Automatic Minimum Text Age Estimation Using Neologism Data from Hanyu Da Cidian 漢語大辭典

Tilman Schalmey, M.A. (schalmey@uni-trier.de)

Universität Trier

Main idea & concept

Information available in the *Hanyu Da Cidian* 漢語大辭典, enriched with some meta data, can be used to detect an estimated minimum age of a given text based on neologisms. The „newer“ the words in the text, the newer the text itself.

Technologies used

-› PHP; Posix Regular Expressions
used for conveniently parsing the *Cidian* locally and storing the data into a

-› mySQL database
using a relational database with tables for the character and word entries of the dictionary, plus an enriched list of books cited in it.

-› Java
using an existing tokenizer like Paoding‘s Knives or IK Analyzer to achieve a splitting of the input text into words.

Main Steps of Preparation

1) creating the database
:: [step 1] :: The full text of the *Hanyu Da Cidian* is parsed. Entries for single characters (table A) are marked with an asterisk * character and word entries (table B) are in 【】 brackets, making it easy to do shallow text analysis using regular expressions.

:: [step 2] :: A list of books cited in the dictionary (C) is created. It has to be enriched with information on the time of the text‘s creation.

:: [step 3] :: The word entries are searched for cited books and the oldest is taken as benchmark for the words‘ appearance as a neologism.

2) Evaluating tokenizers
As there are no blanks for word separation, tokenizing is not trivial, especially for Literary Chinese.

3) Creation of a browser based tool

Relational Database Model (Simplified)

character	entry
*朝	1［zhāo 坐幺］［《廣韻》陟遥切，平宵，知。］亦作“晁 2”。“鼂 2”的被通假字。1.早晨。《易·坤》：“臣弑其君。子弑其父，非一朝一夕之故，…

Table A
character entries

word	character	pinyin	entry	cited book
【朝三暮二】	*朝	zhāo	比喻主意多变。《西游记》第五九回：“似師父朝三暮二的，…	西游记
【朝三暮四】	*朝	zhāo	1.《庄子·齐物论》：“狙公賦芧，曰：‘朝三而暮四。’衆狙皆怒。曰：‘然則	庄子
【朝 2 士】	*朝	cháo	1.古代官名。掌外朝官次和刑狱等。参阅《周礼·秋官·朝士》。	周礼
【朝 2 大夫】	*朝	cháo	1.古代官名。《周礼·秋官·朝大夫》：“朝大夫掌都家之國治。	周礼

Table B
word entries

book	year	dynasty	author
西游记	1550	明	吴承恩
庄子	-327	战国	庄周
周礼	-13	汉	uncertain

Table C
cited books

Simplified Usage Example

Input string
石油所出不一。國朝正德末年，嘉州開鹽井，偶得油水，可以照夜，其光加倍。

Tokenized (using Paoding‘s Knives)
石油|所出|不一|國|朝正|正德|末年|嘉州|州開|開鹽|鹽井|偶|油水|照|夜|光|加倍

Output
The newest word detected within the text is 石油, probably first mentioned in *Mengxi Bitan* 梦溪笔谈, written by 沈括 Shen Kuo (1031–1095) around 1088, so we can assume the text is at least from 宋 Song (960–1279) or newer.

Key Problems and Limitations

- There is no safe way to estimate the *maximum* age of a given text.
- Completing the data for the cited works table (55,403 texts) is a large manual effort.
- Words can have many different meanings, so there is a chance that the „oldest“ use of the word is estimated as too old, which means a good algorithm has to be found to take that into account. Word class detection would help, but is very difficult, especially for Literary Chinese.

Outlook

- Easier detection of forgery based on the minimum age estimate
- Data on texts with high neologism rates give us clues on the reception and on language creativity
- The data can also help to… open for suggestions!

特里尔大学汉学系