

INDUSTRY CONNECTIONS REPORT



INDIAN LANGUAGE RESOURCES— EVALUATION SUBCOMMITTEE REPORT

Authored by

IEEE SA Evaluation Subcommittee

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, *Sarathi.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

Copyright © 2023 by The Institute of Electrical and Electronics Engineers, Inc.

All rights reserved. 18 September 2023. Printed in the United States of America.

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

IEEE is a registered trademark in the U. S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated. All other trademarks are the property of the respective trademark owners.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit <http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

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

ABSTRACT.....	5
1. INTRODUCTION	6
2. SCOPE	7
3. APPROACH.....	8
3.1. STAGE 1: SURVEYING OF EXISTING STANDARDS	8
3.2. STAGE 2: SCOPING.....	8
3.3. STAGE 3: RUBRIC OF EVALUATION STANDARDS	8
3.4. STAGE 4: SPECIFIC USE CASES.....	8
3.5. STAGE 5: INDUSTRY SURVEY.....	9
3.6. STAGE 6: GAP IDENTIFICATION AND RECOMMENDATION	9
4. EVALUATION STANDARDS	9
4.1. STANDARDS FOR MEASURING FUNCTIONAL ACCURACY	14
4.2. EVALUATION STANDARDS ACROSS LANGUAGE, DOMAIN, AND USERS.....	15
4.3. EVALUATION IN INDUSTRY	16
4.4. EVALUATION DATASETS	16
5. CASE STUDY ON MACHINE TRANSLATION.....	17
5.1. CHALLENGES IN MT EVALUATION	18
5.2. MT EVALUATION APPROACHES.....	19
5.2.1. AUTOMATIC EVALUATION	19
5.2.2. HUMAN EVALUATION.....	21
5.3. DATASETS	22
5.4. GAPS IN MT EVALUATION WITH REFERENCE TO INDIAN LANGUAGES.....	23
6. GAPS AND RECOMMENDATIONS	25
6.1. GAPS IN EVALUATION METRICS	25
6.2. GAPS IN EVALUATION DATASETS	29
6.3. GAPS IN EVALUATION PROCESSES	30
6.4. CONCLUSION	31
7. REFERENCES.....	32
APPENDIX A.....	39
A1. QUESTIONNAIRE PRESENTED TO INDUSTRY EXPERTS AND PRACTITIONERS.....	39
A2. QUESTIONNAIRE PRESENTED TO INDUSTRY EXPERTS AND PRACTITIONERS.....	41

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.

¹The 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	<i>Precision, recall, F-score</i> <i>Accuracy, RUC, AUC, specificity-sensitivity</i>
Sequence labeling	POS tagging, named entity recognition and span detection	<i>Accuracy, weighted precision, recall, F-scores.</i>
Regression	Answer/essay grading	<i>Mean squared error, quadratic weighted kappa</i>
Language generation	Transliteration MT Summarization Question-answering Language modeling	<i>Exact match, character-edit distances, phonetically weighted edit distances</i> <i>Recall-oriented understudy for gisting evaluation (ROUGE), bilingual evaluation understudy (BLEU)-1, BLEU-4, exact match, F1 (correct words retrieved) Perplexity</i>