**List of ongoing projects pertaining to TDIL Division**

| S. No. | Document Title (Name of the Ongoing project) | Effective date of the Document / Start date (Administrative Approval date) | Document Definition | Detailed Guidelines the Document (Major Deliverable) | Category of the Document |
|---|---|---|---|---|---|
| 1. | NLTM - Bilingual OCR | 23.03.2020 | R&D Project Document | OCR for Kannada with 96% accuracy. Desktop as well as mobile version. | Un-classified |
| 2. | NLTM - Project Management Unit (PMU) | 23.03.2020 | R&D Project Document | Monitoring of activities at various institutions, two grand-challenges and one workshop in each of the States/UTs. | Un-classified |
| 3. | Web Internationalization and Standardization Initiative' – Phase II. | 15.09.2017 | R&D Project Document | Gap reports covering some of the work items as below will be made available to MeitY: Identify gaps in Readymade Counter Styles, Identify gaps in Pronunciation Gap Analysis, Identify gaps in Common Locale data Repository, Identify gaps in Character Model for the World Wide Web : String Matching, Machine Translation acceptance methodology, Multilingual Dictionaries, Standards for transliteration, Web Payments, Text Layout requirement for Arabic script, Script Grammar, Unicode Technical Report, Internationalization Tag Set (ITS) | Un-classified |
| Sub Projects under **National Language Translation Mission (NLTM) : BHASHINI'** | | | | | |
| 4. | OCRs and Applications in Indian Languages' under the Project titled 'National Language Translation Mission | 03.02.2022 | R&D Project Document | 1. APIs and technology for public use through a web-based delivery platform as envisaged by the NLTM also co-hosted at IIIT Hyderabad. This will be done for all prominent Indian Languages and all popular character recognition modalities (such as printed, handwriting and scene text)<br><br>2. Data, Annotated data, standards, and public release of the datasets for enabling research and development in the broad | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| | (NLTM) : BHASHINI' | | | space of Indian language OCRs. A portion of the data collection and annotation will be carried out as part of the project. (Additional data will be collected in collaboration with DMU or other agencies suggested by NLTM.)<br><br>3. Manpower trained in this specific domain and catalyzing further technology development outside academia in the future.<br><br>4. 96% accuracy for printed documents across 13 scripts, 94% for handwritten documents across 13 scripts, 92% for scene and video text for 13 scripts. | |
| 5. | Indian Language to Indian Language Machine Translation' under the Project titled 'National Language Translation Mission (NLTM) : BHASHINI'' | 15.02.2022 | R&D Project Document | 1. **Translation Technologies**<br><br>i. English-IL and Indian to Indian Language Machine Translation system (11 language pairs [English<->Hindi, English<->Telugu, Hindi <-> Punjabi, Telugu, Urdu, Gujarati, Kannada, Odia, Kashmiri, Sindhi and Dogri], 22 MT systems)<br><br>i. Domain adapted MT systems for chosen domains [Governance; Educational Content in the fields of Science and Technology (Biology, Chemistry, Physics, Environmental Science, Computer Science Engineering, Electrical Engineering, Mechanical Engineering), Law, Economics, Management; Health Care (Consent Forms and Information Sheets, Awareness and Pharma); Judiciary (Case Files); Agriculture and Food Security] and language pairs; for developing efficient NMT systems approximately 70k parallel corpora for each domain in each language pair is required<br><br>2. **Corpora**<br><br>i. Domain specific parallel corpora for 2 domains and domain dictionaries for chosen language pairs for chosen domains (800k parallel corpora)<br><br>i. Annotated data for chosen domains (Total 180K annotated corpora for chosen languages and domains)<br><br>3. **Benchmarks for MT** | Un-classified |

<table>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Technologies<br>i. Benchmark standards and guidelines for ILs<br>i. Benchmark data, Methods and Evaluation for MT and MT tools<br>4. Engineering<br>i. API Gateway for Machine Translation engines and utilities<br>i. Productizing IL-IL MT Technologies<br>5. Workshops and Challenge rounds for building ecosystems consisting of language experts and technology developers</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Collecting datasets and benchmarks for building Indian Language Technology’ under the Project titled ‘National Language Translation Mission (NLTM) : BHASHINI’</td>
<td>17.02.2022</td>
<td>R&D Project Document</td>
<td>
<table>
<tr><th>Task</th><th>Languages</th><th>Pretraining tauto)</th><th>Training (semi-auto)</th><th>Fine-Tuning</th><th>Benchmark</th></tr>
<tr><td>MT<br>(sentences)</td><td>MR<br>LR</td><td>10 billion tokens<br>(combined 22 lang.)</td><td>1,000,000</td><td>100,000<br>50,000</td><td>10,000<br>10,000</td></tr>
<tr><td>ASR<br>(hours)</td><td>MR<br>LR</td><td>1000<br>100</td><td>1000</td><td>100<br>100</td><td>100<br>100</td></tr>
<tr><td>TTS<br>(hours)</td><td>MR<br>LR</td><td>-<br>-</td><td>-<br>-</td><td>40<br>40</td><td></td></tr>
<tr><td>OCR<br>(documents)</td><td>MR<br>LR</td><td>-<br>-</td><td>100,000<br>100,000</td><td></td><td>10,000<br>10,000</td></tr>
</table>
</td>
<td>Un-classified</td>
</tr>
</table>

| | | | | | |
|---|---|---|---|---|---|
| | | | | <table><tr><td>OCR<br><br>(scene)</td><td>MR<br><br>LR</td><td>-<br><br>-</td><td>100,000<br>100,000</td><td></td><td>10,000<br><br>10,000</td></tr><tr><td>SA<br><br>(sentences)</td><td>MR<br><br>LR</td><td>l0billiontokens<br><br>(combined 22 lang.)</td><td>100,000</td><td>10,000</td><td>10,000<br><br>10,000</td></tr><tr><td>QA<br><br>(questions)</td><td>MR<br><br>LR</td><td>10 billion tokens<br><br>(combined 22 lang.)</td><td>100,000</td><td>10,000</td><td>10,000<br><br>10,000</td></tr><tr><td>NER<br><br>(sentences)</td><td>MR<br><br>LR</td><td>10 billion tokens<br><br>(combined 22 lang.)</td><td>100,000</td><td>10,000</td><td>10,000<br><br>10,000</td></tr></table><br>Table 1: List of deliverables containing different types of data for each of the five fundamental technology blocks. MR stands for mid-resource languages and includes (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, Urdu). LR stands for low-resource languages and includes (Bodo, Dogri, Kashmiri, Korikani, Maithili, Manipuri, Nepali, Sanskrit, Santali, Sindhi). Note that only the fine-tuning data and benchmark data will be manually created/curated/verified. All other data (pretraining and training) will be automatically curated from the web. Further all data curated from the web will be released with rules identified by NLTM. | |
| 7. | Language Communicator Tool for End Users' under the Project | 14.02.2022 | R&D Project Document | A. Language Communicator Tool: Hindi - Tamil and English<br>a. An API for authoring tool for representing the semantic representation of any language (in this project for Hindi)<br>b. An aggregate of multilingual | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| | titled 'National Language Translation Mission (NLTM) : BHASHINI' | | | generator platform for Tamil and English generators<br>B. System Description paper, USR Guidelines | |
| 8. | 'English to Indian Language [Hindi, Marathi, Gujarati, Odia, Kannada & Malayalam] and vice versa Machine Translation system' under the project titled 'National Language Translation Mission (NLTM):BHASHINI' | 24.02.2022 | R&D Project Document | The outcome would be a text-to-text Machine Translation system from English to Hindi, Marathi, Gujarati, Odia, Kannada and Malayalam languages and vice versa<br><br>Machine Translation solutions as API/ REST services will be used for further integration to different language-related projects and research works.<br><br>Models developed will be available as web REST service implementation ULCA open API | Un-classified |
| 9. | Discourse Integrated Dravidian Language to Dravidian Language Machine Translation (DL-DiscoMT) under the Project titled 'National Language Translation Mission (NLTM): Bhashini. | 02.03.2022 | R&D Project Document | • A platform for handling Discourse and Conversation<br>• A text to text Machine Translation system from Hindi to Tamil, Tamil to Hindi, Kannada, Malayalam and Telugu Bi-direction systems. Incorporating discourse information in NMT and Sampark.<br>• Leaderboard platform for Evaluation<br>• Machine Translation solution as API/services which can be used for integrating with SSMT systems and by end users. | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| 10. | 'Speech technologies in Indian languages' under the Project titled 'National Language Translation Mission (NLTM): BHASHINI' | 18.02.2022 | R&D Project Document | Setting up standards for data collection, curation, archival, using best practices and benchmarks adapted for the Indian language.<br>1) ASR<br>• ASR systems in Indian English, Tamil Hindi, Telugu, Bengali, Gujarati, Marathi, Assamese, Kannada, Malayalam, Odia, Punjabi (Tonal language), Bodo (Low Resource language) and Manipuri<br>• Total ASR Corpus size for above languages: 30000 hours<br>• 8,000 hours of NPTEL Indian English Technical data curation.<br>2) TTS<br>• TTS systems in Hindi, Tamil, Indian English, Marathi, Bengali, Malayalam, Telugu, Assamese, Kannada, Gujarati, Odia, Rajasthani, Bodo, Manipuri, Urdu, Punjabi, Kashmiri, Konkani<br>• Total TTS Corpus size for above languages: 1360 hours<br>• Develop voice search and voice assistant in Indian English and Hindi. | Un-classified |
| 11. | 'Speech Datasets and Models for Tibeto-Burman Languages (SpeeD-TB)' under the Project titled 'National Language Translation Mission (NLTM) : BHASHINI' | 22.02.2022 | R&D Project Document | The aims and objectives and complete deliverables of the project are as listed below -<br>• To build a transcribed speech dataset of approximately 200 hours each in 6 Tibeto-Burman languages - Bodo (mainly spoken in Assam), Meetei (mainly spoken in Manipur), Chokri (mainly spoken in Nagaland), Kok Borok (mainly spoken in Tripura), Nyishi (mainly spoken in Arunachal Pradesh) and Toto (mainly spoken in West Bengal)<br>• To develop a phone set for each of the languages under study.<br>• To build a language model for the languages under consideration | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| | | | | here.<br>• To build a baseline ASR system for each of the above languages.<br>• To make the dataset and pre-trained and fine-tuned models publicly available through Bhashini / ULCA and also other platforms and sources including GitHub and other appropriate repositories and server under CC-By 4.0 license (for dataset) and AGPL v3 (for the model). | |
| 12. | 'An Interpretable Unified Framework for Text-to-Text Translation among Indian Languages using Sanskrit-based Interlingua Representation' under the Project titled 'National Language Translation Mission (NLTM) : BHASHINI' | 25.03.2022 | R&D Project Document | **T2T translation systems** (Language pairs corresponding to Sanskrit, Hindi, Kannada): The interlingua-based translation models for improved interpretability and faithfulness: API and Web-interface. The models are expected to be more accurate than the available open-source models at ULCA platform, or the Indic-trans platform.<br><br>**Linguistically rich annotated data for the 3 languages**: 40k sentences per language, which will be annotated as per the interlingua annotation scheme with the help of the available morphology tools. The data will be released under CC-BY 4.0 license and can be used for any purpose by all on ULCA. | Un-classified |
| 13. | 'VIDYAAPATI: Bidirectional Machine Translation Involving Bengali, Konkani, Maithili, Marathi, and Hindi' under the Project titled | 30.03.2022 | R&D Project Document | • Bidirectional MT system: Hindi - Bengali, Konkani, Maithili, Marathi.<br>• Mobile App, Web-service, and APIs of the MT systems.<br>• Linguistic Resources:<br>  • Domain-wise size:<br>  • Governance and Policy including Judiciary: 50%<br>  • Education: 30%<br>  • Rest (Science and Technology, Healthcare, Agriculture, Climate, Tourism, | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| | 'National Language Translation Mission (NLTM) : BHASHINI' | | | etc.): 20%<br>• Parallel corpora for each language pair (Hindi - X, where X is one of Bengali, Konkani, Maithili, Marathi) of approximate size 25K parallel sentences will be created. (This is as per MEITY's instruction; 10% of the data will be created by the consortium and 90% by the DMU)<br>• MW, NE, and POS tagged corpus of approximate size 25K sentences of each language among Bengali, Konkani, Maithili, and Marathi.<br>Open-source: code, data, and models will be available to the community for development and utilization. The source code will be released under the license AGPLv3 or Mozilla-v2. The Data will be released under CC-BY 4.0 license. The data and the models will be uploaded to the ULCA.<br>Evaluation metrics and framework.<br>Deployment strategy in language technology | |
| 14. | 'ISHAAN: A System for Bidirectional Machine Translation Between 1) English and Assamese, Bodo, Manipuri, Nepali 2) Manipuri and Hindi 3) Assamese and Bodo' under | 30.03.2022 | R&D Project Document | • Bidirectional MT systems:<br>• English - Assamese, Bodo, Manipuri, Nepali<br>• Hindi - Manipuri<br>• Assamese - Bodo<br>Mobile App, Web-service, and APIs of the MT systems.<br>Linguistic Resources:<br>• Parallel corpora for each language pair (English - 4 North-East Indian Languages) (Approximately 25K parallel sentences for English-X, Hindi-Manipuri, and Assamese-Bodo, where X is one of | Un-classified |

| | | | | | |
|---|---|---|---|---|---|
| | the Project titled 'National Language Translation Mission (NLTM) : BHASHINI' | | | Assamese, Bodo, Manipuri, Nepali)<br>• Domain-wise size:<br>  • Governance and Policy including Judiciary: 50%<br>  • Education: 30%<br>  • Rest (Science & Technology, Healthcare, Agriculture, Climate, Tourism, etc.): 20%<br>• Multiwords (MW), Named Entity (NE) and Part-of-speech (POS) tagged corpus of size approximately 25K sentences for each North-East Indian language<br>Open-source: code, data, and models will be available to the community for development and utilization. The source code will be released under the license AGPLv3 or Mozilla-v2. The Data will be released under CC-BY 4.0 license. The data and the models will be uploaded to the ULCA.<br>Evaluation metrics and framework.<br>Deployment strategy in language technology. | |