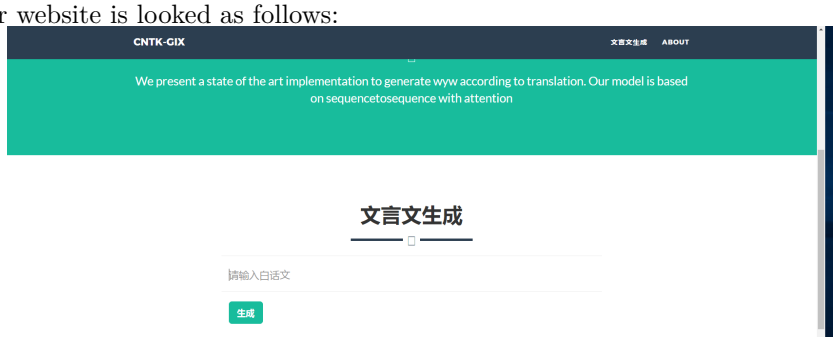
GIX-Project final report

Jianyu Zhou, MSRA Intern

Mar. 6th, 2017
v-zjiany@microsoft.com

1. Introduction

This project is to build a machine translation model based on CNTK. To be more specific, we built a system to translate modern Chinese to ancient Chinese. Our demo is available on http://172.22.80.58:5024 (only addressable in LAN). Our website is looked as follows:



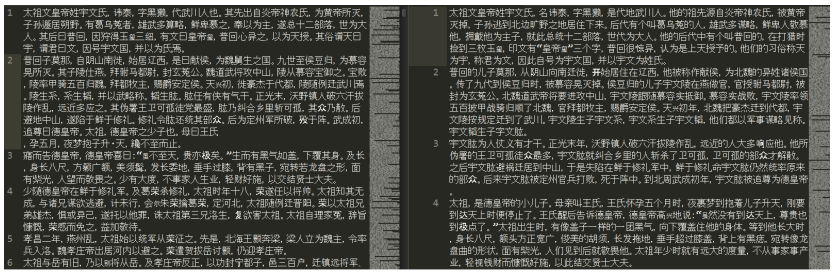
2. Individual Work Description

My main work in this project is to collect and process data. All my work related to this project is available on our server(msraml-linux01):

- data: /home/v-zjiany/GIX-project/data/modern&classical_chinese_translation_data/
- source code: /home/v-zjiany/GIX-project/code/modern_ancient_alignment/

In addition, we build a tool that can automatically generate moder and ancient translation pairs. The tool's usage is as follows:

- **First**, we suppose that you have got an ancient Chinese books and its corresponding modern edition. They should be processed to the following format:



Notice: No level of alignment is required, but books should be divided in to paragraphs, and **one line should represent one paragraph.**

- **Second,** run `book2sentence_aligned.py` with three parameters.

  ```
  python book2sentence_aligned.py <your ancient book path> <your corresponding modern boo
  ```

  Notice:

  - This tool generate sentence pairs with max length of 20 Chinese characters by default. If you want to change the max length, please modify the variable `max_length` in `book2sentence_aligned.py`
  - Output will be found in your given output path. You should found two text files, `origin_xx` and `trans_xx` in your given output path. This two files should have the same number of lines.

3. Impressions with MSRA

First, I'm really grateful to my mentor Fei Gao and Qiwei Ye, they are so kind and just like my senior in campus. I really hope to learn more from them and keep them in contact.

Second, I'd like to say that MSRA's working environment just comforts me well. If had a chance, I will surely choose MSRA to continue my further research. Though one month is a short time for intern, what MSRA gave me is impressive. More importantly, I've polished my academic skills in this month and enriched my working experience.

Finally, thanks to GIX, who offered us the wonderful chance to work there. I really hope to come back to MSRA some day and make more contributions in the future.