中 文 信 息 学 报
第 18 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 1

文章编号:1003 - 0077(2004)01 - 0020 - 06

语料库中熟语的标记问题①

安 娜,刘海涛,侯 敏

(北京广播学院 应用语言学系,北京 100024)

摘要:熟语是自然语言中普遍存在的语言现象。本文分析了国内现有语料库对熟语的标注方式,发现这种方式对语料库的进一步加工是有问题的。为了在语料库标注阶段把熟语问题处理好,本文从信息处理的角度将熟语中的成语、惯用语、歇后语、习用语、专门语以及缩略语归为固定语的范畴,进而提出根据固定语的语法功能给定词性标记,再根据它们的词汇特征给定词汇范畴标记的双层标记法,这样在一定程度上解决了熟语的语料库标注问题。

关键词:人工智能;自然语言处理;熟语;固定语;标注;语料库

中图分类号:TP391 **文献标识码**:A

Tagging of the Idiom in the Corpus

AN Na, LIU Hai-tao, HOU Min

(Applied Linguistics Department, Beijing Broadcasting Institute, Beijing 100024, China)

Abstract: Idiomaticity is a common phenomenon in natural languages. This paper analyses some known means of tagging the idiom in Chinese corpus. These tagging methods are problematic for the further syntactic tagging and parsing of corpus. To find a suitable solution for application in natural language processing, the authors introduce a new concept "fixed expression", which consist of idioms, customary usages, two-part allegorical sayings, terms and abbreviations. These fixed expressions have the same grammatical function as common words, thus we can tag them according to their function in text and give suitable vocabulary category of fixed expressions. This is called two-level tagging method. The proposed solution is useful to build a parsed corpus as knowledge source of NLP.

Key words: artificial intelligence; natural language processing; idiom; fixed expression; tagging of corpus; parsed corpus

1 引言

在建设传媒语言语料库的过程中,我们根据对语料库加工的通行做法,先对原始语料作词性标注。词性标注之后,当我们开始进行句法层次的加工时,发现目前的熟语标记存在一定的问题。我们在对生语料做词性标注时,采用的是北京大学计算语言学研究所与北京大学中文系联合开发的分词标注系统。该系统把熟语中的成语标注为 i,习用语标注为 l,另外把简称略语标注为 j。北大系统是目前国内使用较多的一个系统。带着这个问题我们考察了国内外其他一些语料库词性标注体系,发现国内系统基本上都采用类似北大系统的做法,如,在国家语

① **收稿日期**:2003 - 06 - 08
基金项目:国家广电总局社科项目资助(bw0220),国家语委科研规划项目资助(YB105 - 61A)
作者简介:安娜(1979 —),女,在读硕士研究生,目前主要研究方向为语料库语言学.

委语言文字应用研究所计算语言学研究室制定的信息处理用现代汉语词类标记集规范中,把习用语标记为 i(习用语包括成语、惯用语、谚语、格言等),把简称和略语标记为 j。由于词性标注是对语料库进行加工的第一阶段,是随后要做的句法标注和语义标注的基础,因此,如果按照上述的方法对熟语进行标注,显然就很难对语料库作进一步的句法标注。因为不论是采用 PS(短语结构)还是 DS(依存结构)的句法模型,都没有办法处理如 i、l、j 之类的标记。

以下是我们从语料库中抽取的几个例子:

(1) a. 你/r 看/v 不/d 懂/v 通知单/n 倒/d 也/d 算/v 了/y, /w 就/d 怕/v 成/v 了/u 井底之蛙/i, /w 自大/a 自夸/v 。/w

b. 我/r 觉得/v 台湾/ns 同胞/n 到/v 大陆/n 来/v 旅游/v 已经/d 畅通无阻/i 。/w

(2) a. 我/r 图/v 啥/r 呢/y ,/w 我/r 就/d 图/v 我们/r 国家/n 的/u 安定团结/l 。/w

b. 一般来说/l, /w 一个/m 国家/n 的/u 货币/n 汇率/n 可以/v 反映/v 该国/r 的/u 经济/n 发展/v 的/u 基本/a 情况/n 。/w

(3) a. 国际/n 足联/j 提出/v 的/u 意见/n 很/d 适合/v 中国/ns 国情/n 。/w

b. 美国/ns 在/p 反恐/j 的/u 过程/n 中/f 不得不/d 加强/v 多边/b 合作/v 。/w

显然,i、l、j 只是一种词汇类型,在语法功能上它们还具有不同的性质,那么,在语料库的词性标注阶段,我们应该给它们打上合适的词类标记。

2 熟语的本质与信息处理用“语”的要求

熟语是自然语言中普遍存在的现象,因此任何语料库的加工都很难避开这个问题。从语言分析的角度看,熟语也是许多大牌的语言形式化理论都颇伤脑筋的东西。“熟语”是汉语语言学界广泛使用的一个术语,语言学家一般认为熟语具有以下主要属性:

(1)熟语是语言中定型的词组和句子,使用时一般不能任意改变其组织,包括成语、谚语、格言、歇后语等。(《辞海》1979)

(2)词汇当中,除了许多独立运用的词以外,还有一些固定词组为一般人所经常使用的,也作为语言的建筑材料和词汇的组成部分,这些总称熟语。熟语的范围相当广,包括惯用语、成语、歇后语、谚语、格言等。(胡裕树 1998)

(3)熟语是结构形式和语义容量上大于词,在习用性、现成性和定型性上同于词,在造句功能上大体相当于词的词组或短语,它包括成语、谚语、歇后语、惯用语等。(武占坤 1986)

(4)熟语可以是种种比词大的用语,包括成语、惯用语、专名语、谚语、名言等等。有些熟语单位在使用时,一般只作为句子成分而出现在语句中,作用和词相当;有些熟语单位一般不充当句子成分,而是作为独立的语句出现。(刘叔新 1995)

(5)熟语是个涵盖面很大的概念,它既指成语、惯用语、歇后语、专名语、专门用语等词汇单位,又指谚语、格言、警句等非词汇单位。(周荐 1994)

从上面各家对熟语的解释中可以看出,它在结构形式上具有定型性,是大于词的单位,这一认识人们是共同的;但在功能及范围的认识上并不相同。有人认为它是造句材料,功能等同于词,如(2)、(3);有人认为它应分为两种,有些是词汇单位,有些是非词汇单位,如(4)、(5);在

范围上,它应包括成语、惯用语、歇后语、谚语、格言等是大家都认可的,但有些人认为还应该包括专名语和专门用语,如(4)、(5)。可见语言学家们对熟语的功能和范围还没有形成统一的看法。

从信息处理的角度看,汉语的词汇处理已成为现阶段中文信息处理应用领域(汉字输入、汉语语音识别及合成、全文信息检索及文本自动分类、文本自动校对、机器翻译等等)的主要支撑平台,几乎没有什么技术可以超越这个平台而存在。既然词汇不能超越,熟语也无法避开,语言学家又没有拿出一个定论,我们怎么办?面对语料库建设这样一个颇具规模的语言工程,我们无法等待。而且退一步说,即使真的有一天语言学家们拿出了个统一的意见,如果是面向人际交际的,也未必适合语言工程建设的需要。所以,语言学家长期议而不决的问题,我们现在就必须从工程的角度找到一个合适的解决方案,哪怕一下子还不能做到尽善尽美。此外,词汇中主要包括"词"和"语"这两种单位。为避免理论上的纠缠,"词"的问题,我们已采取信息处理用"词"——"分词单位"的说法来解决,(孙茂松 1995)那么"语"的问题,我们同样也可以采取信息处理用"语"——"固定语"的说法来解决。

国外语言学界也遇到类似的问题。德国学者 Sabine Fiedler 在一部有关习语的专著中,将习语定义为:"一种多词的语言单位,常为习惯用法,具有相对固定的句法 - 语义结构。语言的使用者惯于将它作为一个整体来用,以增强语体效果"。(Fiedler 1999) 她的这一定义基本反映了国外学者对于熟语的看法。为了避免不必要的概念之争,一般在语料库语言学和自然语言处理领域,国外学者将这种有关习语、固定搭配的语言现象统称为"多词组合"(multi-words),也有人把它们称作"带空格的词"。可惜"带空格的词"这一极具挑战性的说法,在汉语中无法使用。"多词组合"指的是在语料库的词性标注阶段几个词在一起担当同一种词类功能的现象。但这一名称对汉语也不适用,因为汉语中这种"语"的构成成份未必都是"词"。

鉴于上述认识,从自然语言信息处理的实用性原则出发,我们把所谓熟语分为三类:

1. 固定语。包括成语、惯用语、习用语、歇后语、专门用语这些词汇单位,另外还要加上缩略语。虽然没有人把缩略语归在熟语中,但它既不是词,在进行词性标注时又要作为一个整体来处理,它的内部不再分析,这些语法性质与上述的成语、惯用语等单位完全相同。因此,在语言信息处理中,可以把它们看作是同一类型。

2. 专名语。之所以把专名语分出来,单独作为一类,是因为在标注时要做不同的处理。在专名语的处理上我们同意北大分词标注系统的意见,既做内部分析,又有整体标注。如"北京大学"应标注为"[北京/ns 大学/n]nt"

3. 谚语、格言、警句等非词汇单位。它们往往体现着一个具体的判断,表达一个明确的思想,甚至是简单的推理,一般不充当句子成分,而是作为独立的语句出现。因此在进行词性标注时,只对其具体的组成单位作标注,整体不做处理,像普通的句子一样。当然,由于它们有着特殊的表达功能,可以在句末打上相关的标记,以便满足语体风格研究的需要。但这已不属于语法层面的标注了。

看来,目前在语料库词性标注中存在问题较大的是第一类:固定语。下面我们集中讨论固定语的标注问题。

3 固定语的标注

固定语虽然在构成形式上大于词,但它意义的整体性、结构的凝固性以及在语言使用中的重复共现性都决定了它在句法分析上具有终极单位的特点,是句法分析的最小单位,在句子结构中的功能相当于一个词。既然如此,那么就该给它应得的语法属性标记,即词性标记。为了

给固定语一个合适的标记,首先遇到的就是词性的判定问题。对于词来说,划分词类、判定词性的依据是词的语法功能,即词在组合中的分布特征。那么,这一原则同样适用于固定语。因此,我们应该在具体语料中,对固定语的句法功能进行全面深入的调查,根据它的分布特征,确定它的语法属性,即词性(尽管这一术语不太准确,但我们似乎又没有必要造出一个新的术语"语性",于是姑且用之)。各类固定语的语法功能是不一样的。下面具体分析。当然,固定语的词性定下来之后,需要修改相应的词表,以便标注系统可以在具体的标注过程中使用。

3.1 成语

成语是一种相沿习用、具有书面语色彩的固定语,它在句子里的语法作用相当于"实词"。凡是实词能够充当的句子成分,它都可以充当。但是由于它特殊的表达作用,从使用频率来看,在句子中经常是作谓语和定语,其次是做状语、补语,而较少作主宾语。根据句法功能和分布来看,成语绝大多数具有"动词性",其次是"形容词性"、"名词性"、"副词性"、"区别词性"。名词性成语是根据成语自身的句法功能即主要做主宾语和它的指称性语义特征划分的,如"赤子之心"、"雕虫小技"、"丰功伟绩"。副词性成语和区别词性成语的确定主要根据其句法功能,只能充当状语的成语为副词性成语,如"长此以往"、"三番五次"、"挨家挨户";只能作定语的成语为区别词性成语,如"莫须有"。由于名词性、副词性和区别词性成语充当句子成分的能力比较弱,所以这三类成语的词性比较容易确定。而动词性成语和形容词性成语的界限却比较模糊,由于大部分动词性成语不具备动词能够带宾语的典型特征,大部分形容词性成语也不具备形容词能够受程度副词修饰的典型特征,因此,成语中只要可以带宾语的就可以确定为动词性成语,只要可以受程度副词修饰的就可以确定为形容词性成语。对那些划分界限模糊的动词性成语和形容词性成语,我们主要依据它内部的构成理据和语义特征,如"全力以赴"、"持之以恒"、"安然无恙",就构成来看以动作性成分为主,主要表达的是动作意义,就确定为动词;"错落有致"、"津津有味"、"娓娓动听"就构成来看以形容词性成分为主,主要表示一种性质或状态,就确定为形容词。显然,这种划分难免有见仁见智的地方。好在形容词性成语和动词性成语在句子中的功能比较相近,所以标注为"动词性"或"形容词性"并不影响下一步的句法分析。

上述标准也适合对下面各种类型的分析。

3.2 惯用语

惯用语是指人们口语中习惯运用的短小定型的固定语。在意义的整体性上,它们比成语是"有过之而无不及"。惯用语在句子中主要做谓语中心,其次作主宾语,很少作附加成份。根据句法功能和分布来看,大多数惯用语具有"动词性",如"打天下"、"戴高帽",这和它们的结构有关,大多数惯用语都是动宾结构的,一般作谓语;其次是"名词性",如"定心丸"、"大锅饭";还有"形容词性"的惯用语,可作谓语、定语,但一般不作状语,如"耳朵软"、"肝火盛",这类惯用语一般都是主谓结构,充当谓语的一般是形容词;"副词性"惯用语比较少,如"手把手"。

与成语不同的是,惯用语在口语中还常拆开来使用,这时就很难再把它看成是一个整体,应分开标注。在这点上,它和汉语中的离合词有些相像。如:

(4)昨天他在领导那儿　碰/v　了一个不大不小的　钉子/n　。

3.3 歇后语

歇后语是我国人民在生活实践中创造的一种特殊语言形式,由近似于谜面、谜底两部分组成的带有隐语性质的口头用语。谜面往往是一种形象化的说明,谜底才是表达的重点,所以应根据谜底的语法功能确定它的词类。如"名词性"歇后语,"巴掌心里长胡须——老手";"动词性"歇后语,"蛤蟆跳井——不懂";"形容词性"歇后语,"偶像面前磕头——毕恭毕敬"。值得注

意的是,有些常用的歇后语,人们在使用时,往往会只说出谜面,歇去谜底,但表达的重点还是谜底,这样的歇后语就有必要在词表中存入两种形式:歇去谜底的和谜底、谜面俱全的,以保证标注的准确率。

3.4 缩略语

缩略语是由原词语简缩而成的,那么,一般说来,缩略语与原词语的语法功能也是相同的。就语法功能来说,缩略语大多数是名词性的,如"北约"、"文革";其次是动词性的,如"监控"、"征管";也有区别词性的,如"远南"、"经贸"。

3.5 专门语

专门语指的是一些带有行业色彩的术语,它们明显地由几个词组成,但表达一个概念,使用上有一定的凝固性。专门语大部分是名词性的,如"神经衰弱"、"西安事变"、"夕阳产业";也有动词性的,如"粗放经营"、"扩大再生产"、"软着陆"。

3.6 习用语

习用语指的是"使用频率很高,但形成的历史尚短,也没有典故"(俞士汶)的固定语。这一类固定语成分很杂,可以说,归不到别的类别中去的固定搭配都可以收罗在这里。这决定了它的语法属性也很杂,既有实词的性质,如:动词性的,"公诸于世"、"顺其自然";形容词性的,"各种各样"、"不冷不热";副词性的,"不管怎样"、"毫不犹豫";名词性的,"数九寒冬"、"歪门邪道";也有虚词的性质,如连词性的"比如说"、"由此看来"。

还有一点应该说明,由于"语"的长度大于"词",内涵深而外延窄,所以它在多义性和兼类性上都明显地要好于"词",即多义和兼类现象都要比词少得多。但是,同一语言片断,也可能兼有固定语与自由短语两种性质,如例(5)、例(6)。这是在标注时应该注意的。

(5) a. 这是一个非常先进的 企业管理/n 软件。

b. 这样可以把 企业/n 管理/v 得更加好。

(6) a. 他们的下属就不怕惹恼了老板,有被 穿小鞋/v 甚至"炒鱿鱼"的危险?

b. 有了鞋店的这位好心的售货员,这位大脚青年就再也不用担心 穿/v 小/a 鞋/n 了。

上面谈了固定语的词性标注问题。我们说,固定语的语法作用相当于词,但是其表达作用决不仅限于此,它的更积极的作用是它的修辞功能。在建设传媒语言语料库的过程中,我们发现,主持人的语言各有风采,有的非常典雅,富有文采;有的亲切自然、平易近人;还有的生动幽默,引人入胜。熟语在其中扮演了一定的角色。因此,对熟语仅仅作出词性标注还无法满足多层次语言研究的需要,我们还需要把成语、惯用语、歇后语、专门语、缩略语等加上相应的词汇范畴标记,这对语体风格的研究无疑是非常重要的。

为了证实我们的观点,我们又参考了国外语料库的标注方法。下面是我们从采用CLAWS标注的BNC(British National Corpus)中选取的几个例子,从中可以看出国外语料库一般是如何处理"多词组合"(习语)的(例子选自BNC2 POS-tagging Manual,2000):

<w AV0> of course (副词)

<w PRP> according to (介词)

<w NN1> persona non grata (复合名词)

<w CJS> except that (连词)

以上这些"多词组合"在功能上一般被视为一个词,故只指定一个标记。国外语料库对习语的标注处理,也证实了我们想法和做法的合理性。

另外,在修改本文的过程中我们又看到了俞士汶等先生的论文《北大语料库加工规范:切

分 词性标注 注音》,发现我们对熟语的认识与北大的新标注体系有一致的地方,但在处理上却存在着不同。首先,在熟语库的建设上,我们除了建有北大有的成语库、习语库、简称略语库、专名库外,还增添了惯用语库、歇后语库、专门语库、谚语格言库,这样就有了一个比较完整的熟语库系统;其次,在具体标注时,我们不是一次标注,而是采用二层标注的方法。对于固定语,首先给出词性标记(v、a、n、d 等),为下一步句法分析打下基础;进而再给出词汇范畴标记(成语:i、缩略语:j、习用语:l、惯用语:gy、歇后语:xh、专门语:zm 等),以满足多层次语言研究的需要。如:

(7)你把自己的芯片说得 天花乱坠/ a/ i,谁知道这里面是什么东西?

(8)每次领导来的时候,他就知道 偶像面前磕头 ——毕恭毕敬/ a/ xh。

二层标注的好处是灵活、方便,在进行不同目的研究时,可以各取所需,使得句法分析简明,减少冗余信息,而且便于进行各种不同层次的检索和统计。

4 结语

语料库正在成为语言学家研究语言的得力工具。而对于计算语言学家来说,语料库不仅仅是工具,还是一种建立大规模真实语言处理系统的知识库(刘海涛 1992)。在 Anne Abeille 新编的“Treebanks:Building and Using Parsed Corpora”(Kluwer Academic Publishers,2003)一书中,绝大多数的论文作者都是计算语言学家,这说明计算语言学需要一个具有句法、语义标注的语料库作为语言信息处理的知识源。毫无疑问,词性标注只是进一步加工语料库的基础,对语料库的加工绝不可能仅仅给出一个词性标注就算完事。语言中的固定语成分,因为一般无法从构成固定语的词的字面意思,得出固定语的意思;构成固定语的词义,只在固定语内有效;固定语在句法上是不合格的,所以无法用标准的词汇和句法规则来分析(Riehemann 2001)。鉴于此,国外有学者将“多词组合”问题与“歧义”问题并称为自然语言处理的两大难题。本文为了满足计算语言学研究对语料库的要求,提出将固定语进行二层标注的方法。首先在语料库词性标注过程中将其句法功能明确标示出来,以满足句法分析的需要;进而再作词汇范畴的标注,以满足语体研究的需要。应该承认,我们的这种做法还只是固定语处理的一个最基本的阶段。随着对语料库加工的深入,我们在进行有关语义、语篇的标注时,还需对固定语问题作进一步的探讨。

参 考 文 献:

[1] 刘叔新. 汉语描写词汇学[M]. 北京:商务印书馆,1990.

[2] 黄昌宁,李涓子. 语料库语言学[M]. 北京:商务印书馆,2002.

[3] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000.

[4] 胡裕树. 现代汉语[M]. 上海:上海教育出版社,1998.

[5] 武占坤. 现代汉语读本[M]. 北京:北京语言学院出版社,1986.

[6] 孙维张. 汉语熟语学[M]. 吉林:吉林教育出版社 ,1989.

[7] 周荐. 熟语的经典性与非经典性[J]. 语文研究. 1994(3).

[8] 丁信善. 语料库语言学的发展和研究现状[J]. 当代语言学. 1998(1).

[9] 孙茂松,张磊. 人机并存,“质”“量”合一 —谈谈制定信息处理用汉语词表的策略[J]. 语言文字应用. 1997,(1).

(下转第 41 页)

9.3,可以推想 $\sum^2$ - CW2SSCH 生成的压缩模型的平均熵约为 7.4。

6 结语

本文研究了基于一种新的全文索引模型 —— $\sum^2$ 邻接矩阵模型的文本压缩模型的生成算法。我们知道基于不定长单词的压缩模型的压缩效率高于基于字符的压缩模型,但是它的最优符号集的寻找算法是 NP 完全问题,本文提出了一种基于贪心算法的较优解的寻找方法,这种方法将文本的 $\sum^2$ 邻接矩阵索引作为统计数据的来源,避免了大量基础数据的在线计算,因此提高了算法效率。初步实验表明算法效率在实际的全文检索系统中是可行的。

参 考 文 献:

[1] K T Lua, A Minimum Entropy Approach for Chinese Text Compression [J]. Computer Processing of Chinese & Oriental Languages, 1995,9(2):155 - 161.

[2] Ian H. Witten,张仲颖,等译. 海量数据管理 ——文档和图像的压缩和索引 [M]. 北京:科学出版社,1996 年 8 月第一次印刷.

[3] 国家语言文字工作委员会,国家标准局. 现代汉语字频词典 [M]. 1992.

[4] 刘源,等. 现代汉语词频词典 [M]. 1992.

[5] 周强. 基于语料库和面向统计学的自然语言处理技术 [J]. 计算机科学,1995,22(4):36 - 40.

[6] 夏莹,等. 基于统计的汉字识别文本自动后处理方法 [J]. 模式识别与人工智能. 1996 年 6 月,9(2).

[7] 胡运发. 另一种全文数据模型 —邻接矩阵模型 [R]. 复旦大学技术报告,1999,4.

[8] 胡运发. 扩展的 $\sum^2$ 邻接矩阵模型 —小膨胀比的全文数据模型 [R]. 复旦大学技术报告,1999.8.

[9] 周水庚,胡运发,关佶红. 基于邻接矩阵的全文索引模型(英文) [J]. 软件学报, 2002,13(10).

[10] 陶晓鹏,胡运发. 文本压缩技术在全文检索系统中的应用 [R]. 1999,6. 复旦大学技术报告.

(上接第 25 页)

[10] 孙茂松, 邹嘉彦. 汉语自动分词研究中的若干理论问题[J]. 语言文字应用. 1995,(4).

[11] 刘海涛. 结构化语言知识库在自然语言处理中的应用[J]. 情报科学. 1992,(5).

[12] 俞士汶. 信息处理用现代汉语词语分类体系介绍[J]. 计算语言学教学参考资料(内部使用).

[13] 俞士汶, 段慧明, 朱学锋, 常宝宝. 北大语料库加工规范:切分 ·词性标注 ·注音[J]. 汉语语言与计算学报. 2003,(6).

[14] 俞士汶,等. 现代汉语语法信息词典详解(第二版) [M]. 北京:清华大学出版社.

[15] 吕叔湘,等. 语法研究入门[M]. 北京:商务印书馆,1999.

[16] Fiedler, Sabine. . Plansprache und Phraseologie[M]. Frankfurt am Main: Peter Lang. 1999.

[17] Garside, Roger; Geoffrey Leech and G. Sampson. The computational analysis of English[M]. London/ New York: Longman. 1987.

[18] Leech, Geoffrey; Roger Garside; Michael Bryant.. Claws4: The Tagging Of The British National Corpus[C]. Proceeding of COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics. 1994.

[19] Riehemann, Susanne 2001, A constructional approach to idioms and word formation[D], Ph. D thesis, Stanford.