

基于混合模型的中文命名实体抽取系统的研究与实现

王睿, 张洁, 张由仪, 于禛, 姚天昉

(上海交通大学计算机科学与工程系, 上海市 200030)

摘要: 本文首先总结分析了中文命名实体抽取的研究现状, 认为存在分词、领域和方法三个方面的问题需要解决。随之, 作者提出了相应的解决方案: 利用规则, 对机器分词后的文本进行修正; 提出“群山”模型, 对不同领域制定不同的语言学规则; 统计学方法和语言学方法结合, 对不同命名实体采用不同的方法等。根据实验结果, 本文得出以下结论: 分词的错误将严重影响到最终的抽取结果; 领域规则的应用可以提升抽取效果; 不同方法的有机结合比采用单一方法要好。

关键词: 分词; 领域; 统计学方法; 语言学方法

Research and Implementation on Chinese Name Entity Extraction System based on a Hybrid Model

Wang Rui, Zhang Jie, Zhang Youyi, Yu Zhen, Yao Tianfang

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200030)

Abstract: After summarizing and analyzing the state of the art on Chinese name entity extraction, we emphasize that three fundamental problems of that, including word segmentation, domain, and method, should be solved. Then we brought forward corresponding solutions: using rules to correct errors in texts after word segmentation; establishing specific rules for different domains based on a new “Mountain Chain” model; and combining statistical with linguistic method for treating different kinds of Name Entity separately. According to the experimental results, we can conclude: word segmentation errors will affect on the final results greatly; domain-specific rules were helpful to improve the extraction; and combination of diverse methods was better than a single one did.

作者简介: 王睿, 男, 汉族, 本科, 主要研究领域为信息检索、问答系统; 张洁, 女, 汉族, 本科; 张由仪, 女, 汉族, 本科; 于禛, 女, 汉族, 本科; 姚天昉, 男, 副教授, 硕士生导师, 主要研究领域为计算语言学和语言技术

key words: Word Segmentation; Domain; Statistical Method; Linguistic Method

1 引言

当今社会,随着信息时代的到来,如何从海量文本(非结构化)信息中,快速准确地找到我们需要的信息越来越受到人们的关注。对于信息检索(Information Retrieval)技术的研究已经成为一个很重要的领域^[1]。传统的方法是依赖关键字检索,但是,关键字检索并不能真正准确有效地获取人们所需要的知识。因此,人们期望引入自然语言的技术来解决这个问题。

在引入自然语言技术的信息检索中,命名实体抽取(Name Entity Extraction)又是一个重要的组成部分。因为人们往往把一些信息放在这些命名实体中,再通过这些命名实体之间的关系来表达知识。本文只讨论命名实体抽取这一部分。

结合中文信息处理的一些特征,我们总结了一下中文命名实体抽取需要讨论的问题,有以下三个:

分词的问题

分词是任何中文信息处理的一大难题。当前,对这一问题的处理有三种办法:不分词、利用自动分词系统和人工分词。不进行分词的系统往往出现在早期的一些自然语言处理系统中,效果不是很好;自动分词又可以分两种,一是利用单独的分词系统,二是将分词与后续系统结合;人工分词是为了研究后续技术而暂时避开这一难题的权宜之计。

领域的取舍

现行的自然语言处理系统一般有开放领域(Open Domain)的和限制领域(Domain-specific)两种。前者准确率比较低,难度大,但应用范围广,需求量大;后者准确率高,难度较小,但适用范围有限。

方法的选择

在方法的选择上,主要分为语言学方法和统计学方法两种。前者一般是基于规则(Rule-based)的,利用语法、语义知识、上下文信息进行命名实体的识别;后者则根据人们用词的频率或语境中出现的概率做大量的统计,总结出规律,给出最有可能的结果。

本文结构如下:第二部分介绍本文研究时需要用到的两个相关系统;第三部分、第四部分和第五部分为文章的主要内容,分别讨论了前面提到的中文命名实体抽取的三个问题;第六部分介绍了本系统的架构与实现;第七部分是测试结果与分析;最后是本文的结论以及未来的工作。

2 相关系统与技术

中国科学院计算技术研究所多年研究基础上,耗时一年研制出了基于多层隐马尔可夫模型(multi-layer Hidden Markov Model)的汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)^[2],该系统的功能有:中文分词;词性标注;未登录词识别。分词正确率高达 97.58% (最近的 973 专家组评测结果),基于角色标注的未登录词识别能取得高于 90% 召回率,其中中国人名的识别召回率接近 98%,分词和词性标注处理速度为 31.5KB/s。

本文将此系统作为自动分词的基本系统,同时我们使用了该系统与命名实体抽取有关的词性(part-of-speech)标注。

本文还使用了一个简单的基于支持向量机(Support Vector Machine)的文本分类系统^[3],主要是希望以此来区分出不同领域的文本。为了尽量减少文本分类的错误对命名实体抽取结果的影响,我

们允许用户指定文本领域，以取得更好的抽取效果。

3 分词的影响

早期的一些信息处理系统往往没有进行分词。利用词库匹配和一些上下文规则取得了一些成果，但效果并不是非常好。一是本身词库匹配无法解决分词歧义的情况；二是规则往往只能确定命名实体的一个边界，另外一个边界则很难控制。于是，人工分词语料和机器自动分词系统便应运而生。

当前评测结果最好的自动分词系统是中科院计算所研制开发的汉语词法分析系统 ICTCLAS。由于目前系统的分词的正确率还没有达到 100%，因此我们有必要在此讨论一下分词错误对命名实体抽取的影响。

- 1) 李文/和/在/操场/上…… (正确: 李文和/在/操场/上……)
- 2) 他们/同/上海/、/沈阳/方面/达成/协议/。(正确: 他们/同/上海/、/沈阳/方面/达成/协议/。)
- 3) 德/比赛/的/战况/如下…… (正确: 德比赛/的/战况/如下……)

通过上面几个例子我们可以发现，1)中的“李文和(人名)”、2)中的“上海”和 3)中的“德比赛”都将无法正确识别出来，因此，在这里，分词的错误将直接导致命名实体抽取的错误。而由于后续步骤都是基于分词和标注系统工作的，它们无法知晓分词错误，即使能够抽取正确也只是靠运气猜对的。

我们为了减少分词错误对命名实体抽取的影响，在对文本分词和标注完成之后加入修正规则进行纠错。规则举例如下：

- 4) (n)/和/(prep) -> (n 和)/(prep) (其中n为名词; prep为介词)
- 5) (n 和)/(n) -> (n)/和/(n) (同上)

当然，规则也不可能做到 100%正确。我们应用这些规则进行纠错的同时，新的错误也随之诞生。比如规则 4)如果应用到以下的例子中便会产生错误：

- 6) 李文和于峥…… (“于”可以作为介词也可以作为姓)
- 7) 李文和向荣华…… (同上)

对于规则 5)我们在测试中文体育语料时，有时候姓名会连续出现，即中间没有连词或者标点符号，这时便有可能产生错误。当然，可以暂时认为这是不规范的文本，不作处理；但是，这样的问题还是存在的。

由此，我们发现，制定的修正规则并不能完全修正分词的错误，甚至会带来一些新的错误，这正是自然语言处理复杂的地方。于是，我们尝试在后续步骤当中再加入规则进行二次修正，比如在抽取完命名实体以后利用命名实体的信息进行修正。因为同样的词在不同的上下文中会有不同的分词结果，比如：

- 8) 尤文图斯对抗AC米兰的比赛正在进行…… (“尤文图斯”被识别成组织名称)
- 9) 尤文图斯站在领奖台上…… (“尤文图斯站”被识别成地名)

最终，我们希望通过修正规则来提高命名实体抽取的效果。同样的工作在参考文献^[4]中已经做过，作者用体育领域内的文本做了测试说明分词错误的修正提高了命名实体抽取的效果。我们的实验扩展到其他几个常见领域，并且做了不修正分词、修正分词和人工分词三方面的数据比较，详见 7 测试结果与分析。

4 领域的影响

基于领域的系统往往规则制定得比较细致，充分利用上下文信息，对于领域内的一些特殊词法、句法分析得比较透彻，能够取得不错的效果；而开放领域的系统一般只能采用笼统的规则，当处理到具有专业知识的文本时就显得力不从心。统计学方法的引入似乎回避了领域的问题，将所有领域的问题一起解决，但实际上是将领域的问题转变成为训练语料的选取问题，并没有实质性的提升。

我们先来分析一下领域对命名实体抽取可能产生的影响：

一词多义

这里的“一词多义”指的是一个词在不同领域内可能表示不同的意思，比如：

10) 在对大连中路无法打开局面的情况下，申花队只有采取边路进攻。（“大连中路”指大连足球场的中间区域）

11) “……黄花忠魂，以励来兹”。（“黄花”指“黄花岗七十二烈士”）

12) 黄蜂大胜爵士。（“黄蜂”和“爵士”均为NBA队名）

如果单单利用词库匹配恐怕很难抽取正确，只有“对症下药”，根据不同的语境采用不同的规则才行。也许有人提出把这两条规则都加入规则库，一是会存在效率问题，二是无法解决一些矛盾的情况，比如：

13) 老鹰大战雄鹿。（“老鹰”和“雄鹿”均为NBA队名；也可以均指动物）

语料选取

根据前面的例子 13)，我们知道“老鹰”、“雄鹿”这样的词既可以在动物领域作为动物的名称，也可以在体育或者其他领域作为组织名称的简写；而句式结构上是完全一致的。这样，识别结果就将取决于给予机器训练的语料中哪种情况出现得多。

为了解决上面的问题，我们采用了称为“群山”结构的规则库，如下图所示：

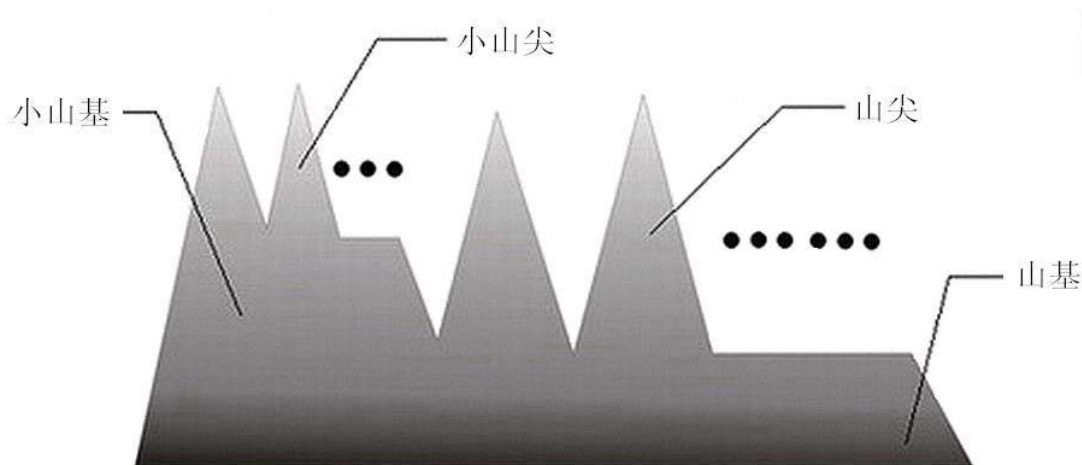


图1 领域解决方案图

图1中，“山基”部分代表的是基础的规则（Baseline），对应于人类思想中比较常见、浅显的知识，不具有专业性，放在任何地方都表达了同一个意思，相互之间没有歧义；“山尖”部分代表的是特殊的规则（Speciality），对应于那些领域内的专业知识，这些词语或用法往往出现在各自的领域中呈现出不同于一般的含义，甚至只出现在那个领域中；“山尖”的大小表示的是领域范围的大小，比如体育领域还可以细分为足球领域、篮球领域等。这时，大的“山尖”就成为“小山基”，代表了体育领域内的基础知识，而“小山尖”就是足球知识、篮球知识等。

这样，我们就把文本分成了许多领域，同时这些领域的范围大小是可以选择的。通过几个基本领域的测试，我们希望看出细分领域带来的抽取效果提升。当然，领域是不可能无限细分下去的，因此我们需要在抽取效果和领域范围上找到一个平衡点。但这一部分已经超出本文讨论的范围，因此7测试结果与分析部分只给出了不同领域范围的抽取效果比较。

5 多种方法的结合

在自然语言处理的发展史上，语言学派和统计学派从两个独立的角度（分别代表了构成计算语言学的两个学科）对自然语言进行了各种方法的处理^[5]。获得的成果大致相当，但过程是交替上升的。而且对于不同的自然语言，两种方法所取得的效果是有所区别的。

然而人类对周围世界的认识是综合应用各种方法的，包括图象、声音、气味等。就算局限在对文本的认识，人类对语言的理解也不拘泥于一种方法，而往往是综合应用词汇、上下文信息、生活经验、别人的建议等各方面信息和知识，才最终在脑中形成一个概念或认识。由此，我们应该考虑将不同的方法进行融合，以测试是否可以取得更好的效果，而不是寄希望于一种方法解决所有的问题。

在前面的分词阶段，我们应用了中科院的分词系统，它是基于多层隐马尔可夫模型的统计学方法。于是，后面的命名实体抽取，我们将采用语言学的方法，对切好词的文本应用语法和语义规则进行处理。

另外，对于不同的命名实体我们将采用不同的方法。这是因为不同的命名实体具有的特点不同，上下文的词法、语法、语义特征也都不同。我们从人类理解自然语言文本的角度出发，尽量让计算机综合应用各种方法，以期获得更好的效果。

针对人名、地名和组织机构名我们分别采用了以下一些方法的结合：

人名 汉字在姓和名中出现的概率^{[6][7]}+上下文规则^[8]

地名 地名库^[9]+上下文规则^[10]

组织机构名 组织机构名^[11]、地名库+已经识别出的地名、人名+上下文规则

最后，我们将应用一些消除歧义的规则，进一步确定各个命名实体。

下面将分类给出一些规则的举例：

人名的识别规则^[12]

14) 老、小等称呼用词+姓

15) 名或者姓名+率领、指责、购买等行为动词

地名的识别规则

16) 在、位于等介词或者已识别地名+其他名词+体育场、足球场等场所类名词

17) 在、位于等介词或者赴、赶赴等动词+地名缩写词

组织结构名的识别规则

18) 已经识别出的地名+宣布、离开等行为动词

19) 人名+其他名词+公司、协会等组织指示词

消除歧义的规则

由于我们在每条抽取规则中加入了权值，这里消除歧义就是比较每个词分别作为不同命名实体的权值的大小（默认值为0），值最大的那种命名实体即为抽取结果。比如：

20) 里昂是法国的一个城市。

21) 里昂是一支实力不凡的球队。

22) 里昂是法国里昂的一支实力不凡的球队。

“里昂”作为地名的规则给予的权值为 50，因为在词法分析层面即可识别出；“里昂”作为组织名称（法国的一个足球俱乐部）给予的权值是 60，因为这需要用到语义信息。对于 20)、21)和 22) 的识别，我们初步定三个“里昂”均为地名，然后看句末（标点符号前）的名词（“球队”和“城市”），以此识别出后两句中“里昂”为组织结构名。

6 系统的架构与实现

本文用到的测试系统^{[13][14]}对文本处理的流程图如下：

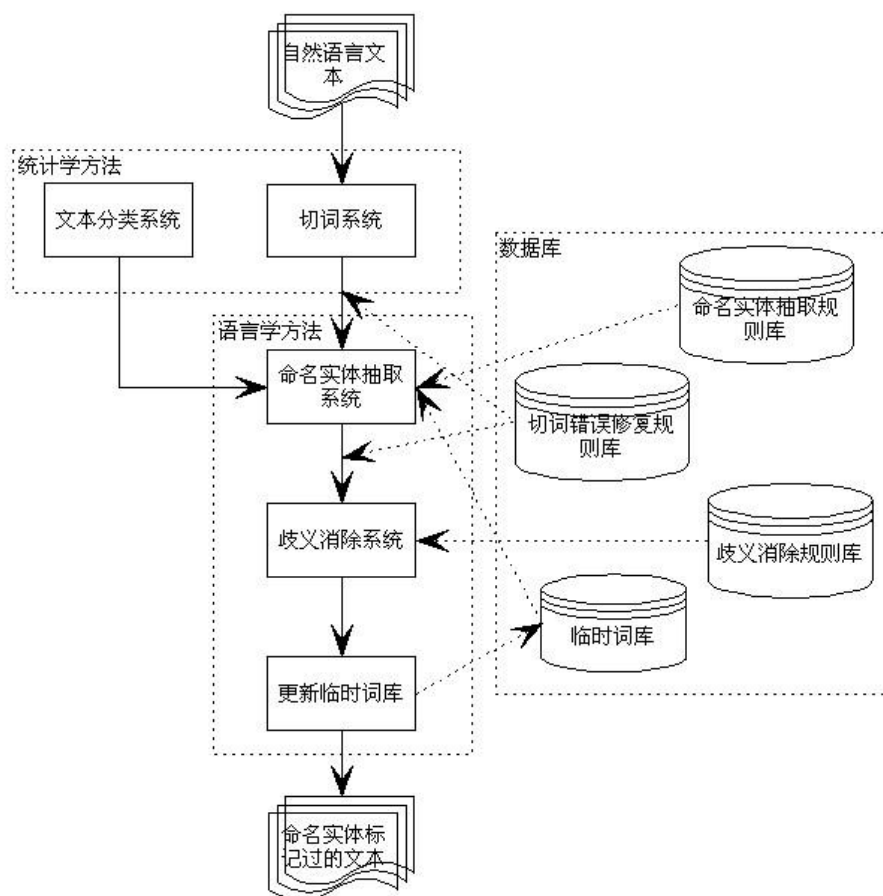


图 2 系统流程图

说明：

1. 临时词库暂时保存系统抽取出的命名实体，主要有两个作用：一是修正部分分词的错误，比如同一个词在不同地方分词结果不同；二是提高效率，因为同一个命名实体多次出现的概率很高。
2. 统计学方法首先使用主要是因为分词方面，统计学方法比语言学方法做得好很多。至于其他的结合方式是否更好，在本文中尚未讨论，我们只是猜想也许迭代式的应用两种方法可能会取得更好的效果。

对于具体实现部分，我们的规则数据结构举例如下：

23) 0100 013 50#T:ns@1/T:n@1/W:俱乐部@1

其中, 0100 是分领域编号, 位数越往后分类越细致, 这里是体育领域; 013 是领域内的编号; 50 是规则的权值; # 是分隔符, 后面表示正式的规则; : 也是分隔符, 后面表示初步识别出的不同类型, 这里 ns 表示地名; T 表示词类; @n 表示 n 个分词单位; W 表示具体的词。

7 测试结果与分析

我们根据上面三部分讨论的问题, 作了三组实验。采用的测试语料来自: <http://cn.yahoo.com>, 其中各个领域的语料都来自于该网站的一级目录。测试语料共计约 5 万字, 其中抽取命名实体共计 2885 个 (1145 个人名、753 个地名以及 987 个组织机构名)。

评价指标

评价的指标包括召回率 (Recall 值)、精确度 (Precision 值) 和 F 值 (F-Measure) 三个指标^[1]。

Recall 值反映了正确识别的命名实体占应当属于这一类的命名实体的比例, 召回率越高, 说明系统能够找出这一类别的命名实体的能力越强。公式如下:

$$Recall = \frac{Correct}{Require}$$

其中, Correct 是系统正确识别出的命名实体数目; Require 是指这一命名实体的总数目。

Precision 值反映了正确分类的问句占分入这一类中的问句的比例, 精确度越高, 说明系统一旦判断出问句的类别就比较准确。定义为:

$$Precision = \frac{Correct}{Fact}$$

其中, Correct 是系统正确识别出的命名实体数目; Fact 是系统总共识别出的命名实体数目。

为了综合评价系统的性能, 通常还计算精确度和召回率的加权几何平均值, 即 F 值, 它的计算公式如下:

$$F - Measure = \frac{(\beta^2 + 1.0) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

其中, β 是精确度和召回率的相对权重, β 等于 1 时, 二者同样重要; β 大于 1 时, 精确度更重要一些; β 小于 1 时, 召回率更重要一些。

本文取 β 为 1, 即两者同样重要。

具体实验情况如下:

分词实验

我们将系统分别应用于机器分词后的文本、经过修改规则修改过的文本以及人工分词后的文本, 然后得到三种命名实体抽取结果的 F 值如下:

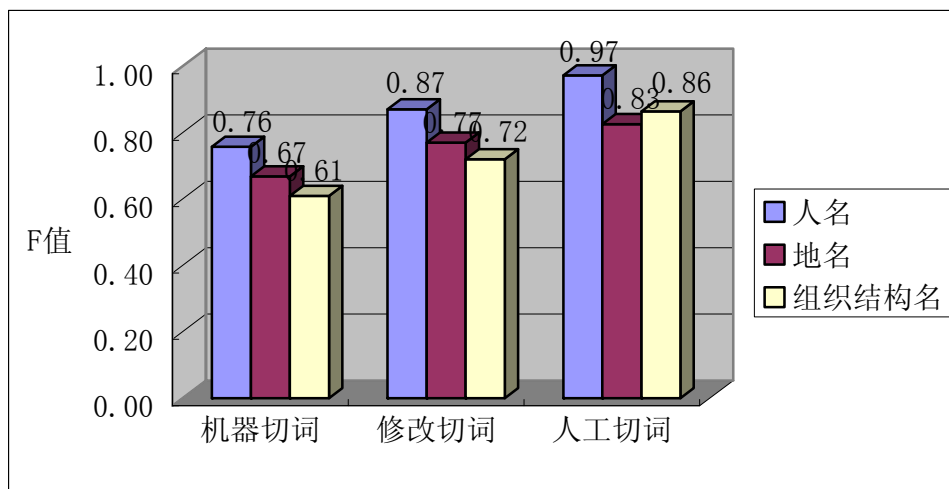


图3 分词实验结果图

从图3中我们可以看出，经过分词规则的修改，各个命名实体抽取的F值得到了一定的提高，但仍然没有人工分词得到的正确率高。错误举例如下：

24) 凯日/nr 曼妙/a 传/n 古德约翰逊/nr (正确: 凯日曼/nr 妙/a 传/v 古德约翰逊/nr)

25) 年薪/n 加/j 奖金/n 达/v 340万/m 美元/q (正确: “加”不应该是缩写词j, 而是动词v)

领域实验 (基于人工分词)

首先，我们比较一下加入领域规则抽取与不加领域规则抽取的结果：

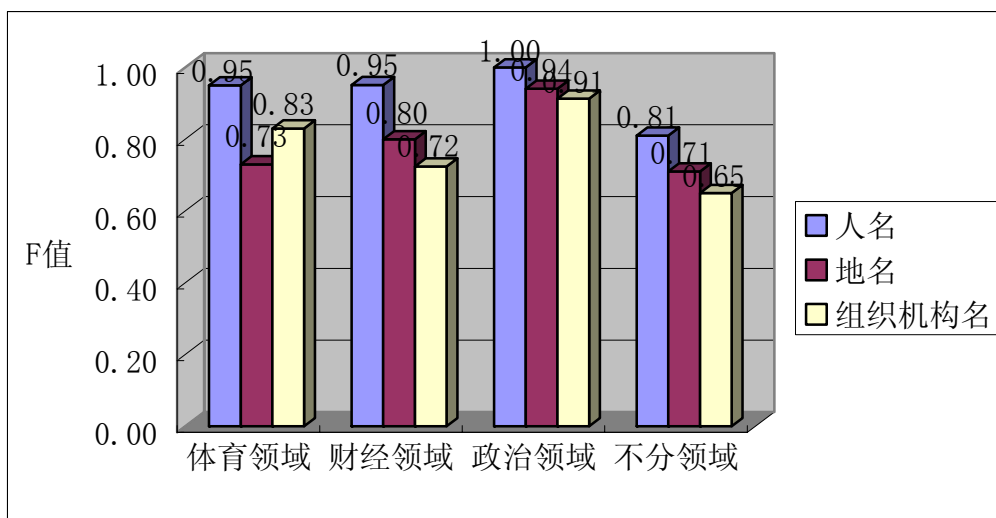


图4 领域实验结果图一

分别将前三个领域的抽取结果与最后一栏进行比较，可以看出加入领域规则提高了三种命名实体抽取的效果；同时，三个领域的提升幅度不尽相同，说明不同领域表现出的“领域性”并不是相同的，这里政治领域最为明显。

为了进一步研究领域划分细致程度的问题，我们又将体育领域分成足球领域与篮球领域，针对性地加入规则，得到以下实验结果：

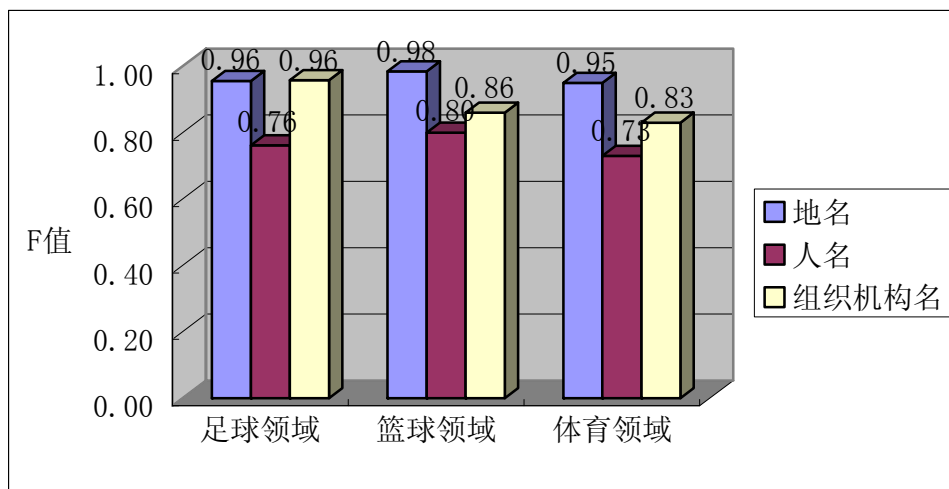


图5 领域实验结果图二

划分以后的抽取结果还是有所提升的, 但并没有前面那么显著, 原因有两个: 一是 F 值越高越难以有所提升; 二是细分领域, 新加入规则有限。但如果按照足球篮球这样的领域细致程度来分领域, 一般的文本分类系统恐怕并不能达到很好的分类效果, 同时划分出的领域数量也将很大。

总体实验

最后, 我们给出所有实验的平均结果:

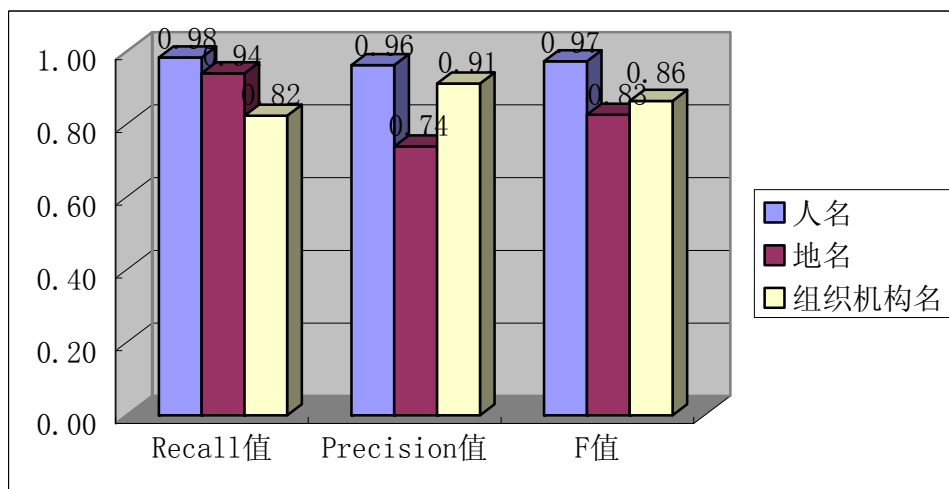


图6 总体实验结果图

人名的抽取效果最好; 地名的 Precision 值较低说明系统找到的错误地名比较多; 组织机构名的 Recall 值比较低, 因为组织机构名的组成比较复杂, 情况也非常多, 歧义也多, 但 Precision 值高说明识别的正确率还是很高的。

8 结论与未来的工作

本文探讨了中文命名实体抽取存在的三个重要问题: 分词问题、领域问题和方法问题。在分析了常见现象后, 用具体实验检测了文章中提出的一些想法, 得到以下结论:

- 分词的错误将直接影响到命名实体抽取的效果；分词与命名实体抽取结合是有助于提升两者的效果的^[15]，因为中文分词是命名实体抽取的基础，而命名实体抽取的结果又可以修正分词的错误。参考文献^[16]中也提到用语义信息提升分词的效果。
- 确定领域，应用“群山”模型结构的规则库是有助于提升命名实体抽取效果的。
- 不同命名实体抽取所应用的方法不应当局限于一种；同时，混合应用统计学和语言学的方法可以取得比单一方法更好的效果。

结合以上结论，我们认为可以在以下两个方面进行进一步研究：一是分词与命名实体抽取交替迭代式地进行如何实现；二是在抽取效果与领域细致程度中寻求一个平衡点，找到切实可行效果也令人满意的划分方法。再进一步，对不同方法的不同组合进行探索和讨论，最终找出最合适的方法组合，期待能最终解决中文命名实体抽取的问题。

参考文献：

- [1] 李保利、陈玉忠、俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10): 1-5.
LI Baoli, CHEN Yuzhong, YU Shiwen. research on information extraction: a survey[J]. Computer Engineering and Applications, 2003, 39 (10): 1-5.
- [2] ZHANG Huaping, LIU Qun, CHENG Xueqi, et al. Chinese lexical analysis using hierarchical hidden Markov model. Sapporo© Japan, July, 2003, 63-70.
- [3] Thorsten Joachims. text categorization with support vector machines: learning with many relevant features©. Dortmund, 27. November, 1997.
- [4] YAO Tianfang, Ding Wei and Gregor Erbach. correcting word segmentation and part-of-speech tagging errors for Chinese named entity recognition. in Günter Hommel and SHENG Huanye (Eds.): the internet challenge: technology and applications. Kluwer Academic Publishers. Dordrecht, the Netherlands. Oct. 2002.
- [5] Ron Cole, et al. survey of the state of the art in human language technology (web edition). Cambridge University Press and Giardini, 1997.
- [6] 张锋, 樊孝忠, 许云. 基于统计的中文姓名识别方法研究[J]. 计算机工程与应用, 2004, 10: 53-54.
ZHANG Feng, FAN Xiaozhong, XU yun. the research on Chinese names recognition method based on statistics[J]. Computer Engineering and Applications, 2004.10, 53-54.
- [7] 刘秉伟, 黄萱菁, 郭以昆等. 基于统计方法的中文姓名识别[J]. 中文信息学报, Vol. 14, No. 3: 16-24, 36.
LIU Bingwei, HUANG Xuanjing, GUO Yikun, et al. statistical Chinese person names identification[J]. Journal of Chinese Information Processing, Vol.14, No.3: 16-24, 36.
- [8] 孙茂松, 黄昌宁, 高海燕等. 中文姓名的自动辨识[J]. 中文信息学报, 第9卷, 第2期: 16-27.
SUN Maosong, HUANG Changning, GAO Haiyan, et al. identifying Chinese names in unrestricted texts[J]. Journal of Chinese Information Processing, Vol.9, No.2: 16-27.
- [9] 黄德根, 岳广玲, 杨元生. 基于统计的中文地名识别[J]. 中文信息学报, Vol.17, No.2: 36-41.
HUANG Degen, YUE Guangling, YANG Yuansheng. identification of Chinese place names based on statistics[J]. Journal of Chinese Information Processing, Vol.17, No.2: 36-41.
- [10] 谭红叶, 郑家恒, 刘开瑛. 基于变换的中国地名自动识别研究[J]. 软件学报, Vo l. 12, No. 11: 1608-1613.
TAN Hongye, ZHENG Jiaheng, LIU Kaiying. research on method of automatic recognition of Chinese place name based on transformation[J]. Journal of Software, Vol.12, No.11: 1608-1613.

- [11] 郑家恒, 张辉. 基于 HMM 的中国组织机构名自动识别[J]. 计算机应用, Vol. 22, No. 11: 1-2, 25.
ZHENG Jiaheng, ZHANG Hui. recognition of HMM-based Chinese institution terms[J]. Computer Applications, Vol.22, No.11: 1-2, 25.
- [12] 张跃, 姚天顺. 基于结合性自动识别中文姓名[J]. 小型微型计算机系统, Vol.18, No. 10: 43-48.
ZHANG Yue, YAO Tianshun. combination based for distinguishing Chinese name automatically[J]. Mini-Micro Systems, Vol.18, No.10: 43-48.
- [13] 谭红叶, 郑家恒, 刘开瑛. 中国地名自动识别系统的设计与实现[J]. 计算机工程, 2002 年 8 月: 128-129, 270.
TAN Hongye, ZHENG Jiaheng, LIU Kaiying. design and realization of Chinese place name automatic recognition system[J]. Computer Engineering, 2002.8: 128-129, 270.
- [14] 张辉, 徐健. 中国组织机构名自动识别系统的设计与实现[J]. 电脑开发与应用, 第 15 卷, 第 1 期: 5-9.
ZHANG Hui, XU Jian. design and implementation of automatic recognition method system of Chinese institution terms[J]. Computer Development & Applications, Vol.15, No.1: 5-9.
- [15] SUN Jian, GAO Jianfeng, ZHANG Lei, et al. Chinese named entity identification using class-based language model. Coling2002©.
- [16] K. Liu. 2001. research of automatic Chinese word segmentation. In Proc. of International Workshop ILT&CIP 2001 on Innovative Language Technology and Chinese Information Processing. Science Press, Beijing, China.