Project abstract

Colby Goettel

November 15, 2014

1 Basic problem

There doesn't appear to be an easy way to automate machine translation (MT) and formatting in LaTeX. This project will build a wrapper around an MT engine (to be decided, see below Technologies section) that will provide an easy, extensible interface for translating a text file from a specified language to another, aligning the text via LaTeX, and outputting to a PDF.

2 Overall approach

The wrapper will all be a standalone program written in Perl with the following flags:

- `--input-file` (required): the name of the input file
- `--output-file`: if specified, the output file will be named according to the option here; otherwise, the output file will be named the same as the input file.
- `--debug`: this command is for debugging purposes and will cause the output to be written to STDOUT.
- `--in-lang` (required): the two-letter code of the input language. The system will be tested for `en`. Extensible to other languages.
- `--out-lang` (required): the two-letter code of the output language. The system will be tested for `fr`. Extensible to other languages.
- `--alignment`: the type of alignment desired. Options will include `sentence` (default), `paragraph`, `none`.
- `--help`: this will display a help prompt. If the required fields are not entered, this will be displayed.

3 Technologies

Gary gave a presentation on MT and bitext alignment. This project will use one of the MT programs talked about (need his slides to figure out which one) and LaTeX for alignment. The main criteria for an MT program is that it can be used on the terminal. The output format isn't a huge concern because parsing text isn't too difficult.

4 Tools

- Wrapper: Perl
- Makefile: Bash and make macros
- MT: still needs to be selected.
- Alignment: LaTeX

5 Knowledge sources and corpora

The corpora used will be selections from the Bible. The main purpose of this project is to potentially aid one of my other projects, translating and typesetting the Bible: `https://github.com/cgoettel/bible/`.

6 Evaluation

The system should be evaluated based on its ability to easily translate a file and produce a correctly formatted output file. Correctly formatted, in this case, means that the output file aligns the parallel texts accurately and according to the user's choice (sentence, paragraph, or none).

7 References

- `http://ctan.org/pkg/parallel`
- `https://github.com/cgoettel/bible/`
- `https://github.com/cgoettel/mt-alignment-wrapper`