大规模中文文本处理

吴立德等 著

复旦大学出版社

内 容 提 要

本书以中文为对象，结合作者的研究工作，系统地介绍了大规模真实文本信息计算机处理的理论和方法。讨论的中心问题是如何从大量文本中迅速有效地找到所需信息。具体内容包括自动分词、自动标注词性和语义、快速句法分析、向量空间表示等方法，以及这些方法在文本自动索引、分类、检索和摘要中的应用。

本书可作为有关大学高年级学生、研究生以及科研和软件开发人员学习大规模文本信息处理的教材或参考书。

前　言

目前，随着 Internet 网和光盘等大容量存储技术的迅速发展，人们已从信息缺乏到信息过多，乃至淹没于大量信息之中。因此，如何迅速、有效地从大量信息中找到所需的信息已成为一个十分迫切需要解决的关键问题。这同时也是一个十分困难的问题。

尽管与多媒体信息相比，文本信息显得比较平凡，但它无疑仍是人们用于记载信息和进行通信的很重要的媒体。同时，它也是人们最熟悉，研究得最多和最成熟的媒体。因此，如何从大量文本信息中迅速有效地找到所需的信息就是既十分重要,又最有可能首先取得突破并获得实际应用的一个领域。著名期刊《BYTE》曾有文章预测：这类软件将继目前的文字处理软件、表处理软件等成为未来最重要的五种软件之一。

本书系统地介绍如何对大规模文本进行索引、分类、检索，甚至自动作出摘要等的理论和方法。所介绍的内容大多是在 80 年代末、90 年代初发展起来的，目前国内尚无全面介绍这些内容的书籍。

国外的研究大多以西方文字，特别是英文为对象的。而中文与英文有着许多差别。因而将国外的有关成果应用于中文并不完全是直接的，而是需要进行许多研究的。本书的介绍以中文为对象，是我们在 863 高科技计划智能计算机主题和国家自然科学基金资助下，进行多年工作的一个小结。

本书由吴立德教授主持编写。第一章概论由吴立德编写，第二章介绍中文信息处理特有的分词，由黄萱菁编写，第三章介绍自动标注方法，由吴立德、季祥编写，第四章介绍快速句法分析方法，由叶丹瑾编写，第五章介绍向量空间表示方法，由王文欣编写，第六章应用，包括自动索引、分类、检索和摘要，由韦雄观、黄萱菁编写。最后由吴立德、黄萱菁负责全书的统稿。

我们希望本书能对有志于大规模文本信息处理这一极有前途领域的研究及开发人员、研究生和大学生有所裨益，使他们能在本书中获得这方面的系统的理论和方法。

我们的研究先后得到国家自然科学基金，863 高科技计划智能计算机主题的资助，也得到清华大学黄昌宁教授、东北大学姚天顺教授、原中软总公司董振东教授和吴蔚天教授，以及许多同行的支持和帮助，在此一并致以深深的谢意。

书中内容，遗漏和错误难免，敬请批评指正。

目　录

第一章 概 论

1.1 自然语言处理

1.1.1 中文信息处理与自然语言处理

中文信息处理，或更一般的，自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，包括中文、英文、俄文、日文、德文、法文等等，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

语言是人类区别其他动物的本质特性。在所有生物中，只有人类才具有语言能力。人类的多种智能都与语言有着密切的关系。人类的逻辑思维以语言为形式，人类的绝大部分知识也是以语言文字的形式记载和流传下来的。因而，它也是人工智能的一个重要，甚至核心部分。

用自然语言与计算机进行通信，这是人们长期以来所追求的。因为它既有明显的实际意义，同时也有重要的理论意义：人们可以用自己最习惯的语言来使用计算机，而无需再花大量的时间和精力去学习不很自然和习惯的各种计算机语言；人们也可通过它进一步了解人类的语言能力和智能的机制。

实现人机间自然语言通信意味着要使计算机既能理解自然语言文本的意义，也能以自然语言文本来表达给定的意图、思想等。前者称为自然语言理解，后者称为自然语言生成。因此，自然语言处理大体包括了自然语言理解和自然语言生成两个部分。历史上对自然语言理解研究得较多，而对自然语言生成研究得较少。但这种状况近年来已有所改变。

无论实现自然语言理解，还是自然语言生成，都远不如人们原来想象的那么简单，而是十分困难的。从目前的理论和技术现状看，通用的、高质量的自然语言处理系统，仍然是较长期的努力目标，但是针对一定应用，具有相当自然语言处理能力的实用系统已经出现，有些已商品化，甚至开始产业化。典型的例子有：种数据库和专家系统的自然语言接口、各种机器翻译系统、全文信息检索系统、自动文摘系统等。

中文信息处理是自然语言处理的一部分，是研究如何用中文与计算机进行通信的。中文是世界上使用人数最多的一种自然语言。它与自然语言处理中研究得最多的的英文和其他西方语言相比，既有许多共同之处，可以相互借鉴，但也有许多差别，需要深入地加以研究和解决。正如陈力为教授所指出的【1.1】：

- 西方语言为拼音文字，而汉语是表意文字；

- 西方的书面语言，词与词之间有空格。而汉语的词与词之间无空格。于是词的切分问题就成了计算机处理汉语的首要问题；
- 西方语言的同音词很少，而汉语的同音词很多；
- 西方语言多有形态变化，而汉语缺少形态变化；
- 汉语的语法尚未形成规范化，而且人们习惯于非规范化的语法。于是语义研究的重要性比西方语言重要得多；
- 汉语的自动（计算机）处理是多学科和跨学科的研究工作，特别需要计算机科学和语言学的密切结合，而且要依靠长期积累的语言学的研究成果。但我国语言学界多着重汉语教学，对象是人，不是机器。因此，对其丰硕的研究成果要经过改造、深化、量化，甚至要从头开始。要清醒地认识到它的艰巨性，要持续不懈地抓下去。

1.1.2 自然语言处理的基本问题

本段中我们将说明为什么自然语言处理，即实现人机间自然语言通信，或实现自然语言理解和自然语言生成是十分困难的。造成困难的根本原因是自然语言文本和对话的各个层次上广泛存在的各种各样的歧义性或多义性（ambiguity）。

一个中文文本从形式上看是由汉字（包括标点符号等）组成的一个字符串。由字可组成词，由词可组成词组，由词组可组成句子，进而由一些句子组成段、节、章、篇。无论在上述的各种层次：字（符）、词、词组、句子、段，……还是在下一层次向上一层次转变中都存在着歧义和多义现象，即形式上一样的一段字符串，在不同的场景或不同的语境下，可以理解成不同的词串、词组串等，并有不同的意义。下面我们来看一些例子。通过它们可以使我们对自然语言处理的困难有更深入的理解。

（1）切词中的歧义【1.2】：即由字到词时的歧义现象，这是中文信息处理中独有的，首先需要解决的问题。

它有两种不同的类型：固有歧义和组合歧义。

固有歧义是指根据不同语境所出现的分词歧义，这里有所谓的“2＋1”和“1＋1”问题。

- “2＋1”问题：一个三字字段，可以是三音节词，也可以是双音节词“＋”单音节词的组合。例如：

  物理学是一门基础科学。　　2＋1＝3

  物理学起来很难。　　2＋1≠3

- “1+1”问题：是指在不同的语境下，有的两字字段，可以是多音节词，也可以是单音节词“＋”单音节词的组合。例如：

  将来的上海将有严重的污染。　　1＋1＝2

  他将来上海。　　1＋1≠2

组合歧义切分是指某个字串，它本身并不组成一个词，但是它在不同语境的条件下，产生不同的组合切分。它也是在不同语境的情况下语义分词的问题。例如：“的确切”。

他的确切地址在这儿。

这块肉的确切得不错。

这是中文处理中非常突出的问题，几乎找不到一个通用的办法来解决这种歧义现象。我们将在第二章中详细讨论这个问题。

（2）词的歧义：这在所有的自然语言中都存在，中文也不例外，甚至更甚。它有两种不同的含义。

一是词性歧义，即一个词有多种词性，也称为兼类。例如下面两个例句中的“学习”一词，一为名词，一为动词。

汉语学习十分重要。

他们努力学习汉语。

另一个是词义的歧义，即一个词有多种词义，如下面两个“红”词，一个表示一种颜色，一个表示“革命的”。

红花。

红军。

这里要指出的是，单独一个“红”字，其词义是有歧义的。但在一定的语境的条件下，歧义实际上是消失了。即“红花”只能是“红色的花”，而不是“革命的花”，同样“红军”也只能是“革命的军队”，而不是“红颜色的军队”。但之所以能做到这一点，是因为我们知道，“花”是无所谓“革命的”等，即有这方面的知识。因此，如果要使计算机系统也能消解这种歧义，它也必须具有这些知识和适当推理的能力。

在更复杂的场合，如：“红旗”，这时两种理解都是可能的：“红颜色的旗子”和“革命的旗帜”，例如：

红旗和绿旗。

高举先烈的红旗。

因此，要消解歧义，就需要考察更多的语境和需要更多的知识，并进行更复杂的推理。尽管这些知识都是常识，其推理过程对人来说也并不复杂和困难。但由于在实用的自然语言处理系统中，涉及的词数以万计，甚至十万计，因此，这些知识的总量是十分可观的，而且还很难进行有效的表示。要在计算机内建立、维护并有效地使用这样的知识库，从而使计算机系统具有充分的消解歧义的能力，还有许多工作要做。

我们下面还要进一步介绍更复杂的歧义现象，它们的处理相应地也要求考察更多的语境，利用更多的知识和更复杂的推理。

（3）结构歧义：即由词组成词组乃至句子时，由于其组成的词或词组间可能存在不同的语法或语义关系而出现的（潜在）歧义现象。

文献【 1.3, 1.4 】中对这一现象有十分精辟的论述，本段中下面的例子均取自文献【 1.4 】。

- “ VP ＋的＋是＋ NP ”型歧义结构：

  其中的 VP 是一个双向动词，“ VP ＋的”作主语，“是＋ NP ”作谓语，整个格式是一个主谓结构。由于主语部分的“ VP ＋的”既可以是施事，又可以是受事，因而产生歧义。例如：

  “反对｜的｜是｜少数人”。

  既可理解为“提反对意见的是少数人”，又可理解为“所反对的是少数人”。

  但并非所有这种结构类型都会产生歧义，而是需要具有一定的条件。条件不满足时不

产生歧义。例如“反对｜的｜是｜战争”。因为“战争”作为无生命的事物，不会主动去反对什么东西，“反对的”不可能是施事，而只能是受事，歧义消失了，即“反对的是战争”只能理解为“被反对的东西是战争”。不过，由此也可看出即使在后一种场合，实际上能消解歧义，但要计算机系统能做到这一点，计算机系统中必须具有有关“战争”等的知识和进行上述推理的能力，这些都是相当困难的。

- “N1 + N2 + N3”型歧义结构：

  这里 N1， N2， N3 都是名词，可能产生两种不同的理解：

  “（N1 + N2）+ N3” 和 “N1 +（N2 + N3）”

  例如：

  北欧｜语言｜研究会。

  既可理解为“北欧语言的研究会”，又可理解为“北欧的语言研究会”。

- “ADJ.+ N1 + N2”型歧义结构：

  这里 ADJ.是形容词， N1， N2 是名词，可能产生两种不同的理解：

  “（ADJ.+ N1）+ N2”和“ADJ.+（N1 + N2）”

  例如：

  小｜学生｜词典。

  既可理解为“小学生用的词典”，又可理解为“小的学生词典”

- “VP + N1 +的+ N2”型歧义结构：

  其中 N1 作为 VP 的宾语，述宾结构“VP + N1”加上“的”之后，作为名词 N2 的定语，整个结构是一个定中结构。但 N1 又可以与“的”结合在一起作 N2 的定语，“N1 +的+ N2”，一个名词词组再作为 VP 的宾语，整个结构是一个述宾结构。因此产生了如下的潜在歧义：

  “（VP + N1 +的）+ N2” 和 “VP +（N1 +的+ N2）”

  例如：

  咬死了｜猎人｜的｜狗。

  可以理解为“咬死了一只猎人的狗”，又可理解为“一只把猎人咬死了的狗。”

- “VP + ADJ.+的+ N”型歧义结构：

  其中的 ADJ.可以作为 VP 的宾语，述宾结构“VP + ADJ.”再加上“的”之后作名词 N 的定语，整个结构是一个定中结构。但是， ADJ.也可以加上“的”之后作名词 N 的定语，“ADJ. +的+ N”整个名词词组作为 VP 的定语，整个结构是一个述宾结构，即有如下潜在的歧义：

  “（VP + ADJ.+的）+ N” 和 “VP +（ADJ.+的+ N）”

  例如：

  喜欢｜干净｜的｜小孩。

  可以理解为“某一个喜欢干净的小孩”，又可理解为“喜欢某一个干净的小孩。”

- “N1 +的+ N2 +和+ N3”型歧义结构：

  由于连词“和”管辖领域的不同，有如下潜在的歧义：

  “（N1 +的+ N2）+和+ N3” 和 “N1 +的+（N2 +和+ N3）”

  例如：

衣服 | 的 | 袖子 | 和 | 口袋。

可以理解为“衣服的（袖子和口袋）”，也可理解为“（衣服的袖子）和口袋。”

- “N1 ＋和＋ N2 ＋的＋ N3 ”型歧义结构：

同样由于连词“和”管辖领域的不同，产生了如下的潜在歧义：

“ N1 ＋和＋（ N2 ＋的＋ N3 ）”和 “（ N1 ＋和＋ N2 ）＋的＋ N3 ”

例如：

桌子 | 和 | 椅子 | 的 | 腿。

既可理解为“桌子和（椅子的腿）”，也可理解为“（桌子和椅子）的腿”。

- “ V ＋ N1 ＋ N2 ”型歧义结构：

其中 N1 和 N2 可以分别作 V 的宾语，形成双宾语结构， N1 又可以作 N2 的定语，组成“ N1 ＋ N2 ”的名词词组作 V 的宾语形成述宾结构。

例如：

赠 | 意大利 | 图书。

既可理解为“赠图书给意大利”，也可理解为“赠送与意大利有关的图书。”

- “数量结构＋ NP1 ＋的＋ NP2 ”型歧义结构：

其中“数量结构”可以限定 NP1，作 NP1 的定语，又可以限定“ NP1 ＋的＋ NP2 ”，因而产生如下的潜在歧义：

“（数量结构＋ NP1 ）＋的＋ NP2 ”和 “数量结构＋（ NP1 ＋的＋ NP2 ）”

例如：

三个 | 学校 | 的 | 实验员。

既可理解为“（三个学校）的实验员”，也可理解为“三个（学校的实验员）。”

- “ VP ＋数量结构＋ NP ”型歧义结构：

其中“数量结构”可以作 VP 的补语，又可作 NP 的定语，因此产生如下的潜在歧义：

“（ VP ＋数量结构）＋ NP ” 和 “ VP ＋（数量结构＋ NP ）”

例如：

发了 | 三天 | 工资。

既可理解为“花三天发工资”，又可理解为“只发了三天的工资”。

我们将在第四章中讨论解决这些歧义的一些方法。

（ 4 ）指代和省略中的歧义：

- 指代中的歧义指的是代词（如我，你，他等）和代词词组（如“这一点”，“那件事”等）所指的事件可能存在歧义，如：

老师给大家讲了一个动人的故事，这使大家很激动。

这里的“这”既可理解为“老师讲故事”这件事，也可以理解为“动人的故事”。

- 省略中的歧义。自然语言中，经常有省略，该省略的不省略反显得罗嗦，但有时也会由此产生可能的歧义。例如：

他说不清楚。

这里既可理解为“他说得不清楚”，也可以理解为“他说他不清楚”的省略。

（ 5 ）更复杂的情形：

- “言外之意”

例如当别人看见你带着手表，却问你：

你知道现在什么时候吗？

他实际上是希望你告诉他现在的时间。

- 由于场景的不同，同样的话可以有不同的意义。例如：同样的一句问话：

你知道到南京路怎么走吗？

（Ⅰ）如果上述问话是一个行人问警察，那是真的问路，希望得到的回答是诸如，“向前走一条马路，再向右转，一直走就到了。”之类的话。

（Ⅱ）设想你和你的朋友，在上海旅游，计划逛南京路。你的朋友正在看地图，这时你讲的上面那句话的意思其实是想知道你的朋友是否真的知道怎么去南京路，而并不是具体的走法。

（Ⅲ）如果上述问话是一个行人问出租车司机，这时的含义是要出租车司机将他送到南京路。

- 由于讲话人的地位不同，同样的话可以有不同的意义。例如，看如下的三组对话。

（Ⅰ）将军：“我想要一个牛肉汉堡。”

随从：“是，长官。”

（Ⅱ）孩子：“我想要一个牛肉汉堡。”

母亲：“今天不行，明天中饭吧。”

（Ⅲ）在某学校的食堂中，

员工1：“我想要一个牛肉汉堡。”

员工2：“喔，我也想要一个，这里的伙食太差了。”

这里第一种情形实际上是一种命令，第二种情形是一种请求，而第三种情形，是一种（很可能是无法实现的）愿望。

从以上的介绍中，不难看出自然语言中存在着大量的歧义现象。一般情况下，它们中的大多数都是可以根据相应的语境和场景的规定而得到解决的。也就是说，从总体上说，并不存在歧义。这也就是我们平时并不感到自然语言歧义，和能用自然语言进行正确交流的原因。但是一方面，我们也看到，为了消解歧义，是需要极其大量的知识和进行推理的。如何将这些知识较完整地加以收集和整理出来；又如何找到合适的形式，将它们存入计算机系统中去；以及如何有效地利用它们来消除歧义，都是工作量极大且十分困难的工作。这不是少数人短时期内可以完成的，还有待长期的、系统的工作。

以上说的是，一个中文文本或一个汉字（含标点符号等）串可能有多个含义。它是自然语言理解中的主要困难和障碍。反过来，一个相同或相近的意义同样可以用多个中文文本或多个汉字串来表示。在词义这一层次上，即所谓的同义词，如

爸爸，父亲，爹爹，老子，爸，爹

诊脉，切脉，把脉，详脉，按

至于在句子这一层次上，则更多了。如：

我借给他书。

他向我借书。

书是我借给他的。

书是他向我借的。

在更高层次上，如篇章上，给出一个题目，可以写出成千上万的作品。

因此，自然语言的形式（字符串）与其意义之间是一种多对多的关系。其实这也正是自然语言的魅力所在。试想每一句话都只有一种意义，每一种意义都只有一种表示方法，那不是太贫乏、太枯燥了吗？幽默没有了，文学更没有了！

但从计算机处理的角度看，我们必须消除歧义，而且有人认为它正是自然语言理解中的中心问题，即要把带有潜在歧义的自然语言输入转换成某种无歧义的计算机内部表示【1.5】。

歧义现象的广泛存在使得消除它们需要大量的知识和推理，这就给基于语言学的方法、基于知识的方法带来了巨大的困难，因而以这些方法为主流的自然语言处理研究几十年来一方面在理论和方法方面取得了很多成就【1.6】，但在能处理大规模真实文本的系统研制方面，成绩并不显著。研制的一些系统大多数是小规模的、研究性的演示系统。

1.2 自然语言处理中的新趋势

1.2.1 新趋势的特征

大约90年代开始，自然语言处理领域发生了巨大的变化。这种变化的两个明显的特征是：

（1）对系统输入，要求研制的自然语言处理系统能处理大规模的真实文本，而不是如以前的研究性系统那样，只能处理很少的词条和典型句子。只有这样，研制的系统才有真正的实用价值。

（2）对系统的输出，鉴于真实地理解自然语言是十分困难的，对系统并不要求能对自然语言文本进行深层的理解，但要能从中抽取有用的信息。例如，对自然语言文本进行自动地提取索引词，过滤，检索，自动提取重要信息，进行自动摘要等等。

在【1.7】中，更进一步将上述特征，细化成如下几种特征：

（1）由句子到文章：以往的自然语言处理系统多数都是只用细心选择过的少量例句来进行实验，而现在要处理数以百万计的真实的文本（即报纸等多种出版物上直接收录的文本）。这种处理深度虽然不够，但针对特定的任务还是有实用价值的。

（2）由完全的语法分析到部分语法分析：由于真实文本的复杂性（其中甚至有不合语法的句子），对所有句子都要求完全的语法分析几乎是不可能的。同时，由于具体文章数量极大，还有处理速度方面的要求，因此，目前的多数系统往往不要求进行完全的分析，而只进行必要的部分分析。

（3）由语言学到统计学：从方法上说，以往的系统主要依赖语言学的理论和方法，而新研制的系统同时还依赖于对大量文本的统计性质分析。统计学的方法在新研制的系统中起了很大作用。

（4）由较窄的领域到很宽的领域：以往的系统往往只能针对某一较窄的领域，例如只适合分析去饭店的场景对话等。而现在的系统则可适用于很宽的领域，甚至是与领域无关的，即系统工作时并不需要用到与特定领域有关的领域知识。

（5）由学院式评价到性能评价：对系统的评价不再是只用少量几个人为设计的典型例子，而是根据系统的应用要求，用真实文本进行较大规模的、客观的、定量的评价。不仅要注意系统的质量，同时也要注意系统的处理速度。

（6）由“故事”到新闻报道：这是针对自然语言处理研究的历史的。历史上曾有许多工作讨论如何深入理解短故事（包括故事中出现的人物，他们的意图等）。现在的系统则要求能分析新闻报道中的多种短消息和长故事。

（7）由原始文章到“排版过的”文章：以前的系统处理的文本一般是“纯”文本，不包含任何“排版”信息，而现在由于要求处理真实文本，而且许多这类文本都是经由字处理系统或排版系统处理过的，因而含有相应的排版信息，就自然提出了这种要求。

同时，由于强调了“大规模”，强调了“真实文本”，下面两方面的基础性工作也得到了重视和加强。

（1）大规模真实语料库的研制。大规模的经过不同深度加工的真实文本的语料库，是研究自然语言统计性质的基础。没有它们，统计方法只能是无源之水。

（2）大规模、信息丰富的词典的编制工作。规模为几万，十几万，甚至几十万词，含有丰富的信息（如包含词的搭配信息）的计算机可用词典对自然语言处理的重要性是很明显的。这一点在介绍歧义性的过程中就可以看得很清楚了。

1.2.2 美国的情况

在美国，自然语言处理研究中产生上述方面的变化的原因主要有两个。

（1）长期以来基于知识的方法，所需知识太多，而且目前的技术短期内还无法将它们收集、整理出来，装入计算机并能有效地应用它们，所以除继续努力外有必要另辟新途。

（2）美国政府的资助部门（如美国国防部的 ARPA 等）的推动。ARPA 早就对语音识别进行了长期的资助，现在他们要求将语音（口头语）方面的研究与语言（书面语）方面的研究集成起来。而在语音识别的研究中，大规模语音库统计方法（如隐马尔可夫模型）等已取得了很大的成功，从而推动了在自然语言处理研究中大规模语料库、统计方法的使用，并取得了许多进展。反过来又提高了美国政府对别的语言和机器翻译的兴趣，也导致由政府资助在宾州大学建立了语言学数据所（Linguistic Data Consortium），专门从事大规模语料库的研制和开发，在新墨西哥州立大学建立了词典研究所（Consortium for Lexical Research），专门从事大容量词典的编制工作。建立这两个机构的主要目的是将过去在这两方面的工作加以集成总结，以形成较完整、统一的词典和语料库。它们一方面可供各地的研究者共同使用，甚至进行国际间的合作（要支付适当费用）；另一方面，也是更重要的，是可为客观、定量地评价和改进实用的各种自然语言处理系统，诸如数据库的自然语言接口、机器翻译系统，从大规模文本中提取的信息系统，自动摘要系统等提供基准。

下面再简单介绍美国的消息理解会议（Message Understanding Conference 简记 MUC）的情况以加深对上述变化的认识。

MUC 是一个开始两年一度，后来一年一度的会议,它得到 ARPA 的资助。这个会议除了像别的学术会议一样交流论文外，还组织对各个单位的系统进行评测。

以 1991 年第三届 MUC 为例。参加单位与系统有 15 家，见表 1.1 。

表 1.1 第三届 MUC 参加单位

缩写	单位全称
ADS	先进决策系统公司
BBN	BBN 系统与技术公司
GE	通用电器研究与发展中心
GTE	GTE 管理系统公司
Hughes	Hughes 研究实验室
ITP	智能文本处理公司
LSI	语言系统公司
MDESC	McDonnell Douglas 电子系统公司
NYU	纽约大学
SRI	Stanford 研究所
Synch/UMD	Synchronetics 公司和马里兰大学
Umass	麻省大学
UNL/USL	内不拉斯加－林肯大学和路易斯西南大学
Unisys	Unisys 先进信息技术中心

规定的系统任务是：从外国（即非美国）通讯社的电讯文本中提取出九个拉丁美洲国家的恐怖活动事件的重要信息。这里的恐怖活动限定为有政治目的带威胁性的活动。但如被攻击的是恐怖主义者、游击队、军队和警察，则不认为是恐怖活动，而认为是游击战，这样实际上增加了任务的难度，即要求系统有区别它们的能力。系统的输入（即消息或电讯文本）的一个例子如下：

TST2-MUC3-0069
BOGOTA, 7 SEP 89 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [MARIBEL OSORIO] [TEXT] MEDELLIN CONTINUES TO LIVE THROUGH A WAVE OF TERROR. FOLLOWING LAST NIGHT'S ATTACK ON A BANK, WHICH CAUSED A LOT OF DAMAGE, A LOAD OF DYNAMITE WAS HURLED AGAINST A POLICE STATION. FORTUNATELY NO ONE WAS HURT. HOWEVER, AT APPROXIMATELY 1700 TODAY A BOMB EXPLODED INSIDE A FAST-FOOD RESTAURANT.

A MEDIUM-SIZED BOMB EXPLODED SHORTLY BEFORE 1700 AT THE PRESTO INSTALLATIONS LOCATED ON [WORDS INDISTINCT] AND PLAZA AVENUE. APPROXIMATELY 35 PEOPLE WERE INSIDE THE RESTAURANT AT THE TIME. A WORKER NOTICED A SUSPICIOUS PACKAGE UNDER A TABLE WHERE MINUTES BEFORE TWO MEN HAD BEEN SEATED. AFTER AN INITIAL MINOR EXPLOSION, THE PACKAGE EXPLODED. THE 35 PEOPLE HAD ALREADY INJURED; HE WAS

THROWN TO THE GROUND BY THE SHOCK WAVE. THE AREA WAS IMMEDIATELY CORDONED OFF BY THE AUTHORITIES WHILE THE OTHER BUSINESSES CLOSED THEIR DOORS. IT IS NOT KNOWN HOW MUCH DAMAGE WAS CAUSED; HOWEVER, MOST OF THE DAMAGE OCCURRED INSIDE THE RESTAURANT. THE MEN WHO LEFT THE BOMB FLED AND THERE ARE NO CLUES AS TO THEIR WHEREABOUTS.

要求的输出是按指定的格式将消息中的每一个恐怖活动事件的有关信息提取出来。例如上述消息的期望输出为表 1.2 。

表 1 2 消息的期望输出

0. MESSAGE ID	TST2-MUC3-0069
1. TEMPLATE ID	1
2. DATE OF INCIDENT	(06 SEP 89) / (06 SEP 89-07 SEP 89)
3. TYPE OF INCIDENT	ATTACK
4 CATEGORY OF INCIDENT	? TERRORIST ACT
5 PERPETRATOR: ID OF INDIV(S)	-
6. PERPETRATOR ID OF ORG(S)	-
7. PERPETRATOR. CONFIDENCE	-
8. PHYSICAL TARGET: ID(S)	"BANK"
9. PHYSICAL TARGET: TOTAL NUM	1
10. PHYSICAL TARGET: TYPE(S)	FINANCIAL": "BANK"
11. HUMAN TARGET: ID(S)	-
12. HUMAN TARGET: TOTAL NUM	-
13. HUMAN TARGET: TYPE(S)	-
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	-
16. LOCATION OF INCIDENT	COLOMBIA: MEDELLIN(CITY)
17. EFFECT ON PHYSICAL TARGET(S)	SOME DAMAGE: "BANK"
18. EFFECT ON HUMAN TARGET(S)	-
0. MESSAGE ID	TST2-MUC3-0069
1. TEMPLATE ID	2
2. DATE OF INCIDENT	07 SEP 89
3. TYPE OF INCIDENT	BOMBING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"TWO MEN" / "MEN"
6. PERPETRATOR: ID OF ORG(S)	-