中文句子边界检测

中文句子边界检测相对因为容易的多。中文句子绝大多数以问好“？”、感叹号“！”和句号“。”结束；而英文句号"."的用法比较多，除了表示句子结束还可以表示人名分割符、缩写甚至小数点，需要仔细区分。而中文处理中分词是分析的重点，英文单词之间由空格隔开，所以无需分词。

实例分析

The Stanford Natural Language Processing Group 有一个开源中文分词器，见 http://nlp.stanford.edu/software/segmenter.shtml。

对 stanford 中文分词程序 stanford-chinese-segmenter-2008-05-21 的源码进行分析后发现，其断句功能由类 ChineseDocumentToSentenceProcessor 完成（源代码在本文件夹里）

该类中定义了一个字符 HashSet，存储用到的句子分界符：

```
private static Set<Character> fullStopsSet = new
HashSet<Character>(Arrays.asList(new Character[]{'\u3002', '\uff01', '\uff1f',
'!', '?'}));
```

其中 '\u3002', '\uff01', '\uff1f'分别代表全角句号“。”、全角感叹号”！‘和全角问号“？”，后面两个是半角感叹号“!”和半角问号”?”。

代码后面还有一句注释："not \uff0e . (too often separates English first/last name, etc.)" 解释了没有加入半角句号的原因：经常用作分隔英语的名和姓等。

注:stanford-chinese-segmenter 的输入文件以 UTF-8 编码，在进行断句前会先 normalization。

下面是类 ChineseDocumentToSentenceProcessor 中的断句代码，分析结果见注释

```
 int lastCh = -1;
 for (Character c : content) { /*content 是准备被断句的字符串*/
   /* sentenceEnd 是一个标记变量，标记一个句子是否结束 */
   if (sentenceEnd == false) {
       /* 若当前字符 c 是句子分界符号表里的字符*/
       if (fullStopsSet.contains(c)) {
           sentenceString += newChar; /* 将当前字符加入句子 */
           sentenceEnd = true;  /* 将句子结束标记设置为真 */
     } else {
         sentenceString += newChar;  /* 将当前字符加入句子 */
     }
   } else { // 若句子结束标志为真
     if (rightMarkSet.contains(c)) {
       sentenceString += newChar;
       // EncodingPrintWriter.out.println("  Right mark char", "UTF-8");
     } else if (newChar.matches("\\s")) {
       sentenceString += newChar;
     } else if (fullStopsSet.contains(c)) {
```

```
        sentenceString += newChar;
      } else { // otherwise
        if (sentenceString.length() > 0) {  /* 继续分析 content 中的下一个句子*/
          sentenceEnd = false;
        }
        sentenceString = removeWhitespace(sentenceString, segmented);
        if (sentenceString.length() > 0) { /*若当前句子非空，将其加入句子列表 */
          sentenceList.add(sentenceString);
        }
        sentenceString = new String(); /* 新建一个句子 */
        sentenceString += newChar;
      }
    }
    lastCh = c.charValue();
  } // end for (Character c : content)
  return sentenceList
```

李劼 Joker Lee
Nov, 2009
jokerleee@gmail.com