ISMLA Multilingual Session 2: Greedy Tokenization of Chinese

Johannes Dellert

Tübingen University

November 15, 2017

Plan

1. Chinese Writing
2. Tokenizing Chinese
3. Exercise 01: A Greedy Tokenizer For Chinese

Chinese Writing: Some Basics

- in Old Chinese, character = syllable = word
- about 200 ancient characters (**radicals**) are actually pictographs:
  目 “eye”, 耳 “ear”, 月 “moon”, 心 “heart”, 木 “tree”
- about 400 **ideogrammic compounds** (combinations of radicals):
  古 “old” (十 “ten” 口 “mouths”), 尿 “urine” (尸 “body” 水 “water”)
- 90 percent of characters are **phono-semantic compounds**
  of a radical giving a semantic hint, and a phonetic component:

考	kǎo	“to examine”	只	zhǐ	“only”
拷	kǎo	“to flog”	扺	zhǐ	“to initiate”
栲	kǎo	“mangrove”	枳	zhǐ	“trifoliate orange”

- internal structure irrelevant for automated processing

Chinese Writing: Syllables and Words

- about 1,200 possible syllables for about 7,000 characters in use
  $\Rightarrow$ much homophony, in some instances more than a hundred characters are pronounced identically (e.g. 几 机 积 鸡 基 讥 饥 绩 are only the most common of 117 characters pronounced *jī*)
- for disambiguation, syllables with similar meaning are very frequently combined to yield words in modern Mandarin:
  - both 保 *bǎo* and 护 *hù* mean "to protect",
    but in a good EN $\rightarrow$ ZH dictionary you will find 保护 *bǎohù*
  - the standard word for "forest" is 森林 *sēnlín*, a compound of words for a larger forest or jungle 森 and a smaller forest or grove 林
- prosodically motivated restrictions often help parsing:
  - disyllabic equivalents (VV) of many monosyllabic verbs (V) exist
  - VV N with a monosyllabic noun is typically avoided for verb+object
  - V N, V NN, and VV NN are fine
  - 护林 2.640.000, 保护森林 411.000, 护森林 43.700, 保护森 7.760

Chinese Writing: Pinyin

- 拼音 **Pīnyīn**: the official Latin transcription in mainland China
- includes spacing between words (= presupposes tokenization)
- accurately represents Mandarin phonology
- slightly counterintuitive usage of some Latin characters:
  - j = [dɕ]
  - q = [tɕʰ]
  - x = [ɕ]
  - s = [s], sh = [ʂ]
  - z = [dz], zh = [dʐ]
  - c = [tsʰ], ch = [tʂʰ]
  - h = [χ]
- commonly used to represent Chinese in Latin-alphabet languages (but in a toneless version)

Tokenizing Chinese: The Challenge

Reasons why Chinese tokenization is challenging:

- no spaces between words
- no morphological marking of word classes
- some very frequent characters are plurifunctional:
  - 等 *děng* as a verb means "to wait"
  - as part of nominal compounds, it means "rank"
  - at the end of an enumeration, it means "etc."
- gap strategy for relative clauses and similar constructions enhance syntactic flexibility, increasing the ambiguity in POS assignment

Tokenizing Chinese: The Greedy Approach

- basic idea: if neighboring symbols can be found together in a dictionary, we are very likely to have found the correct dictionary entry
- a symbol which was not found in connection with any neighboring symbol is likely to represent a monosyllabic word
- $\Rightarrow$ basic procedure: look ahead, consume the maximal sequence of symbols that is still an entry in the dictionary
- works surprisingly well with a dictionary of frequent forms
- ceases to work once the dictionary becomes **too** high-coverage, e.g. if it attempts to also capture classical style

Tokenizing Chinese: Example

Successful example: “Let me work quietly.”

让	ràng	VT	let
我	wǒ	PRN	I
安静	ānjìng	A	quiet
地	de	PRT	[Adverb]
工作	gōngzuò	VI	work
。		PNT	.

Failed example: “I like tasty things.”

我	wǒ	PRN	I
爱好	àihào	N	hobby
吃	chī	VT	eat
的	de	REL	[relativizer]
东西	dōngxi	N	thing
。	.	PNT	.

Tokenizing Chinese: Existing Methods

Tokenizing Chinese is a major field of research:

- Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning (2005): **A Conditional Random Field Word Segmenter**. 4th SIGHAN Workshop on Chinese Language Processing.
- Chu-Ren Huang, Petr Šimon, Shu-Kai Hsieh, Laurent Prévot (2007): **Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification**. ACL '07.
- Pi-Chuan Chang, Michel Galley and Chris Manning (2008): **Optimizing Chinese Word Segmentation for Machine Translation Performance**. StatMT '08.

Exercise 01: A Greedy Tokenizer for Chinese

Basic layout of the exercise:

- in your type system, define a `ChineseToken` type extending `Annotation` with feature `pinyin`, `cat` and `gloss` (all with `String` values)
- write an annotator which in one pass through the document, always consumes the longest chunk of characters starting at the current position it can find in the dictionary, and annotates this chunk with a `ChineseToken` object based on information found in a dictionary
- test the method on a dataset of five snippets from the Chinese Wikipedia we provide

Exercise 01: The Dictionary

Lexical lookup is already implemented by `CmnEngDictionary`:

- initialization: `new CmnEngDictionary()`
- use: `List<SimpleDictionaryEntry> lookup(String orthForm)`

A `SimpleDictionaryEntry` contains the following fields:

- `orth`: orthography in Chinese characters
- `prnc`: pronunciation in Pinyin
- `category`: basic part-of-speech information (not important)
- `glosses`: a list of English glosses defining the meaning(s)

Exercise 01: Hints for the Implementation

- the Wikipedia snippets are distributed as an archive (unpack it into an otherwise empty data directory)
- `CmnEngDictionary` is provided in the JAR file, which needs to be added to the build path
- for debugging (and interpreting the result), we recommend to print the contents of each `SimpleDictionaryEntry` in a tab-separated format to the console (c.f. examples on slide 8)