Report on German-English (de-en) Datasets for Neural Machine Translation

---

Overview

This report provides a detailed comparison of various datasets used in Neural Machine Translation (NMT) research for the German-English (de-en) language pair. The datasets covered include OPUS100, Europarl, and the series of WMT datasets from 2014 to 2020. The comparison includes the size of the datasets, key data sources, and the domains of the test sets.

---

1. OPUS100

Dataset Description: OPUS100 is a curated subset of the OPUS collection, containing parallel sentences from various domains. It aims to provide high-quality, large-scale multilingual datasets for NMT.

Approximate Size:

- Approximately 1 million sentence pairs for the German-English language pair.

Key Data Sources:

- **Tatoeba:** Volunteer-contributed sentence translations.
- **TED Talks:** Transcripts and translations of TED Talks.
- **GlobalVoices:** News articles translated by Global Voices.
- **Wikipedia:** Parallel sentences extracted from Wikipedia.
- **Subtitles:** Movie and TV subtitles.

Test Set Domain:

- Mixed domains including news, spoken language, and general text.

---

2. Europarl

Dataset Description: Europarl is a parallel corpus extracted from the proceedings of the European Parliament, covering multiple languages including German and English.

Approximate Size:

- Approximately 2 million sentence pairs for the German-English language pair.

Key Data Sources:

- **European Parliament Proceedings:** Official records of debates and speeches.

Test Set Domain:

- Primarily political and formal language used in parliamentary sessions.

---

3. WMT Datasets (2014-2020)

The WMT datasets are annual datasets released by the Workshop on Machine Translation, which include training, validation, and test sets for multiple language pairs.

WMT14

Approximate Size:

- Approximately 4.5 million sentence pairs.

Key Data Sources:

- **Europarl**
- **Common Crawl**
- **News Commentary**
- **Web Crawled Corpus**

Test Set Domain:

- News

WMT15

Approximate Size:

- Approximately 4.5-5 million sentence pairs.

Key Data Sources:

- Similar to WMT14 with additional cleaned data.

Test Set Domain:

- News
- Biomedical

WMT16

Approximate Size:

- Approximately 5.9 million sentence pairs.

Key Data Sources:

- Expanded web crawled corpus.
- Back-translated data.

Test Set Domain:

- News
- IT

WMT17

Approximate Size:

- Approximately 5.9-6 million sentence pairs.

Key Data Sources:

- Diverse sources including Wikipedia.

Test Set Domain:

- News
- Challenging sets for robustness

WMT18

Approximate Size:

- Approximately 5.85 million sentence pairs.

Key Data Sources:

- Expanded with ParaCrawl.

Test Set Domain:

- News
- Robust tasks

WMT19

Approximate Size:

- Approximately 36 million sentence pairs.

Key Data Sources:

- More comprehensive inclusion of ParaCrawl data.
- Web-crawled data.

Test Set Domain:

- News
- Domain adaptation

WMT20

Approximate Size:

- Approximately 40 million sentence pairs.

Key Data Sources:

- Additional ParaCrawl data.
- Domain-specific data.

Test Set Domain:

- News
- Challenging scenarios for robustness

Summary Table

WMT Year	Approximate Size (Sentence Pairs)	Key Data Sources	Test Set Domain
2014	4.5 million	Europarl, Common Crawl, News Commentary	News
2015	4.5-5 million	Similar to WMT14, additional cleaned data	News, Biomedical
2016	5.9 million	New sources, back-translated data	News, IT
2017	5.9-6 million	Diverse sources including Wikipedia	News, challenging sets
2018	5.85 million	Expanded with ParaCrawl	News, robust tasks
2019	36 million	More web-crawled data, ParaCrawl	News, domain adaptation
2020	40 million	More ParaCrawl, domain-specific data	News, challenging scenarios

Conclusion

The datasets OPUS100, Europarl, and WMT14-20 provide diverse and extensive resources for training and evaluating NMT models for the German-English language pair. The WMT datasets have progressively increased in size and diversity, incorporating new data sources to improve robustness and domain adaptation. Europarl provides a specialized corpus for formal political language, while OPUS100 offers a mix of various domains, making it suitable for broad applications.

These datasets collectively contribute to advancing the state of NMT by providing comprehensive benchmarks and training resources, helping researchers to develop and evaluate more accurate and robust translation models.