Parallel Corpus Extraction from CommonCrawl

Herve Saint-Amand
Jason Smith
Magdalena Plamada

Overview

- CommonCrawl: 60 TB web corpus hosted on Amazon's Simple Storage Service (S3)
- Costly to download, but free to access through Elastic MapReduce
- Our project: write code to find parallel sentences from multilingual websites in this data

General Strategy

- Identify multilingual websites through URLs:

  http://europa.eu/index_de.htm
  http://europa.eu/index_en.htm

- Map these websites to a language-independent URL:

  http://europa.eu/index_*.htm

- Align the websites and produce parallel data

Prior Work

- STRAND (Resnik and Smith, 2003):
  - Find potentially bilingual websites through URLs/links
  - Align HTML tags and text nodes
  - Extract parallel sentences from aligned text nodes
- Gale and Church (1993) used for sentence alignment

Hadoop Implementation

MapReduce basics:

- Mapper: Executed on each data point, produces 0 or more <key, value> pairs
- Output of the mappers are sorted by key
- Reducer: Executed on each key (and all values associated with that key), produces the final output

Hadoop Implementation

- Mapper:
  - Checks URL for language codes/other identifiers

    example.com/en/index.html
  - Runs language ID on the text of the webpage to check if the URL match was correct
  - Produces a <key, value> pair:

    <example.com/*/index.html, (full webpage)>

Hadoop Implementation

- Reducer:
  - Receives all webpages whose URLs mapped to the same language-independent string:

    example.com/en/index.html

    example.com/fr/index.html

    ...
  - Aligns the HTML structure of all pairs of pages
  - If the pages align well enough, run sentence segmentation, sentence alignment, and produce parallel sentences

Pilot Experiment

- Tested a baseline implementation on a ~15GB subset of the full data
- Checked for webpages in a few languages:
  - English
  - German
  - French
  - Spanish

Results

Language pair	Sentences	SL words	TL words
FR-EN	46,496	187,476	170,590
FR-ES	40,275	103,112	101,795
EN-ES	30,673	72,575	74,567
EN-DE	27,957	106,430	101,823
FR-DE	22,774	83,818	75,100
DE-ES	17,367	47,176	55,888

- 78% precision for **FR-EN** and **EN-DE**

Conclusions

- Extraction from CommonCrawl is cheap and fast
  - Simple strategy: one MapReduce iteration
- Workflow in place to extract from the full dataset, or from a local copy of a small subset