IEEE SA
STANDARDS
ASSOCIATION

INDUSTRY CONNECTIONS
REPORT

INDIAN LANGUAGE RESOURCES—

EVALUATION SUBCOMMITTEE REPORT

Authored by

IEEE SA Evaluation Subcommittee

IEEE

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

ACKNOWLEDGMENTS

Special thanks are given to the following reviewers of this paper:

Ajay Srinivasamurthy, *Amazon*
Akash Singh, *Sarthi.ai*
Amitava Das, *Wipro*
Babita Shinde, *CDAC*
Bharat Gupta, *TDIL*
C. V. Jawahar, *IIIT Hyderabad*
Prasenjit Majumder, *DAIICT*
Sriram Ganapathy, *IISc*
Vijay Kumar, *TDIL*
Vinay Namboodiri, *IIT Kanpur*

The authors also thank the following individuals for their insightful comments:

Nikesh Garera, *Flipkart*
Mitesh Khapra, *IIT Madras*
Anoop Kunchakuttan, *Microsoft*
Pratyush Pandey, *IIT Madras*
Shourya Roy, *Flipkart*

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

PDF: STDVA26426 979-8-8557-0065-7

Find IEEE standards and standards-related product listings at: http://standards.ieee.org.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association ("IEEE SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder.

This Work is published with the understanding that IEEE and the IC Com members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

EVALUATION SUBCOMMITTEE REPORT

ABSTRACT

This report summarizes the findings and recommendations of the evaluation subcommittee of the IEEE prestandardization effort on standards for Indian language resources and evaluation for speech and language technology. Evaluation is a crucial aspect of any system. There are several generic challenges for the evaluation of speech and language technology, especially due to the ambiguous nature of human languages, as well as the interactive nature of many of these systems. Furthermore, there are additional challenges specific to Indian languages and the Indic context. This report summarizes the current metrics and datasets used for the evaluation of various technologies and identifies gaps that need to be addressed. Since there were dedicated subcommittees on speech, text, script, and accessibility, therefore, this report does not go into the details of various evaluation issues specific to these subareas. Instead, here the more generic issues that cut across these subdomains are discussed. A case study is also presented on machine translation (MT).

1. INTRODUCTION

Evaluation is a crucial aspect of any software system, be it an end-user-facing system or an intermediate tool or model that is being used to build other end-to-end systems. In the context of speech and language technology (SLT), evaluation includes three distinct parts—the evaluation protocol or process, the evaluation test bench (or test data), and evaluation metrics. All these three components require standardization because ideally, any evaluation process and its findings should have the following properties:[1]

- **Independent**—Of the system architecture, datasets, and development process.
- **Intentional**—The rationale for an evaluation and the decisions to be based on it should be clear from the outset.
- **Transparent**—The process of evaluation and the metrics should be transparent.
- **Reproducible**—Evaluation processes should be reproducible, and the same system/model evaluated on the same test set should ideally result in the same (or very similar) values for the standardized evaluation metrics.
- **Impartial**—Evaluation should be carried out keeping in mind the specifications of the system, such as the languages and domains where it is supposed to work on, and not on an arbitrary set of test cases.
- **Of high quality**—All evaluations should meet minimum quality standards defined by the Evaluation Office.
- **Timely**—Evaluations should be timely (and therefore, the standards—including protocols, metrics, and test sets, should also be available in a timely manner).
- **Used**—Evaluation should be useful for the industry, including the companies building the system, or using those provided by a third party. It should also be useful to researchers who wish to invent better (according to the evaluation standards) systems, and to the end user (when applicable) who ideally should be able to estimate the personal utility of a system based on evaluation reports.

The above guidelines for ideal evaluation calls for well-thought-out standards of metrics, datasets, and protocols of evaluation. In this report, the authors present their analysis of the existing standards for evaluation of SLT and the gaps and challenges, in the specific context of Indian languages.

[1] The United Nations Development Programme (UNDP) Norms for Evaluation sets out a detailed list of ideal features that an evaluation process of development program should follow. Many of the following points have been taken from the UNDP Norms for Evaluation with appropriate adaptation to the context of SLT.

The authors begin by specifying the scope of this analysis (Section 2), the approach taken (Section 3), followed by the overview of their findings (Section 4). Since it is impossible to present an analysis of every task in language and speech processing, they take one important use case—machine translation (MT) (Section 5) through which they illustrate the various aspects of standards and availability of resources for these tasks and highlight the gaps. Finally, in Section 6, the gaps are summarized and several recommendations are made that are related to the standardization of evaluation practices in SLT for Indian languages.

2. SCOPE

This report will cover the following aspects of evaluation:

1) Survey of existing evaluation metrics and standardized approaches and protocols, including automatic, semiautomatic, and manual approaches, for evaluation of the following language processing systems that are end-user-facing, as well as developer-facing.
    a) Text processing: MT, information retrieval, information extraction, dialogue and conversation systems, question answering, and summarization.
    b) Multimodal language processing.
2) Survey of existing evaluation testbenches across the 22 scheduled languages of India that will enable one to test language systems in an automated fashion.
3) Identify gaps in both #1) and #2) above.

The following aspects, though within the scope of evaluation in general, were not investigated:

1) Specifics of evaluation of speech processing, basic text processing (morphology, parsing, parts of speech (POS), sentiment, etc.), and script/input method evaluation. These were covered by the respective subcommittees. Here, on the other hand, the authors report the more general aspects of the evaluation of speech and text processing systems that cut across specific tasks and applications.
2) Evaluation of quality of datasets—such as noise in a corpus, interannotator agreement metrics, etc. These could be covered by respective subcommittees while documenting existing resources.
3) Reporting of numbers from evaluation drives and studies.
4) The authors primarily cover the evaluation of functional aspects (i.e., accuracy type metrics) of the

systems. Although the important nonfunctional aspects (speed, user interface aspects, developer-friendliness, scalability, etc.) are documented, they may not be discussed in depth. Certain aspects (like fairness and transparency) might be discussed in depth, if the subcommittee feels that these are important aspects of language systems for which standards are required.

3. APPROACH

The authors took a six-stage approach as follows in the following subsections.

3.1. STAGE 1: SURVEYING OF EXISTING STANDARDS

During this phase, the existing metrics and datasets for evaluation in different areas of SLT were surveyed. Evaluation exercises carried out by the government, industry, and academia, which are publicly known, were also listed and studied.

3.2. STAGE 2: SCOPING

The findings were exchanged with other subcommittees, and it was decided that since the evaluation of SLT is a very large area, it would not be possible to list the metrics and standards for each task in SLT. Instead, the standards for specific speech and text processing tasks will be discussed in the respective subcommittee reports. It was decided that the evaluation subcommittee will take a broader perspective on standards and datasets for evaluation and focus mostly on the gaps that exist in these areas, especially in the context of Indian languages.

3.3. STAGE 3: RUBRIC OF EVALUATION STANDARDS

Based on the findings of stage 1, the authors came up with a rubric of evaluation standards—various aspects of evaluation; this helped them to further define the scope of this subcommittee.

3.4. STAGE 4: SPECIFIC USE CASES

Taking MT as a use case, the authors conducted a detailed study of the availability of metrics and gaps in the area. MT was chosen as a use case because it covers a large range of issues representative of SLT.

Furthermore, the evaluation metrics—both human and automatic—for MT are largely debated, and there are interesting Indian language-specific issues that have been raised as well. Therefore, this case study provides a holistic idea of what are the various challenges in SLT evaluation.

3.5. STAGE 5: INDUSTRY SURVEY

In parallel, a survey was conducted with several industry experts and practitioners to understand the practical challenges in the evaluation of user-facing SLT systems that were being built in the industry. These were conducted through an initial online survey form (Appendix A1) followed by in-depth interviews with experts who indicated that they were available and willing.

3.6. STAGE 6: GAP IDENTIFICATION AND RECOMMENDATION

Based on the findings of the stages, the authors finally came up with a set of gaps and recommendations. Some of these gaps are generic and apply across the tasks and subareas of SLT, whereas a few gaps are more specific to certain areas but important enough to be discussed.

4. EVALUATION STANDARDS

In this section, the authors discuss various aspects of evaluation. Usually, by "evaluation of an SLT system" one means the analysis of the functional accuracy of the system, which shall also remain the focus of the current discussion. However, they are aware that the evaluation of a system, typically, is a much broader exercise that encompasses issues like the speed and throughput of the system, its privacy, and security aspects. During the interviews with industry experts and practitioners, several of these aspects came to the forefront. Based on their analysis, they propose a holistic rubric for evaluation, which is summarized in Table 1.

TABLE 1 **Aspects of SLT system evaluation**

Both	End user	Developer
Functional accuracy	Privacy	Robustness
Performance–time*	Fairness	Scalability
Security	Explainability and accountability	Memory footprint*
Licenses and pricing*		Language/domain/task-independence
Importance: Obligatory (green), recommended (yellow), ideal (blue), and optional (white).		Documentation
*Metrics are obvious, and standardization might not be necessary.		Code management

Evaluation usually can be done from the perspective of the end user or the developer. Some aspects are common to both, while certain aspects of evaluation are specific to the target group. Therefore, in Table 1, the authors present these three aspects of evaluation—end-user-centric, developer-centric, and those relevant for both under three columns. One would immediately evaluate several of these aspects, such as memory footprint, performance time, documentation, and code-management, which are generic to any software system. They did not come across any case or issue that requires special treatment for SLT. Hence, here they will passingly mention those aspects and devote more space for discussion on aspects that require specific standards or datasets for SLT.

The authors elaborate on each of these following aspects.

- Functional accuracy refers to the evaluation of a system purely based on the function it is supposed to fulfill. For instance, a POS tagger's functional accuracy or evaluation depends on how well it can do POS tagging. This is the most crucial and commonly discussed aspect of the evaluation of Natural Language Processing (NLP) systems. Functional accuracy could be computed automatically or through manual or semiautomatic processes. There are various metrics, both qualitative and quantitative to specify the functional accuracy of a system. Section 4.1 discusses various aspects of functional accuracy and lists standard metrics and datasets.
- Performance—Time is also a practically important aspect of any system, which might depend on not only the NLP model but also other things including hardware and network latencies. The standard

metrics of evaluation are as follows:

  - Response time—The time elapsed between the issuance of a query (or input) to the system by a user, and the display of the output by the system.
  - Throughput—The amount of information (input units) processed by the system per unit of time.

  Note that the response time might be adversely affected when the system is operated at its peak throughput. Since these factors are common to all software systems, these will not be discussed any further.

- The memory footprint of a system determines its usefulness, especially portability on mobile phones. Often large models cannot be operated from small devices. They require to be run on the cloud. Therefore, from the user's perspective, such a system would require continuous internet access. This is an important concern, especially in India, because of the following two reasons:

  - Many Indian users may not have access to high-end phones with large in-phone memory.
  - Internet access might be expensive or unavailable to several users.

  On the other hand, SLT systems are becoming larger and larger, leading to more accurate systems. This creates an interesting tradeoff between the accuracy and affordability of the systems for the end users. The authors believe that the memory footprint of a system/model is, therefore, an important parameter on which they should be evaluated. More specifically, performance gain per unit of memory footprint for a model (such as F-score-change per megabyte of a model size) should be an important standard of evaluation.

- Scalability refers to aspects such as how many users can access the system, or how many requests can be handled by the system simultaneously. In the specific context of SLT, it could also refer to how large a database of documents a search engine can index and search within a response-time budget, or how many languages a single model can process (for instance, in the context of current massively multilingual models such as multi-lingual BERT (mBERT) [1][2]). The dimensions of scale—users, documents, languages, etc.—vary widely for systems; but once defined, the metric is of scalability becomes obvious. Therefore, this is not discussed further here, except for noting that scalability has a

[2] Numbers in brackets refer to references listed in Section 7.

strong dependence on memory-footprint and performance-time, and the three factors need to be considered simultaneously.

- Security is an extremely important aspect of any software system. The basic requirements for security for an SLT system are similar to those of any software system. International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) 27001, part of the growing ISO/IEC 27000 family of standards, is an Information Security Management System (ISMS) standard, of which the last revision was published in October 2013 by the ISO and the IEC. Its full name is ISO/IEC 27001:2013—Information technology—Security techniques—ISMSs—Requirements. ISO/IEC 27001 formally specifies a management system that is intended to bring information security under explicit management control and is applicable to SLT system security standards as well.

- Privacy and data protection: Since SLT systems often directly interface with users, they typically have access to user data. Therefore, an important aspect of the evaluation of these systems includes whether the system is compliant with the privacy and data protection laws as well as standards in the respective geography. These standards typically also cover language data collection and data usage policies. Note that these laws and standards are typically applied based on the region where the system is being used rather than the language. Two typical examples of existing laws and standards are as follows:

    - General Data Protection Regulation [European Union (EU)] 2016/679 (GDPR) is a regulation in EU law on data protection and privacy in the EU and the European Economic Area (EEA).

    - California Consumer Privacy Act is a bill passed in 2018 that protects the data privacy rights of the residents of California, USA.

- Robustness: The authors define the robustness of an SLT system by its ability to handle varied (possibly ill-formed, such as due to spelling and grammatical errors) inputs, and its ability to handle various boundary conditions. Note that the varied inputs also must be from the domain/language that the system is supposed to handle. Measuring the robustness of SLT systems is a difficult problem. The authors believe the recently proposed checklist approach [2] provides a technique to test the robustness of the SLT system. However, the checklist is limited to text-based classification tasks. It is to be seen how this can be extended to other kinds of tasks and for speech. Further, the current approach to check the listing of text-based systems is ad hoc. An important research agenda could be the standardization of the checklist for a large number of text and speech tasks and making the checklists public.

- Language/domain/task independence: SLT systems typically perform well for a domain, and they are built for a single task and language. However, from a developer's perspective, it is important to know whether and to what extent an approach or a system is generalizable to other languages, domains, and tasks (LDTs). More specifically, for multilingual models, it is important to know how well they serve the different LDTs. Standard metrics to characterize these features would be of great importance for the proper evaluation and adoption of the models.

- Fairness: SLT systems must be fair along different dimensions such as race, gender, sexual orientation, age, geographical-region, etc., and their intersections. Fairness can be defined variously, but more often it refers to equal or equitable performance of a system on inputs that contain and/or was generated by users from different groups along these dimensions. For instance, along the gender axis, fairness could be defined as (1) equal or comparable performance on input speech or text from people of different genders and (2) equal or comparable performance on input text or speech examples, where all except the gender of the lexical items are same. The fairness of the SLT system is a hot topic of debate as machine learning systems are often biased toward majority classes and show incorrect or objectionable outputs for underrepresented minorities. See Blogdett et al. [3] and references therein. Since this is an evolving area, metrics and standards of fairness do not exist, and this gap will be discussed in detail in Section 6.

- Explainability and accountability: Like fairness, explainability and accountability of AI systems have become an important point of discussion, especially because deep-learning-based models, which are popular for their good performance, are often nontransparent. The outputs of these systems are difficult to explain, and consequently, the systems are also difficult to debug (from a developer's perspective) and use to make real-world decisions (from a user's perspective, such as in flagging content as offensive). There are no standards or metrics of explainability. The authors believe that this is an area that, again like fairness, will draw a lot of attention in the near-future from researchers and policy makers. Some of the specific gaps related to explainability are discussed in Section 6.

- Documentation, code-management, and licensing: The general software standards for the documentation apply here, and the authors do not foresee any specific issues that might arise in the context of language and speech. Hence, this is not discussed further.

4.1. STANDARDS FOR MEASURING FUNCTIONAL ACCURACY

As mentioned at the beginning of this section, functional accuracy is the most reported aspect of evaluation. These are also the metrics that researchers and system builders typically optimize for. Hence, these metrics are well-developed. At a high level, the functional metrics for a task are chosen based on the nature of the task. For instance, classification tasks, such as sentiment analysis, offensive or hate speech detection, and natural language inferencing, are measured using precision, recall, and F-score, and variants of these, while ranking tasks such as document retrieval, use mean average precision (MAP), mean reciprocal rank, normalized discounted cumulative gain (nDCG), etc., for evaluating the model performance.

Table 2 summarizes the common metrics used for the functional evaluation of some standard problem paradigms, along with popular SLT tasks related to these paradigms. This is not meant to be an exhaustive list; rather, it is for exemplifying the strong connection between paradigms and evaluation metrics, which in turn implies that for most tasks that can be casted into one of these popular paradigms, one does not need to design special metrics of evaluation. Complex structured prediction tasks like parsing and morphological analysis, and language generation tasks such as transliteration and MT are exceptions, where task-specific metrics are used.

TABLE 2 Common evaluation metrics used for SLT tasks based on problem paradigms

Problem Paradigm	Example Tasks	Evaluation Metrics
Classification	Sentiment/emotion detection Bias/offensive/hate speech detection Natural language inferencing, entailment	*Precision, recall, F-score* *Accuracy, RUC, AUC, specificity-sensitivity*
Sequence labeling	POS tagging, named entity recognition and span detection	*Accuracy, weighted precision, recall, F-scores.*
Regression	Answer/essay grading	*Mean squared error, quadratic weighted kappa*
Language generation	Transliteration MT Summarization Question-answering Language modeling	*Exact match, character-edit distances, phonetically weighted edit distances* *Recall-oriented understudy for gisting evaluation (ROUGE), bilingual evaluation understudy (BLEU)-1, BLEU-4, exact match, F1 (correct words retrieved)* *Perplexity*

Problem Paradigm	Example Tasks	Evaluation Metrics
Ranking	Document retrieval/web search, answer retrieval, recommendation	*Precision, recall, F-measure, B-pref* *MAP, mean reciprocal rank, 11-point interpolated average precision, nDCG* *Expected reciprocal rank for graded relevance*
Clustering	Topic modeling	*Purity, precision-recall, and index*
Distribution similarity	Subjective judgments in annotation	*KL divergence, mutual information, cross-entropy*
Set-based similarity	Keyword extraction	Jaccard similarity, DICE
Structured prediction	Parsing Morphological analysis	There are task-specific metrics.
Retrieval/extraction	Entity and event extraction Relation extraction Ontology	Coverage, coupling, cohesion Longest common subsequence, ROUGE, BLEU
Dialog system		Task completion rate

4.2. EVALUATION STANDARDS ACROSS LANGUAGE, DOMAIN, AND USERS

An important aspect of the investigation carried out during this effort was whether and which of the evaluation standards are sensitive to language, domain, and users. While functional metrics (as described in Table 2) are mostly independent of these factors, there are certain aspects of evaluation that might require special attention. Here are a few examples:

- Language-specific metrics:
  - Transliteration metrics based on edit-distance might depend on language-specific character-substitution cost matrix.
  - Metrics for morphological analysis might be sensitive to language-specific phenomena.
- Domain-specific metrics:
  - Document retrieval in the legal domain is more sensitive to recall than precision. Also, in the absence of ground truth, estimated precision and estimated recall are used instead of true recall and precision. https://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW16.pdf
- User-specific metrics:

- Readability level might be an important metric for MT when the target user is children or semiliterate people.

4.3. EVALUATION IN INDUSTRY

Industry often follows different evaluation practices, where the primary metric is based on user satisfaction. User satisfaction can be assessed through direct feedback or indirect signals like clicks or time-spent. A second set of metrics that the industry is often interested in are DAU and MAU—daily and monthly active users (of a system), respectively. One standard process followed by the industry to assess the usefulness of a new technology is A/B testing, where A is the existing system and B is the new technology. A random set of users receive B, whereas the remaining user still receives A. The feedback, both direct and indirect, is measured for these two sets of users. If the feedback from those using B is more favorable, then A is eventually replaced by B. These methodologies are independent of the task, language, domain, and other factors, and are therefore useful. But they also require fully developed deployable systems. Note that A/B testing and user feedback might not capture the fairness and explainability aspects of a system.

During their interviews, the authors also found that almost all companies heavily rely on standard metrics of evaluation (such as those summarized in Table 2) during the model-building phase and would benefit from publicly available metrics and standard test datasets for the evaluation of various SLT tasks across languages and domains.

4.4. EVALUATION DATASETS

Datasets are an essential component of all evaluation activities. While it is not directly part of a standard, the availability of "standard datasets" for various tasks is useful in a fair comparison of systems. To this end, shared tasks and public datasets, and leaderboards have been especially useful in fostering research. In Indian languages, it is noted that there has been a series of shared tasks run by FIRE[3] on information retrieval and extraction. Similarly, there are International Conference on Natural Language Processing (ICON)-shared tasks.[4] WMT-shared tasks on MT (Appendix A) have several Indian languages, and NEWS-shared tasks have released data for transliteration.

[3] Forum for Information Retrieval Evaluation (irsi.res.in)

[4] ICON 2020: AI-NLP-ML Group, IIT Patna

Some of the existing multilingual test benches like XNLI,[5] XGLUE[6] [4], and XTREME[7] [5] cover a few Indian language datasets for a few tasks. GLUECoS[8] and LINCE[9] benchmarks have code-mixed datasets for English–Hindi and English–Nepali. AI4Bharat catalogue[10] of Indian language resources lists several other publicly available datasets for Indian languages.

5. CASE STUDY ON MACHINE TRANSLATION

Translating the text from a source language into a target language or between any pair of languages by computers has been one of the earliest goals in computational linguistics; the method is well known as MT. Given the diverse interest of the MT community, research and development activities in MT have been pursued in many directions. Some of the core tasks in MT are listed next.

- Generic MT
- Domain adaptation of MT
- Similar language MT
- Low resource MT
- Document level MT
- Robustness of MT to noisy text
- Multimodal MT

There exist several approaches to MT as listed next.

- Translation memory (TM)
- Rule-based (linguistic) MT (RBMT)
- Example-based MT (EBMT)
- Statistical MT (SMT)
- Neural MT (NMT)
- Hybrid MT

[5] XNLI (nyu.edu)
[6] XGLUE (microsoft.github.io)
[7] GitHub - google-research/xtreme: XTREME is a benchmark for the evaluation of the cross-lingual generalization ability of pretrained multilingual models that covers 40 typologically diverse languages and includes nine tasks.
[8] GLUECoS (microsoft.github.io)
[9] ritual.uh.edu/lince
[10] GitHub - AI4Bharat/indicnlp_catalog: A collaborative catalog of resources for Indian language NLP

The performance of MT systems can be evaluated using both automatic and human evaluation. In 5.2, various evaluation strategies are described followed in MT evaluation. Section 5.2 also mentions the different evaluation campaigns and shared tasks organized to assess the state-of-the-art and progress in MT on standard testbenches. Finally, the gaps in MT evaluation regarding Indian languages are described.

Human evaluation is the de facto standard in MT evaluation. However, human evaluation needs expertise in the source language, target language, as well as domain expertise. Moreover, it is expensive and very slow. Automatic MT evaluation is very fast, and it helps tracking the progress of MT systems development. However, it requires ground truth produced by human translators. Different automatic MT evaluation metrics and human evaluation criteria are mentioned in the following sections.

5.1. CHALLENGES IN MT EVALUATION

The following are the challenges in MT evaluation:

- Like MT, evaluation of it is a bilingual task that demands the knowledge about source and target language.
- Language being ambiguous in nature raises the same difficulty in MT evaluation as it does in the case of the MT system. Evaluating the semantics of a source sentence in the target makes it challenging.
- It is difficult to get a generalized MT evaluation system as a single solution to all based on its dependency on the approaches and types of MT systems.
- The dependency on domain or language pairs may cause biasness in estimation on the new instance.
- The acceptance of translation quality and its subsequent evaluated score is exceptionally subjective.
- Knowledge, both in terms of human resources and as a data repository for MT evaluation, is always limited/biased or skewed which may question the trust in its correctness.

Some issues on the MT evaluation methods are discussed by Ananthakrishnan et al. [6] such as (1) operational—how much savings in time or cost an MT system brings to a process or application; (2) declarative—how much of the source is conveyed by the translation (fidelity) and how readable it is (intelligibility); and (3) typological—what linguistic phenomena are handled by the system. Operational and declarative methods are by definition of the black box kind [7], while typological methods may evaluate both intermediate and final outputs.

5.2. MT EVALUATION APPROACHES

The performance of MT systems can be evaluated using both automatic and human evaluation. In the following section, various evaluation strategies followed in MT evaluation are described.

Human evaluation is the de facto standard in MT evaluation. However, human evaluation needs expertise in the source language, target language, as well as domain expertise. Moreover, it is expensive, tiresome, and very slow. Automatic MT evaluation is very fast, and it helps tracking the progress of MT systems development. However, it requires ground truth produced by human translators. Sometimes it also lacks correlation with human judgments.

Sometimes, obtaining the reference translation is time-consuming and very expensive which motivated the researchers to start working on the translation quality estimation approach without using the reference translation [9], [10], [11]. Although, when reference translation is given, then the translation evaluation metrics are still preferred in both the approaches, i.e., automatic and human evaluation processes. Different automatic MT evaluation metrics and human evaluation criteria are mentioned in the following.

5.2.1. AUTOMATIC EVALUATION

- Untrained automatic evaluation metrics:
    - n-gram overlap metrics:
        - F-score.
        - BLEU [12].
        - NIST (Przybocki and Martin [13] ; Doddington [14]).
        - ROUGE (Lin, 2004).
        - Metric for evaluation of translation with explicit ordering (METEOR) (Lavie et al., 2004 [15]; Banerjee & Lavie, 2005 [16]).
        - Harmonic mean of enhanced length penalty, precision, n-gram position difference penalty, and recall (HLEPOR) (Han et al., 2013 [17]).
        - Rank-based intuitive bilingual evaluation score (RIBES) (Isozaki et al., 2010 [18]).
        - Consensus-based image description evaluation (CIDEr) (Vedantam et al., 2014 [19]).
    - Distance-based metrics:
        - Edit distance-based metrics:
            - WER word error rate (WER).

      - Multireference WER (mWER) (Ali et al., 2015 [20]).
      - All reference WER (aWER) (Tomás et al., 2003 [21]).
      - Translation edit/error rate (TER) (Snover et al., 2006 [22]).
      - Improved TER (ITER) (Panja & Naskar, 2018 [23]).
      - CDER (Leusch et al., 2006 [24]).
      - CharacTER (Wang et al., 2016 [25]).
      - Extended Edit Distance (EED) (Stanchev et al., 2019 [26]).
    - Vector similarity-based evaluation metrics:
      - MEANT 2.0 (Lo, 2017 [27]).
      - YISI (Lo, 2019 [28]).
      - Word mover's distance (WMD) (Kusner et al., 2015 [29]).
  - Semantic similarity models used as evaluation metrics:
    - Semantic textual similarity (STS) (Agirre et al., 2016 [30]).
    - Paraphrase identification (PI) (Kauchak & Barzilay, 2006 [31]).
    - Textual entailment (TE) (Padó et al., 2009 [32]).
  - Syntactic similarity-based metrics:
    - Using constituent labels and head-modifier dependencies (Liu and Gildea, 2005 [33]).
    - Using shallow parsers (Lo et al., 2012).
    - Reference dependency-based automatic evaluation metric (RED) (Yu et al., 2014, 2015 [34]).
- Machine-learned/tunable evaluation metrics:
  - Large-scale pretrained language models (PLMs):
    - OpenAI's generative pretrained transformer 3 (GPT-3) [45].
    - Google's Bidirectional Encoder Representations from Transformers (BERT) [1].
    - Embeddings from Language Models (ELMo) [46].
    - Google's XLNet [47].
    - Google's ALBERT [48].
    - Universal Language Model Fine Tuning (ULMFiT) [49].
    - Facebook's RoBERTa [50].
    - BLEURT (Sellam et al., 2020 [72]) [meta evaluation].

- NN-based used evaluation metrics:
    - Continuous space deep neural network (CSDNN) by Kreutzer et al. [54].
    - QUETCH system using Theano [55].
    - MLP architecture with lookup table layer and nonlinear activation function tanh by Collobert et al. [75].
    - Recurrent neural network (RNN) approach by Kim and Lee (2016).
    - Paetzold and Specia [56] introduced SimpleNets.
    - NMTScorer by Mareček et al. [57] for MT evaluation by exploiting long short-term memory (LSTM) models with attention to its core.
- Diagnostic evaluation: Automatic MT evaluation metrics apply a system-level single score for the entire test set, it does not tell anything about the strengths or weaknesses of an MT system. Diagnostic evaluation is carried out on linguistic checkpoints, and it gives linguistic unit-specific fine-grained evaluation scores:
    - Woodpecker[11] [39].
    - DELiC4MT[12] [8], [34], [36], [37], [38].

5.2.2. HUMAN EVALUATION

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

At the first level, MT human evaluation techniques can be classified as black box or glass box. Black box techniques consider only the output of the system, whereas glass box techniques look at the internal components of the system and the intermediate outputs.

Standard approaches to human evaluation are as follows:

- Automatic Language Processing Advisory Committee (ALPAC)[13] approach: Intelligibility and fidelity.
- ARPA[14] approach: Fluency and adequacy.

[11] https://www.microsoft.com/en-us/download/details.aspx?id=52447
[12] https://github.com/antot/DELiC4MT
[13] http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Automatic_Language_Processing_Advisory_Committee_.28ALPAC.29

- Postediting effort:

  Postediting effort [52] is measured by the number of keystrokes and time spent on producing a "correct" translation from MT output. Postediting effort considers word-to-pause ratio and average pause ratio, in addition to time spent and the number of keystrokes. Bach et al. [53] used it to highlight the translated words, phrases, and sentences that require revisions.

- Cognitive load (CL) measures:

  CL estimation in the translation domain [51] is particularly interesting.

  - Due to the parallel activation of two languages, reading for translation imposes more demand on the working memory than reading within a single language.
  - Very bad MT proposals that are still very easy to post edit due to the simplicity of the segments, or the contrary situation, a very high MT quality where spotting the error can remain difficult and induce a high CL.
  - Features that are typically included in CL estimation:
    - Time-based features.
    - Text-based features.
    - Sensor-based features.
    - Combination of the above three.
    - Nine-scale subjective evaluation.

5.3. DATASETS

There are several workshops and shared tasks for MT evaluation exercises through which parallel data have been created in various Indian languages for training as well as testing of MT systems. In Appendix A, a comprehensive list of these datasets and evaluation exercises are presented.

[14] http://en.wikipedia.org/wiki/Evaluation_of_machine_translation#Advanced_Research_Projects_Agency_.28ARPA.29

5.4. GAPS IN MT EVALUATION WITH REFERENCE TO INDIAN LANGUAGES

Nevertheless, due to several reasons, the quality of the translation produced by the MT system is still not perfect, and therefore, the end user is unable to trust a particular translation. MT evaluation can efficiently detect the errors involved in translation and conclude the overall quality of the MT system. Though not much, research work has grown up at a fast pace for MT and its evaluation for Indian language. The quality labeling used for each instance of the corpus on a five-point scale is the score annotated manually by using Heval, a subjective human evaluation metric for Hindi language proposed by Joshi et al. [58]. A study on evaluating MT evaluation's NIST metric for English–Hindi language MT was presented by Tomer (2012) and evaluating MT evaluation's BLEU metric for English–Indian language MT by Sinha and Tomer [59].

Ananthakrishnan et al. [6] argue that BLEU is not appropriate for the evaluation of Indian language MT systems that produce indicative (rough) translations. The paper criticizes the BLEU score for its (1) intrinsically meaningless score; (2) admits too many variations; (3) admits too little variations; (4) an anomaly—more references do not help; and (5) poor correlation with human judgments.

Detailed points on the gap in MT evaluation are explored further as follows:

1) Existing metrics cannot capture the relative flexibility in word ordering in Indian languages.
    a) Dependency relation-based MT evaluation can capture this.
2) Existing MT evaluation metrics do not consider morphology while measuring the similarity between the hypothesis and reference(s). This is a crucial gap in MT evaluation since the majority of the Indian languages are morphologically rich.
    a) For example, বলেছিলাম [(I) had said] → বল + ◌ে + ছ + ◌ি + ল + ◌াম, vs. বললাম [(I) said] → বল + ল + ◌াম.
    b) Although METEOR considers synonym matches, it does not consider morpheme matching. Moreover, METEOR considers synonym matches only for English.
    c) Subword level MT evaluation together with word level matching can perhaps capture morphology to an extent.
3) Some punctuations are critical in evaluating MT since misplacing those punctuations might totally alter

the meaning. However, automatic MT evaluation metrics are not capable of differentiating these cases. For example:

- এখানে জঞ্জাল ফেলবেন না, ফেললে জরিমানা হবে [Don’t litter here, otherwise you will be fined.]
- এখানে জঞ্জাল ফেলবেন, না ফেললে জরিমানা হবে [Litter here, otherwise you will be fined.]

4) MT evaluation does not support domain-specific terminology evaluation, which is very important, particularly from the industry perspective, as well as in critical translation domains like translation of patents, legal documents, etc. For example:
    - Pick your Samsung smartphone.
    - Pick your Apple iPhone.
5) Existing evaluation metrics do not consider document-level formatting tags.
6) Evaluating document level MT:
    a) Anaphora evaluation (evaluating pronouns and their agreements with respect to honorifics and person information).
    b) Coherence and cohesion.
    c) Discourse (maintaining the flow of the document including anaphoras).
7) Existing MT evaluation metrics do not consider the source text while evaluating the hypothesis; all the metrics essentially consider the match between the reference and the hypothesis. To properly mimic human evaluation, automatic metrics should also consider the source text.
8) Existing MT systems are not robust enough to handle the code-mixed text scenario abundant in the Indian context.
9) Lack of proper tools/interface for human evaluation.
10) Lack/scarcity of benchmark datasets (both training and testing).
11) Lack of enthusiasm/interest in the industry in MT metrics, benchmarking datasets, and organizing shared tasks in MT.

With the availability of data and computational resources, the need arises to get MT evaluation metrics that would be able to overcome the additional difficulty originating from the complexity to model the user and changes in data domain and language pairs. An approach is needed for estimating the translation quality by targeting the robustness of the end user and domain changes using multiple machine learning and deep learning algorithms for evaluating the MT output quality with or without using the gold reference translation.

6. GAPS AND RECOMMENDATIONS

During this study, the authors overserved several gaps in the current metrics, datasets, and processes for the evaluation of speech and language technology. The gaps are categorized into the following types:

- Generic gaps are those that were observed to be applicable across many systems and technologies. These gaps are the primary focus.
- Technology-specific gaps have also been observed, mostly in the context of datasets, domains, and languages. These will be cursorily mentioned.

6.1. GAPS IN EVALUATION METRICS

As discussed in the previous sections, depending on the type of task, most SLT systems have well-developed standard evaluation metrics, both automatic and semiautomatic, that are applicable across all languages. In other words, for most technologies, as far as measurement of function performance is concerned, standard evaluation metrics already exist and are applicable to Indian languages and context. However, certain aspects of evaluation have not received sufficient attention and/or some of these aspects have been recognized only recently by the community. Hence, there are no standard metrics available, and it is believed that these are important gaps that should be bridged. These cases are listed below.

- Fairness/transparency/explainability:

  As discussed in Section 4.1, fairness, transparency, and explainability are important aspects of any deployed system. Language can easily be used to directly or indirectly express ideas or statements that are biased or ambiguous. For instance, the output of a translator can have gender bias, when the source language pronoun is gender neutral, but the target has gendered pronouns. Yet another aspect of biased treatment can spring from the unequal performance of a machine across languages, user groups, or regions. Furthermore, deep-learning-based systems are often nontransparent, and it is difficult to explain the decisions made by such systems. Therefore, it is of interest to all the stakeholders—users, developers, and policymakers that deployed systems must be:

- Evaluated on its fairness, transparency, and explainability.
- The system provides some minimal fairness guarantees.

Currently, there are no standard metrics to characterize and evaluate the types and extent of bias and nontransparency of a system. The committee believes that this is a serious gap that needs to be addressed earnestly.

- Demography-sensitive usability metrics.

Most evaluation metrics do not consider the fact that different user groups might have different needs. In other words, the present metrics are insensitive to demography-specific needs. This could be problematic. For instance, let us say an FAQ chatbot deployed on a government portal to answer questions regarding certain policy documents might receive favorable assessment when evaluated using a set of metrics, such as retrieval accuracy or communication fluency. However, if this chatbot is to be used by semiliterate rural users, the chatbot might be completely ineffective because of its use of highly technical language. Similarly, translation tools should be evaluated keeping in mind the target user. Translations targeted at children versus adults should be evaluated differently. While most developers and policymakers understand these issues, none of the metrics directly address these issues. Instead, it is assumed that the evaluation of domain-specific test data or on specific user groups will take care of the problem. However, this poses two problems:

- It is left to the technology builder to decide on the metrics or datasets for the domain or target user groups, which in turn, leads to variability in evaluation processes leading to incomparable performance reports.
- An off-the-shelf tool or model or dataset may not be equally useful for building technology for a specific user group. This is because the evaluation reports are user-agnostic, which in effect means it is based on certain assumptions of the user demographics.[15]

User demography can be modeled as an interaction of several axes such as age, gender, level of education/literacy, familiarity with language technology, socio-economic class, physical and cognitive abilities, etc. Ideally, evaluation metrics should be able to give a demography-specific performance,

[15] This is also related to the issue of fairness in the following sense: Often the evaluation is carried out keeping in mind a fluent and educated user, typically from urban areas and mid to upper socio-economic groups.

and similarly, testbenches and models should also be standardized to reflect user-demography-specific applicability.

- Accessibility:

  Yet another extremely important usability criterion is that of accessibility. Ideally, there should be standardized evaluation metrics for measuring the accessibility of a system. While many such standard metrics exist in the space of physical accessibility (see the chapter on accessibility for details), there are gaps in evaluation standards for measuring the neuro-cognitive accessibility of the SLT systems (see Dalton [68], Rapp et al. [69], and Motti [70] for accessibility challenges for neurodiverse users). For instance, conditions such as dyslexia, cerebral palsy, attention deficit disorder, and Alzheimer's disease can create severe barriers for a user in accessing a standard SLT system. There are not many readily available evaluation metrics to measure the neurodiverse accessibility of SLT systems.

- Phenomenon-specific metrics:

  The committee also identified certain language-specific phenomena that calls for special attention while designing evaluation metrics.

  - Code-mixing or code-switching refers to the mixing of more than one language in the same conversation or utterance and is extremely common in multilingual societies. Most studies on code-mixing assume that the evaluation of code-mixed datasets is enough to estimate the performance of a system on code-mixing [64]. However, the nature and complexity of code-mixing across these datasets are variable, which leads to the widely variable performance of the systems across datasets. Standardization of evaluation metrics, therefore, is necessary to measure the performance of the systems at a more fine-grained level. One suggestion has been to measure code-switching points [65]. This is a phenomenon where it is apparent that more standard metrics of evaluation would be useful.
  - Romanization refers to the presentation of Indic language text in roman script. This is extremely common in social media and other user-generated content [67]. In many cases, multiple scripts are mixed in the same piece of text (Sequeira et al., 2015). While evaluation metrics exist for the normalization and transliteration of romanized text, there are no standard metrics for the evaluation of SLT models on romanized or mixed texts. It is not clear currently

whether separate metrics are required, but as in the case of code-mixing, in this case too, it is likely that standard metrics are not sufficiently nuanced and fine-grained to capture the deficiencies of a model for romanized or mixed-script texts.

- Language models and language generators are used across many SLT tasks. Off-the-shelf language models such as BERT [1], RoBERTa (Liu et al., 2020), Turing,[16] Z-code,[17] T5,[18] and their multilingual counterparts are useful for a variety of language processing tasks. Language generators such as GPT3 [45] and its precursors are similarly used in many systems. It is well understood that intrinsic evaluation of LMs through perplexity is not ideal as it weakly correlates to the accuracy of the system that is built on top of the LM. Instead, the popular methodology is to evaluate the LMs on a bunch of end-tasks, which are presented as popular testbenches such as XTREME [5] and XGLUE [4]. It is important to understand that the testbenches are largely ad hoc and have many flaws or gaps. Furthermore, the current method of averaging performance across tasks and languages for comparing models and building leaderboards has also been criticized [62]. Yet another criticism of these models is that they have representational biases [3]. An important gap in evaluation standards today is to come up with a set of comprehensive and reliable metrics to measure the usefulness of the LMs across tasks, domains, and languages. It is also important to be able to measure and quantify their biases.

There are also specific tasks and domains of application, where there is a need for standard evaluation metrics. Some of these are discussed in the specific chapters on speech, text, and script processing. Here, a few more domain-specific examples are highlighted. This is not an exhaustive list and is meant to serve as an exemplar for such gaps.

- Evaluation standards for offensive and biased content for chatbots. As chatbots become ubiquitous, they must be evaluated for the potential to generate offensive and biased content which, in turn, can potentially harm the user [63]. What is offensive and/or biased is often culture-specific, and therefore, domain and culture cannot be ignored while coming up with such evaluation standards.
- In information retrieval, measuring the marginal relevance of pages is still a challenge. There is no universally accepted metric for the evaluation of marginal relevance.

[16] Turing-NLG: A 17-billion-parameter language model by Microsoft - Microsoft Research
[17] A holistic representation toward integrative AI - Microsoft Research
[18] GitHub - google-research/multilingual-t5

- Processing of texts in biomedical and legal domains has received a lot of attention. The standard metrics of evaluation are often inadequate for these domains. Special metrics have been proposed by researchers; however, there is a need for standardization.
- Indic languages are morphologically productive, and decades of research have addressed the issues of morphological analysis and generation. Nevertheless, the metrics for the evaluation of these systems are ad hoc and variable. Standardization of morphological processing evaluation metrics is important and recommended.

6.2. GAPS IN EVALUATION DATASETS

Datasets created for training are almost always useful and used for testing. Therefore, the datasets described in the text and speech processing sections also serve a dual purpose as evaluation benchmarks. Alongside, there are also datasets that have been created specifically as an evaluation testbench, and at times, several such datasets are combined to form a comprehensive test set for a task or domain or phenomenon. Here, the authors do not intend to survey and make a comprehensive list of available datasets for Indic languages. Instead, they refer to the following websites and portals that already list a variety of datasets created by different agencies.

- The data distribution portal by LDC-IL: भारतीय भाषा भाषावैज्ञानिक डाटा संकाय | Linguistic Data Consortium for Indian languages (ldcil.org).
- Datasets created by the Government of India through www.tdil-dc.in.
- Language Resource Centre (cdac.in)—datasets created by CDAC.
- AI4Bharat IndicNLP portal—a crowdsourced catalog of Indian language resources that are publicly available: https://github.com/AI4Bharat/indicnlp_catalog.

Through analysis of these portals, as well as through interviews,[19] the authors arrived at the following observations:

- Unlabeled text corpus is available in around 20+ Indic languages in moderate to large volume.
- Labeled data for testing single sequence classification problems, such as sentiment analysis, is available for around ten Indic languages.
- Data for sequence labeling and structured prediction is available in around five or so Indic languages.
- There is very little data (only adequate for testing) available for high-level tasks such as multisentence

[19] Many of these points came up during a candid discussion with Prof Mitesh Khapra, Dr. Anoop Kunchakuttan, and Prof Pratyush Kumar from AI4Bharat.

classification. For instance, there is only one test set for natural language inferencing.

- There are some datasets on code-mixing (tasks—language identification, POS tagging, parsing, question answering, and sentiment) in English–Hindi and to a lesser extent in code-mixing between other Indian languages.
- Transcribed speech datasets are available in around ten Indic languages, and large-scale datasets (100+ h) are available in only a few Indic languages.

Note that many of these datasets are not available freely or publicly. And, even for the publicly available datasets, they may not have a license that allows commercial usage. Thus, the following four primary gaps in evaluation datasets for Indic languages include:

1) High-level tasks such as NLI, dialog intent, summarization, and question-answering.
2) Domain-specific datasets such as in biomedical or legal domains.
3) Multimodal datasets, such as for image captioning.
4) Datasets in Indic languages beyond the top ten (by population) for basic tasks as well as all the above.

6.3. GAPS IN EVALUATION PROCESSES

Our interaction with industry experts leads them to believe that there is very little standardization of evaluation processes for SLT systems. Most companies follow their proprietary evaluation processes, which often rely on a hybrid approach with five distinct phases:

1) Evaluation of publicly available datasets using standard metrics (whenever evaluation).
2) Evaluation of domain-specific user data created or collected by the company.
3) User evaluation before deployment.
4) Evaluation under tentative deployment (such as A/B testing).
5) Continued evaluation (both qualitative and quantitative) for systems under production and use.

Clearly, stage 1 relies completely on publicly available standards of evaluation; stage 2 might utilize standard metrics and processes, but not data; stage 3 rarely uses standardized processes; and stages 4 and 5 follow proprietary processes and metrics. One of the issues that came up during their interviews with industry experts is that measurement of success for stages 4 and 5 is often quite obvious and technology independent.

For instance, daily and monthly active users, and direct feedback from users (both quantitative and qualitative). Often, there is very little correlation between standard metrics and these latter metrics of success. This is not surprising as user engagement depends on several factors beyond the model's accuracy. However, it would be interesting to study whether it is possible to design metrics that have better correlations with the practical measures of success.

Ribeiro et al. [2] have proposed a checklisting approach to testing of SLT systems inspired by software testing approaches. They express a fundamental concern regarding static testbench or test set-driven approach to evaluation. It is often observed that systems or methods quickly adapt to training datasets and perform well on standard testbenches. However, that does not mean that the new systems are fundamentally better than older ones in processing language because they might have simply learned better correlations in the training data, or even worse, overfitted to the data. Instead, they propose one should test a system for a set of fundamental capabilities, which could be predefined for a system or task. For instance, a sentiment analyzer must have the ability to process negation or apply temporal reasoning or world knowledge when applicable.

In this approach, a checklist is created for every task which is organized into capabilities. Each capability has a bunch of templates that test the capability; the templates, in turn, are used to generate test examples. Checklisting provides a novel and insightful way to evaluate the SLT system. The committee believes that checklists for a multitude of SLT tasks for a set of Indian languages will improve the state of evaluation of Indic SLT. These will serve as simple yet informative testbenches that complement the current set of evaluation datasets and testbenches.

6.4. CONCLUSION

Evaluation standards are extremely important and useful for the development of SLT in a language. Apart from traditional functional metrics of evaluation on static testbenches, there are alternative methodologies as well as dimensions for evaluation. Several of these aspects, such as fairness and bias estimation of a system, and checklisting-based testing of SLT are recent advances in the field. These are expected to mature over the coming years.

In the context of Indic language SLT, apart from the dearth of evaluation datasets, there are also phenomena and demography-specific evaluation metrics that require special attention. Many of these cases, such as code-mixing and romanization, extend far beyond Indic languages. Standardization of evaluation of practices paying due attention to these phenomena will have far-reaching benefits worldwide.

7. REFERENCES

The following sources have been referenced within this paper or may be useful for additional reading:

[1] Devlin, J., et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” Oct. 2018, *arXiv:1810.04805*.

[2] Ribeiro, M. T., et al., “Beyond accuracy: Behavioral testing of NLP models with CheckList,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–11.

[3] Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of ‘Bias’ in NLP,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–23.

[4] Liang, Y., et al., “XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation,” 2020, *arXiv:2004.01401*.

[5] Hu, J., et al., “XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization,” 2020, *arXiv:2003.11080*.

[6] Ananthakrishnan, R., P. Bhattacharyya, M. Sasikumar, and R. M. Shah, “Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU,” in *Proc. ICON,* Hyderabad, India, Jan. 2007.

[7] White, J. S., “Approaches to Black Box MT Evaluation,” in *Proc. Mach. Transl. Summit V*. 1995, pp. 1–10.

[8] Naskar, S. K., et al., “A framework for diagnostic evaluation of MT based on linguistic checkpoints,” in *Proc. 13th Mach. Transl. Summit Asia–Pacific Assoc. Mach. Transl. (AAMT)*, Xiamen, China, Sep. 2011, pp. 529–536.

[9] Specia, L., et al., “Predicting machine translation adequacy,” in *Proc. Mach. Transl. Summit*, vol. 13, 2011, pp. 1–8.

[10] Specia, L., et al., “Findings of the WMT 2018 shared task on quality estimation,” in *Proc. 3rd Conf. Mach. Transl., Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, 2018, pp. 689–709.

[11] Pighin, D., M. González, and L. Màrquez, “The UPC submission to the WMT 2012 shared task on quality estimation,” in *Proc. 7th Workshop Stat. Mach. Transl*. Montreal, QC, Canada: Association for Computational Linguistics, 2012, pp. 127–132.

[12] Papineni, K. S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, 2001, pp. 311–318.

[13] Przybocki, Mark, and Alvin Martin. 2000 NIST Speaker Recognition Evaluation LDC2001S97. Web Download. Philadelphia: Linguistic Data Consortium, 2001.

[14] Doddington, George, *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*, doddington@nist.gov.

[15] Lavie, A., and Denkowski, M.J. The METEOR metric for automatic evaluation of machine translation. Machine Translation 23, 105–115 (2009). https://doi.org/10.1007/s10590-009-9059-4

[16] Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

[17] Han, Aaron Li-Feng, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013. A Description of Tunable Machine Translation Evaluation Systems in WMT13 Metrics Task. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 414–421, Sofia, Bulgaria. Association for Computational Linguistics.

[18] The RIBES (Rank-based Intuitive Bilingual Evaluation Score) from Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh and Hajime Tsukada. 2010. "Automatic Evaluation of Translation Quality for Distant Language Pairs". In Proceedings of EMNLP. https://www.aclweb.org/anthology/D/D10/D10-1092.pdf

[19] Vedantam, Ramakrishna et al. “CIDEr: Consensus-based image description evaluation.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 4566-4575.

[20] Ali, Ahmed, Walid Magdy, and Steve Renals. 2015. Multi-Reference Evaluation for Dialectal Speech Recognition System: A Study for Egyptian ASR. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 118–126, Beijing, China. Association for Computational Linguistics.

[21] Tomás, Jesús, and Casacuberta, Francisco. (2003). A quantitative method for machine translation evaluation. 10.3115/1641396.1641401.

[22] Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers, pp. 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

[23] Panja, Joybrata and Sudip Kumar Naskar. 2018. ITER: Improving Translation Edit Rate through Optimizable Edit Costs. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 746–750, Belgium, Brussels. Association for Computational Linguistics.

[24] Leusch, Gregor, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 241–248, Trento, Italy. Association for Computational Linguistics.

[25] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 505–510, Berlin, Germany. Association for Computational Linguistics.

[26] Stanchev, Peter, Weiyue Wang, and Hermann Ney. 2019. EED: Extended Edit Distance Measure for Machine Translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 514–520, Florence, Italy. Association for Computational Linguistics.

[27] Lo, Chi-kiu. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In Proceedings of the Second Conference on Machine Translation, pp. 589–597, Copenhagen, Denmark. Association for Computational Linguistics., 2017.

[28] Lo, Chi-kiu. YiSi—a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 507–513, Florence, Italy. Association for Computational Linguistics.

[29] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K. (2015). From Word Embeddings To Document Distances. Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research 37:957-966. Available from https://proceedings.mlr.press/v37/kusnerb15.html.

[30] Agirrea, Eneko, Carmen Baneab, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Rada Mihalceab, German Rigaua, Janyce Wiebef. SemEval-2016 Task 1: Semantic Textual Similarity,

Monolingual and Cross-Lingual Evaluation. Available from https://aclanthology.org/S16-1081.pdf

[31] Kauchak, D., R. Barzilay. Paraphrasing for automatic evaluation, Proceedings of the human language technology conference of the NAACL, main conference (2006), pp. 455–462.

[32] Padó, Sebastian & Galley, Michel & Jurafsky, Dan & Manning, Christopher. (2009). Textual entailment features for machine translation evaluation. 37-41. 10.3115/1626431.1626437.

[33] Liu, Ding and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation* and/or *Summarization at the Association for Computational Linguistics Conference* 2005.

[34] Yu, Hui, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. "RED: A Reference Dependency Based MT Evaluation Metric." International Conference on Computational Linguistics (2014).

[35] Quang, L. N. and A. Toral, "Diagnostic evaluation of MT with DELiC4MT," in *Machine Translation Marathon*. Edinburgh, U.K., Sep. 2012.

[36] Balyan, R., S. K. Naskar, A. Toral, and N. Chatterjee, "A diagnostic evaluation approach targeting MT systems for Indian languages," in *Proc. Workshop Mach. Transl. Parsing Indian Lang. (MTPIL)*, Mumbai, Dec. 2012, pp. 61–71.

[37] Toral, A., et al., "DELiC4MT: A tool for diagnostic MT evaluation over user-defined linguistic phenomena," *Prague Bull. Math. Linguistics*, vol. 98, pp. 121–131, Oct. 2012.

[38] Naskar, S. K., et al., "Meta-evaluation of a diagnostic quality metric for machine translation," in *Proc. 14th Mach. Transl. Summit, Nice*, K. Sima'an, M. L. Forcada, D. Grasmick, H. Depraetere, and A. Way, Eds., Sep. 2013, pp. 135–142.

[39] Zhou, M. et al., "Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points," in *Proc. 22nd Int. Conf. Comput. Linguistics (COLING)*, Manchester, U.K., 2008, pp. 1121–1128.

[40] Evaluation of Machine Translation. (Jan. 9, 2014). Available: http://en.wikipedia.org/wiki/Evaluation_of_machine_translation.

[41] Goyal V. and G. S. Lehal, "Evaluation of Hindi to Punjabi machine translation system," 2009, *arXiv:0910.1868*.

[42] Josan, G. S., and G. S. Lehal, “A Punjabi to Hindi machine translation system,” in *Proc. 22nd Int. Conf. Comput. Linguistics, Demonstration Papers*, 2008, pp. 157–160.

[43] Correa, N., “A fine-grained evaluation Framework for machine translation system development,” in *Proc. MT Summit IX*, 2003, pp. 1–8.

[44] Van Slype, G., “Critical study of methods for evaluating the quality of machine translation,” Prepared Commission European Communities Directorate Gen. Sci. Tech. Inf. Inf. Manag. Brussels, Luxembourg, Tech. Rep., BR 19142, 1979.

[45] Brown, T. B., et al., “Language models are few-shot learners,” 2020, *arXiv:2005.14165*.

[46] Peters, M., et al., “Deep contextualized word representations,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 2227–2237.

[47] Yang, Z., et al., “XLNet: Generalized autoregressive pretraining for language understanding,” 2019, *arXiv:1906.08237*.

[48] Lan Z., et al., “ALBERT: A lite BERT for self-supervised learning of language representations,” in *Proc. ICLR*, 2020, pp. 1–17.

[49] Howard, J. and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 328–339.

[50] Liu, Y., et al., “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, *arXiv:1907.11692*.

[51] Herbig, N., et al., “Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation,” *Mach. Transl.*, vol. 33, pp. 1–25, Jun. 2019. Available: https://www.dfki.de/fileadmin/user_upload/import/10278_MT_Journal___Special_Issue_on_Human_Factors_AAM_version.pdf.

[52] Vela,M., et al., “Improving CAT tools in the translation workflow: New approaches and evaluation,” 2019, *arXiv:1908.06140*.

[53] Bach, N. F. Huang, and Y. Al-Onaizan, “Goodness: A method for measuring machine translation confidence,” in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol*., vol. 1. Portland, Oregon: Association for Computational Linguistics, Jun. 2011, pp. 211–219.

[54] Kreutzer, J., S. Schamoni, and S. Riezler, “Quality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation,” in *Proc. 10th Workshop Stat. Mach. Transl*.

Lisboa, Portugal: Association for Computational Linguistics, 2015, pp. 316–322.

[55] Bergstra, J., et al., “Theano: A CPU and GPU math compiler in Python,” in *Proc. 9th Python Sci. Conf.*, 2010, pp. 3–10.

[56] Paetzold, G. and L. Specia, “SimpleNets: Quality estimation with resource-light neural networks,” in *Proc. 1st Conf. Mach. Transl*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 812–818.

[57] Mareček, D., et al., “CUNI experiments for WMT17 metrics task,” in *Proc. 2nd Conf. Mach. Transl.*, vol. 2. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 604–611.

[58] Joshi, N., I. D. H. Mathur, and A. Kumar, “HEval: Yet another human evaluation metric,” *Int. J. Natural Lang. Comput. (IJNLC)*, vol. 2, no. 5, pp. 21–36, 2013.

[59] Sinha, D. and N. Tomer, “Evaluating machine translation evaluation’s BLEU metric for English to Indian language machine translation,” *Int. J. Comput. Sci*., vol. 1, no. 6, pp. 48–58, Aug. 2012.

[60] Shen, L., A. Sarkar, and F. J. Och, “Discriminative Reranking for Machine Translation,” in *Proc. HLT-NAACL*, 2004, pp. 177–184.

[61] Priyanka, J. et al., “Evaluation of automatic Text Visualization Systems: A Case Study,” in *Proc. 5th Int. Conf. Adv. Mach. Learn. Technol. Appl. (AMLTA)*, in Springer Series of Advances in Intelligent Systems and Computing, vol 1141. Singapore: Springer, 2021, pp. 25–37, Doi: 10.1007/978-981-15-3383-9_3

[62] Choudhury, M. and A. Deshpande, “How linguistically fair are multilingual pre-trained language models?” in *Proc. AAAI*, 2021, pp. 12710–12718.

[63] Wambsganss, T., et al., “Ethical design of conversational agents: Towards principles for a value-sensitive design,” in *Proc. 16th Int. Conf. Wirtschaftsinformatik (WI)*, 2021, pp. 1–17.

[64] Sitaram, S., et al., “A survey of code-switched speech and language processing,” 2019, *arXiv:1904.00784*.

[65] Barman, U., “Automatic processing of code-mixed social media content,” Doctoral dissertation, Dublin City University, School of Computing, Dublin, Ireland, 2019.

[66] Sowmya, V. B., et al., “Resource creation for training and testing of transliteration systems for Indian languages,” in *Proc. LREC*, 2010, pp. 1–6.

[67] Sequiera, R., et al., “Overview of FIRE-2015 shared task on mixed script information retrieval,” in *Proc. FIRE Workshops*, vol. 1587, 2015, pp. 19–25.

[68] Dalton, N.S., 2013. Neurodiversity & HCI. In CHI'13 Extended Abstracts on Human Factors in Computing Systems (pp. 2295-2304).

[69] Rapp, A., et al., “Designing mobile technologies for neurodiversity: Challenges and opportunities,” in *Proc. 21st Int. Conf. Hum.-Comput. Interact. Mobile Devices Services*, Oct. 2019, pp. 1–5.

[70] Motti, V. G., “Designing emerging technologies for and with neurodiverse users,” in *Proc. 37th ACM Int. Conf. Design Commun.*, Oct. 2019, pp. 1–10.

[71] Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pp. 74–81, Barcelona, Spain. Association for Computational Linguistics.

[72] Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892, Online. Association for Computational Linguistics.

[73] Kim, H., & Lee, J. H. (2016, August). Recurrent neural network-based translation quality estimation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (pp. 787–792).

[74] Tomer, Neeraj and Deepa Sinha (2012). Evaluating NIST Metric for English to Hindi Language Using ManTra Machine Translation Engine. International Journal of Academy Research Computer Engineering and Technology-IJARCET. 1. 365–369.

[75] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011, August). Natural language processing (almost) from scratch. Journal of machine learning research (pp. 2493–2537).

APPENDIX A

A1. QUESTIONNAIRE PRESENTED TO INDUSTRY EXPERTS AND PRACTITIONERS

Products from your company related to Speech and NLP that you can tell us about. From the following, please tick/mention all that are applicable:

- Machine translation.
- Speech recognition and synthesis.
- Dialogue systems and conversational agents/chatbots.
- Search and information retrieval.
- Information extraction and question answering.
- Text analytics (including text classification, topic and sentiment detection, fake news detection, and so on).
- E-commerce applications and recommendation systems.
- Other.

Evaluation practices in your organization:

For any of the abovementioned products, what testing and evaluation protocols does your company follow? Tick all that apply.

- Evaluation of static testbench.
- Evaluation of dynamic/evolving testbench.
- Manual evaluation.
- A/B testing.
- Continuous evaluation.
- Other.

Do you use standard evaluation metrics (such as BLEU, nDCG, etc.) or have proprietary metrics developed within your organization?

- Yes, we use standard evaluation metrics.
- No, we use proprietary metrics.
- Other.

What fraction of the total product development cycle is spent on evaluation?

How do you decide whether your model is ready for production?

How long does it typically take in your setup for a piece of language technology s/w to do the movement from "lab to land"?

Are there incentives for following good s/w development and documentation processes in your organization?

- Yes, there is a satisfactory and effective incentivization process in place.
- Yes, there are incentivization processes, but I believe we can do better.
- Yes, but they are not quite effective.
- No.
- Other.

Do you rely on publicly available evaluation testbenches and/or standards (such as WMT for MT, XNLI, or GLUE)?

- Yes, we rely only or mainly on public testbenches.
- Yes, whenever available.
- Yes, but we do not depend on public testbenches for making any critical decisions.
- Rarely or never.
- Other.

Gaps and challenges

Do you think the evaluation protocol followed in your company for the offline models truly reflects their online performance?

- Yes, to a good extent.
- Yes, to some extent.
- Rarely.
- Never.
- Other.

What are the primary challenges that you face while evaluating speech or NLP models?

Follow up:

Will you be willing to participate in a 60-min one-on-one conversation with us to discuss some of these points further?

We are interested in understanding critical-to-quality (CTQ) parameters that you use in testing your systems; would you be interested in talking about them?

Information privacy statement for the survey:

The information provided here will not be shared beyond the committee members. However, we will share our consolidated findings in the final report, where we will mention the names of the companies that we have gathered information from, without referring to any piece of information to a company or an individual. If we wish to do so, we will ask for your consent at a later point, which you will have the right to decline.

We need your e-mail to be able to contact you later for an interview, clarifications, or consent. The e-mails will not be used for any other purposes and will be deleted beyond after this survey is over (no later than September 2021).

A2. QUESTIONNAIRE PRESENTED TO INDUSTRY EXPERTS AND PRACTITIONERS

There are several well-known workshops and shared tasks organized by different MT forums such as WMT, WAT, IWSLT, ICON, NIST, etc., to encourage researchers and developers to investigate ways to improve the performance of their systems for diverse languages including morphologically rich languages, resource poor and resource rich languages, etc. Some of the evaluation campaigns that involve Indian languages are mentioned in the following table:

Institute/ Workshop/ Conference	Exercise Name	Date	Links/ References	Description
Task: Machine Translation in WAT				
WAT 2020	Indic Tasks	December 4, 2020	Report	Indic task: Odia–English, Bengali/Hindi/Malayalam/Tamil/Telugu/Marathi/Gujarati–English Multimodal: English–Hindi
WAT 2019	Indic Task	November 3–4, 2019	Report	Indic task: Hindi–English, Tamil–English, and English–Hindi

Institute/ Workshop/ Conference	Exercise Name	Date	Links/ References	Description
WAT 2018	Indic Task	December 3, 2018	Report	Indic languages multilingual tasks: Bengali/Hindi/Malayalam/Tamil/Telugu/Urdu/Sinhalese–English
WAT 2017	Mixed Domain Subtask	November 27, 2017	Report	Mixed domain subtask: Hindi–English and Hindi–Japanese.
WAT 2016	Mixed Domain Subtask	December 12, 2016	Report	Mixed domain subtask: Hindi–English and Hindi–Japanese
Task: Machine Translation in WMT				
WMT 2020	Shared Task: Machine Translation of News	February–July 2020	Report	Tamil–English Parallel data: Wiki Titles v2, WikiMatrix, PMIndia v1, Tanzil v1, The NLPC UOM En-Ta corpus and glossary (v1.0.3), The CVIT corpora (PIB and MKB), and The UFAL EnTam corpus. Monolingual data: News crawl, Common Crawl, and Wiki dumps.
WMT 2019	Shared Task: Machine Translation of News	January–May 2019	Report	Gujarati–English Parallel data: Wiki Titles v1.
WMT 2014	Shared Task: Machine Translation	June 2014	Report	English–Hindi and Hindi–English Parallel data: Wiki Headlines and HindEnCorp.
Task: Similar Language Translation				
WMT 2020	Shared Task: Similar Language Translation	April–November 2020	Report	Hindi–Marathi
WMT 2019	Shared Task: Similar Language Translation	February–June 2019	Report	Hindi–Nepali
Task: Transliteration				
NEWS 2018	Shared Task on Named Entity Transliteration	April–June 2018	Whitepaper, Report	English → Hindi, Tamil, Kannada, and Bangla
NEWS 2016	Shared Task on Transliteration of Named Entities	February–May 2016	Whitepaper, Report	English → Hindi, Tamil, and Kannada

Institute/ Workshop/ Conference	Exercise Name	Date	Links/ References	Description
NEWS 2015	Shared Task on Transliteration of Named Entities	February–May 2015	Whitepaper, Report	English → Hindi, Tamil, and Kannada
NEWS 2012	Shared Task on Machine Transliteration	2012	Whitepaper, Report	English → Hindi, Tamil, and Kannada
NEWS 2011	Shared Task on Machine Transliteration	2011	Whitepaper, Report	English → Hindi, Tamil, and Kannada
NEWS 2010	Shared Task on Machine Transliteration	2010	Whitepaper, Report	English → Hindi, Tamil, Kannada, and Bangla
NEWS 2009	Shared Task on Machine Transliteration	2008	Whitepaper, Report	English → Hindi, Tamil, and Kannada
Task: Metric Task				
WMT 2020	Shared Task: Metrics	September 18–20, 2020	Report	
WMT 2019	Shared Task: Metrics	August 1–2, 2019	Report	
WMT 2018	Shared Task: Metrics	October 31–November 1, 2018	Report	
WMT 2017	Shared Task: Metrics	September 7–8, 2017	Report	
WMT 2016	Shared Task: Metrics	August 11–12, 2016	Report	
WMT 2015	Shared Task: Metrics of Machine Translation Quality	September 17–18, 2015	Report	
WMT 2014	Shared Task: Metrics	June 26–27, 2014	Report	

RAISING THE WORLD'S STANDARDS

3 Park Avenue, New York, NY 10016-5997 USA   http://standards.ieee.org

Tel.+1732-981-0060 Fax+1732-562-1571