Translation Benchmark Guidelines

AI4Bharat

November 2022

Task 1: Source sentence verification, correction and domain classification

For this task, we have 2 different subsets of data:

1. Wikipedia Data
2. Other sources

There is a possibility that sentences might have inconsistencies in the structure or grammar as the data for different subsets have been mined from the web. Furthermore, the domain of the sentence is automatically labeled according to the domain of the web url the content was extracted from. However, this might not always be a correct assumption in all cases due to the presence of cross-reference urls to other sources.

Therefore, we expect the task 1 to be a blend of different subtasks:

1. Source Sentence Verification & Correction
    a. We want all the sentences verified and corrected in cases where inconsistency is detected rather than eliminating those sentences as we need to ensure additional constraints in terms of length and domain diversity.
    b. Please note that no sentence should be “marked as corrupt and discarded”
2. Domain Verification
    a. In this subtask, we want the annotators to select the most appropriate domain from the available options given the source sentence and context.
    b. We explicitly avoid exposing the automatic labels to the annotators to avoid any potential confounding biases.
    c. In the later stages, we will assess the agreement between the annotated labels and the automatic labels.

d. The source sentences which have direct match will be directly forward for further tasks.
e. We will again send the remaining source sentences for a review exposing the annotated and automatic labels where the annotators have to retain the most appropriate label suitable for source sentence and context.

Task 2: Translating the verified source sentences

Below we describe the guidelines to be used while translating sentences from source language. These guidelines are partly inspired from similar guidelines prepared by LDC for the BOLT Chinese-English translation task.

2.1.1. General Principles

2.1.1.1. The translation in the target language must be faithful to the text in the source language in terms of both meaning and style. The translation should mirror the original meaning as much as possible while preserving grammaticality, fluency, and naturalness.

2.1.1.2. To the extent possible, the translation should have the same speaking style, tone or register as the source. For example, if the source is polite, the translation should maintain the same level of politeness. If the source is rude, excited, or angry, the translation should convey the same tone.

2.1.1.3. The translation should contain the exact meaning conveyed in the source text and should neither add nor delete information. For instance, if the original text uses Modi to refer to Honorable Prime Minister Narendra Modi, the translation should not be rendered as Prime Minister Modi, Narendra Modi, etc. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.

2.1.1.4. All sentences should be spell checked and reviewed for typographical errors before submission.

2.1.1.5. While writing sentences in Indian languages, the official native script of the language should be used as mentioned below:

2.1.1.5.1. Bangla script for Bengali, Assamese

2.1.1.5.2. Devanagari script for Bodo, Dogri, Hindi, Konkani, Maithili, Marathi, Nepali, Sanskrit, Sindhi

2.1.1.5.3. Gujarati script for Gujarati

2.1.1.5.4. Gurumukhi script for Punjabi

2.1.1.5.5. Kannada script for Kannada

2.1.1.5.6. Malayalam script for Malayalam

2.1.1.5.7. Meitei Mayek script for Manipuri

2.1.1.5.8. Odia script for Odia

2.1.1.5.9. Ol Chiki script for Santali

2.1.1.5.10. Perso-Arabic script for Kashmiri, Urdu

2.1.1.5.11. Tamil script for Tamil

2.1.1.5.12. Telugu script for Telugu

2.1.2. Named Entities

2.1.2.1. Named entities in English which have a well accepted conventional translation in the regional language should be translated using this conventional translation. For example, Indian Institute of Technology would be translated as “भारतीय प्रौद्योगिकी संस्थान” in Hindi.

2.1.2.2. If a well accepted conventional translation of the English named entity does not exist in the target language, then the named entity should be transliterated. For example, “Pope Francis” should be translated as “पोप फ्रान्सिस” in Hindi.

2.1.2.3. In all cases, avoid inventing translations of named entities in the target language if they do not exist already. Use transliteration instead.

2.1.2.4. The above rules are language specific, and it is possible that an English named entity gets translated in one Indian language and transliterated in another. The key deciding factor is the presence or absence of a well accepted conventional translation of that named entity in that Indian language.

2.1.3. Code Mixing and Borrowing

2.1.3.1. In everyday usage it is common to use code-mixing (e.g., in spoken conversations, English terms are commonly mixed with the native language by many native speakers in India). Such code-mixing is acceptable while translating informal content, such as everyday conversations and voice commands. However, such code-mixing should be avoided while translating formal content.

2.1.3.2. Similarly, many words have been borrowed from English into Indian languages and are now nativised (e.g., train, computer, internet, etc). The translators may use such code mixed and borrowed terminology in the translation if it is a more well accepted term in the target language than a pure translation in the target language (e.g., we refer to “संगणक” as the pure translation of computer as opposed to “कम्प्यूटर” which is also well accepted) . However, this may not be imposed very strictly as some variety in the corpus is also desired (e.g., some sentences which use the pure translation “संगणक” and some which use the borrowed word “कम्प्यूटर”). Note that this is applicable to both formal as well as informal content.

2.1.4. Errors in the source sentence

2.1.4.1. Factual errors in the source sentence should be retained as it is. For example, if the source sentence says “Ranveer Singh and Alia Bhatt starrer Brahmastra will release in theaters today” then the translation should also contain this factual error and not correct it to Ranbir Kapoor. Such factual errors may also be present in descriptions of historical events (e.g., dates may be incorrect or alternative versions of events may exist). Similarly, such factual errors may also be present in scientific theories which are disputed/controversial. In all such cases, the translators should produce a translation which is faithful to the given source sentence.

2.1.4.2. Spelling mistakes in the source sentence should be corrected. If the source sentence has severe grammatical errors then it should be discarded. If the source sentence belongs to formal content and has minor grammatical errors then such errors should be corrected. If the source sentence belongs to informal content (e.g., everyday conversations) and contains minor grammatical errors then such errors may be retained as it is as they are a reflection of everyday usage of the language.

2.1.5. Numbers and Units

2.1.5.1. Numbers in the translation should either be spelled out in full or written as digits, according to how they appear in the source text.

2.1.5.2. It is acceptable to use English numerals instead of their equivalents in the regional language. However, we leave this choice to the language experts with the understanding that this choice should be consistent across sentences (i.e., either use English digits in all sentences or regional digits in all sentences).

2.1.5.3. Roman numerals in English should be retained as it is in the target language.

2.1.5.4. Big numbers (upto million), should be translated using the conventions of the target language. For example, 700 million should be translated as 70 करोड़ as opposed to 700 मिलियन. Very large numbers, such as billion and trillion could be translated as it is (e.g., 7 billion should be translated as 7 बिलियन). However, alternative translations containing regularly used terms in the target language are also acceptable if they are popular and well accepted in the target language. This would ensure some diversity in the corpus.

2.1.5.5. For units of measurement that may differ between English and Indian languages (for example "miles" v/s "kilometers" or “gallons” v/s “liters”), the translators should produce a translation which retains the units as mentioned in the source sentence. For example, “3 miles” should be translated as “3 मील” and not “4.8 किलोमीटर” (even though kilometer is a more popular/acceptable unit in India).

2.1.6. Dates

2.1.6.1. Dates in the translation should either be spelled out or written as digits, according to how they appear in the source text. For example, 17 January 2022, would be translated as “17 जनवरी 2022” and not “17-01-2022”.

2.1.6.2. Dates written in numeric format (mm-dd-yyyy, dd-mm-yyyy, dd-mm-yy, etc) should be translated as they occur in the source sentence. For example, the English date “01-09-2022” should simply be translated as “01-09-2022” in Hindi.

2.1.6.3. The year should be translated using 4 digits or 2 digits depending on how it appears in the source sentence. For example, “01-09-21” should be translated as “01-09-21” in Hindi and not as “01-09-2021” (even though the latter translation has no ambiguity).

2.1.7. Technical terms

2.1.7.1. For translating technical terms, translators should refer to the class 1 to class 12 books in the native language provided by NCERT, NIOS of state boards. The translators should refer to the translation dictionaries prepared by the Commission for Scientific and Technical Terminology (CSTT) and TDIL for different domains (Science, Engineering/Technology, Medical Science, Humanities, Social Sciences, Agricultural Science, Veterinary Science).

2.1.7.2. If a technical term does not have a native translation in NCERT, NIOS or other state board textbooks or in CSTT dictionaries or if the translation in the CSTT dictionary is too archaic/academic, then it should be transliterated into the target language. Note that in many languages, such as Sanskrit, Santali which have limited Western influence, a large number of terms in English will have to be transliterated. For example, terms like “penalty shootout” do not have a well accepted native translation in Sanskrit. While a Sanskrit term for “penalty shootout” can be coined, it may not sound natural in the context of the sport. Hence, it is acceptable to transliterate such terms.

2.1.7.3. Acronyms in Roman script should be mapped character by character, with periods between mapped characters on the Indian language side. E.g. If the English sentence contains CSTT, the Hindi side equivalent would be सी. एस. टी. टी.
Acronyms like NASA and NREGA which have an accepted convention in the target language those should be written as नरेगा and नासा and NOT एन.आर.ई.जी.ए. and एन.ए.एस.ए.

2.1.7.4. Acronyms in Roman script should be mapped character by character, with periods between mapped characters on the Indian language side. However, if the form without periods is an acceptable and widely used loan word in the target language, you can drop the period. E.g. If the English sentence contains pH, the Hindi side equivalent would be पीएच.

Task 3: Conversation Translation

Intro & Issues

The project involves translation of spoken text, with content mostly taken from functional language scenarios (day-to-day life & regular activities)

- The spoken text thus ranges from casual through informal to semi-formal, and covers both face-to-face and telephonic conversations
- The spoken text, however, is laid out in writing and not audio, and therefore available only for reading and not listening
- This is particularly difficult to handle since we generally tend to translate any text given as a script in a written style, while CT content needs to be rendered in a spoken style
- The variation between written Vs spoken styles is a universal phenomenon, and very pronounced in a few languages like Tamil, giving rise to the diglossia problem (a wide gap between both)
- The variations occur mainly at the following levels
    - Sound/spelling
    - Extent of Transliteration or load-word usage (code-mixing)
    - Phrasing
    - Vocabulary/word-choice
    - Sentence/utterance construction, both partial & overall
- The goal of the translation in CT is to capture and convey the content as a spoken text in spoken style with its necessary features and the right level of formality and tone

Approach & Procedures

Given the nature, goals and challenges, the following method/techniques are suggested, esp. in avoiding written-style translation

- Please use Sense-for-Sense translation
    - Expression structure would vary, sometimes considerably, between spoken Vs written
- Read aloud the source text and imagine listening to it, getting a feel of speech

    - Please imagine a conversation setting where you could hear only (like a radio broadcast), not see (as in TV shows)
    - A thumb rule would be to ‘speak’ the translation & then write it
    - And, to imagine as if we are writing a script for someone to speak
- It is recommended and preferable to pair up as translators A & B and work together, one reading the text and the other translating for half the file and then reversing the roles
- Mark DD-PPP to get the thought process on track and get the right usage, formality & tone
    - DD refers to Domain and Discourse (text type, I.e., dialogue)
    - PPP refers to Purpose, People and Presentation of communication for the content
    - Presentation can be face-to-face or telephonic, etc.
    - It can be marked at the end in the working Excel sheet, as shown in the example below
- MT suggestion will not be available
    - Using any MT is far from desirable since it doesn’t help in spoken language expressions
- Regular feedback by the reviewer showing points for improvement will help refine the quality

DD-PPP: Domain - Education (school)// Discourse - dialogue// Purpose - inquiry about admission procedure// People - parent & school PRO// Presentation - telephonic conversation

General Guidelines

- The conversation is going to range from casual through informal to semi-informal
- Use a semi-rural to semi-urban setting to reduce the proportion of transliteration (code-mixing)

Checklist Points based on Criteria for Spoken Language Utterances

S. No.	Criteria
1.	Sentence Construction - Overall
2.	Sentence Construction - Partial
3.	Word choice
4.	Phrasing
5.	Transliteration Judgments
6.	Spelling (Written style)

Spoken language peculiarities and some recommendations (depends on language):

- Tends to use 'active' voice for the most part (including impersonal active, which means 'the doer' is not explicitly expressed)
- Keep the last round of checking & tweaking/polishing for a CT file to your next session (on the same day or next day) and then finalize it by reading through the entire dialogue flow
- Use a hyphen for suffixing an abbreviation or transliterated word if you suspect ambiguity otherwise
- Redundancy is very well allowed in spoken expressions, where required
- For numerals, keep them as they are, without writing in words, if not required

Note:
In some cases, where the gender information of one of the speaker of the conversation is not explicit in the English version, then:-

1. If the source English sentence has gender-neutral reference / forms, the preferred choice would be to have gender-neutral reference / forms in the translation in the target Indic language.
2. If having gender-neutral reference / forms is not possible / not very natural in the target language, then please consider translating with the assumption of the feminine gender.

Task 4: Review of NMT Benchmark

- In this task, we expect the reviewers to assess the faithfulness and naturalness of the sentences to be in accordance with the norms in your language. We follow the same guidelines as Task 2 in this document.
- Furthermore, we expect the reviewers to directly make the necessary corrections to satisfy the above and accept the sentences without sending it to the annotators for revision.
- If you find a lot of translations done by any specific annotator are not upto standards with the guidelines then please bring this to our notice to take the necessary steps to balance the review workload for the future tasks.