale	Description
Grade 0	**Unintelligible**	Nothing or almost nothing of the message is comprehensible. The meaning of the sentence is not understandable.
Grade 1	**Basely Intelligible**	A part only of the content is understandable representing less than 50% of the message. The general idea is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choice
Grade 2	**Fairly Intelligible**	The major part of message passes, despite some inaccuracies; one can understand the information to be conveyed.
Grade 3	**Very Intelligible**	All the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language.

Fidelity:-

Definitions:

1. Fidelity is a measure of how much information is retained in the translated sentence compared to original source sentence.
(http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Automatic_Language_Processing_Advisory_Committee_.28ALPAC.29)

2. *T. C. Halliday* defines fidelity as, *measurement of correctness of the information transferred from source language to the target language.*

3. G. Van Slype defines fidelity as, *subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation*.

Following 4-point scale has been defined by *G. Van Slype* for fidelity in which original source sentences are also given to evaluator with its corresponding target translation sentences.

Grade	Scale	Description
Grade 0	**Completely unfaithful**	Doesn't make any sense
Grade 1	**Barely faithful**	less than 50% of the original source information passes in the translation
Grade 2	**Fairly faithful**	more than 50% of the original source information passes in the translation
Grade 3	**Completely faithful**	Completely faithful or almost completely faithful

Generally the fidelity rating should be equal to or lower than Intelligibility.

3.3.1.3 4-Point Scale By Google Translation

Following is the 4-point scale used by Google for evaluation of English→French.

Grade	Scale	Description
Grade 0	**Poor**	None of the content is translated well
Grade 1	**Fair**	Only some of the content is translated well
Grade 2	**Good**	Most of the content is translated well
Grade 3	**Excellent**	All the content is translated well

- **Evaluator's Selection Criteria**

1. For each language pair evaluation, 3 to 5 evaluators should be used.
2. Evaluator should have knowledge of both source language as well as target language.

3. No evaluator should be a linguist.

3.3.2 Diagnostic Evaluation

Human evaluation is a major part of machine translation evaluation process, where MT output is given to evaluators and asked them to evaluate it on the provided criteria (subjective).

However, only giving rating to the MT output will not give any constructive feedback to MT developers. ~~So~~ To provide qualitative feedback/assessments diagnostic evaluation of MT output is important, wherein various linguistic features can be checked. This diagnostic evaluation can help MT developers to overcome the shortcomings of the system by focusing on particular module or component. In diagnostic evaluation focus is on identifying the weaknesses or errors of the system.

3.3.2.1 Major Areas for Linguistic Evaluation

Along with the subjective (human) evaluation, in-depth evaluation of MT system's ability to handle various linguistic processes is also important. The aim of the linguistic / diagnostic evaluation is to test whether the translation system can handle morphology, syntax & lexis when translating from source language to target language. Following are the major areas targeted for linguistic evaluation.

1. Orthography
2. Morphology & Morphotactics
3. Syntax
4. Semantics
5. Pragmatics
6. Lexicology
7. Norms Adherence
8. Ill formed data
9. Transliteration
10. Named Entity Recognition
11. Discourse

A. Orthography

Major areas to be focused are

1. Misspellings: Spelling correction- strenght for strength

   dias for dais

   जांच for जाँच

2. Spelling variants: organise vs. organize

   हिंदी vs. हिन्दी

   गरदन vs. गर्दन

3. Acronyms: भाजपा, IT : Information Technology or Income Tax

4. Misused Terms: Loose for lose

B. Ill-Formed Data

This comprises data which has

1. Mixed Characters: अप्रैल ka mahina सूरज का ताप तेज है।
2. Control Characters
3. Ill formed ISCII syllable: वर्षा एेक निवेदन
4. Use of hidden characters: ZWJ & ZWNJ

C. Morphology & Morphotactics

1. Languages such as Urdu, Hindi and Punjabi show less Morphotactics and are less agglutinative in their writing system. These are Type1 languages. Movement within these languages is relatively easy although Urdu because of its writing system and absence of short vowels does lead to a large degree of homonymic collision demanding disambiguation. Tamil to Telugu and vice-versa will be relatively easier than moving from Telugu to Hindi.

2. Marathi, Bangla and for that matter Gujarati, Assamese, Odia display a certain degree of agglutination in their writing systems. These are Type2 languages. Movement between Type1 and Type 2 languages demands a certain degree of parsing and lemmatization. Given the cognate nature of the languages, a certain amount of positive transfer is possible. However transfer from Hindi to Marathi and vice-versa will imply a large degree of syntactic and morphological movements.

3. Dravidian languages: Tamil, Malayalam & Telugu display a marked degree of agglutination with an extremely complex Morphotactics. These come under Type3. Moving from Type 1 to Type 3 demands the greatest amount

of analysis. Given the highly agglutinating nature of these languages and the low Morphotactics in Hindi which is analytical in its writing system, a large amount of parsing and analysis will be called for.

4. Suffix ordering

   EN→MR From this very table: tebl+aa+var+caa+c in Marathi

   EN→GJ From the boy's side (with stress): chokr+aa+vaaLaa+o+maaN+thi+ya

5. Geminated sandhi common to Dravidian

   HN->MAL maram+il -> marattil

6. Case: Handling case in grammars, where case is a surface phenomenon e.g. Hindi, Malayalam and languages where case is at a deep structure level. e.g. English

7. Morphological Variant: e.g. Urdu joining the helper verb or keeping the helper verb distinct. *kiyaagayaa* written together instead of separately.

8. Homonymic collision and Disambiguation as exemplified Urdu: The same word in Urdu can be read either in more than one way.

This OR That:اس

And OR Direction (n)اور

D. Syntax

Some sample instances are provided below.

1. Structure of N.P. and V.P. as well as N.P + V.P. in source and target. Concord and agreement between various words: e.g. the little girl came → choTI laRKI aayI
2. Handling of word order: SVO → SOV for English to Indian Languages
3. Use of correct negation: *mat and nahin in Urdu.*
4. Anaphora and cataphora interpretation from source to target language.

   e.g. It is well known that Rajasthan has forts.

   The dog came. Its tail was wagging
5. Translation of complex and compound sentences

e.g. The ship which is here goes to Goa.

The boy's aunt who is here is my neighbor.

6. Translation of verbal mood and aspect, verb phrase, Tense-mood-aspect-PNG-Register complex, auxiliary, modal/helper verb, Explicator/vector verbs.

7. Voice

    - Active vs. passive
    - completive vs. Incompletive

      e.g. 1. The door is shut; The door is shut by him

      2. One should not speak in this manner.

8. Conditional - If structure

   If I were……If I win…If he had gone….

9. Negation: e.g. He has NOT come….He does NOT know.

10. Standalone sentences as headers: Go go Goa…Scenic beauty of God's own country.

11. parenthesized sentences: The food (veg, meat) is remarkable

12. Noun Paradigm

    - Noun phrase along with determiners, pre & post determiners
    - Participle qualifiers & Gerundive
    - Adjective phrase
    - Adverbial phrase

E. Semantic & Lexis

1. Ambiguity

    - Lexical or grammatical ambiguity: e.g. मैंने दौड़ते हुए घोड़े को देखा ।
    - Semantic Ambiguity: e. g. American head seeks arms. Here "head" can be interpreted as a noun meaning either chief or the anatomical head of a body. Also "arms" can be interpreted as a plural noun meaning either weapons or body parts.
    - Structural Ambiguity: e.g. Stolen painting found by tree. It can be interpreted as either "A tree found a stolen painting (which is wrong)" or "A person found a stolen painting near a tree."

2. Idioms: e.g. पर्वतीय प्रदेश की यात्रा के कारण अंग-अंग टूटता है।

3. Proverb: e.g. बंदर क्या जाने अदरक का स्वाद।

4. Polysemy:

   - Lexical

     i. e.g. उसने शतरंज की चाल चली।

     ii. इस साईकिल की चाल तेज है।

     iii. वह उस व्यक्ति की चाल मे आ गया।

   - Grammatical

     i. संज्ञा- यहाँ एक आता है तो एक जाता है।

     ii. सर्वनाम- सभी लड़के आ गये पर एक नहीं आया।

     iii. विशेषण- मेरी एक बात तो सुन लो।

F. Pragmatics

Pragmatics is used in a loose sense, here as essentially dealing with problem or rhetoric the major areas to be tested are listed below.

1. Doublets: Doublets are created in Hindi for stylistic effects. A simple verb is juxtaposed to a victories form as in samples given below.

   - सराहना- सराहना करना
   - लड़ना- लड़ाई करना
   - त्यागना- त्याग देना

2. Styles and figures of speech: "Alankaara" of different kinds dealing with personification, simile, metonymy, metaphor especially in Urdu.

3. Affective speech acts: Exclamatory rhetoric: Translation of exclamative discourse

   - Vocative: हे भगवान! क्या सुहाना सफर है !
   - Surprise: अरे क्या!

- Plaudits: खूब! वाह!
- Joy: वाह-वाह!

4. Topic & Commentary: Theme & Rhyme
    - Object fronting: e.g. गिलास तोडा सिता ने, डाँट पड़ी राम पर।
    - Displacement of object to SVO position normally with Det + Noun: e.g. मैं समझ गया उसकी चाल।
    - Placement of verb in first position: effect of irony: e. g. कर चुकी आप मेरी मदद!

G. Discourse

1. Direct Speech: .e.g.उसने एक ही शब्द में कहा, वाह!क्या बात है!
2. Reported Speech: e.g. उन्होंने कहा है कि अमेरिका में कुछ भी मुमकिन है।
3. Free Reported Speech:
    - Absence of markers of reported speech
    - Absence of punctuation markers of direct speech
    - Absence of presence of the क marker

    e.g. मुलायम सिंह ने कहा कि कलाम एक जाने माने और सम्मानित वैज्ञानिक है।

H. Lexicology

1. Borrowing: A word taken directly from another language. e.g. "strawberry" in target language.
2. Calque: Source language foreign word/phrase is translated and incorporated into target language free verse.
3. Transliteration & NER: Named entity recognition plays an important role in language analysis. The name entity hierarchy is divided into three major classes: name, time & numerical expression. The task of NER is to identify the class in which an entity falls. e.g. ताज महल, चार मिनार
4. Translation of "function words" such as "for" e.g. Rajasthan is famous for its castles
5. Translation of phrasal verbs: go in for, go for, go into
6. Divalence of POS: yellow (Adj, noun, verb)

I. Norms

1. Spelling Norms: Compliance with spelling norms of the respective target languages.

   E.g. Urdu: Imlaanaamaa

   Bangla: Bangla Akademi

2. Storage Norms: Compliance with Unicode

Below is the sample template for error analysis, which evaluator is suppose to fill after evaluation.

Template For Linguistic Evaluation									
Name of Evaluator: Domain: Source Language: Target Language:									
Type of Sentences	Number of sentences	NE errors	Transliteration errors	Idioms & Phrases	Wrong word choice (ambiguity)	Misspelling			
Simple Sentences									
Complex Sentences									
Compound Sentences									

3.3.2.2 Error Analysis

Along with subjective & automated evaluation, in-depth analysis of MT system needs to be carried out like which are the most prominent source of errors, etc. We have to analyze errors by investigating error patterns where translations are incorrect, so this error analysis is important for diagnostic evaluation. So we have to identify all types of errors in the translated text along with their frequencies. For this types of errors need to be classified and defined clearly, so the root cause of the problem can be identified.

e.g. if majority errors are of wrongly translated words, then it signifies that bilingual dictionary need more focus and wrong word choice is a problem of word sense disambiguation.

Such diagnosis / error analysis will help to improve quality of MT system. Different types of errors can be as below for error analysis of MT system

1. Missing words
2. Untranslated words
3. Wrong choice of words (ambiguities)
4. Wrong translation
5. Addition of words
6. Deletion of words
7. Punctuation
8. Acronyms & abbreviations

Below is the sample template for error analysis, which evaluator is suppose to fill after evaluation.

Template For Error Analysis									
Name of Evaluator: Domain: Source Language: Target Language:									
Type of Sentences	Number of sentences	Un translated Words	Missing Words	Addition of Words	Deletion of Words	Wrong choice of words (ambiguities)	Wrong Translation	Punctuation	Acronyms & Abbreviations
Simple Sentences									
Complex Sentences									
Compound Sentences									

- **Evaluators' Selection Criteria**

1. For each language pair evaluation, 2 evaluators should be used.
2. Evaluator should have knowledge of both source language as well as target language.
3. Evaluator should be a linguist.

3.4 Limitations of Automatic Evaluation

Human evaluation of MT system is expensive and time consuming activity. It can take months to complete the evaluation process, which involves human labor, which cannot be reusable. Contrary to it automatic evaluation is useful which is quick, inexpensive and mostly language independent. It is useful wherein frequent evaluations are required.

Human judgment is the benchmark for assessing automatic metrics as humans are the end-users of any translation output. And it correlates highly with human evaluation.

Following are the most widely used automatic evaluation metrics

- BLEU
- METEOR
- NIST

However, it is well known that these automatic MT evaluation metrics have limitations and due to these limitations these metrics are not applicable for Indian languages machine translation evaluation.

1. Automatic measures e.g. BLEU, METEOR, NIST are not diagnostic as these are based on measures of string similarity. These metrics do not provide feedback on the ability of MT systems to translate various aspects of the language.

2. Word Order: These most popular metrics e.g. BLEU, NIST, PER, TER do not work well when we have to evaluate translation among distant language pairs like English to Indian languages. As English has different word order than Indian languages. English has word order SVO and Indian languages are SOV and all the Indian languages need special attention to word order in translation, otherwise meaning of the sentence changes completely, which often lead to misunderstanding and incomprehensibility. To translate the sentence just word by word is wrong, it should be grammatically correct.

3. Multiple correct translations: In case of BLEU, only exact match is considered and synonyms are not considered. All the Indian languages are morphologically rich; there can be multiple correct translations for single input sentence just word to word matching may lead to wrong results and make evaluation process harder. Multiple correct translations may differ in word choice or word order choice also.

   METEOR considers it, but it uses syntactic parsers, synonym databases, stemming. However very often these resources are available only for few languages or include training and optimizing models.

4. BLEU is N-gram precision based metric and does not care about the untranslated words in candidate translation.

5. BLEU has poor co-relation with the human judgment.

6. No single automatic metrics can perform well for all the Indian language pairs. To work METEOR exceptionally well for all the language pairs, it needs huge corpus and a large repository of linguistic tools & resources.

7. WER is one of the first automatic metrics used to evaluate MT systems, which is the standard evaluation metric for Automatic Speech Recognition. There are multiple correct translations for any given sentence. These correct translations may differ not only in word choice but in word order also.WER does not allow reordering of words where the word order of translated sentence can be different from the word order of source sentence, even though it is correct translation. A word that is translated correctly but in wrong location is penalized as a deletion (in the output location) and an insertion (in the correct location).

   Automated MT evaluation metrics are mainly the tools for consistent MT system development and may not have that much to do with real quality of translation.

   Culy & Riehemann (2003) state this:"Final important point is a reminder that the n-gram metrics are really document similarity measures rather than the translation quality measures".

CHAPTER 4

TEST DATA SELECTION CRITERIA

4 Test Data Selection Criteria

MT system translates the text from one natural language to another. These MT systems can be web-based, desktop-based or mobile-based with different development approaches. But irrespective of the nature of the MT system, evaluation of MT is important and test data is the first need of evaluation. Hence test data creation is very important, though it is very tedious, time consuming and expensive task; which requires specialized support from linguist.

To evaluate the performance of MT system first we have to identify the domain for which MT system is being developed, along with the source and target language of system.

4.1 Domain Specific MT system

In domain specific MT systems, all the resources like dictionaries, rules are restricted to the targeted domain. Frequency of occurrences of words from that particular domain is always high and domain specific system will always perform better for targeted domain. Test data to be used for evaluation should be domain specific to check the performance of MT system, for out of domain test data, system may not work as expected.

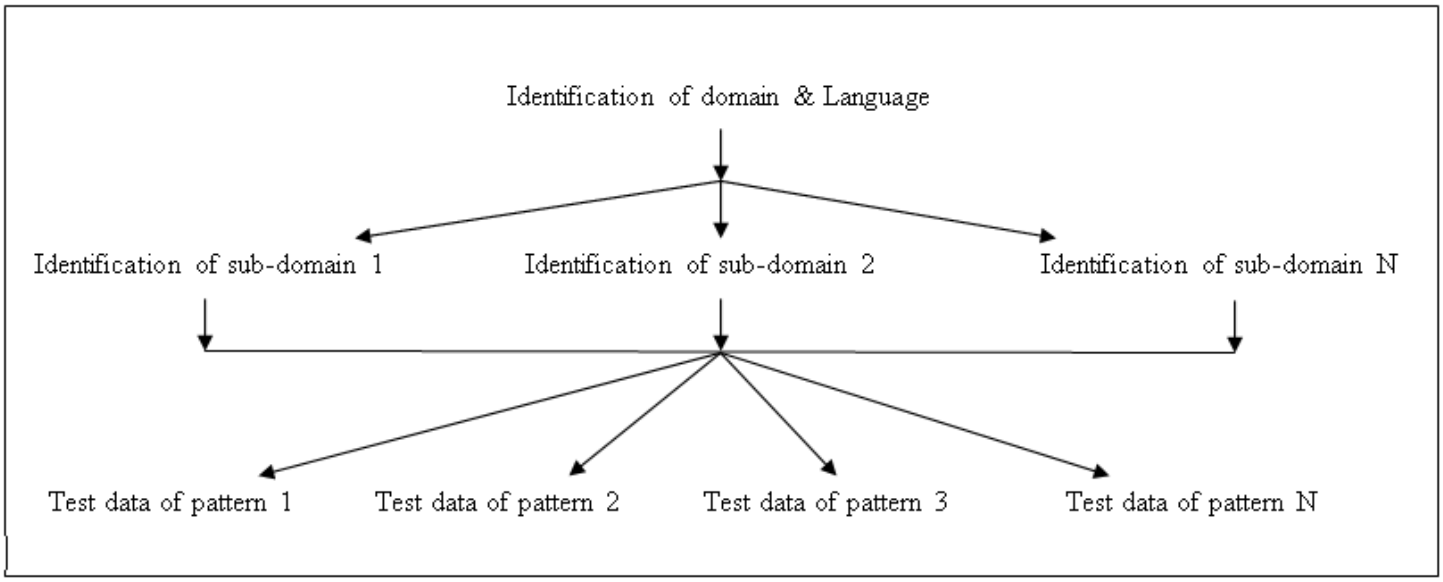
4.2 General Domain MT system

For general domain MT system, any type of input can be given and MT system is expected to provide good translation. General domain may cover major domains like tourism, health, politics, children's stories, recipe books, law, sports, etc. While collecting the test data for general domain MT system, it should cover all the subjects along with source & target language features.

4.3 Test data

To evaluate the quality of MT system, it is advised to collect the test data domain wise as well as different pattern wise, so that it will give exact measure of the MT system on which domain MT system is performing better. Also while designing the test data, different patterns and linguistic features need to consider. For this linguist's help can be taken, because it should be representative of the grammatical patterns of language.

Broad level data categories are as provided below.



While collecting domain wise test data, its major sub-domain need to be considered carefully and data should be collected from each sub domain.

e.g. In tourism domain, major sub-domains need to consider like major sub-domain is Ecotourism and its sub domains are wildlife, hill stations, the desert tourism and other attractions of nature, Or if major sub-domain is Adventure tourism then its sub domains are trekking, paragliding, river rafting, skiing, jeep safari, jungle safari and likewise.

4.3.1 Different Patterns of sentences

Various patterns of sentences should be used in test bed creation to analyze how well the MT system can handle each grammatical & linguistic pattern of a language. This will give the useful analysis regarding quality of MT system.

Testing of MT system is gradual process and for this test data should be designed accordingly. Following are the major patterns focused for test data creation for evaluation of MT system with sample examples.

1. **Simple Sentences**

   आज का मौसम सुहाना है।

   The sun sets in the west.

2. **Compound Sentences**

   एक बालक खेल रहा था और एक बालिका पढ़ रही थी ।

   I went to market and purchased a T.V.

3. **Complex Sentences**

जो टुरिस्ट कुर्सी पर बैठा है उसे बुलाओ।

उन्होंने जिन टुरिस्टों को बुलाया है उनको मैं नहीं जानता।

रॉबर्ट, जो आज खजुराहो में है, जल्दी ही यहाँ आ रहा है।

जो पर्यटक आया, वह अंग्रेज है।

जो पर्यटक बाहर घूम रहा है, उसे बुलाओ।

जब हम केरल में थे तब बहुत वर्षा हुई थी।

अकबर बादशाह जहाँ रहते थे वहाँ बड़ा आलिशान महल बनवाया गया।

जहाँ तक पर्यटक बंबई में है वहाँ तक हम भी उनके साथ रहेंगे।

जैसा दाम वैसा काम।

पर्यटक चाहता है कि गाइड उसे मिले।

यदि आप चाहे तो हम जुम्मा मस्जिद की सैर करेंगे।

4. **Adjective Phrase**

अच्छा लड़का आया।

अच्छा समझदार लड़का आया।

उसने अच्छी समझदार खूबसूरत लड़की से शादी की।

बहुत अच्छा लड़का आया।

पर्यटक गाइड से छोटा है।

गाइड सबसे बड़ा है।

5. **Adverbial Phrase**

जब वह कश्मीर आ रहा था तब बारिश हो रही थी।

अमरनाथ की यात्रा में यहाँ कौन-कौन मिलने आने वाले हैं ?

घाटी में जैसा ही अंधेरा बढ़ा पर्यटक घबरा गए।

अमरनाथ यात्रा में लगभग सौ आदमी सम्मिलित हुए।

6. Affective

हे भगवान ! यह क्या हुआ !

ओह ! सूर्यास्त का यह दृश्य कितना सुंदर है।

यात्री ने कहा, वाह जनाब, इस दृश्य का तो जवाब नहीं।

हाय रे भाग्य! यात्रा में यह कैसा संकट आया।

7. Ambiguity

यात्रा में आम खाना आम बात है।

गाइड की चाल को पर्यटक समझ गए।

8. Discourse

गाइड ने कहा, "अब हम अगले साल फिर मिलेंगे।"

रामलाल ने कहा, "मैं आगरा जरूर जाऊँगा।"

"अक्षरधाम मंदिर बहुत सुंदर है", गोपाल ने कहा, "मैं अगले वर्ष उसे देखने जाऊँगा"।

9. Figures of speech

उस पर्यटक का चेहरा कमल-सा सुंदर है।

यात्री का हृदय सिंधु-सा विस्तृत था।

पर्यटक ने अपने बालक के बारे में कहा, यह जीवन-निशि का शुभ्र सवेरा है।

10. Infinitive Structures

गाइड कहानी बता रहा है।

पर्यटक ने सहयात्री को डूबने से बचाया।

राम गाना सीख रहा है ।

11. Interrogative

क्या पर्यटक सो गया?

पर्यटक हॉस्टल में रहते है कि होटल में?

सिकंदराबाद एक्सप्रेस कब छूटनेवाली है?

12. Krudanta forms

दौड़ना सेहत के लिए अच्छा है।

यात्रा में हमारे साथ श्याम का आना लाभदायक था।

13. Misspelled words

यात्रा में गिरि/गिरी पर चढ़ना पड़ता है।

शरदी के/सरदी के दिन ऊनी कपड़े पहनने चाहिए।

14. Named Entity Measures

यात्री 2 घंटे 20 मिनट और 40 सेकेंड चलता रहा।

पुणे से मुंबई 190 किलोमीटर दूर है।

मुंबई महाराष्ट्र की आर्थिक राजधानी है।

15. Noun Phrase

मुसाफिर आया।

गाइड, जल्दी जाओ।

चार यात्री आए और दस गए।

ऐन वक्त पर टिकट का किराया तीन गुना देना पड़ा।

16. Particles and clitics

पर्यटक अपने होटल जा ही रहा था जब उसका फोन कॉल आया।

अब तक की यात्रा के बाद यह तो कुछ और ही हो रहा है।

हमारी यात्रा में वहाँ भी (नैनीताल) तुम मिल सकते हो।

17. Negation

यात्रा में स्टेशन पर खुला रखा खाना न खाना।

नए प्रदेश की यात्रा में अपरिचित का कहना न मानना।

सिंहगड़ पर इस से आगे न चलिए।

18. Passive Voice

यात्रा में चोरों के हाथ पर्यटक लूटा गया।

यात्रा में पर्यटक साँप द्वारा काटा गया।

ड्राइवर द्वारा यात्रा में गाड़ी ठीक समय पर छोड़ी गई थी।

19. Poetic

कश्मीर का सुहाना सफर हम जिंदगी भर भूल नहीं सकते।

संकटों को पार करते यात्री आखिर श्रीनगर पहुँच ही गए, सच ही है हिम्मत-ए-मर्दा तो मदद-ए-खुदा।

यात्रा में घूमते हुए पर्यटक ने लगे हाथों अपना निजी काम भी कर लिया।

20. Semantics polysemy

1) धर्म

नियम — सहृदय व्यक्ति को प्रभावित करना कविता का मूल धर्म है।

कर्तव्य — सबकी सेवा करना मनुष्य का धर्म है।

संस्कार — मनुष्यों का धर्म नष्ट होता जा रहा है।

2) चाल

चालाकी — उसने शतरंज की चाल चली।

गति — इस साईकिल की चाल बड़ी तेज है।

धोखा — वह उस व्यक्ति की चाल में आ गया।

चरित्र — इस लड़की की चाल ठीक नहीं है। चाल चली सादा, निभे बाप—दादा।

21. Spelling variants

यात्रा के कारण रमेश दस दिन कालेज/कॉलेज नहीं गया।

यात्री का भाई उससे बस स्टैन्ड/स्टैंड पर मिला।

यात्रा में उसने हैट/हॅट पहनी।

22. Verbal Phrase

पर्यटक दक्षिण भारत-यात्रा करने जाएगी।

कश्मीर यात्रा में पर्यटक शिकारा नौका खेते रहे।

तुम हिमालय की यात्रा में औरों के साथ चल रही है।

पर्यटक बीच-बीच में पानी पी रही थी।

23. Idioms & Phrases

अंधा क्या चाहे दो आँखें — जब आवश्यक और इच्छित वस्तु बिना किसी प्रयत्न के अचानक मिल जाय तब यह कहावत लागू होती है।

कई दिनों से बेकार रमेश से अशोक ने पूछा कि तुम क्या मेरे ऑफिस में नौकरी करोगे और रमेश ने तुरंत उत्तर दिया वाह भाई अंधा क्या चाहे दो आँखें।

अँगूठा चूमना - चापलूसी करना।

आजकल विद्वानों को भी रूपये के लिए धनिकों के अँगूठे चूमने पड़ते हैं।

24. Days of week

कुछ यात्री मंगलवार के दिन व्रत रखते है।

आमतौर पर पहले और तीसरे शनिवार के दिन सरकारी कार्यालय बंद रहते हैं।

25. Focus and Thematisation

गिलास तोड़ा सीता ने डाँट पड़ी नीतु पर।

गलती की यात्री ने, डाँट पड़ी गाइड़ को।

पेन चुराया यात्री ने, डाँट पड़ी व्यवस्थापक को।

शराब पी पर्यटक ने, डाँट पड़ी व्यवस्थापक पर।

देख लिया यात्रियों ने गाइड का साहस।

CHAPTER 5

CONCLUSION

Conclusion

This document contains different maximum possible evaluation techniques for Indian languages MT evaluation and issues associated with it. Irrespective of MT development approaches, evaluation of MT system is important. Both human evaluation (subjective) and automatic evaluation (quantitaive) have some advantages and disadvantages.

Human evaluation is expensive, slow and not reusable and it has inherent problem of inter-annotator agreement. On the other hand automatic evaluation is quick, inexpensive and mostly language independent. But MT evaluation metrics are mainly the tools for consistent MT system development and may not have that much to do with real quality of translation. Hence preference is given to human evaluation along with the error analysis. Error analysis will help to identify the issues associated with individual modules in MT system. This will help to improve the quality of individual modules and overall MT output quality.

The most commonly used human evaluation methods are "fluency & adequacy" and "Intelligibility & fidelity".

Fluency shows that text is well formed according to the rules of target language & adequacy evaluates how much information is transferred between the original and translation.

Intelligibility shows comprehensibility and clarity of the translation which may be affected by grammatical errors, mistranslation, untranslated words & on the other hand fidelity evaluates how much information is retained in the translated sentence compared to original source sentence.

The main idea of automatic evaluation is to compare machine translation output with reference; however Indian languages have more than one single correct output, so only comparison will not work in evaluation of MT. Hence along with subjective evaluation, error analysis needs to carry out.

For this subjective evaluation, the translators as well as moderators/evaluators needs to be chosen with great care and a workshop be organized for the moderators especially to make them aware of the complexities of evaluation.

Also this document provides guidelines and an approach to collect testing data for MT system. Documentation of this process would help us in future to relate how the test database was generated. The main purpose is to test MT system with all the possible parameters. This document will always be in updating mode as new data type introduced in future would be added.

References

References

1. Evaluation of machine translation [Online]. Available: http://en.wikipedia.org/wiki/Evaluation_of_machine_translation.[January 9, 2014]

2. Sangal Rajeev et al. Machine Translation System:Shakti[Online]. Available: http://gdit.iiit.net/~mt/shakti/ , 2003.[January 20,2014]

3. Word error rate [Online]. Available: http://en.wikipedia.org/wiki/Word_error_rate. [January 28,2014]

4. Martin Thoma. Word Error Rate Calculation [Online]. Available: http://martin-thoma.com/word-error-rate-calculation. 2013. [February 4, 2014]

5. Sara Stymne, Machine Translation Evaluation [Online].Available: http://www.ida.liu.se/labs/nlplab/gslt/mt-course/mteval-sarst.pdf. [February 10,2014]

6. Evaluation of machine translation [Online]. Available: http://www.translationdirectory.com/articles/article1814.php . 2008. [February 27,2014]

7. White, John S. "Approaches to black box MT evaluation." *Proceedings of Machine Translation Summit V*. 1995.

8. Flanagan, Mary. "Error classification for MT evaluation." *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. 1994.

9. Goyal, Vishal, and Gurpreet Singh Lehal. "Evaluation of Hindi to Punjabi machine translation system." *arXiv preprint arXiv:0910.1868* (2009).

10. Josan, Gurpreet Singh, and Gurpreet Singh Lehal. "A Punjabi to Hindi machine translation system." *22nd International Conference on on Computational Linguistics: Demonstration Papers*. Association for Computational Linguistics, 2008.

11. Project EuroMatrix : Statistical and Hybrid Machine Translation Between All European Languages. Saarland University (USAAR), University of Edinburgh (UEDIN), Charles University (CUNI-MFF), CELCT, GROUP Technologies, MorphoLogic. Saarbrucken, Germany. 2007

12. Sinha, R. M. K., and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system." *MT Summit IX, New Orleans, USA* (2003): 494-497.

13. Balyan, Renu, et al. "A diagnostic evaluation approach for english to hindi MT using linguistic checkpoints and error rates." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2013. 285-296.

14. Joshi, Nisheeth, et al. "HEVAL: Yet Another Human Evaluation Metric." *arXiv preprint arXiv:1311.3961* (2013).

15. Finch, Andrew, Young-Sook Hwang, and Eiichiro Sumita. "Using machine translation evaluation techniques to determine sentence-level semantic equivalence." *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. 2005.

16. Minnis, Stephen. "Constructive machine translation evaluation." *Machine Translation* 8.1-2 (1993): 67-75.

17. Parton, Kristen. *Lost and Found in Translation: Cross-Lingual Question Answering with Result Translation*. Diss. Columbia University, 2012.

18. Zhou, Ming, et al. "Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points." *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008.

19. Specia, Lucia, et al. "Predicting machine translation adequacy." *Machine Translation Summit*. Vol. 13. No. 2011. 2011.

20. Correa, Nelson. "A fine-grained evaluation Framework for machine translation system development." *MT Summit IX*. 2003.

21. Van Slype, Georges. "Critical study of methods for evaluating the quality of machine translation." *Prepared for the Commission of European Communities Directorate General Scientific and Technical Information and Information Management. Report BR* 19142 (1979).

22. Goyal, Vishal. *Development of a Hindi to Punjabi Machine Translation system*. Diss. Ph. D. Thesis, Punjabi University, Patiala, 2010.

23. Mauser, Arne, Sasa Hasan, and Hermann Ney. "Automatic Evaluation Measures for Statistical Machine Translation System Optimization." *LREC*. 2008.

24. Specia, Lucia, V. Nunes Maria das Graças, and Mark Stevenson. "Exploiting rules for word sense disambiguation in machine translation." *Procesamiento del lenguaje natural, nº 35 (sept. 2005); pp. 171-178* (2005).

25. Kalyani, Aditi, et al. "Assessing the Quality of MT Systems for Hindi to English Translation." *arXiv preprint arXiv:1404.3992* 2014.

26. Song, Xingyi, Trevor Cohn, and Lucia Specia. "BLEU deconstructed: Designing a better mt evaluation metric." *Proceedings of the 14th International conference on Intelligent text processing and computational linguistics CICLING*. 2013.

27. Isozaki, Hideki, et al. "Automatic evaluation of translation quality for distant language pairs." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

28. Liu, Chang, Daniel Dahlmeier, and Hwee Tou Ng. "Better evaluation metrics lead to better machine translation." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

29. Callison-Burch, Chris, and Miles Osborne. "Re-evaluating the role of BLEU in machine translation research." *In EACL*. 2006.

30. Brew, Chris, and Henry S. Thompson. "Automatic evaluation of computer generated text: a progress report on the TextEval project." *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994.

31. Condon, Sherri, et al. *Evaluation of machine translation errors in English and Iraqi Arabic*. MITRE CORP MCLEAN VA, 2010.

32. Fishel, Mark, et al. "Automatic translation error analysis." *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2011.

33. Vilar, David, et al. "Error analysis of statistical machine translation output."*Proceedings of LREC*. 2006.

34. Stymne, Sara, and Lars Ahrenberg. "On the practice of error analysis for machine translation evaluation." *LREC*. 2012.

35. Ananthakrishnan, R., et al. "Some issues in automatic evaluation of english-hindi mt: more blues for bleu." *ICON* 2007.

36. Joshi, Nisheeth, Hemant Darbari, and Iti Mathur. "Human and Automatic Evaluation of English to Hindi Machine Translation Systems." *Advances in Computer Science, Engineering & Applications*. Springer Berlin Heidelberg, 2012. 423-432.

37. Pierce, John R., and John B. Carroll. "Language and machines: Computers in translation and linguistics." 1966.

38. Doherty, Stephen, Sharon O'Brien, and Michael Carl. "Eye tracking as an MT evaluation technique." *Machine translation* 24.1 (2010): 1-13.