中文人称名词短语单复数自动识别

郎 君[1] 秦 兵[1] 刘 挺[1] 李正华[1] 李 生[1]

摘 要 名词短语的单复数信息在共指消解中是必不可少的特征. 与英语不同, 中文属于汉藏语系, 名词本身不能明显体现单复数信息, 需要借助其所在的名词短语来进行体现. 本文在自动内容抽取 (Automatic content extraction, ACE) 语料上抽取得到人称名词短语的单复数信息, 分别采用了基于规则和机器学习的方法来进行人称名词短语的单复数自动识别. 基于规则的方法, 在一些知识资源的基础上定义了规则模板库, 每条规则采用槽和槽值的方法来进行体现; 机器学习方法采用最大熵模型组合考察了词形、词性、词义、数量关系等特征. 两种方法分别达到了 48.24 % 和 87.48 % 的正确率. 实验结果显示, 基于规则的方法能够保证精确率而不能保证召回率, 机器学习的方法可以更好地完成单复数信息的识别任务.

关键词 人称名词短语, 单复数, 机器学习
中图分类号 TP391

Number Type Recognition of Chinese Personal Noun Phrase

LANG Jun[1] QIN Bing[1] LIU Ting[1] LI Zheng-Hua[1] LI Sheng[1]

Abstract Number type is absolutely a necessary feature for co-reference resolution. Different from English, Chinese, belonging to Sino-Tibetan language family, cannot reflect number information directly by nouns themselves. However, the problem can be tackled by virtue of noun phrase. This paper presents two methods of number type recognition of Chinese personal noun phrase and their tests on ACE 2005 corpus. The first one is rule-based, which defines the template rules based on some knowledge resources, employing some slots and slot values. The other one is machine learning method, with maximum entropy model on features of word, pos, word sense, and quantitative relation. The two methods reached total accuracies of 48.24 % and 87.48 %, respectively. Experimental results showed that the rule based method could ensure the precision but the recall, while the machine learning method managed the number type recognition task.

Key words Personal noun phrase, number type, machine learning

名词短语的共指消解是指在篇章中判断哪些名词短语指向现实世界中的同一实体[1]. 它是自然语言处理中的一项核心任务, 在很多领域中都有重要应用, 例如自动问答、机器翻译、自动文摘和命名实体识别等. 近年来, 美国国家标准技术研究所 (NIST) 组织的自动内容抽取 (Automatic content extraction, ACE) 系列评测中, 名词短语的共指消解一直作为评测任务之一, 吸引了越来越多的研究人员的关注.

在名词短语的共指消解研究中, 单复数特征具有非常重要的作用[2−3]. 最早的基于句法分析消解方法中, 单复数信息被作为基本的属性来完成先行语的过滤和筛选[4−5]. 例如, “一群人来到张华的家中, 说是他请他们请吃饭.” 这个句子中在消解 “他们” 时, 前面的 “张华” 和 “他” 都是单数, “一群人” 是复数; 从而 “张华” 和 “他” 被过滤掉, 而 “一群人” 被初步筛选出来. 目前共指消解的主流是基于语料库统计的方法, 单复数信息总是被用来作为一个必不可少的特征[6−7], 例如候选先行语和当前名词短语的单复数是否一致等.

张黎指出: 在以英语为代表的印欧语系中, 语言具有数量范畴, 数的形态标志一方面表示事物的数量特征, 同时也是与语法相关联的, 在一些情况下可以替代一些对名词所指事物的数量的陈述. 而属于汉藏语系的中文, 其名词的数与句法范畴无关, 不承担结构功能, 只承担表示事物数量特征的表义功能. 因此可以说, 数在印欧语言中是语法范畴, 在汉语中属语用范畴. 数的表义作用在汉语中通过多种多样的句法结构关系和语用手段显示出来, 使用数量结构直接作名词的定语限定其数的属性就是其中一种, 还可以用其他手段显示出来[8].

英语上指代消解的研究中对于单复数信息可以用一些简单的规则获得. 例如单词末尾是否含有 -s, -es, 以及查看一些特定的单复数词汇枚举列表等. 中文指代消解中, 很多研究工作都采用手工标注名词短语的单复数信息[9−10]. 手工标注的方法对于完成特定的实验有所帮助, 但对完成一个实用的指代消解系统却是不能采用的.

收稿日期 2007-05-09 收修改稿日期 2007-09-20
Received May 09, 2007; in revised form September 20, 2007
国家自然科学基金 (60575042, 60503072), 国家高技术研究发展计划 (863 计划) (2006AA01Z145) 资助
Supported by National Natural Science Foundation of China (60575042, 60503072), National High Technology Research and Development Program of China (863 Program) (2006AA01Z145)
1. 哈尔滨工业大学信息检索研究室 哈尔滨 150001
1. Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001
DOI: 10.3724/SP.J.1004.2008.00972

近来, Wang 将单复数信息分为单数和复数两类, 采用如下 5 条规则来自动获取[11]:

1) 复数: 机构名, 地名;

2) 单数-I: 人名;

3) 单数-II: 一些特殊词语, 如头衔“总统”等;

4) 单数-III: 一些符合下列模式的短语: (这 | 那 | 该 | 某 | 一) + [位 | 名] + X + 人称名词;

5) 其他情况都被视为复数.

类似地, 李国臣等也采用了基于规则的方法来进行识别[12]. 他们将单复数属性主要分为单数、复数和无单复数三类. 这一属性根据明显的单复数搭配词语进行识别, 例如含有 “们、一大群、许多、每个” 等, 表示并列关系的名词短语归入复数类别, 其他无法识别的划入无单复数类别, 如 “职工和学生”、“人们” 和 “有些幸存者” 属性为复数类. 其他为无单复数类.

本文借鉴前人的工作, 专门研究了如何对中文人称名词短语进行单复数类型的自动识别. 对于中文, 依靠一些搭配模板能够确定一定数量的单复数信息, 结合多种知识库构建了尽量充分的槽模板和相关的模板槽值. 在 ACE 2005 评测语料上抽取得到了单复数类型的语料库. 在这个语料库的基础上尝试采用机器学习的方法, 将单复数类型判定看成典型的分类问题, 结合名词短语的词形、词性、词义、数量关系等多种特征, 自动构建单复数类型的识别器. 通过实验对比发现, 人工总结规则很难保证规则的完备性, 而机器学习的方法能够获得较高的正确率.

本文内容组织如下: 第 1 节是人称名词短语单复数信息的相关定义, 详细说明了单复数信息的类别和常见情况; 第 2 节介绍如何根据多种知识源构建尽量完备的单复数识别模板; 第 3 节详细介绍单复数信息自动识别的机器学习框架以及需要考虑的各种特征; 第 4 节是实验情况以及结果分析; 最后是结论和展望.

1 基本定义

ACE 任务之一的实体检测与识别 (Entity detection and recognition, EDR) 需要识别出文章中的实体及其类型, 并且将表示现实世界同一事物的实体合并到一起. 实体 (Entity) 被定义为现实世界中存在的对象或对象的集合, 需要识别的实体共有七类[1], 包括设施名 (Facility, FAC)、行政区划 (Geopolitical entity, GPE)、人名 (Person, PER) 等. 对于每类实体需要识别出相应的子类别. 其中 PER 的子类别从 ACE2005 开始被确定为: 单数 (Individual, Indiv), 复数 (Group), 不确定 (Indeterminate, Indet). ACE 的定义中每个实体包含了在文章中多次出现的表示同一事物的出现 (Mention), 每个 Mention 又包含 Extent 和 Head, 其中 Extent 表示名词短语, Head 表示当前名词短语的核心词. 我们把 PER 类型的 Entity 下的各个 Mention 称为人称名词短语. 如表 1 所示, 该人称实体中含有两个单数的 Mention.

ACE 2005 语料中的单复数信息全是人工标注的, 表 2 是一些标注的单数、复数、不确定的例子.

表 1 ACE 中实体示例

Table 1 An example of ACE entity

```
<entity ID="CBS20001006.1000.0874-E6" TYPE="PER" SUBTYPE="Individual" CLASS="SPC">
    <entity_mention ID="CBS20001006.1000.0874-E6-30" TYPE="NAM" LDCTYPE="NAM" LDCATR="FALSE">
        <extent><charseq START="342" END="343">普京</charseq></extent>
        <head><charseq START="342" END="343">普京</charseq></head>
    </entity_mention>
    <entity_mention ID="CBS20001006.1000.0874-E6-31" TYPE="NOM" LDCTYPE="NOM" LDCATR="TRUE">
        <extent><charseq START="337" END="341">俄罗斯总统</charseq></extent>
        <head><charseq START="340" END="341">总统</charseq></head>
    </entity_mention>
    <entity_attributes>
        <name NAME="普京"><charseq START="342" END="343">普京</charseq></name>
    </entity_attributes>
</entity>
```

[1] http://www.nist.gov/speech/tests/ace/ace07/doc/ace07-evalplan.v1.3a.pdf 中有详细说明

表 2 ACE 中单复数类型示例

Table 2 Examples of ACE number types

类型	Head	Extent
单数	普京	普京
	市长	台北市长
	市民	一名台北市民
	他	他
复数	歌手	蔡振南以及陈冠华等歌手
	企业家	德国 11 名企业家
	民族	民族
	他们	他们
不确定	人	有人
	领导	工会领导
	人士	不明人士
	别人	别人

2 构建规则模板库

在自然语言处理中, 很多系统都是基于规则的方法. 对于单复数识别而言, 总结规则是很困难的, 因为需要考虑的情况太多. 我们在之前研究[11−12]的基础上, 参考了相关词典和知识资源[13−15], 采用槽模板的方法总结制定了 9 条启发式规则, 如表 3 所示 (表中 “( )” 表示出现一个即可,“[ ]” 表示出现或者不出现,“{ }” 表示槽).

表 3 单复数识别的 9 条规则

Table 3 Nine rules for number type recognition

编号	类型	槽模板
1	单数	Head 前包含: ({这_那_某}、{一}、{序数词})+[{个体量词}]
2	复数	Head 前包含: {集合量词}
3	复数	Head 前包含: {些 _ 多}、{大量}
4	复数	Head 前最后一个数词大于一
5	单数	Head 是: {单独人名}
6	单数	Head 是: {我_你_他_她_谁_自己_称人_自称}
7	复数	Head 以 {们_等_辈_行}结尾
8	不确定	Head 是: {人家}
9	不确定	其他

表 3 中的前 4 条规则都是在考察 Extent 中 Head 前面部分的内容, 随后 4 条规则考察 Head 部分, 最后 1 条是把前 8 条规则不能判断的情况都默认为不确定类型. 9 条规则在进行单复数判断时是依次执行的, 即只有当前规则不能判断了, 才启用下一条规则.

表 3 中涉及到的相关槽值如下:

{这 _ 那 _ 某}=这、此、那、哪、彼、其、该、某、某某、别、另、最.

{一} =一、壹、1、幺.

{序数词} =元、首、冠、魁、初、头、长、伯、次、其次、亚、仲、叔、季、第 +(数词).

{个体量词}=个、名、位、员、号、只、口、任、届.

{集合量词}=族、群、帮、帮子、伙、批、班、拨、支、股、派、列、排、簇、对、双、种、类、众、家、民、层、阶级、党、委、会、人马、队、组.

{些 _ 多} =些、多.

{大量}=别的、另外的、其他、其它、其他的、其它的、全部、全体、一切、所有、全、各、各位、万千、大量、大批、巨量、广大、浩大、无数、有的是、成千上万、不少、成百上千、上百、为数不少、不在少数、不计其数.

{我 _ 你 _ 他 _ 她 _ 谁 _ 自己 _ 称人 _ 自称}=我、咱、俺、本人、个人; 你、您; 他、她、彼、伊、伊人; 自己、自家、自个儿、自身、自我、自、独自; 谁、谁个、哪个、何人; 同志、阁下、先生、小姐、太太; 鄙人、小弟、兄弟、不才、小人、老朽.

{们 _ 等 _ 辈 _ 行} =们、等、辈、行、之流、之辈.

{人家} = 人家、别人、他人、旁人、人.

3 机器学习框架以及相关特征说明

人工总结规则的方法在早期的自然语言处理中得到了大量的应用, 一个典型的例子就是机器翻译. 但是实际经验告诉我们, 人工编写规则是一件非常繁琐的事情, 需要大量的人力和时间, 而且总结出来的规则容易出现相互矛盾的现象, 同时规则之间不同的执行顺序也容易导致出现不同的判断结果.

在人工标注好单复数类型的语料上, 可以尝试采用机器学习方法来自动学习相关的经验. 单复数识别是一个典型的分类问题, 可以采用的机器学习模型有很多, 例如贝叶斯、神经网络、决策树、SVM、最大熵等. 针对各种自然语言处理任务, 现在普遍认为在这些模型中效果最好的是 SVM 和最大熵. 对于具体的问题和特征空间, 经过多次尝试不同的 Kernel 函数和其他参数设置, SVM 往往能够获得最好的效果. 但是对于单复数类型识别, 本文选用最大熵模型, 主要出于以下两点考虑:

1) 最大熵方法具有控制细微结果, 不作未经验证的假设, 模型简单、易于理解等特点[16];

2) 该算法能够使得我们更多地关注在实验中选用各种不同特征的组合上, 而不是算法本身.

对于每个 Mention 的 Extent 和 Head, 本文采用语言技术平台 (Language technology platform, LTP)[2]处理得到相应的分词、词性标注、词义消歧、依存句法分析等结果. LTP 是一套基于 XML 的开放式中文语言处理平台, 目前集成了包括词法、词义、句法、语义、篇章分析等 10 项中文处理核心技术, 可以方便地对文本进行各种自然语言处理[17]. 基于这些处理结果, 本文选取了各种特征, 详细说明如下.

词语特征: 对于单复数识别词形、词性、词义都是值得考虑的对象. 一些特殊词语, 例如上节中的 {集合量词}、{大量}等都可以在词形上直接判断. 事实上, 上节中一些槽中的词语的词性很多都是相同的, 可以加入词性特征覆盖更多词语. 词义是对当前 Extent 名词短语中各个词语进行词义消歧后得到的《同义词词林》[15] 中的词义代码, 如表 4 所示实例.

表 4　Extent 词义消歧结果示例

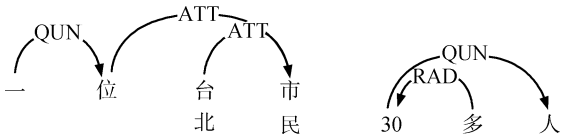
Table 4　The example of Extent′s word sense disambiguation result

词语	词义代码	词义解释
30	Null	未知
多	Dn05	半_概数_若干
位	Dn08	数量单位
各	Ed61	这个_那个_某个_各个_其他_何
党派	Di10	团体_派别
立法委员	Ae02	工人

Head 特征: 核心词在单复数识别中也能起到一定的作用, 很多时候直接判断核心词就能得到相应的单复数类别. 例如“我们”、“大家”、“他”、“他们”等. 经过统计发现, ACE 语料中 92.5 % 的 Head 包含单个词语, 98.3 % 的 Head 包含一个或者两个词语. 为了能够覆盖 Head 中绝大多数的词语, 这里选取 Head 中的前两个词语作为相关特征, 对于只有一个词语的 Head, 对应的第二个词语为“空” (Null). 每个词语都需要提取相应的词语特征.

Extent 特征: 在一些较长的 Mention 中, 除去 Head, Extent 常常包含一些特殊的数量模板型词汇, 如表 3 中前 4 条规则所示. 统计 ACE 中抽取到的单复数语料, 发现 39.3 % 的 Mention 中 Head 之前还有词语. 考虑到如表 3 中所示的数量模板涵盖的词汇数量不超过三个, 这里选取 Extent 的前四个词语作为特征库考虑的对象.

Qun 特征: 中文名词不具有直接表示单复数信息的语义功能, 经常需要借助一些数量结构来进行体现. 在 Extent 中如果能够直接抽取出相应的数量关系的成分, 就能够很方便地进行单复数信息的识别. 汉语依存句法分析可以对句子进行详细的分析[18], 能够自动获得数量关系 (Quantitive relation, Qun), 而且数量关系的识别精确率和召回率较高, 分别达到了 96.17 % 和 94.75 %. 两个具体的数量关系如图 1 所示.



(a) 实例 1　(b) 实例 2

(a) Example 1　(b) Example 2

图 1　数量依存关系示例

Fig. 1　The examples of Qun dependency relation

Len 特征: 在语料中, 统计发现 Extent 的词语数量较少的 Mention 有时单复数信息不好识别. 这个问题对于不确定类型比较突出, 例如 (Head = 人, Extent = 人); 而词数较多的 Mention 的单复数信息相对容易识别. 为此, 引入 Extent 的长度特征 (Length, Len), 即 Extent 包含的词语个数.

上述特征可以体现在一个三维特征空间中, 如图 2 所示. 其中 x、y 轴上分别至少使用 1 个特征, 即分别有 7 种特征使用组合; z 轴上 Len 使用与否均可. 所以, 在整个特征空间中共有 $7 \times 7 \times 2 = 98$ 种组合.

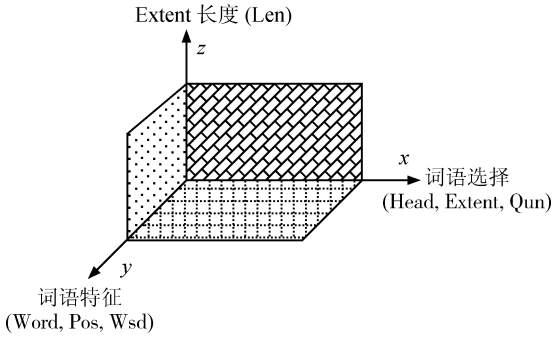


图 2　机器学习框架中的三维特征空间

Fig. 2　Three dimensions of feature space for machine learning framework

4　实验设计及结果分析

4.1　语料说明及评价指标

ACE 2005 年的训练语料中共有 6 501 个 PER 实体, 每个 Entity 下的 Mention 都对应于当前所在

[2]http://ir.hit.edu.cn/demo/ltp/

Entity 的单复数信息. 抽取全部的语料, 得到了如表 5 所示的单复数信息语料, 其中单数类型所占比例最大, 达到 61.44 %.

表 5 ACE 上抽取得到的语料分布

Table 5 The distribution of ACE number type corpus

Indiv	Group	Indet	Total
8 895	4 691	891	14 477

对于单复数识别的实验结果, 本文采用整体正确率 (Total accuracy, TA) 来衡量总体的实验效果, 采用精确率 (Precision) 和召回率 (Recall) 来衡量每种单复数类别以及每条规则的识别效果. 三种指标定义如下:

$$\text{TA} = \frac{\text{单复数类型正确识别的 Mention 个数}}{\text{全部 Mention 的个数}} \quad (1)$$

$$\text{Precision} = \frac{\text{单复数类型正确识别的 Mention 个数}}{\text{试图识别类型的 Mention 个数}} \quad (2)$$

$$\text{Recall} = \frac{\text{单复数类型正确识别的 Mention 个数}}{\text{语料中对应类型的 Mention 个数}} \quad (3)$$

4.2 基于总结规则的方法

我们采用第 2 节中确定的规则以及相应的槽值进行了详细的实验, 结果如表 6 和表 7 所示. 表 6 中采用联立矩阵展示了各种类型的详细识别情况. 其中 "r-" 表示实验判断得到的结果; "g-" 表示标准语料中的真实情况; 对应的 Recall 表示当前行对应单复数类型的召回率; Precision 表示当前列对应的单复数类型的精确率; 右下角的 TA 表示整体正确率. 表 7 中, 每行数据表示在执行过程中当前规则覆盖的实例个数; 数据前面的 "*" 表示当前规则判断默认得到的类别.

表 6 基于规则的方法各种类别的实验结果

Table 6 The experimental result of all kinds of number types by rule-based method

	r-Indiv	r-Group	r-Indet	Recall
g-Indiv	4 402	223	4 270	49.49 %
g-Group	166	1 744	2 781	37.18 %
g-Indet	27	27	837	93.94 %
Precision	95.80 %	87.46 %	10.61 %	TA = 48.24 %

从表 6 可以看出, 9 条规则总体的正确率只有 48.24 %, 单数、复数的精确率较高, 但召回率较低; 同时不确定类型的召回率较高、精确率偏低. 说明 9 条规则对于不确定类型的判断能力很弱. 在详细统计每条规则覆盖能力的表 7 中, 前 7 条规则的精确率都很高, 但是平均覆盖的实例数量却较少; 规则 8 是精确提炼的判断不确定类型的规则, 但从效果看来, 很多情况还是被误判. 前面 8 条规则不能判断的情况, 都被规则 9 判断为不确定. 从结果看来, 这样的规则是不恰当的, 前面 8 条规则覆盖的实例数量毕竟有限.

表 7 基于规则的方法各条规则的详细结果

Table 7 The detailed experimental results of each rule

Rule	r-Indiv	r-Group	r-Indet	Precision
1	*410	131	12	74.14 %
2	33	*60	4	61.86 %
3	30	*234	3	87.64 %
4	152	*855	8	84.24 %
5	*3 320	24	5	99.13 %
6	*672	11	10	96.97 %
7	8	*595	12	96.75 %
8	50	160	*112	34.78 %
9	4 220	2 621	*725	9.58 %

4.3 基于机器学习的方法

实验选用了 Zhang Le 的最大熵工具[3], 采用了其中的默认参数设置. 实验中采用了 5-Fold 交叉验证的方法来对比各种特征组合. 实验中涉及的特征可以分为三个维度: 词语的选择 (Head 词语、Extent 词语、Qun 词语), 词语特征的选择 (词形、词性、词义), Extent 词数的选择. 表 8 展示了三个维度下各种特征组合的详细结果. 表格中数据都是 5-Fold 的平均正确率. 每个单元格中的上下两个数据分别是不采用 Len 特征和采用 Len 特征的实验结果. Len gain 表示当前行或者当前列使用 Len 特征得到的平均正确率的增益. 表格中右下脚数据表示各种组合下使用 Len 特征得到的平均增益.

4.3.1 整体实验结果数据分析

表 8 (见下页) 显示, 单纯使用 Head 词形本身能够达到的正确率是 79.98 %. 也就是说中文人称名词短语的核心词能够体现一定的单复数类型信息, 当然这种信息是体现在对语料的统计学习上的.

表 8 中正确率最高的是 87.48 %, 对应的参数设置是 Head + Extent + Qun/Word + Pos + Wsd/No

[3] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

表 8　98 种特征组合的实验结果

Table 8　The experimental result of 98 kinds of features combinations

No len/with len	Head only	Extent only	Qun only	Head+Extent	Head+Qun	Extent+Qun	Head+Extent+Qun	Len gain
Word only	**79.98%**	83.95%	68.18%	85.84%	86.45%	84.77%	86.03%	0.74%
	83.89%	84.46%	68.43%	85.96%	86.20%	85.20%	86.24%	
Pos only	66.73%	69.93%	65.77%	70.65%	66.94%	70.10%	70.91%	0.70%
	66.98%	70.12%	65.91%	71.15%	70.28%	70.30%	71.22%	
Wsd only	76.47%	75.24%	66.73%	75.91%	77.42%	75.20%	76.10%	0.38%
	76.55%	75.10%	67.02%	76.60%	77.14%	76.36%	76.96%	
Word+Pos	84.13%	85.77%	68.34%	86.62%	86.60%	86.28%	87.16%	0.00%
	83.92%	85.85%	68.40%	87.10%	86.45%	85.74%	87.44%	
Word+Wsd	84.20%	84.95%	68.32%	86.25%	86.66%	85.27%	86.94%	-0.02%
	84.24%	84.60%	68.41%	86.16%	86.52%	85.40%	87.08%	
Pos+Wsd	77.23%	77.16%	66.86%	78.12%	77.77%	77.27%	78.11%	0.23%
	77.02%	77.14%	67.07%	78.26%	78.39%	77.59%	78.61%	
Word+Pos+Wsd	84.43%	85.92%	68.34%	86.95%	86.79%	85.89%	**87.48%**	**-0.11%**
	84.26%	86.04%	68.34%	86.92%	86.41%	85.90%	87.15%	
Len gain	0.53%	0.06%	0.15%	0.26%	0.40%	0.25%	0.28%	**0.27%**

len. 这组参数设置说明, 选用涉及到的全部词语, 并且每个词语选用全部的特征可以取得最好的实验效果. 同时参数设置显示 Len 特征并没有对最好的实验效果起到作用.

4.3.2　特征空间中三个维度的相关分析

在词语特征方面 (表 8 中的不同行) 可以发现, 词形、词性、词义三个特征单独使用时词形优于词义, 词义优于词性. 三个特征混合使用时词形 + 词性最好, 词性 + 词义表现最不好. 三个特征全部使用时效果比使用词形 + 词性稍好一些.

我们分析出现这种情况是因为中文中常用的词语词形有几万个, 而词性只有几十个, 涉及到的词义代码有一千多个. 因此可以说, 词形、词义、词性三种特征的粒度越来越大. 当使用机器学习方法对上万的训练语料来进行单复数类型识别时, 自然是粒度越小的特征效果越好. 当然, 词形、词性、词义三者对于词语的描述是不同侧面的, 因此混合使用三个特征时效果肯定比只使用单个特征的效果要好.

在词语的选择方面 (表 8 中的不同列), Head、Extent、Qun 三种特征单独使用时 Extent 优于 Head, Head 优于 Qun. 三个特征混合使用时 Head + Extent 和 Head + Qun 的整体效果接近, 都优于 Extent + Qun. 三种特征都用时效果达到最好. 我们分析出现这种情况的原因有如下两点:

1) 在进行特征设计时, 选用了 Extent 前面的四个词语, Head 前面的两个词语, Qun 的两个词语. 事实上 60.7% 的 Mention 中 Extent 完全等于 Head, 而 Qun 关系在整个语料中只出现了 1 655 次, 仅占 Mention 总数的 11.43%. 所以出现单独使用时 Extent 最好, Qun 最差. 需要说明的是, 虽然只有 11.43% 的 Mention 带有 Qun 关系, 仅用 Qun 特征的实验正确率可以达到 68.18%, 高出了默认为单数类型的正确率 61.44% (表 5 所示). 这是因为最大熵算法会在带有 Qun 关系的 11.43% 的 Mention 上主要利用 Qun 特征来获得很高的准确率, 在剩下的 88.57% 的不带有 Qun 关系的 Mention 上会采用默认为单数类型的处理方法, 从而得到表 8 中显示的仅用 Qun 关系的实验正确率 68.18%.

2) 由于 Qun 涉及的词语多数都出现在 Extent 的前部, 因此当混合使用三个特征时, Extent + Qun 比其他两种混合获得的信息要少, 从而效果最差. 至于 Head + Extent 和 Head + Qun 的效果差不多, 说明在拥有 Head 的前提下, Extent 和 Qun 事实上是差不多的.

在选用 Extent 长度特征方面, 从表 8 中可以看出, 总体来说使用 Len 能平均提高整体正确率 0.27%. 对于不同的词语选择, 使用 Len 总是得到

正值的增益. 对于不同的词语特征选择, 当使用单特征时 Len 能够取得正值的增益, 当使用混合特征时 Len 就不一定能够取得增益了, 甚至在使用全部的词语特征时出现了 −0.11 % 的增益, 即增加 Len 特征起到了相反的作用. 这说明, 对于单复数识别, 并不是特征越多效果越好. 在训练实例数量确定的情况下, 选用更多的特征就会构建更大的特征空间, 从而在使用机器学习方法时会造成训练不充分的情况, 为此, 对于这种情形, 特征选择是应该考虑的问题.

4.3.3 一组实验数据分析

为了详细分析各种单复数类别的实验效果, 这里详细列出了正确率最高的一组实验中的 Fold 1 作为测试语料时的实验结果, 如表 9 所示.

表 9 正确率最高的一组实验数据: Fold 1

Table 9 The Fold 1 detailed to the highest accuracy

Fold 1	r-Indiv	r-Group	r-Indet	Recall
g-Indiv	1673	83	23	94.04%
g-Group	113	799	26	85.18%
g-Indet	47	55	76	42.70%
Precision	91.27%	85.27%	60.80%	TA = 88.01%

从表 9 可以看出, 单数的精确率和召回率都达到 90 % 以上, 复数的精确率和召回率都达到 80 % 以上, 但是不确定类型的精确率和召回率却都很低. 这说明不确定类型确实很难判断. 其实人工判断人称名词短语的单复数类型时, 面对这种不确定类型, 想要迅速判定也是有困难的. 例如, 下面两句中“人家”分别是不确定和单数类型[4].

人家 说 这件 事情 是 真的 .

你 慢点 走 , 人家 脚 扭 了 .

面对这种情形, 单纯依靠 Extent 和 Head 内能够抽取得到的各种信息是很难判断单复数类型的. 一种可能解决方法就是考察 Mention 周围更多的上下文信息, 但是如何考察还有待深入研究.

5 结论和展望

本文处理的单复数信息采用 ACE 的定义方法, 专门针对 Person 类型的 Mention 进行. 每个 Mention 下面包含名词短语和核心词. 中文属于汉藏语系, 名词本身不具备体现语义的功能, 对于单复数, 多数核心词不能起到界定作用, 同一个核心词在不同的名词短语中单复数类型可能会不同, 因此需要借助核心词所在的名词短语来帮助识别单复数类型.

单复数信息的识别可以根据一些人工设定的规则来完成, 常见的中文上进行单复数识别都采用这种方法. 本文在前人工作的基础上, 总结了更加丰富和尽量完备的 9 条规则在 ACE 语料上进行了单复数类型的自动识别. 结果显示对于单数和复数类型, 能够取得很好的精确率, 但是召回率却很低. 基于规则的方法判定不确定类型的能力很弱, 出现的大量错误都是这个上面产生的. 这种方法具有不完备性.

为了能够更好地解决单复数信息的自动识别, 本文采用最大熵方法, 在 LTP 处理结果上详细对比了词语选择、词语特征选择、Extent 长度选用与否等相关的各种特征组合. 结果发现词语选择上, 当选用 Head 时, Extent 和 Qun 具有相似的作用; 在词语特征选择上, 词形、词性和词义是对词语不同粒度的表示, 也是不同侧面的描述, 三者一起选用时可以达到最佳的效果. Extent 的长度特征总体来说对于单复数识别能起到一定的作用, 但是实验结果显示, 达到最好识别效果的特征配置里不选用 Len 较好. 实验结果显示, 选用 Head 的前两个词语、Extent 的前四个词语、Qun 关系对应的起止词语, 每个词语选用词形、词性、词义, 不选用 Extent 的长度特征时能够达到最好的识别正确率 87.48 %.

通过实验发现, 如果想更加准确地确定名词短语的单复数类别, 需要考虑分析更多上下文, 例如所在的句子. 采用最大熵来进行单复数识别是为了方便地进行各种特征的组合对比, 也可以采用其他的机器学习方法来进行单复数类型的识别. 下一步, 我们将会考虑如何应用短语所在句子来对不确定类型进行更加精确的确定, 同时也会采用其他方法来进行对比实验, 例如在机器学习识别结果的基础上采用错误驱动的方法来增强识别的结果.

[4]两个句子中的 Mention 的 Head 和 Extent 都是“人家”

References

1 Ng V. Shallow semantics for coreference resolution. In: Proceedings of International Joint Conference on Artificial Intelligence. Hyderabad, India: AAAI, 2007. 1689−1694

2 Mitkov R. Anaphora Resolution: the State of the Art, Technical Report, University of Wolverhampton, Wolverhampton, UK. 1999

3 Wang Hou-Feng. On anaphora resolution within Chinese text. *Applied Linguistics*, 2004, (4): 113−119
(王厚峰. 汉语篇章的指代消解浅论. 语言文字应用, 2004, (4): 113−119)

4 Mitkov R. Robust pronoun resolution with limited knowledge. In: Proceedings of the 36th Annual Meeting on Association for Computational Linguistics. Montreal, Canada: Association for Computational Linguistics, 1998. 869−875

5 Wang Hou-Feng, He Ting-Ting. Research on Chinese pronominal anaphora resolution. *Chinese Journal of Computers*, 2001, **24**(2): 136−143

(王厚峰, 何婷婷. 汉语中人称代词的消解研究. 计算机学报, 2001, **24**(2): 136−143)

6 Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 2001, **27**(4): 521−544

7 Luo X Q, Ittycheriah A, Jing H Y, Kambhatla N, Roukos S. A mention-synchronous coreference resolution algorithm based on the bell tree. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2004. 135−142

8 Zhang Li. The realization and related discussion of number category in Chinese nouns. *Chinese Language Learning*, 2003, (5): 28−32
(张黎. 汉语名词数范畴的表现方式. 汉语学习, 2003, (5): 28−32)

9 Wang Zhi-Qiang. Research on Chinese Coreference Resolution and Its Related Technologies [Ph. D. dissertation], Beijing University of Posts and Telecommunications, 2006
(王智强. 汉语指代消解及相关技术研究 [博士学位论文], 北京邮电大学, 2006)

10 Wang De-Liang. Anaphora Resolution in Chinese from the Centering Perspective [Ph. D. dissertation], Beijing Normal University, 2006
(王德亮. 基于向心理论的汉语回指消解研究 [博士学位论文], 北京师范大学, 2006)

11 Wang H F, Mei Z. An empirical study on pronoun resolution in Chinese. In: Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico: Springer-Verlag, 2004. 213−216

12 Li Guo-Chen, Luo Yun-Fei. Chinese pronominal anaphora resolution via a preference selection approach. *Journal of Chinese Information Processing*, 2005, **19**(4): 24−30
(李国臣, 罗云飞. 采用优先选择策略的中文人称代词的指代消解. 中文信息学报, 2005, **19**(4): 24−30)

13 Dong Da-Nian. *Xiandai Hanyu Fenlei Cidian*. Shanghai: Publishing House of an Unabridged Chinese Dictionary, 1998
(董大年. 现代汉语分类词典. 上海: 汉语大词典出版社, 1998)

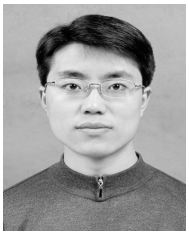
14 Dong Zhen-Dong, Dong Qiang. Construction of a knowledge system and its impact on Chinese research. *Contemporary Linguistics*, 2001, **3**(1): 33−44
(董振东, 董强. 知网和汉语研究. 当代语言学, 2001, **3**(1): 33−44)

15 Mei Jia-Ju, Zhu Yi-Ming, Gao Yun-Qi, Yin Hong-Xiang. *Tongyici cilin (Second Edition)*. Shanghai: Shanghai Lexicographical Publishing House, 1996
(梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 同义词词 (第二版). 上海: 上海辞书出版社, 1996)

16 Li Su-Jian, Liu Qun, Zhang Zhi-Yong, Cheng Xue-Qi. Method of maximum entropy model for language processing. *Computer Science*, 2002, **29**(7): 108−110
(李素建, 刘群, 张志勇, 程学旗. 语言信息处理技术中的最大熵模型方法. 计算机科学, 2002, **29**(7): 108−110)

17 Lang Jun, Liu Ting, Li Sheng, Zhang Hui-Peng. LTP: an XML-based open language technology platform. In: Proceedings of the 25th Anniversary of the Chinese Information Processing Society of China, Beijing: Tsinghua University Press, 2006. 561−572
(郎君, 刘挺, 李生, 张会鹏. 基于 XML 的开放式语言技术平台: LTP. 中国中文信息学会成立二十五周年学术学会. 北京: 清华大学出版社, 2006. 561−572)

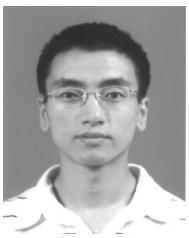
18 Liu Ting, Ma Jin-Shan, Li Sheng. Chinese dependency parsing model based on lexical governing degree. *Journal of Software*, 2006, **17**(9): 1876−1883
(刘挺, 马金山, 李生. 基于词汇支配度的汉语依存分析模型. 软件学报, 2006, **17**(9): 1876−1883)



郎 君 哈尔滨工业大学博士研究生. 主要研究方向为信息抽取、共指消解、机器学习. 本文通信作者.
E-mail: bill_lang@ir.hit.edu.cn
(**LANG Jun** Ph. D. candidate at Harbin Institute of Technology. His research interest covers information extraction, coreference resolution, and machine learning. Corresponding author of this paper.)

秦 兵 哈尔滨工业大学副教授. 主要研究方向为自然语言处理、信息检索.
E-mail: qinb@ir.hit.edu.cn
(**QIN Bing** Associate professor at Harbin Institute of Technology. Her research interest covers natural language processing and information retrieval.)

刘 挺 哈尔滨工业大学教授. 主要研究方向为自然语言处理、信息检索.
E-mail: tliu@ir.hit.edu.cn
(**LIU Ting** Professor at Harbin Institute of Technology. His research interest covers natural language processing and information retrieval.)



李正华 哈尔滨工业大学硕士研究生. 主要研究方向为自然语言处理.
E-mail: lzh@ir.hit.edu.cn
(**LI Zheng-Hua** Master student at Harbin Institute of Technology. His main research interest is natural language processing.)

李 生 哈尔滨工业大学教授. 主要研究方向为自然语言处理、信息检索、机器翻译. E-mail: lisheng@ir.hit.edu.cn
(**LI Sheng** Professor at Harbin Institute of Technology. His research interest covers natural language processing, information retrieval, and machine translation.)