Terminology Shared Task @WMT25:  
Data Format and Submission Guidelines

# [Track 1: Sentence-Level Translation](https://drive.google.com/open?id=1cObJDi1MWlKZSJTifxIDUiAfP8bM2eaY)

For each language pair, we have three files (each corresponding to three modes - no/proper/random terminology). The file naming schema is as follows:

{src}{tgt}.{mode}.jsonl, where:

* Src is 2-char lang code of the source language (always en)
* Tgt is 2-char lang code of the target language (en/de/es)
* Mode is 6-char name of the terminology mode: noterm for no terminology, proper and random - for proper and random terminologies, respectively

Each line in the file is a json dictionary, which consists of the following items:

* “en”: str - source sentence
* *For “random” and “proper” modes:*   
  “terms” {str: str} - dictionary of terms (key is source language, value is target language)

### No Terminology (noterm):

**Filename**: ende.noterm.jsonl

**Format of each line:**

{"en": "Request a checkup to perform a health diagnostic and better analyze what was wrong with your data flow.\n"}

### Proper Terminology:

**Filename**: ende.proper.jsonl

**Format of each line:**

{"en": "Request a checkup to perform a health diagnostic and better analyze what was wrong with your data flow.\n", "terms": {"check": "Kontrolle"}}

### Random Terminology:

**Filename**: ende.random.jsonl

**Format of each line:**

{"en": "Request a checkup to perform a health diagnostic and better analyze what was wrong with your data flow.\n", "terms": {"wrong": "falsch"}}

## Submitted files format

You are expected to modify the jsonl files by adding a new item to each line:

* tgt: str - where tgt is a 2-char lang code of the target language, and str is the translation of the sentence.

You should submit each language pair and each mode separately; as for the filenames, you should prepend the initial filename with the name of your system. IMPORTANT: please name your system consistently! If you want to submit multiple systems, please do so in separate submissions!

You should submit all your files in a single (not nested) folder, the same way as the input files are stored.

**Filenames**: {system}.ende.{terminology mode}.jsonl

**Format of each line**: same as in corresponding inputs, with additional key “de” for each line. For example, for “random” mode, it will look like:

{"en": "Request a checkup to perform a health diagnostic and better analyze what was wrong with your data flow.\n", "terms": {"wrong": "falsch"}, "de": "Fordern Sie eine Untersuchung an, um eine Gesundheitsdiagnose durchzuführen und besser zu analysieren, was mit Ihrem Datenfluss falsch war."}

# [Track 2: Document-Level Translation](https://drive.google.com/open?id=1a76_mwOBWc9VVSg9VdCEUqFjafV1vuMU)

Both the naming convention and the file structure are organized slightly differently.

We have two translation directions: English->Traditional Chinese and Traditional Chinese -> English. We also have the same three modes of terminology: noterm, proper, random.

Contrary to sentence-level track, we have several files for each direction, each representing a separate document (5 documents for each direction). Within each document, there are several json lines, representing a paragraph within a document, and (for proper and random modes) a document-level dictionary.

NB:

1. since the dictionary is document level, it is the same within a given file, e.g., for 2015.enzh.proper.jsonl, the dictionary for each line would be identical.
2. The dictionary structure is also richer compared to sentence-level translation: for Traditional Chinese -> English direction, we allow multiple translations (mostly due to having both a full translation and abbreviation, e.g. “中國人民銀行”: [“People’s Bank of China”, “PBoC”]). For English -> Traditional Chinese direction, the same homonymy is represented through having multiple keys corresponding to the same value.   
   Therefore, the dictionary format is str: List[str]. It is like that for both directions for the consistency of format, however, in English -> Traditional Chinese direction, length of the list is restricted to 1.

The file naming schema is as follows:

{year}.{src}{tgt}.{mode}.jsonl, where:

* Year is a 4-digit of the report
* Src is 2-char lang code of the source language
* Tgt is 2-char lang code of the target language
* Mode is 6-char name of the terminology mode: noterm for no terminology, proper and random - for proper and random terminologies, respectively

Each line in the file is a json dictionary, which consists of the following items:

* “en”: str - source sentence
* *For “random” and “proper” modes:*   
  “terms” {str: List[str]} - dictionary of terms (key is source language, value is target language)

### No Terminology (noterm):

**Filename**: 2015.enzh.noterm.jsonl

**Format of each line:**

* Each line has an paragraph (up to 2000 tokens) which is a part of the same report
* Key of the line is the source language code (en or zh)

**Example:**

{"en": "This is an example paragraph from terminology MT"}

### Proper Terminology:

**Filename**: 2015.enzh.proper.jsonl

**Example:**

en->zh:

{"en": "This is an example paragraph for terminology MT.", "terms": {"MT": ["機器翻譯"]}}

zh->en:

{“zh”: “這是術語機器翻譯的示例段落。”, “terms”: {"機器翻譯": [“MT”, “machine translation”]}}

### Random Terminology:

**Filename**: 2015.enzh.random.jsonl

**Example:**

en->zh:

{"en": "This is an example paragraph for terminology MT", "terms": {"this": ["這"]}}

zh->en:

{“zh”: “這是術語機器翻譯的示例段落。”, “terms”: {"這": [“this”, “the”]}}

## Submitted files format

**Filenames**: {system}.{year}.enzh.{terminology mode}.jsonl

**Format of each line**: same as in inputs, with additional key “zh” or “en” for each line. For example:

{“zh”: “這是術語機器翻譯的示例段落。”, “terms”: {"機器翻譯": [“MT”, “machine translation”]}, “en”: “Here is an example paragraph of the term machine translation.”}

# Submission Guidelines

1. **If you have multiple systems, submit them separately.**  
   Further instructions work for a single submitted system.
2. **Store all files for each track in a single non-nested folder.**   
   For track 1, it should be 9 files (3 translation pairs x 3 terminology modes); for track 2 it should be 30 files (10 yearly reports x 3 terminology modes)
3. **Pay attention to file naming and inner structure of the files.**   
   For track 1, exact guidelines are [here](#_ra0954nf48lo), for track 2 - [here](#_6l30br2umd49).
4. **Run validation code.**   
   It can be found [here](https://drive.google.com/file/d/1e1t4xV5HWYbLpleHRitrGTI4z7MCkx8t/view?usp=share_link); all non-valid submissions will be desk-rejected.
5. **Submit through our form** [**here**](https://forms.gle/WSZX9obVCvRUQz6Q7)**.**

Good luck!