使用互信息辅助在篇章范围内识别命名实体

郭志立
IBM 中国研究中心
e-mail: guozhili@cn.ibm.com

摘　要：识别命名实体（本文指专有名称、未登录普通词和篇章术语）是中文处理的一个重要问题。本文采用篇章内统计的方法，计算文本文档初步切分后任意两个邻接项(包括词和落单字)的互信息，以此作为判定这两个邻接项是否可能形成新的命名实体的依据。对于可能形成新命名实体的串，继续利用互信息并结合构词法向左右两个方向扩展来确定其边界。最后根据串的内部构造和篇章上下文确定命名实体的类别。

关键词：互信息 专有名称 未登录词

Using Mutual Information in Document Scope to Aid the Identification of Various Named Entites

Abstract: The task of identifying proper names, unknown words and new terms, is an important step in text processing systems. This paper describes a method of using mutual information to collect possible segments as candidates of these three entity types in a document scope; Mutual information values are also utilized to determine a pattern's boundaries. Then the construction and context of each possible entity is examined to determine its type, canonical form and meaning.

Keywords: mutual information. proper name identification, new word identification

1　引言

中文切词是中文处理系统的重要步骤。实用化的系统，比如概念抽取、文本分类、机器翻译，往往要求切词程序在达到一定切词准确率的同时，能够识别篇章内的人名、地名、组织机构名、缩略语、商标品牌、篇章术语、未登录的普通词等等。如果不识别这些命名实体，切词结果中将出现真正的单字词和落单字混杂、普通词和专有名称混杂的情况，对后续的处理造成障碍。

命名实体识别是目前公认的中文自动切词的难点。针对这一难点的解决方法大概可分为两类：一类是专门解决特定的命名实体，另一类是用统计或学习的方法对各种命名实体做一揽子处理。前一类方法一般通过构造相关字表和词表来分别识别类似词和词组的命名实体，例如用中国人名姓氏表、中国人名用字表和常用中国人名词表来识别中国人名【文献 1、2】，用欧美人名译音字表和中国地名字表分别识别译音人名和地名【文献 3】，用大学名称后缀和常用字词表来识别大学名称【文献 4】。后一类方法一般需要对大规模语料进行人工标注，然后从熟语料中进行学习以获得不同颗粒度的知识【文献 5】。同前一类方法

相比，后一类方法只需要加工足够规模的语料，但它获得的知识较难转变为可以让人理解的方式，因此较难接受语言学专家已有的知识。

本文引入互信息来考察连续的字（或词）在篇章内结合的紧密程度，以此搜集可能的模式作为候选的命名实体，并根据模式自身的频率分布、上下文、和其构成计算可信值。最后根据语言知识判定命名实体的类别：人名、地名、组织机构名、缩略语、商标品牌、篇章术语、未登录的普通词。这样既充分利用字词在篇章内的概率分布，又有效地利用了语言知识。

本文的另一优点是在搜集候选的命名实体主要依靠字词的概率分布，不考虑语言知识，而且不必区分命名实体是由落单字构成（大多数的人名、地名、缩略语和未登录普通词），还是由词语组合而成（大多数的为机构名和篇章术语）。

2 文本模式的互信息

2.1. 统计文本片段的频度及其上下文

【文献 6】介绍了一种对大量文本模式进行排序的算法。对 n 个文本模式，该算法的时间复杂度为 O(n*log(n))，空间复杂度为 O(n)。

输入的文档做了篇章结构分析后，首先按照一个通用领域的基本词典进行切词。该词典中所有词条在大规模语料中验证为常用词。在依据词典切词的同时对数字、日期、URL 和电子邮件地址等极高准确率的简单模式做了识别，以减少参与排序的文本模式的数量，加快排序的速度。排序之后的结果如下：

我/国/小将/　常/昊/今天/表现/
/刘/小/光/和/　常/昊/均/在/中盘/
中国/的/…/、/　常/昊/六/段/和/
第/二/大/　城市/釜山

表一：“常/昊”在篇章内的频度及上下文

/的/新/药/"/　糖/脉/康/"/。
预计/明年/"/　糖/脉/康/"/的/产值/
，/"/　糖/脉/康/"/的/推广/
……　……
/化学/合成/降/　糖/药/对/糖尿病/

表二：“糖/脉/康”在篇章内的频度及上下文

从排序后的序列很容易计算任意长度文本模式的频度，比如第一个例子中单个切分单元“常”和“昊”的频度都是 3，由这两个切分单元构成的更长模式串“常昊”的频度也是 3。从排序序列中还可以很容易地获得各文本模式向右的所有延展情况。我把该排序算法略做增强，就可同时容易地获得各文本模式向左的所有延展情况，这样所有文本模式的所有上下文就可以很容易地查询到。

2.2. 计算熵和互信息

本文参照【文献 7】利用互信息构造关联网络的方法，把两个文本模式 s 和 t 的互信息

定义为：

$$MI(s,t) = H(s) + H(t) - H(st)$$

其中的 H(*s*)和 H(*t*)分别是文本模式 s 和 t 在文档中各自的熵，H(*st*)是 s 和 t 邻接出现，也即在文档中拼接为长串时的熵。熵的定义为：

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i))$$

由于字词在文档中的出现都是散列的数值，在计算熵时使用的频率 p(x)采用的是 x 在文档的各个段落中的频率。比如一篇对糖尿病新药“糖脉康”的报道，全文共 8 个自然段，文档初切分之后“脉”作为独立的切分单元共出现 7 次，分布在 5 个段落中的出现次数分别为 1、1、2、2、1，则它的概率分别为 0.14、0.14、0.28、0.28 和 0.14，它的熵为 2.24。这种方法计算出来的熵反映了各切分单元继续组成更长的命名实体的能力。熵越大，则切分单元在该篇章内的分布越均匀，它继续组词的能力就越强。下表列出了这篇例子文档中一些模式的熵值：

糖 2.30	“糖脉康”7 次，“降糖药”2 次，不计在“糖尿病”中的 14 次
脉 2.24	全部出现在“糖脉康”中，7 次
康 2.24	全部出现在“糖脉康”中，7 次
新 1.92	“新药”4 次，“人数正以每年七十五万新患者的速度递增”1 次
药 1.58	“新药”4 次，“降糖药”2 次
新.药 1.50	4 次
脉,康 2.24	全部出现在“糖脉康”中，7 次
糖.脉/康 2.24	7 次

表三

互信息则反映两个文本模式 s 和 t 在篇章内结合的紧密程度。如果 s 和 t 的组合在文档内仅仅出现过一次，则 s 和 t 的互信息等于 0，不能通过计算互信息的方法识别与 s 和 t 有关的术语或命名实体。互信息大，则说明其中的某个模式在构词或语义上较强地关联于另外一个模式；互信息小，则说明两个模式彼此较为互相独立，一般不能形成有价值的命名实体。在“糖脉康”这篇例子文档中，“脉”和“康”的总是同时一起使用，它们的互信息值为 2.24，仅低于“糖”和“脉”的互信息值 2.30。

由于互信息考虑了切词单元在篇章内的频度和概率分布，所以用它作为判定切词单元之间相互关联程度的依据，比使用单纯的频度或条件概率更为可靠。实验结果验证了这一点。

3 命名实体识别算法描述

计算所有切分单元的熵和邻接二元切分单元的互信息之后，本文的命名实体识别算法由以下三个步骤：

(1) 选择合适的阈值，将互信息高于该阈值的串作为候选命名实体的种子；

(2) 比较种子本身的互信息和它向左右两端延展之后更长的模式的互信息，决定命名实体的边界；

(3) 为筛选下来的命名实体计算可信值，猜测其类别。

3.1. 构造候选的两元命名实体

本文采用相对的数值作为互信息的阈值(TMI, Threshold Mutual Information)，以决定哪些邻接的切分单元作为候选的二元命名实体。阈值的大小直接影响筛选后所能保留下来的候选命名实体的数量，从而进一步影响能最终识别出来的命名实体。从平衡算法准确率和召回率的目标出发，本文把 TMI 设定为所有二元邻接切分单元的最大互信息的 0.6 倍。

在筛选可能的二元命名实体时，还需要考虑构词法的一些知识。比如“对/糖尿病”的互信息较高，其原因是“对”字在文档中总是同“糖尿病”一起出现，因此在筛选时可以把含有虚词的模式过滤掉。

3.2. 决定命名实体的边界

为了识别由多个切分单元组成的命名实体，本文将前一步获得的二元模式作为“种子”，向左右两个方向进行延展，通过比较延展前后文本模式的频度和互信息的变化，来决定是否接受延展后更长的模式。如果把延展后的模式接受为候选命名实体，则同时决定是否继续保留原来的短模式。

对于已经筛选出来的模式“t_i / t_{i+1}”，由以下方法判定其右边界：

- 如果向右延展后所有模式的频度都为 1，则原模式“t_i / t_{i+1}”停止向右延展，将其本身保留为候选的命名实体。例如表一中的“常/吴”，向右延展后的三个模式“常/吴/今天/”、“常/吴/均/”和“常/吴/六/”都仅仅出现 1 次，则只把“常/吴”这个模式作为候选的命名实体；
- 如果向右延展后所有模式的频度大于 1，则计算它们的互信息 $MI(t_i t_{i+1}, t_{i+2})$。在表二中，计算 MI("糖/脉", "康")。如果 $MI(t_i t_{i+1}, t_{i+2}) < \lambda_1 MI(t_i, t_{i+1})$，（实验后选取 λ_1=0.4），即延展后模式的互信息比延展前减小了许多，则原模式“t_i / t_{i+1}”停止向右延展，将其本身保留为候选的命名实体；如果 $MI(t_i t_{i+1}, t_{i+2}) > \lambda_2 MI(t_i, t_{i+1})$，（实验后选取 λ_2=0.9），即延展后模式的互信息比延展前减小的幅度很微弱或者增大了，则原模式“t_i / t_{i+1}”取消，保留延展后的模式“$t_i / t_{i+1} / t_{i+2}$”并检测其继续向右延展的情况；如果延展后模式的互信息值介乎于 $\lambda_1 MI(t_i, t_{i+1})$ 和 $\lambda_2 MI(t_i, t_{i+1})$ 之间，则同时保留“t_i / t_{i+1}”和“$t_i / t_{i+1} / t_{i+2}$”两个模式，但停止向右延展。

用类似的比较延展前后互信息值的方法确定模式的左边界。除了单纯依靠互信息值的

大小之外，凡遇到标点符号（特别是顿号）就可明显确定候选命名实体的边界。在实际的实验系统中还引入了一些语法知识，比如，介词、结构助词“的”等虚词也往往可以作为边界。

3.3. 确定命名实体的类别和可信度

从服务于实际应用出发，本文把命名实体分为三大类：专有名称（人名、地名、组织机构名、缩略语、商标品牌）、未登录的普通词、篇章术语。确定命名实体的类别时主要依据以下因素：

- 命名实体是由字还是由词组成：人名、地名、缩略语、未登录的普通词一般由字组成，篇章术语和组织机构名一般由词组成；
- 专有名称常用后缀表：目前有地名后缀、机构名后缀和人称称谓表，分别用于地名、机构名和人名的判别；
- 专有名称常用字表和词表：比如在判定人名时，如果某三字模式的首字是中国人名姓氏表中所收录的字、且另两字也在中国人名名字常用字表中，则可肯定地判定这个模式是中国人名；
- 模式在篇章内上下文中有无线索词出现：比如在判定人名时，如果在篇章内该模式曾经同“先生”或“记者”等表示人的称谓或职业的词共同出现，则该模式是人名的可信度就显著增强，这些称谓或职业词汇就担当了线索的作用。机构名称往往冠以其所属地，所以地名词汇也可担当机构名的线索；
- 缩略语主要针对机构名进行猜测：对于截取全称的一部分（例如“四川富益”是“四川富益电力股份有限公司”前部的两个词）或从全称各部分截取字头（例如“川富电”是“四川富益电力股份有限公司”的字头简称）而形成的两类缩略语，都可以很容易地从第 1 步的统计结果中判定出来，所以本文对这两类缩略语都进行了识别。这些缩略语和全称形成同指关系；
- 不能明确判定类别的模式若全部由单字组成，则被归入未登录的普通词，否则归入篇章术语。篇章术语往往是文档的关键概念。

判定模式所属类别的过程还可辅助确定该模式的可信度。如果某模式有很强的证据支持其属于某一类别，则它作为命名实体的可信度也可因此而大大加强。把此项数值、该模式在文档内的频度、以及模式在延展过程中互信息值的变化情况综合起来，就可初步确定每个模式被识别为命名实体的可信度。

4 实验结果

本文的算法可以从识别命名实体的能力和判定命名实体类别的能力两个方面进行评价。其中提高识别命名实体的准确率和召回率是引入互信息的主要目的。

测试过程使用的是美国宾州大学的中文树库，树库规模为 18 万汉字。其中的专有名称（包括人名、地名和机构名）共 9700 个，没有细分类。本文算法自动识别出来的人名、地名、缩略语和未登录普通词基本上与树库中的“词”对应，篇章术语和组织机构名与树库

中的“短语”对应。

在计算准确率时，自动识别出来的命名实体只要与树库中的任一短语层次匹配，该命名实体即被认为是识别正确。这样得到的准确率为 87%。在计算召回率时，由于互信息的方法只对识别篇章中出现多次的模式有效，所以仅考虑出现多次而且在树库中切分为词(即短语结构的最内层)的命名实体，这样算出的召回率为 92%。考虑全部命名实体则召回率为 42%，结合原有基于字表和规则的识别方法之后，召回率可提高到 80%。

实验系统的速度，在 Pentium III 500 和 512MB 内存的 PC 机上可以达到 200KB/秒，根据基本词典的切分过程大致消耗 2/3 的时间，术语和专有名称的识别大致消耗 1/3 的时间。这个速度可以达到使用化的需求。

5 结论

引入互信息作为构造命名实体候选集和确定命名实体边界的参考依据，并结合已有的利用字表和规则识别命名实体的方法，可以提高识别命名实体的准确率和召回率，从而为实用的语言处理系统提供稳健的基础。如何将这个方法与基于大规模语料的机器学习的方法有机结合起来，有待今后的进一步研究。

参考文献

【1】 孙茂松等：中文姓名的自动辨识。《中文信息学报》1995 年第 2 期。

【2】 季姮等：基于反比概率模型和规则的中文姓名自动辨识系统。《自然语言理解和机器翻译》，黄昌宁、张普主编，清华大学出版社 2001 年。

【3】 谭红叶等：中国地名自动识别方法研究。《计算语言学文集》，黄昌宁、董振东主编，清华大学出版社 1999 年。

【4】 张小衡：中文机构名称的识别与分析。《中文信息学报》1997 年第 4 期。

【5】 Baluja S, Mittal V, Sukthankar R. Applying machine learning for high performance named-entity extraction. Proc. Pacific Association for Computational Linguistics. 1999.

【6】 陈小荷：自动分词中未登录词的一揽子解决方案。《语言文字应用》1999 年第 3 期。

【7】 Butte AJ, Kohane IS. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairewise Entropy Measurements. Pacific Symposium on Biocomputing (PSB 2000).