NMT training sets and resources

1. WMT (Conference on Machine Translation) Corpora

The WMT corpora are among the most widely used datasets for machine translation research. They are updated annually and include a variety of language pairs and domains.

- **WMT 2021**:
  - **Language Pairs**: English-German, English-Russian, English-Chinese, etc.
  - **Data**: News articles, subtitles, and other text types.
  - **Size**: Varies by language pair, but typically includes millions of sentence pairs.
  - **Source**: WMT 2021
- **WMT 2020**:
  - **Language Pairs**: Similar to WMT 2021.
  - **Data**: News articles, subtitles, etc.
  - **Size**: Varies by language pair.
  - **Source**: WMT 2020

2. Europarl Corpus

The Europarl corpus is a large parallel corpus extracted from the proceedings of the European Parliament. It is widely used for training NMT models due to its high quality and coverage of multiple languages.

- **Language Pairs**: English-French, English-German, English-Spanish, etc.
- **Data**: Official documents and speeches.
- **Size**: Up to 2 million sentence pairs per language pair.
- **Source**: Europarl Corpus

3. UN Parallel Corpus

The UN Parallel Corpus is derived from the United Nations documents and covers a wide range of languages and topics.

- **Language Pairs**: English-Arabic, English-Chinese, English-Russian, etc.
- **Data**: Official UN documents.
- **Size**: Varies by language pair.
- **Source**: UN Parallel Corpus

4. IWSLT (International Workshop on Spoken Language Translation) Corpus

The IWSLT corpus is designed for spoken language translation and includes transcriptions of spoken dialogues and presentations.

- **Language Pairs**: English-French, English-German, English-Chinese, etc.
- **Data**: Spoken dialogues, presentations.
- **Size**: Smaller than WMT but high quality.
- **Source**: IWSLT Corpus

5. Tatoeba Corpus

The Tatoeba corpus is a collection of sentences and their translations contributed by volunteers. It is useful for training models on a wide variety of languages and domains.

- **Language Pairs**: Over 300 languages.
- **Data**: Sentences and their translations.
- **Size**: Varies by language pair.
- **Source**: Tatoeba Corpus

6. MultiUN Corpus

The MultiUN corpus is derived from United Nations documents and is similar to the UN Parallel Corpus but with a focus on multilingual alignment.

- **Language Pairs**: English-Arabic, English-Chinese, English-Russian, etc.
- **Data**: Official UN documents.
- **Size**: Varies by language pair.
- **Source**: MultiUN Corpus

7. News Commentary Corpus

The News Commentary corpus is a collection of news articles translated into multiple languages. It is part of the WMT datasets but is also available separately.

- **Language Pairs**: English-French, English-German, English-Spanish, etc.
- **Data**: News articles.
- **Size**: Varies by language pair.
- **Source**: News Commentary Corpus

8. OpenSubtitles Corpus

The OpenSubtitles corpus is derived from movie subtitles and is useful for training models on conversational language.

- **Language Pairs**: Over 60 languages.
- **Data**: Movie subtitles.
- **Size**: Large, with millions of sentence pairs.
- **Source**: OpenSubtitles Corpus

9. JW300 Corpus

The JW300 corpus is derived from the Jehovah's Witnesses' publications and is useful for training models on religious and formal language.

- **Language Pairs**: Over 100 languages.
- **Data**: Religious texts.
- **Size**: Varies by language pair.
- **Source**: JW300 Corpus

10. OPUS (Open Parallel Corpus)

OPUS is a collection of multiple parallel corpora from various sources, covering a wide range of languages and domains.

- **Language Pairs**: Over 200 languages.
- **Data**: Various types of texts.
- **Size**: Varies by language pair.
- **Source**: OPUS Corpus

Tips for Dataset Preparation

1. **Preprocessing**: Tokenize the text, remove punctuation, and normalize case.
2. **Cleaning**: Remove duplicates, near-duplicates, and noisy data.
3. **Balancing**: Ensure the dataset is balanced across different domains and topics.
4. **Validation**: Use a separate validation set to tune hyperparameters and avoid overfitting.

These datasets provide a solid foundation for training LSTM-based NMT models. Depending on your specific use case, you may also consider domain-specific datasets or augmenting these datasets with additional parallel corpora.