Third International Chinese Language Processing Bakeoff

Data Download

The page supports the download of training materials for the Third International Chinese Language Processing Bakeoff sponsored by SIGHAN. The data has been available from April 17, 2006.

Word Segmentation Task

Four corpora are available for this bakeoff:

Corpus	Encoding	Zip Archive	Tar.gz Archive	Annotation guidelines
Traditional Chinese				
Academia Sinica	Unicode/Big Five Plus	Zip	tar.gz	PDF
City University of Hong Kong	HKSCS Unicode/Big Five	Zip	tar.gz	PDF
Simplified Chinese				
Microsoft Research	gb18030/Unicode	Zip	tar.gz	Doc
University of Pennsylvania/University of Colorado	CP936/Unicode	Zip	tar.gz	HTML

Named Entity Recognition Task

There are three corpora available for this task:

Corpus	Encoding	NE Types	Zip Archive	Tar.gz Archive	Annotation Guidelines
Traditional Chinese					
City University of Hong Kong	HKSCS Unicode/Big Five	PER, LOC, ORG	Zip	tar.gz	PDF
Simplified Chinese					
Microsoft Research	gb18030/Unicode	PER,ORG,LOC	Zip	tar.gz	PDF
Linguistic Data	CP936/Unicode	PER,LOC,ORG,GPE	Converts to Co-NLL		html

Consortium			and XML fomrats format		