Global Under-Resourced MEdia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D1.3 – Initial release of project data

Nature	Report	**Work Package**	WP1
Due Date	30/06/2020	**Submission Date**	30/06/2020
Main authors	Felipe Sánchez-Martínez (UA)		
Co-authors	Barry Haddow (UEDIN)		
Reviewers	Alexandra Birch (UEDIN)		
Keywords	language resources, corpora, machine translation		
Version Control			
v0.8	**Status**	1st Draft	08/06/2020
v1.0	**Status**	Final	25/06/2020
v1.1	**Status**	Minor update	03/07/2020

Contents

Abstract

This deliverable briefly describes the data released during the execution of the first half of the GoURMET project. In particular, we provide the links to the different corpora that we have crawled from the Internet and used, in addition to the corpora that were already available, for the development of the translation models for the language pairs addressed to date.

1 Introduction

During the execution of the first 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 eight languages addressed to date. We dedicated out efforts to those languages for which there was not enough parallel corpora available. As a result of the eight languages addressed to date, we crawled data for all of them but Bulgarian, for which there was enough bilingual resources already available.

Next section briefly describes each corpus and provides the link from which it can be downloaded. These link 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 *D1.2 Initial progress report on data gathering and augmentation*.

2 Corpora

2.1 English–Swahili parallel corpus and Swahili monolingual corpus

Parallel and monolingual corpora obtained by crawling a collection of 3,751 websites containing documents in English and Swahili.

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

The preparation of this corpus is described in Sánchez-Martínez et al. (2020).

2.2 English–Turkish parallel corpus and Turkish monolingual corpus

Corpora obtained by crawling a collection of 1,248 websites with documents in English and Turkish.

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

2.3 English–Amharic parallel corpus and Amharic monolingual corpus

These corpora were obtained by crawling a collection of 3,378 websites containing documents in English and Amharic.

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

2.4 English–Kyrgyz parallel corpus and Kygryz monolingual corpus

These corpora were obtained from a collection of 98 websites with documents in English and Kyrgyz.

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

2.5 Kyrgyz–Russian parallel corpus

This corpus was crawled from a collection of 80 websites containing parallel data in Kyrgyz and Russian.

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

2.6 English–Serbian parallel corpus

This corpus was crawled from a collection of 208 websites containing parallel data in English and Serbian.

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

2.7 English–Serbo-Croatian parallel corpora

This corpus was crawled from a collection of 7,876 websites containing parallel data with English in one side and Serbian, Croatian or Bosnian on the other side.

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

2.8 Parallel and monolingual corpora of languages of India

The PMIndia corpus contains parallel and monolingual corpora for 13 Indian languages, including Gujarati and Tamil. The corpora were obtained from the Prime Minister of India's news updates.

- Link for download: http://data.statmt.org/pmindia/

The preparation of this corpus is described in Haddow and Kirefu (2020)

2.9 Monolingual News Crawl

These are crawled and shuffled news crawls from 42 languages, including all the languages studied by Gourmet so far.

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

References

Haddow, B. and Kirefu, F. (2020). PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Sánchez-Martínez, F., Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., Forcada, M. L., Esplà-Gomis, M., Secker, A., Coleman, S., and Wall, J. (2020). An English–Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22th Annual Conference of the European Association for Machine Translation*, pages 299–308, Online Conferece. European Association for Machine Translation.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D1.3 Initial release of project data