Extended Parallel Corpus for Amharic-English Machine Translation

Andargachew Mekonnen Gezmu,
Andreas Nürnberger,
Tesfaye Bayu Bati

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

INTRO

- Amharic, the official language of Ethiopia, is a Semitic language
- Amharic uses a syllabic writing system, Ethiopic
- In Amharic orthography, there is no case difference
- Amharic words are highly inflectional and have a root-pattern morphology
- In this research, we
  - Collected bilingual documents from various sources
  - Aligned sentences by addressing language-specific issues
  - Available at: http://dx.doi.org/10.24352/ub.ovgu-2018-145
  - Trained and evaluated NMT and PBSMT models

DATA SOURCES

1. Newswires
2. Magazines
3. The Bible

RESULTS

Document	Number of sentence pairs
Awake	16491
Watchtower	72512
The Bible	48651
News articles	7710
Total	145364

Dataset	Sentences	English Tokens	Amharic Tokens	English Types	Amharic Types
Test	2500	46154	34689	5842	11644
Validation	2864	53818	39980	6470	13068
Training	140000	2574538	1930220	33589	155824
Total	145364	2674510	2004889	45901	180536

Translation Direction	System	BLEU	BEER	CharacTER
Amharic-to-English	NMT-1K	32.2	0.575	0.536
	NMT-2K	32.2	0.575	0.536
	NMT-4K	32.8	0.577	0.530
	NMT-8K	33.0	0.576	0.527
	NMT-16K	32.9	0.574	0.528
	NMT-32K	32.2	0.570	0.539
	NMT-Word-Based	28.8	0.537	0.588
	SMT-MERT	26.0	0.514	0.629
	SMT-MIRA	23.2	0.494	0.705
English-to-Amharic	NMT-1K	25.5	0.558	0.520
	NMT-2K	25.7	0.554	0.525
	NMT-4K	26.1	0.557	0.517
	NMT-8K	26.4	0.555	0.521
	NMT-16K	26.7	0.555	0.520
	NMT-32K	26.7	0.552	0.523
	NMT-Word-Based	23.0	0.514	0.585
	SMT-MERT	20.0	0.502	0.643
	SMT-MIRA	19.2	0.484	0.704