方法论 独立日 日语日语 新工具 材料编

分词、句子树、语言处理 100 击

OCTOBER 11, 2015

假如你正打算或刚开始学习日文，将会有一个长期困扰你的问题。假如你学习日文已有一段时日，此问题或许仍时不时给你带来不快。它便是：分词。

英文中的词被空格隔开，几乎就不需要分词。中文里虽没有空格，作为最小表意单位的汉字却是天然分开的，查不了词典至少可以先查字典。而到了日文这儿，事情变得相当麻烦，譬如某日文版《小王子》开头讲“蟒蛇吞猎物”的一段。从中可以看出，日文同中文一样并不惯用空格。里面那些并非汉字的文字叫做假名。要命的是，日文中的最小表意单位往往是由**不确定数目**的假名连在一起构成的。想查词典吧，还得先搞清（或猜中）该从哪儿断开，用哪一串假名去查。猜错个几次，便坏了读书的兴致。

有没有办法自动分词呢？答案是：有。

人们试图用计算机为日语自动分词的尝试至少可以追溯至上世纪 1992 年京都大学对日语分词系统「JUMAN」的开发。1996 年，奈良先端科学技术学院（该学院的特点是强调前沿科研，只设研究生院，不设本科生院）在 JUMAN 的基础上开发了「茶筌（Chasen）」并将其开源。其后，JUMAN 和 Chasen 都不断迭代，并衍生出了如今最为优秀的日语分词系统「和布蕪（Mecab）」。

2010 年，日本国立国语研究所和千叶大学的研究者们从报纸、文学作品及博客中抽取了共计 3200 句日文，用 Mecab 配合他们制作的形态素解析辞书 UniDic 对这些句子进行解析，总体分词准确率超过了 99.4% 。如今，在网络上不难找到基于 Mecab 的在线分词工具（其一、其二），粘入日语文本，获得分词后的结果，其中还包含对汉字读音和单词原形的推测。

然而，使用在线分词工具也好，在自己的电脑上安装配置 Mecab 也好，对只想安静自在地读点日文的学习者来说都远不够友好和便利。这自然是希望运用信息技术改善语言学习者体验的人士（包括笔者在内）所需要面对的课题。不过，除了使用现成的工具，你还可以走得更近些，看看在 Mecab 和 UniDic 的背后和周围正发生着什么。

分词任务从属于一个名为自然语言处理的交叉学科。苹果 iPhone 上的 Siri 等语音助理就是基于自然语言处理技术的产物。2010 年，上述奈良先端科学技术学院的乾健太郎副教授来到地处仙台的日本东北大学，创立了那里的自然语言处理研究室。2011 年，日本自然语言处理界的代表人物辻井润一教授从东京大学退休，任微软亚洲研究院首席研究员。311 震灾后，其门下的冈崎直观研究员赴仙台加入了乾教授的研究室，任副教授。

研究室的新人不一定经历过正统的编程训练。为了带他们入门，并使他们养成良好的编程习惯，冈崎老师编写并公开了一套名为语言处理 100 击的练习。重点是，这套循序渐进的练习全部选用极具实用性的题材！如：

- 第 31 题，对夏目漱石的小说《我是猫》进行自动形态素解析，从中抽出所有动词。
- 第 44 题，对《我是猫》中的内容进行自动句子结构解析，并用树状图将其展现出来。
- 第 70 - 79 题，用机器学习自动判断影评中的句子是在赞还是在踩。
- 第 80 - 99 题，以十万余篇维基百科词条为材料，用向量空间法让机器学习单词语义。

对已有轻度编程经验及少量日语基础的人而言，练“ 100 击”既能入门自然语言处理，又能加强日语及编程技能，一石二鸟。对不具备这些基础的同学而言，想从“ 100 击”获益，恐怕还得配上一套手把手的教程。这样的一套教程目前尚不存在，希望以后会有。

‹ Newer Older ›