Goal

Building a machine translation system that is able to do translation between classical Chinese and modern Chinese.

Method

The method will be based on word alignment algorithm from IBM Models. Moses will be used as the decoder.

Data

We use Sima Qian's *Records of the Grand Historian*, also known by its Chinese name *Shiji* (史記), except for the *Table* (表) volumes. The original text contains more than five hundred thousand Chinese characters.

The original text we use comes from several sources. We mainly use the text from Gutenberg Project, which comes as a nice single plain text file[1]. However, about 1,000 characters are corrupted, which we found using the following regular expression:

```
[^\u2E80-\u9FFF，。；：“”‘’《》！？、·「」『』…\r\n]
```

We recovered each corrupted character by cross-referencing Wikisource[2] and the Chinese Text Project[3]. The original text is then converted from traditional Chinese characters to simplified Chinese characters using ConvertZ[4].

The translation we use mainly comes from read126.cn[5]. (Obviously this is not originally where the translation was published, but we cannot find the original source.) There are also corrupted characters in translations, which are harder to deal with. Here's what we did:

1. Try to find the same translation from other sources by Googling the nearby uncorrupted sentence;
2. Try to find other translations by Googling the corresponding original text in classical Chinese;
3. If the corrupted character is a name of some person / location / plant /animal, try to recover by consulting the corresponding original sentence in classical Chinese;
4. Occasionally when all the above fail, try to fill in a translation (as a native speaker).

We also tried to correct the wrong punctuation by looking for consecutive ，。；：！？、• characters. For each case, we checked whether it indicated a missing sentence or was just a typo.

[1] http://www.gutenberg.org/files/24226/24226-0.txt

[2] http://zh.wikisource.org/wiki/史記

[3] http://ctext.org/shiji

[4] http://dl.pconline.com.cn/html_2/1/76/id=495&pn=0.html

[5] http://www.read126.cn/194c6894-51d5-4df3-a4bc-fa1282139f82!c0856342-2132-4498-921c-d81450904044!66ee2899-4ffa-41e9-ae09-126ca0281c65.html

Evaluation

We use the average character-level BLEU score of each paragraph as evaluation metric.

We use character-level BLEU rather than word-level BLEU because according to Li et al. 2011[6], character-level metrics correlate better with human assessment.

We manually went through the corpus, breaking and merging paragraphs to make each translation paragraph align with the corresponding original paragraph.

We considered using sentence alignment algorithms to automatically do this work, but since the performance of such algorithms needs to be evaluated, the validity of our evaluation metric would become questionable. (Many sentence alignment algorithms depend on the length correlation of a sentence and its translation. However, during the translation of a classical Chinese sentence, the translator sometimes needs to add additional context information, potentially translating a short sentence into a fairly long one. This makes the performance of sentence alignment algorithms especially questionable.)

The reason why we use average paragraph BLEU rather than average sentence BLEU is that having tried to manually align sentences of a volume, we found the workload unacceptable.

In the calculation of BLEU, punctuation marks are treated indiscriminately, and are not included in n-grams.

Baseline

Since classical Chinese and modern Chinese share many characters, words and sometimes even grammar, the simplest thing we can do is not doing anything. Here's the result we get:

	Modern to Classical	Classical to Modern
1-BLEU	0.43476	0.45993
2-BLEU	0.11040	0.17935
3-BLEU	0.02617	0.05008
4-BLEU	0.00514	0.01627

As we can see, this do-nothing baseline works better for classical-to-modern translation, possibly because it suffers less brevity penalty.

Among the n-BLEU scores, our do-nothing baseline gets relatively high score for 1-BLEU. This conforms to the fact that classical and modern text share many characters. However, as n goes up to 4, the BLEU scores decrease significantly.

Unfortunately, we don't have a classical- modern Chinese dictionary that can be used to improve the baseline. The best we can do is to replace/remove some most frequent words.

For the modern-to-classical baseline, we replace 之 with 的, 曰 with 说, 至 with 到, 秦/楚/齐 with 秦/楚/齐国. We also removed the following words and characters: 了, 他们, 所以, 军队, 在, 他, 就, 是, 这.

[6] Li, M., Zong, C., & Ng, H. T. (2011). Automatic evaluation of Chinese translation output: word-level or character-level?

For the classical-to-modern baseline, we replace 对曰 with 回答说, 曰 with 说. We remove all 者 before punctuation marks, and then replace the rest with 的. We replace each 之 that are not followed by a punctuation mark or the character 以 with 的.

Here's the BLEU score of our improved baseline:

	Modern to Classical	Classical to Modern
1-BLEU	0.47505	0.47286
2-BLEU	0.13330	0.18782
3-BLEU	0.03426	0.05284
4-BLEU	0.00844	0.016906