browser-based multilingual translation

bergamot

Horizon 2020 Research and Innovation Action
Grant Agreement No. 825303

https://browser.mt

Deliverable 4.1: Parallel Webpage Texts and Metadata

Lead author(s): Mark Fishel (UTARTU)
Contributing author(s): Maksym Del (UTARTU)
Internal Reviewer(s): Nikolay Bogoychev (UEDIN)

Work Package: 4
Type of Deliverable: Report
Due Date: 30 September 2019
Date of Submission: 30 September 2019
Current Version: 1.0

Document History

Version	Date	Changes
1.0	30 September 2019	Original Submission Draft

Executive Summary

One part of the Bergamot project is developing new neural machine translation methods that would be able to improve translation quality based on any additional information about the webpage that is being translated. This additional information should cover the document context and any other features of the text being translated. In order to train and test such translation models we need a dataset with parallel sentences from web pages as well as the corresponding additional information to be used.

Here we describe the first version of this dataset, which is based on the already crawled parallel sentences from the Paracrawl corpora for Estonian-English, Czech-English and Polish-English. We have added the following additional information to this corpora: the titles and descriptions of each webpage (extracted from the HTML head and meta tags), the first 3 headers and first 3 paragraphs of the webpage content, as well as the automatically predicted genre of the webpage, given by an automatic genre classifier of web URLs.

The data collection work package (WP4.1) continues till June 2020, during which time we will add more language pairs and other information to the dataset.

Contents

1 Motivation

Most contemporary neural machine translation (NMT) systems work on sentence level without access to the broader context in which the input sentence appears. Access to additional information such as the document context and text domain/genre can result in better translation quality. The aim of this deliverable was to create a parallel corpus of web page texts together with additional information that can improve translation quality and the capacity of Bergamot's NMT systems to dynamically adapt to the web page at hand.

The initial idea of this work package was to collect web pages together with additional information like HTML meta-data and other document-level context information. Since in the meantime the ParaCrawl corpus[1] has appeared, we re-used its texts instead of crawling new ones, extracted meta-data partially based on the raw HTML code of the webpages behind ParaCrawl (v 3.0) and are now releasing the additional annotation for some of its language pairs.

This deliverable reports on results of work package 4.1 (data collection), lead by UTARTU and with help from UEDIN and the ParaCrawl project.

2 Description

The collected dataset and associated models are available online at:
https://owncloud.ut.ee/owncloud/index.php/s/cRCnrzRrgrGYp4A

The ParaCrawl corpus includes sentence pairs together with the URLs of the source and target documents where the sentences were extracted from. In this version of the meta-data dataset we collect document-level information, thus the additional data is assigned to each unique URL in the selected ParaCrawl data. We cover three Bergamot language pairs in this release: Estonian-English, Czech-English and Polish-English.

The format of the dataset is JSON-Lines, where each line includes

- the web page URL (main line key, string)
- web page genre, assigned automatically with a URL genre classifier
- the web page titles (list of up to 3 strings)
- the web page descriptions (list of up to 3 strings)
- up to first 3 headers (`h1` tags) of the web page
- up to first 3 paragraphs (`p` tags) of the web page

These are described in detail below.

[1] https://paracrawl.eu

2.1 Genre

This field consists of the automatically predicted genre of the web page content, based on the page URL. Details of automatic genre prediction are provided below in Section 3.

2.2 Titles

We extract up to 3 titles from each web page. This includes

- the contents of the `<title>` tag inside the `<head>` section of the web page
- the contents of the Facebook Open Graph title meta-tags[2]
- the contents of the Twitter Summary Card title meta-tags[3]

The various titles are not distinguished and are saved as a list of strings; its length can thus be shorter than 3.

2.3 Descriptions

Similarly to titles we extract up to 3 descriptions from each web page. This includes

- the contents of the `<meta name="description"...>` tag inside the `<head>` section of the web page
- the contents of the Facebook Open Graph description meta-tags[2]
- the contents of the Twitter Summary Card description meta-tags[3]

Again, in case any of these three are missing, the list of description is shorter than 3 and can be 0.

2.4 Headers

Headers include the contents of the first 3 `<h1>` tags inside the `<body>` section of the web page. The intuition behind saving them is that the first headers might be more characteristic of the webpage content.

[2] https://developers.facebook.com/docs/sharing/webmasters/#markup
[3] https://developer.twitter.com/en/docs/tweets/optimize-with-cards

2.5 Paragraphs

Similarly to headers, paragraphs include the contents of the first 3 <p> tags inside the <body> section of the web page. The intuition behind saving them is that the first paragraphs might also be characteristic of the overall webpage content.

2.6 Example

Here we show an example of an entry in our dataset, corresponding to a single URL. For the sake of easier reading line breaks are introduced; these are missing in the dataset and each entry is presented on one line:


```
{"16eur.ee/munkenhof/":

  {"url_category": "Business",

  "header_titles": ["Munkenhof - 16 EUR Hostel"],

  "header_descriptions": [],

  "body_h_tags": ["Choose your language", "To Start Chat",
    "Click The Icon"],

  "body_p_tags": ["\u0420\u0443\u0441\u0441\u043a\u0438\u0439
    suomi eesti english",

    "Old Town Munkenhof Guesthouse is situated in the heart of
    Old Town, Tallinn’s medieval centre. The guesthouse is 100
    meters from the Town Hall Square on the quiet Munga street,
    which means you are just steps away from the city’s best
    restaurants, cultural attractions, and nightlife.",

    "We provide quality accommodation for affordable prices.
    Our guests can choose between comfortable private rooms and
    4- or 6-bed dormitories, all with fresh linen and towels.
    Amenities include a 24-hour reception, cozy ..."]}
}
```

Table 1: Performance of the genre classifier, estimated on a test set of $30,000$ URLs from the DMOZ dataset. The overall accuracy estimate is 0.85.

	precision	**recall**	**f1-score**
Adult	0.98	0.17	0.29
Arts	0.49	0.91	0.63
Business	0.72	1.00	0.83
Computers	0.91	0.95	0.93
Games	0.96	0.93	0.94
Health	0.99	0.95	0.97
Home	0.98	0.87	0.92
Kids	0.93	0.64	0.75
News	1.00	0.56	0.72
Recreation	0.92	0.98	0.95
Reference	0.77	0.91	0.83
Science	0.90	0.95	0.92
Shopping	0.97	0.97	0.97
Society	0.81	1.00	0.89
Sports	0.97	0.92	0.95
Weighted avg	0.89	0.85	0.83

3 Automatic Genre Prediction

Automatic genre prediction is trained on the DMOZ URL classification dataset[4]. This dataset was collected and annotated manually by Mozilla and contains website URLs and genres of these websites. The used top-level genres are listed in Table 1.

We trained a supervised classifier that takes the website URL as input and learns to predict the website genre. The performance metrics are given in Table 1. The `Arts` genre has a low precision (0.49), while `Kids`, `News` and especially `Adult` genres have particularly low recall values (0.64, 0.56 and 0.17, respectively).

URLs from both source-side (non-English) and and target-side (English) were classified and included in the meta-data dataset. We also release the trained prediction model together with an example of Python source code for loading and applying classifier to new URLs.

4 Statistics

Next we present some brief statistics on the collected additional data.

[4] https://www.kaggle.com/shawon10/url-classification-dataset-dmoz/downloads/url-classification-dataset-dmoz.zip/2

4.1 Dataset Sizes

Table 2 presents the number of documents (URLs) present in the collected data and resulting parallel corpus sizes.

Table 2: Resulting sizes of the dataset

	Estonian	Czech	Polish
Nr. of documents	365 042	2 927 638	2 398 165
Nr. of sentences	1 102 851	9 161 516	8 295 489

4.2 Genre Distributions

The frequencies of automatically predicted genres on the three language pairs in our datasets are presented in Figure 1.

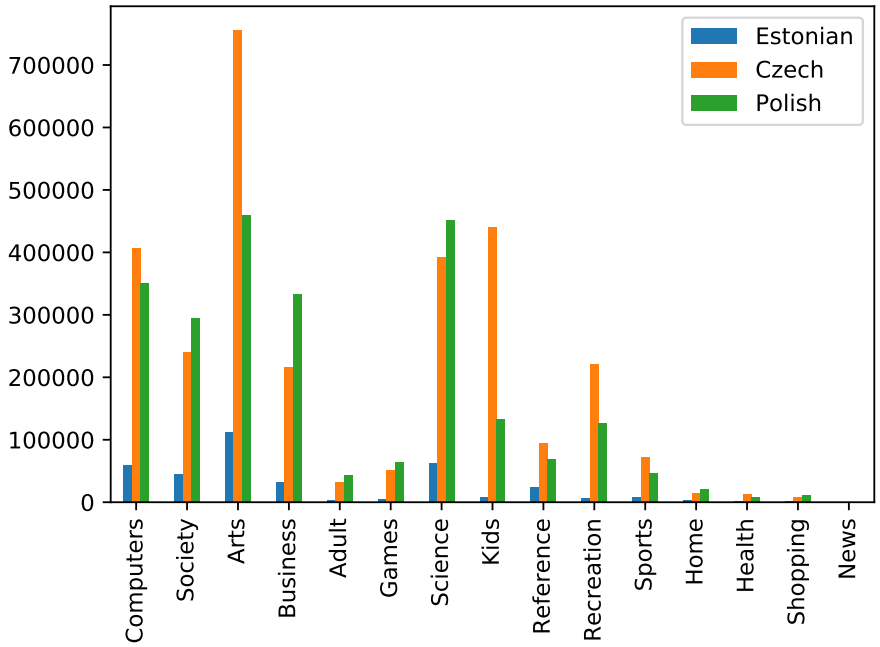

Figure 1: Frequencies of predicted URL genres for Estonian, Czech and Polish ParaCrawl documents

An interesting thing to note is the elevated frequency of the `Arts` category in Czech and Estonian, given that its prediction accuracy is quite low; based on Table 1 half of these might be false-positives.

5 Main Challenges and Next Steps

5.1 Next Steps

The data collection work package (WP4.1) continues till June 2020. In the next 9 months we will add more language pairs and other information to the dataset. This includes

- German-English, French-English and Spanish-English data
- automatic clusters on the level of sentences, based on sentence embeddings
- automatic clusters on the level of documents, based on document embeddings

As part of work package 4.4 on meta-data incorporation into NMT we will make extensive use of the collected data and evaluate, which information helps translate web pages with higher quality and better adapt to each web page and document context.

We will also test if the genre can be predicted more precisely based on other input besides the URL available in the database, to improve the low-scoring category assignment.

5.2 Challenges

The main challenge in the preparation of this dataset was the volume of data that needed to be processed. Extracting parts of the original HTML documents from ParaCrawl required the raw data to be used, which meant processing hundreds of gigabytes of data. More precisely

- the volume of the compressed full set of raw data for Estonian is **153 Gb** (more for other covered languages)
- after decompression and filtering (leaving only the documents with sentences from the parallel corpus in them) the volume of Estonian data was **24 Gb**, Czech data – **134 Gb**, Polish data – **138 Gb**
- for estimating future work we extracted and filtered the French corpus: after filtering the uncompressed size is **9.2Tb**, which poses a challenge from the point of view of both storage, reading/writing and processing

As a result of the volume of data to be processed, fully automatic preparation of the Estonian-English part of the released data took 69 hours. Based on this, similar data for German, French, Spanish and other resource-rich languages will take considerably more time to prepare.