关于1995年度自动文摘评测的总结报告

俞士汶　段慧明

关键词：自动文摘、自动文摘评测

一. 准备阶段与出题指导思想

1995年度自动文摘评测工作与机器翻译评测工作是由同一个小组完成的。一些共同的情况已在机器翻译总结报告中谈了，这里不再赘述。本文集中讨论自动文摘评测的一些特有的问题。

自动文摘是一个新的研究课题。95年的自动文摘评测是第一次。除个别专家外，自然语言处理评测小组内的专家对自动文摘的发展现状也不甚了解，因此一切都要从头开始摸索。

按照863智能机专家组办公室的要求，自然语言评测组中负责自动文摘评测的专家也如期发布了自动文摘测试大纲。这个大纲明确了以下几点：

⑴　自动文摘系统应能将原文的主题思想或中心内容自动提取出来

⑵　文摘应具有概况性、客观性、可理解性和可读性。

⑶　系统适用于任意领域、任意体裁文章的难于限定领域、特定体裁的。

⑷　全自动的系统难于人机交互系统。

⑸　文章使用不同于原文的句子难于取自原文的句子。

由于参加评测的几个系统均无限定领域和体裁的要求，我们选择了有关现代汉语语法与语义的两篇文章作为测试用原文。之所以选择这两篇文章，是基于以下考虑：⑴这两篇文章均出自语言学家之手，且是公开发表了的，文笔流畅，没有不合语法的句子。⑵内容是出题者与参评者都可以理解的，容易就主题思想或中心内容达成共识。⑶篇幅适中，一篇约7000字，一篇约5000字。

由于真实文本中有排版标记(标题、小标题可用不同字号、字体排印)，而这些标记对自动文摘可能是有利用价值的。因此，我们在纯文本文件上对标题、二级标题加了标记。这一点在发布测试大纲时已经说明，自动文摘系统可以利用，也可以不利用。

应该坦率地承认，与机器翻译评测不同，我们在看到由机器作出的实际的文摘之前，对如何评价自动文摘，心中是不太有底的。

二. 测试概况

参加1995年自动文摘评测的有三个系统。

2.1 现场测试

现场测试于1995年1月27日下午进行。三个系统都能按照指定的摘取率(20％与5％)自给定的两篇文章中提取出文摘，同时记录了各个系统的执行时间。

为了试验，又选择了一篇两万字以上的文章作为原文，只有单位1的系统可以完成文摘工作，另外两个系统因内存资源不足而失败。

2.2 专家评分

2.2.1 关于文摘情况的调查

将3个系统对两篇文章所作的文摘与原文作了严格对照，发现以下3个情况：

⑴　3个系统都可以按指定的比率从原文中摘取一部分语句。

⑵　抽取的文摘都是原文中的语句，只有单位2的文摘中剔除了一些中文数字。

⑶　三个系统的文摘几乎完全不相重合。

从第⑶点，专家们能够判断，自动文摘是一项非常困难的研究课题，还有许多工作要做。

2.2.2 评比方法与结果

由于3个结果都是原文中的语句，离散程度又如此之高，直接比较三者实在难以评判谁优谁劣，尽管专家们感觉单位2的结果较好。

俞士汶提出了一个试验方法。由三位专家对5000字的那篇文章作文摘，每人从原文中抽取15句，让这15句尽可能表达原文的中心意思。对照3位专家的抽取结果，3者重合的在50％以上，两两重合的达70％。由此进一步验证了，对同一篇文摘分别抽取的两个较好的文摘应该有一定的重合。

接着，又将3个系统的文摘与专家们的文摘进行比较。单位2的文摘的每一个句子至少与一个专家的文摘相重合，与两个以上专家重合的句子也有7句。而单位1与单位3的文摘中的句子与专家们抽取的句子几乎没有重合。

单位2的文摘删去了原文中的一些中文数字，也产生了副作用。当中文数字不是表示序号而是在句子中作代词使用却又被删除时，语句就不通了。不过专家们认为，这个问题容易解决。

采用上述评比方法，专家们一致认定单位2的文摘的质量优于单位1和单位3。不过单位1处理长篇文章的能力比单位2和单位3要强。

三. 自动文摘的进展

自动文摘在信息猛增的时代的重要性是不言而喻的，与MT不同，这个领域的研究开展得较晚，但这次竟有三个系统参加评测，出乎意料之外，说明中文信息处理学界已将这个研究课题提到了日程上，且取得了成绩。

日本经济新闻社的新闻检索系统，也对新闻报导作文摘。但它只是一刀切地将每篇报导的前100字或200字截取下来作为文

摘，因为新闻报导通常在开头点出五个W(什么时间 什么地点 什么人 做什么事及怎样做的)，同日本经济新闻社的做法相比较，参加评测的三个系统的技术确实有了进步。

参加评测的三个系统都是根据文章的外在特征抽取原文中的部分句子作为摘要，属于机械文摘。这样的系统做得好，也是颇有实用价值的。

增加“理解”的深度可以使文摘更确切地反映文章的中心内容，重新生成句子可以使文摘更简洁。但这些工作的技术难度都很大，尚需作更多的研究。

四. 评测工作小结

由于自动文摘是初次评测，第二节中所叙述的评测方法也是第一次采用，这个方法是否合理，能否得到自然语言处理领域的特别是从事自动文摘研究的专家们的认同，都还需要时间。

如果像现在三个自动文摘系统所做的那样，文摘均由原文的部分句子组成，笔者以为机械文摘的自动评测也是可以实现的。笔者正在构想这样的自动评测系统。当然要实现这个构想，还需付出艰苦的努力。

制订测试大纲的清华大学孙茂松副教授和参加评分的各位专家为这次自动文摘评测都贡献了力量，在此致以谢意。

本文以题目“自动文摘评测报告”发表于《计算机世界》报1996年3月25日，评测专版，P183，内容略有改动