Global Under-Resourced MEdia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D1.5 – Final release of project data

Nature	Report	**Work Package**	WP1
Due Date	31/06/2022	**Submission Date**	31/06/2022
Main authors	Felipe Sánchez-Martínez (UA)		
Co-authors	-		
Reviewers	Wilker Aziz		
Keywords	language resources, corpora, machine translation		
Version Control			
v0.8	**Status**	1st Draft	07/06/2022

Contents

Abstract

This deliverable provides links to the different corpora that we have crawled from the Internet during the execution of the second half of the GoURMET project. These corpora are freely available and can be downloaded from the project webpage (https://gourmet-project.eu/data-model-releases/).

1 Introduction

During the execution of the second half of the GoURMET project we have crawled from the Internet a number of parallel and monolingual corpora for training the translation models for the seven languages addressed since the mid-term evaluation, including the training of an NMT system for a surprise language.

Next section briefly describes each corpus and provides the link from which it can be downloaded. These links are available from the project webpage (https://gourmet-project.eu/data-model-releases/). The README file accompanying each corpus provides additional information on the crawling process. A detailed description of the approaches followed to crawl these corpora can be found in deliverable *D1.4 Final report on data gathering and augmentation*.

2 Corpora

2.1 English–Macedonian parallel corpus and Macedonian monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 408 websites containing documents in English and Macedonian.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-mk.zip

2.2 English–Yoruba parallel corpus and Yoruba monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 14 websites containing documents in English and Yoruba.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-yo.zip

2.3 English–Burmese parallel corpus and Burmese monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 901 websites containing documents in English and Burmese.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-my.zip

2.4 English–Pastho parallel corpus and Pastho monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 152 websites containing documents in English and Pastho.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ps.zip

2.5 English–Igbo parallel corpus and Igbo monolingual corpus

Parallel and monolingual corpora obtained from the Internet Archive.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ig.zip

2.6 English–Hausa parallel corpus and Hausa monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 113 websites containing documents in English and Hausa.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.en-ha.zip

Training data for WMT 2021 crawled from the President of Iran's website.

- Link for download: https://data.statmt.org/wmt21/translation-task/ha-en/khamenei.v1.ha-en.tsv

Developments and test data for WMT 2021:

- Link for download (dev set): https://data.statmt.org/wmt21/translation-task/dev.tgz
- Link for download (test set): https://data.statmt.org/wmt21/translation-task/test.tgz

2.7 Tigrinya monolingual corpus

Monolingual corpora obtained by crawling a collection of 5 websites containing documents in Tigrinya.

- Link for download: http://data.statmt.org/gourmet/corpora/GoURMET-crawled.ti.zip

2.8 Monolingual News Crawl

Monolingual news crawl gathered from online news sources from around the world.

- Link for download: http://data.statmt.org/news-crawl

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D1.5 Final release of project data