

消除语义网络中分词歧义方法研究

陈晓明¹, 李伶俐¹, 梁雄友²

(1 广东司法警官职业学院, 广东 广州 510520; 2 厦门大学 计算机科学系, 福建 厦门 361005)

摘 要: 针对计算机语义网络中交集型和组合型词汇歧义的问题, 通过分析了传统分词方法中存在的缺陷, 提出基于最大概率计算的自动分词歧义方法. 运用上下文语义相关度对产生歧义的词汇进行有效修正, 重新计算切分候选词所产生的有效“费用”, 运用最大概率算法对产生歧义的词汇进行关联程度概率计算, 克服传统分词方法的弊端. 成功地解决交集型歧义、连环交集型歧义、组合型歧义、混合型歧义切分问题, 消除语义网络中的交集型和组合型词汇歧义的影响, 取得了不错的效果.

关键词: 动分词; 消歧; 最大概率法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)03-0178-04

Eliminate Semantic Network Word Segmentation Ambiguity Method Research

CHEN Xiao-ming¹, LI Ling-li¹, LIANG Xiong-you²

(1 Guandong Judicial Police Vocational College, Guangzhou 510520, China;

2 Department of Computer Science, Xiamen University, Xiamen 361005, China)

Abstract: According to the computer in the semantic networks combination of overlap type out the problem of righteousness vocabulary, through the analysis of the traditional word segmentation method defects. Based on the maximum probability calculation method of automatic word segmentation ambiguity. Application for the semantic relatedness produces different meanings to the context of vocabulary, effective correction to the calculation of the segmentation candidates from the word of effective "expenses", the most probability calculation method to produce ambiguity of vocabulary connection degree probability calculation, overcome the shortcomings of the traditional word segmentation method. Successfully solve overlap type righteousness, serial overlap type, combination and mixed segmentation problem, eliminate righteousness in the semantic networks and combination of overlap type out the effect of vocabulary righteousness, and achieved good effect.

Key words: automatic segmentation; disambiguation; algorithm of maximum probability

1 引言

计算机机器语言理解汉语的首要任务就是把连续的汉字串分割成词的序列, 即自动分词. 近二十年来, 汉语自动分词研究取得了很大的成就, 推出了一批有代表性的分词系统, 如清华大学的 SEGTag 系统^[1-3]; 北大计算机语言研究所分词系统等. 同时提出了许多分词算法, 其中有一定代表性的主要有: 最大匹配法(正向、反向、双向)、最大概率法、特

征词库法、邻接约束法、人工神经网络方法、联想-回溯法等. 然而, 汉语自动分词问题仍然是制约中文信息处理发展的瓶颈之一, 歧义切分字段又是影响分词系统切分精度的重要因素, 歧义切分字段又大体分为交集型歧义字段和组合型歧义字段^[4-5].

但是, 目前主流的分词算法对歧义消除问题还存在着弊端, 主要是计算机语义网络中存在着交集型和组合型词汇歧义, 这给相关算法的词义分类带来了不小的难度^[6]. 为了解决这一问题, 本文中, 对

复杂的歧义字段还专门加以细分,并给出了相应的定义——连环交集歧义字段和混合歧义字段.运用概率决策的方法解决这一问题.

2 分词的切分歧义原理及其描述

在人类使用的语言中,存在着一词多意的问题,成为语言的歧义现象,在计算机语义网络中也不例外,歧义是语言结构形式与其自身本体的语义之间产生的一种特殊关系,一种语言结构具有几种结构层次关系,或一个言语表现不同的意图或含义等产生的.语言歧义性探讨在语言教学、翻译理论和实践、文学语言研究等方面,具有一定的作用,可以帮助人们提高理解和运用语言的能力.但是在计算机语义网络中,歧义性的语言会给计算机的执行带来一些负面的影响,因此必须加以消除.所以有了分词算法.在计算机语言中,歧义的种类很多,下面将对典型歧义语言进行介绍.

在分词的切分歧义中,交集型歧义字段、组合型歧义字段、连环交集型歧义字段、混合歧义切分字段等多种分类,下面将对其进行分别介绍.

交集型歧义字段:在字段 AJB 中, $AJ \in W$ 并且 $JB \in W$, 则称 AJB 为交集型歧义字段.其中 A 、 J 、 B 为字串, W 为词表.歧义字段的构成形式除上述 AJB 外还可能有多种结构.对交集型歧义字段中的链长是指:交集型歧义字段中含有交集字段的个数,称为链长.当交集型歧义字段具有 AJB 形式时,只有一个交集字段,称 AJB 链长为 1.当交集型歧义字段具有 $ABCD$ 形式时,共有 ABC 和 BCD 两个交集字段,称 $ABCD$ 链长为 2.链长大于 2 时,依此类推.

组合型歧义字段:在字段 AB 中, $A \in W$, $B \in W$, W 为词表,则称 AB 为组合型歧义字段.此外,组合型歧义字段除 AB 型外,还有 ABC 型或 $ABCD$ 型.

连环交集型歧义字段: Z 是一汉字串,若 Z 可以切分成诸如: $A AB B BC C CD D \dots$, $A \in W$, $AB \in W$, $B \in W$, $C \in W$, $CD \in W \dots$, W 为词表,则 Z 为连环交集型歧义字段.连环交集型歧义字段的特点是链长都为 2 或以上.

混合歧义切分字段:组合歧义字段中含有交集歧义字段或者交集歧义字段中含有组合歧义字段,这种复杂的歧义字段称混合歧义字段.从人民日报作语料库研究得知,组合歧义字段中含有交集歧义的这种混合歧义情况占了大多数,本文将重点解决混合歧义的这种情况.

3 最大概率法分词法

为了解决以上问题,本文提出基于最大概率分词法的语义网络词汇划分方法,通过加入概率决策的概念,完成词汇的准确划分.

3.1 自动分词的统计模型的建立

概率分词算法的第一步是建立有效的模型.根据噪声——信道模型解释:一个词串经过信道传送,由于噪声干扰而丢失了词界标记.那么自动分词就是已知一个汉字串,求跟它对应的有最大概率的词串.即

$$W' = \arg \max_w P(W | Z) \quad (1)$$

由贝叶斯公式知:

$$\begin{aligned} W' &= \arg \max_w P(W | Z) \\ &= \arg \max_w \frac{P(W)P(Z | W)}{P(Z)} \end{aligned} \quad (2)$$

其中, $P(Z)$ 是汉字串的概率,它对于各个候选词串都是一样的,不必考虑. $P(Z | W)$ 是词串到汉字串的条件概率.在已知词串的条件下,出现相应的汉字串的概率是 1 (从词串恢复为汉字串只有惟一的方式).于是仅仅需要考虑的是 $P(W)$, 即词串的概率.因此,上述简化为:

$$W' = \arg \max_w P(W) \quad (3)$$

这就是说,概率最大的词串,便是最佳的词串.而词串概率可以用 N 元语法来求.如果用二元语法,则:

$$P(W) = \prod_{i=1}^n P(w_i | w_{i-1}) \quad (4)$$

其中 $1 \leq i \leq n$, W_0 是一个虚设的串首词.

但若使用二元语法进行分词的话,词到词的概率转移矩阵非常庞大.而一般词表都 4 万词以上,转移概率矩阵规模就有 16 亿以上.因此我们仅用一元语法来求词串的概率:

$$P(W) = \prod_{i=1}^n P(w_i) \quad (5)$$

每个词的概率则可以用频率来估计,因此词表中至少应有词形—词频两个属性.

算法基本思路是:先根据词表把输入串中的所有可能的词都找出来,然后把所有可能的切分路径(词串)都找出来,并且从这些路径中找出一条最佳路径(概率最大的)作为输出结果.

3.2 概率费用转化及最佳路径的计算

为了量化概率统计结果,可将概率转化为费用.因为每个词的概率都是一个很小的正数(小于 1),

如果汉字串较长,最后得到的各种可能的词串的概率都接近 0,无法在机器上表示和比较了.解决办法是求每词概率的对数之和,于是乘法变加法,对数后结果为负数,取反则变为正数,这个正数称作该词的“费用”.显然,对词或词串而言,概率越高则费用越低:

$$Fee(W) = \sum_{i=1}^n -\log P(W) \quad (6)$$

其中 $-\log P(W)$ 是第 i 个词的费用. 在计算出相关的费用之后,需呀 利用动态规划进行最佳路径搜索. 显然,最佳路径是“结合/成/分子/时”,它的总费用为 12.286;而按最大匹配法得到的切分结果“结合/成分/子时”的总费用是:13.451. 这种问题我们可以用多阶段决策的动态规划方法解决.

据多年人民日报语料库作统计,歧义切分现象的出现概率约为 1/110. 其中,交集型歧义切分现象占 86%,而最大概率法是比较有效的解决交集型歧义切分的分词方法. 对组合型歧义切分问题这种分词方法是无能为力的,这是因为任意两个词的费用之和不小于这两个词作为一个整体的费用. 结构分布如图 1 所示.

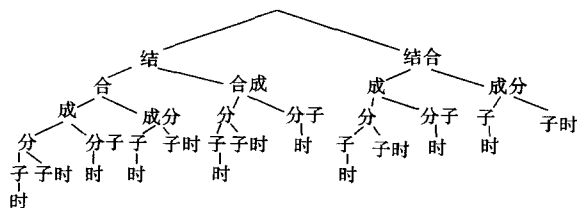


图 1 分词之间的结构分布

3.3 最大概率汉语自动分词算法的步骤

基于最大概率的分词决策算法基本步骤如下:

步骤一 对人名、地名、国名、机构名、简称、可先确定的包括前后缀、数字词短语、重叠词、成语、习惯用语、非汉语字符以及未登陆新词预识别处理

步骤二 对候选词的进行优化选择:在没有覆盖交集型歧义以及连环交集型歧义字段前提下,对组合歧义字段进行最大化覆盖;

候选词需要优化选取的原因是:前面已谈到最大概率法对组合型歧义切分问题是无能为力的,这是因为任意两个词的费用之和不小于这两个词作为一个整体的费用;另外,这种切分法能使路径集相对于全切分路径集数量可以实现几何级下降,句子越长越见效,这是因为利用动态规则寻找最佳路径所需要时间会大大缩短.

步骤三 对候选词进行分类处理

①若不存在歧义字段,则转到步骤六(A句在本文步骤三先看成不存在歧义字段,然后在组合歧义消歧时再考虑);

②若只含有普通交集型歧义字段(链长小于 2),则转到步骤四;

③若含有连环型交集歧义字段,则转到步骤五.

步骤四 处理普通交集型歧义字段

①直接调用最大概率法,进行动态规划切分;

②转到步骤六

步骤五 处理连环交集型歧义切分

①对各候选词条分别计算新的“费用”;

②调用最大概率法,进行动态规划切分;

③继续执行步骤六.

步骤六 处理组合型歧义切分

①若没有发现组合型或混合型歧义,转到步骤八;

②若发现混合型歧义,则转到步骤七;

③将组合型歧义字段全切分,若切分后候选词频每个均不足原切分前组合字段的 1/100,则可以不用考虑切分,直接转到步骤八;

④对与组合型歧义字段相邻的词计算相关度,并取平均值大的一组作最优结果;

⑤转到步骤八.

步骤七 处理混合型歧义切分

①若没有发现组合型或混合型歧义,转到步骤八

②将混合型歧义字段全切分;

③按若其中的交集型歧义字段链长不足 2,则按步骤六处理,否则按步骤五处理.

步骤八 输出切分结果.

通过以上方法,能够准去的对产生歧义的语义网络词汇进行有效的分类,最大限度的消除歧途,提高工作效率.

4 仿真实验

为了验证本文提出方法的有效性,本文采用对比实验的方法,对同一实验数据分别采用不同的算法进行歧义划分. 对比各算法的优劣性. 其中,实验选用的语料库以人民日报 2003 年 1 月的报纸全部内容为基础,以及近 500 篇精心挑选的文学作品以及特选带连环型、组合型、混合型歧义字段句子的文章累计近 400 万汉字. 充分构造歧义环境. 实验运行的硬件环境为普通台式电脑,奔腾 4CPU 处理器,10G 硬盘,2M 内存环境. 软件环境:Windows 2000

Server+VC6.0.

通过对以上数据进行统计,可以看出,在相同约束条件下,本文算法的划分准确度随着复杂程度的增加而增加,效果要明显好于传统方法.

表1 实验统计结果

	句子数	切分正确率
无歧义字段	约25万	100%
含普通交集型歧义字段	约1.5万	90.4%
含组合歧义字段	2015	85.2%
含连环交集型歧义字段	998	83.3%
含混合型歧义字段	1124	99.79%
共计	约27万个句子	89%

当在人名、地名、国名、机构名等专有名词以及未登陆新词的预识别,再依靠词频和词汇语义相关度进行动态规则分词消歧后及二次扫描再消歧后,还是未能100%消除所有组合型、连环交集型和混合型歧义字段.这跟词汇语义的相关度计算标准还不是太精确反映词义相关度很有关系.其实在进行自然语言处理时,仿人脑处理过程其实才是最科学的:人脑在消歧过程其实是一个不断学习积累的过程,而且在这分词过程中往往还结合词性、词义、句型、篇章等不同层次进行分词、词性、词义、句法进行同时消歧.

通过以上的实验数据可以看出,本文提出的算法在消除歧义方面,比传统算法更有优势,而且效率更高.虽然未能完全消除歧义语言环境,但是准确度也达到了令人满意的效果.

5 结束语

针对计算机语义网络中交集型和组合型词汇歧义的问题,本文分析了传统方法中存在的一些缺陷.

提出基于最大概率计算的自动分词歧义方法.通过计算相关的概率,对上下文语义相关度对产生歧义的词汇进行有效修正,重新计算的切分候选词所产生的有效“费用”,运用最大概率计算法对产生歧义的词汇进行关联程度概率计算,克服弊端.成功地解决交集型歧义、连环交集型歧义、组合型歧义、混合型歧义切分问题,消除语义网络中的交集型和组合型词汇歧义的影响,取得了不错的效果.

参考文献:

- [1] 赵伟,戴新宇,尹存燕,等.一种规则与统计相结合的汉语分词方法[J].计算机应用,2004(3):34-36.
- [2] 魏博诚,王爱平,沙先军,等.一种消除中文分词中交集型歧义的方法[J].计算机技术与发展,2011(5):34-36.
- [3] Jeon B, Bae J, Moon W. Automatic recognition of woven fabric patterns by artificial neural network[J]. Textile Res J, 2003,73(7):645-650.
- [4] Kuo C, Shih C, Lee J. Automatic recognition of fabric weave pattern by a fuzzy C-Means clustering Method[J], Textile Res J, 2004,74(2):107-111.
- [5] 王荔,宋胜利,冯佳,等.一种全切分与统计结合的分词系统[J].微电子学与计算机,2009(5):22-24
- [6] 方华,王振华,陆汝占,等.运用改进的分词方法进行外国译名识别的研究[J].计算机仿真,2005(3):135-136.

作者简介:

陈晓明 男,(1963—),副教授.研究方向为数据库技术与信息安全.

梁雄友 男,(1971—),研究生,系统分析员.研究方向为数据库技术、人工智能.

李伶俐 女,(1977—),讲师.研究方向为程序设计与信息安全.