

Guidelines for the Linguistic Data Consortium's Project to Support Automatic Evaluation of Language Translation Technology (Chinese-to-English)

Goal

Our goal is to support the development of automatic means of evaluating translation quality. To this end it is necessary that we have a number of different translations of the same source material. Specifically, we will collect up to ten translations for each of **approximately 100** news stories. This data is provided as an attachment to these guidelines. You have been selected as one of the possible sources to provide these translations. In order to be considered for this work, you must conform to the following guidelines.

The Translation Team

A single translation “team” **must** be used to translate all of the source language data. This team may be:

- 1) A single bilingual translator
- 2) A Chinese dominant bilingual who does initial translation and an English dominant bilingual who proofreads and edits the output of the first translator
- 3) An MT system that does initial translation and a translator who proofreads and edits the output of the translation system
- 4) Either 1), 2), or 3) above, assisted by a translation memory system.
- 5) Some other “team” that we have not anticipated but that we might be willing to entertain

The translation team must not change during translation, and the team must be fully documented. Documentation includes:

- 1) The name (or pseudonym), native language, second languages, age and years of translation experience of the translator(s)
- 2) The order of processing (i.e. the name of the person who performs the first pass, second pass, etc.)
- 3) The name and version number of any translation system or translation memory used
- 4) A description of any additional quality control procedures or other relevant parameters or factors that affect the translation

In some cases, LDC may allow a translation agency to perform more than one translation. To be considered, the agency must propose completely different teams for each of the translations. In this case, the teams must be **completely independent**, with absolutely no communication or resource sharing between teams.

Chinese Source Text

The source data to be translated is **approximately 100** stories comprising a total of approximately 40,000 Chinese characters. Each story has SGML tags added at the beginning and end to aid automatic processing, as follows:

```
<MT_input source_language="Chinese"
doc_ID="NNN">
    {Chinese text to be translated}
</MT_input>
```

In addition, each story is divided into segments, with markers between the segments. These markers are to be preserved in the translation.

English Translation File Format

The English translation of each source story is to be rendered as plain ASCII text, with enclosing SGML tags that preserve the attributes of the original story. In addition, two additional attributes are to be added, namely the target language (English) and a system ID (which serves to identify the translator or translation team and will be assigned by LDC), as follows:

```
<MT_output source_language="Chinese"
target_language="English"
doc_ID="NNN" sys_ID="XXX">
    {English translation}
</MT_output>
```

The translated data is to be organized in the exactly same way as the source data. Each segment of Chinese text should be translated into a corresponding segment of English text. The markers in the source text should be preserved in the English translation. Note that there may be multiple English sentences per segment.

It is recommended that the translator makes a copy of each source file and performs translation directly in the copy by translating the Chinese text into its corresponding English text without altering any segment or paragraph markers. Since most word processors or text editors for Windows automatically add a file extension if “save as” is chosen, an output file with .txt extension will be accepted. But the translator should be consistent in this respect.

Electronic transmission of output translations (as zipped email attachments or ftp) must be used. Paper transmission is not acceptable.

Translation Quality

Translation agencies will use their best practice to produce translations. While we trust that each translation agency has its own mechanism of quality control, we have specific guidelines so that all translations share a common ground. These are:

- 1) The English translation must be faithful to the original Chinese text in terms of meaning and style. The Chinese source text is usually a news story, thus the translation should also be journalistic. The translation should mirror the original meaning as much as possible without sacrificing grammaticality, fluency, and naturalness.
- 2) The translation should be as factual as possible. For example, if the original text uses “Bush” to refer to the US President, the translation should **not** be rendered as “President Bush”, “George W. Bush,” etc. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.
- 3) The translation should also respect the cultural matrix of the original. For example, if the Chinese text uses the phrase “Comrade Jiang Zemin,” the translation should **not** be rendered as “Mr. Jiang Zemin.”

Translation of Proper Names

Proper names should be translated using common practice. This is summarized as follows:

- 1) Whenever a Chinese proper name has an existing conventional translation into English, that translation should be used. For names without an existing translation, Pinyin should be used in most cases. However, some Taiwanese, Hong Kong and overseas Chinese names do not use Pinyin by tradition. For example, the former Taiwanese president should be translated as “Lee Teng-hui,” not “Li Denghui.”

- 2) The order of “last-name first and first-name last” in the source should be preserved. For example, the Chinese president should be “Jiang Zemin”, not “Zemin Jiang.”
- 3) Non-Chinese proper names should be translated as they would be translated into English directly from the original language. This is particularly important for translating Japanese, Korean, and Vietnamese names, and also for non-Han Chinese names such as Tibetan, (Inner) Mongolian, and Uigur names. In the case of an original English name appearing in the Chinese text, the normal English form should be used.
- 4) Lacking preexisting knowledge of how to translate a foreign proper name, the translator should use existing resources (such as information gleaned from the www) to decide on a best translation. Failing this, simply use Pinyin as if the name were a Chinese name.

Names must be translated consistently across all of the documents.

Workflow

Upon approval by the LDC, the first 10 stories should be translated and submitted to LDC for quality assurance before further translation is performed. The documentation of the translation team should also be included. This is to allow the LDC to make sure that the translation is being done according to the LDC’s expected standards. Once this is accomplished, will give the go-ahead to translate the remainder of the first half of the stories to be translated. Agencies that successfully complete this translation task in a timely fashion and following the guidelines will be asked to translate the second half of the stories.

Guidelines

In case these guidelines prove to be unclear, LDC reserves the right to modify them. Agencies will always use the latest version.

Amendment to the Guidelines for the Second Set of Data

The second set of data is now being delivered to you along with this amendment and the questionnaire for collecting translation team information.

There are two subsets for the new data. The first set, compressed in `voa.zip`, has 26 files of broadcast transcripts from the Mandarin service of the Voice of America (VOA). The second set, compressed in `zaobao.zip`, has 27 files, each of which is a news story collected from `www.zaobao.com`, a Chinese news portal based in Singapore. The total Chinese character count of the new data is slightly more than that of the first set of data.

All the files are reformatted in accordance with the translation guidelines. To avoid any file naming confusion, all the files now have the `.txt` extension. The translator should not rename any of the files. Windows operating systems most likely will open them in Notepad by default. If the translator uses a more advanced text editor such as MS Word, be sure to save the files in plain text without changing the file name.

As before, for each translated file, `target_language="English"` and `sys_ID="XXX"` should be inserted into the start tag and the closing tag should be replaced with `</MT_output>`.

Special notes on the VOA files:

- The original VOA files were of human transcripts of VOA Mandarin news broadcasts and were of relatively low quality. Two native speakers of Chinese at the LDC went over each of them and corrected the typos (e.g. an incorrect character for the context) as detected. Some were reapplied punctuations as appropriate. It's by no means that every character is correct and every sentence is grammatical. The translator should exercise their best judgment in interpretation. Due to limitations of the Unix system the LDC uses, certain characters may not be what they are. For example, the character for "rong2" as in the Chinese premier's name uses a different one.
- The VOA files of varying lengths. Most short ones are from the first 5 minute news briefings. There are a few lengthy files, some of which may have non-scripted interviews and thus may be stylistically colloquial. In such cases, the translation should also be colloquial. That conforms to the guidelines.
- The VOA files do not have `-Headline-` tags.

Special notes on the Zaobao files:

- The Zaobao files are very similar to the Xinhua files. The file length is in the range of 350~450 Chinese characters.
- The Zaobao files may have proper names that are very different from those use in the Xinhua files for the same people. Please be aware of this.