SIGHAN

Second International Chinese Word Segmentation Bakeoff

Detailed Instructions

The following comprises the complete description of the training and testing for the Second International Chinese Word Segmentation Bakeoff. By participating in this competition, you are declaring that you understand these descriptions, and that you agree to abide by the specific terms as laid out below.

Training: Description of Tracks

Dimension 1: Corpora

Four corpora are available for this bakeoff:

Corpus	Encoding	Word Types	Words	Character Types	Characters
Traditional Chinese					
Academia Sinica	Unicode/Big Five Plus	141,340	5,449,698	6,117	8,368,050
City University of Hong Kong	HKSCS Unicode/Big Five	69,085	1,455,629	4,923	2,403,355
Simplified Chinese					
Peking University	CP936/Unicode	55,303	1,109,947	4,698	1,826,448
Microsoft Research	CP936/Unicode	88,119	2,368,391	5,167	4,050,469

You may declare that you will return results on any subset of these corpora. For example, you may decide that you will test on the Sinica Corpus and the Beijing University corpus. The only constraint is that you **must not** select a corpus where you have knowingly had previous access to the testing portion of the corpus. A corollary of this is that **a team may not test on the data from their own institution**.

Dimension 2: Open or Closed Test

You may decide to participate in either an open test or a closed test, or both.

In the open test you will be allowed to train on the training set for a particular corpus, and in addition you may use **any** other material including material from other training corpora, proprietary dictionaries, material from the WWW and so forth.

If you elect the open test, you will be required, in the two-page writeup of your results, to explain what percentage of your correct/incorrect results came from which sources. For example, if you score an F measure of 0.7 on words in the testing corpus that are out-of-vocabulary with respect to the training corpus, you must explain how you got that result: was it just because you have a good coverage dictionary, do you have a good unknown word detection algorithm, etc?

In the closed test you may **only** use training material from the training data for the particular corpus you are testing on. **No other material or knowledge is allowed**, including (but not limited to):

1. Part-of-speech information
2. Externally generated word-frequency counts
3. Arabic and Chinese Numbers

4. Feature characters for place ames
5. Common Chinese surnames

Declaration

When you download the training corpora, you will be asked to register and provide various information about your site, including the contact person, and you will be asked to declare which tracks you will participating in.

Format of the data

Both training and testing data will be published in the original coding schemes used by the data sources. Additionally it will be transcoded by the organizers into Unicode UTF-8 (or, if provided in Unicode, into the defacto encoding for the locale.) The training data will be formatted as follows.

1. There will be one sentence per line.
2. Words and punctuation symbols will be separated by spaces.
3. There will be no further annotations, such as part-of-speech tags: if the original corpus includes those, those will be removed.

Licensing

The corpora have been made available by the providers for the purposes of this competition only. By downloading the training and testing corpora, you agree that you will **not use these corpora for any other purpose than as material for this competition**. Petitions to use the data for any other purpose MUST be directed to the original providers of the data. Neither SIGHAN nor the ACL will assume any liability for a participant's misuse of the data.

Testing

The test data will be available for each corpus at the website at 12:00 GMT, July 27, 2005. The test data will be in the same format as described for the training data, but of course spaces will be removed.

You will have roughly two days to process the data, format the results and return them to the SIGHAN website. The final due date/time is:

July 29, 2005, 12:00, GMT

Late submissions will not be scored.

The format of the result **must** adhere to the format described for the training data. In particular, there must be one line per sentence, and there must be the same number of lines in the returned data as in the data available from the site. Segmented words and punctuation must be separated by spaces, and there should be **no further annotations** (e.g. part of speech tags) on the segmented words. **The data must be returned in the same coding scheme as they were published in.** (For example, If you utilize the UTF-8 encoded version of the testing data, then the results must be returned in UTF-8.) Participants are reminded that ASCII character codes may occur in Chinese text to represent Latin letters, numbers and so forth: such codes should be left in their original coding scheme. Do not convert them to their GB/Big5 equivalents. Similarly GB/Big5 codings of Latin letters or Arabic numerals should be left in their original coding, and not converted to ASCII.

The results will be scored completely automatically. The scripts that were used to score will be made publicly available. The measures that will be reported are **precision**, **recall**, and an evenly-weighted **F-measure**. We will also report scores for in-vocabulary and out-of-vocabulary words.

Note: by downloading the test material and submitting results on this material you are thereby declaring that you have not previously seen the test material for the given corpus.

You are also declaring that your testing will be fully automatic. This means that any kind of manual intervention is disallowed, including, but not limited to:

1. Manual correction of the output of your segmentation.
2. Prepopulating the dictionary with words derived by a manual inspection of the test corpus

Results

Results will be provided in two phases. Privately to individual participants by August 5, 2005, then publicly to all participants and to the community at large at the SIGHAN Workshop. By participating in this contest, you are agreeing that the results of the test may be published, including the names of the participants.

Writeup

By electing to participate in any part of this contest, you are agreeing to provide, by August 19, 2005, a two-page writeup that briefly describes your segmentation system, and a summary of your results. In the closed tests you **may** describe the technical details of how you came by the particular results. In the open test you **must** describe the technical details of how you came by the particular results.

The format of the two-page paper must adhere to the style guidelines for IJCNLP-05, except for the two page limit and the submission via the SIGHAN site.

---

tree@sighan.org
Last edited: November 18 2005 11:46:33.