Final Project - Chinese Word Segmentation
Write-up Second Draft
Sean Welleck

Problem Definition

In languages such as English or French, words are typically separated by spaces, making sentence tokenization simple. However, in Chinese (and Japanese Kanji), the written language consists of characters that are not delimited, making word tokenization difficult. A given character may have an independent meaning as a word, and a separate meaning when grouped with other characters. For instance, the character 中 means 'middle', 将 means 'will', and when combined, 中将 means 'lieutenant general'. This can lead to ambiguities when attempting to segment and translate a sentence, since depending on the context, the correct translation of 中将 may be 'middle will', while in another context it may be 'lieutenant general'.

To illustrate with a couple English examples, first consider the character sequence “thesearedaredevils”. This sequence could either be segmented as “these are daredevils” or “the seared are devils”. Even a single segmentation affects the interpretation of a sentence: “iship” can be segmented as “I ship” or “is hip”, leading to the two different sentences “I’ll save the code that I ship” and “I’ll save the code that is hip”. Without word delimiters, Chinese text encounters these ambiguities, making Chinese word segmentation (CWS) an important step when translating from Chinese.

The CWS problem is to transform an input character sequence S without spaces to a sequence S_{seg}, where S_{seg} contains spaces between word segments. Numerous methods have been developed or applied to the problem, such as the Compression-Based algorithm[1], Conditional Random Fields[2], and the Maximum Entropy Model[3]. I also found a relatively simple algorithm called TANGO[4]developed for Japanese Kanji, but applicable to Chinese.

Sources

The textbook gives a very brief motivation of Chinese Word Segmentation on pg. 34.

“A Compression-based Algorithm for Chinese Word Segmentation”[5] give a compression-based algorithm as well as an introduction to the problem with three examples.

---

[1] http://acl.ldc.upenn.edu/J/J00/J00-3004.pdf
[2] http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1091&context=cs_faculty_pubs
[3] http://acl.ldc.upenn.edu/W/W03/W03-1728.pdf
[4] http://www.research.ibm.com/talent/documents/ando-lee-nle03.pdf
[5] http://acl.ldc.upenn.edu/J/J00/J00-3004.pdf

"Improved Statistical Machine Translation by Multiple Chinese Word Segmentation"[6] discusses the impacts of segmentation on translation quality.

"Mostly-Unsupervised Statistical Segmentation of Japanese Kanji Sequences"[7] gives a clear introduction of the Japanese segmentation problem with examples, and a simple algorithm. It mentions that the algorithm could extend to other languages.

"Chinese Segmentation and New Word Detection using Conditional Random Fields"[8] detail using CRFs for segmentation.

Finally, "The Second International Chinese Word Segmentation Bakeoff"[9] discusses a competition held to evaluate CWS systems, and describes the datasets used.

Objective Function

A common evaluation technique that I found was to use Recall, Precision, and F-measure. This is the approach used in the Bakeoff competition, for instance. Since the leaderboard requires a single number, I use a weighted sum of the three, $objective = \frac{1}{3}(recall + precision + F measure)$. The statistics are over the entire test set, specifically:

m = number of correctly segmented words

n = total number of words segmented

N = total number of words in testing truth data

$$recall = \frac{m}{N}$$

$$precision = \frac{m}{n}$$

$$F measure = \frac{2(recall)(precision)}{recall+precision}$$

Another approach would be to use an extrinsic evaluation measure to see how changes in word segmentation influence translation quality. This would involve segmenting sentences, translating those sentences, then using the BLEU score as the objective. To do this, we could use Joshua to train a model using a Chinese-English parallel corpus such as the MultiUN[10] corpus. A portion of the corpus could be held out as a test set. Then, the spaces would be removed from the Chinese portion of the test set. The CWS system could then segment the unsegmented test set, which would then be translated using the model trained with Joshua. Finally we would evaluate the BLEU score of the resulting translations. CWS systems could be compared by substituting in each different system for the segmentation step,

[6] http://www.aclweb.org/anthology/W/W08/W08-0335.pdf
[7] http://www.research.ibm.com/talent/documents/ando-lee-nle03.pdf
[8] http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1091&context=cs_faculty_pubs
[9] http://acl.ldc.upenn.edu/I/I05/I05-3017.pdf
[10] http://opus.lingfil.uu.se/MultiUN.php

running the translation on the newly segmented test set, and evaluating the BLEU score. The 'best' segmentation would be the segmentation that leads to the highest quality translation, as measured by the BLEU score.

Data

A corpus of segmented data will be needed to train a model using a supervised method. To test the model, a set of unsegmented test sentences and segmented "truth" sentences will be needed. The 2005 Bakeoff competition posted the 4 datasets that they used, found at: http://www.sighan.org/bakeoff2005/. I chose the dataset provided by Beijing University. The training data has 1.1M Chinese words that are separated by line into sentences. There is a segmented and corresponding unsegmented version. The testing data consists of 104K Chinese words. Similar to the training set, there is a segmented file and a corresponding unsegmented file.