

基于语义计算的中文歧义字段消歧算法

邓凡¹, 鱼滨²

(1. 西北大学 信息科学与技术学院, 陕西 西安 710069;

2. 西安电子科技大学 计算机学院, 陕西 西安 710071)

摘要:针对中文中歧义字段对中文处理及理解带来的诸多问题提出了一种基于自然语言理解的中文汉字歧义消除算法。对于交集型歧义和组合型歧义,利用《知网》为主要语义资源,以知识图知识表示方法,通过提出的字段消歧算法,对歧义字段以及上下文的语义进行计算,从而选出正确的句子切分方案,达到消除歧义的目的。经过实验数据表明本算法提高了中文歧义字段歧义切分的正确率。

关键词:自然语言理解;交集型歧义;组合型歧义;词义消歧;关联度

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2008)06-0107-04

Disambiguation Algorithm for Chinese Word Segmentation Based on Semantic Calculation

DENG Fan¹, YU Bin²

(1. College of Computer Science and Technology, Northwest University, Xi'an 710069, China;

2. College of Computer, Xidian University, Xi'an 710071, China)

Abstract: In this paper, based on a study of natural language understanding of Chinese characters algorithm to eliminate ambiguity. Aiming at cross ambiguity and combinational ambiguity, using the HowNet as the main semantic resources, using knowledge map's knowledge as representation, using the algorithms mentioned in this paper based on semantic calculations, and calculating the discrepancy field of semantic context, to select the correct sentence cut programs, so as to achieve the objective to remove ambiguities. Experiments show that the algorithm does indeed contribute to improving ambiguity cut to the correct rate.

Key words: natural language processing; cross ambiguity; combinational ambiguity; word sense disambiguation; relevance

0 引言

自然语言处理是计算机领域中的一个重要研究方向,而汉语自动分词是中文信息处理的重要基石。汉语自动分词是指将汉字串切分为正确的词串。汉语自动分词在中文信息处理的许多方向都有很重要的意义。长期制约着汉语自动分词发展的一大难题就是歧义字段的消除。所谓歧义也称歧义切分字段,是句中某个片段存在两种或两种以上的切分形式,歧义字段可能引起错误的切分,歧义现象是自动分词中不可避免的现象。歧义字段切分是影响分词系统切分精度的重要因素,所以提出或改进好的歧义字段消除算法是非常有必要和重要的。因此文中提出了一种基于语义计算的歧义消除算法,实验证明有助于提高消除歧义

字段的正确性。

1 相关知识

在歧义字段消除歧义之前需要将歧义字段识别出来,所以简单介绍一下歧义字段的产生和歧义字段的分类以及一般的识别方法。汉语词语的切分往往存在歧义^[1,2],歧义切分是自动分词不可避免的现象,对歧义的处理严重影响分词的精度^[3]。歧义产生的根本原因是中文文本的书写时的词和词之间没有界限标致,所以一个句子看起来只是一段字串。其次,汉语语素的构词能力和汉语词类的多功能性以及地名的大量存在都是产生和增加歧义字段的原因^[4]。

歧义切分字段从构成形式上主要有:交集型歧义字段和组合型歧义字段。识别交集型歧义字段一般采用双向扫描的方法,即对同一字段分别用正向匹配和逆向匹配方法,如果两种方法的切分结果不同,则认为有交集型歧义。而对组合型歧义则主要以分词词典为

收稿日期:2007-09-06

作者简介: 邓凡(1984-),女,硕士研究生,研究方向为软件工程;
鱼滨,副教授,博士,硕士研究生导师,研究方向为软件工程、Web服务与中间件技术、形式化方法和高可信软件。

依据,或者通过建立组合型歧义字段库的方式对组合型歧义进行识别。

接下来介绍一些知识图知识表示方法,因为在计算关联度时需要用到。文中使用张瑞霞等在《语义的汉语句法分析系统的研究与实现》^[5]中提到的知识表示方法,根据知网中词典的概念项构造词图。具体步骤如下:

首先,对《知网》进行分析,《知网》主要收录的词语主要分为两类,一类是实词,一类是虚词;由于虚词在中文理解中意义不大,因此在计算时没有必要对虚词构造词图。

实词构造词步骤如下:针对《知网》给出的概念以网状形式表示,对一般词语选取其概念项中的主要部分,即对其概念项进行自动抽取得到如下形式:

<DEF> = {<主要义原>}

{<主要义原>: <关系表达式>[, <关系表达式>]}

如对于词语:

眼睛:DEF = {部件: whole = {动物}, PartPosition = {眼}}, 其词图如图 1 所示。

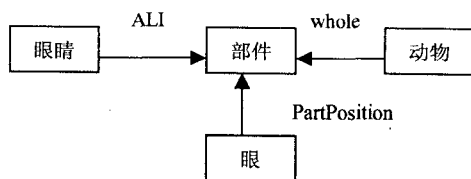


图 1 “眼睛”的词图

其中“{”和“:”之间的为“主要义原”,建立中心义原结点(如图 1 中的“部件”);为“关系表达式”中的“次要义原”建立非中心义原结点(如“眼睛”);为“关系表达式中的“主要义原”建立非主要义原结点(如“动物”和“眼”);在非中心义原结点和中心义原结点之间建立从非中心义原结点到中心义原结点的弧(如图 1 中的弧“ALI”、“whole”、“PartPosition”),弧的关系类型为响应“关系表达式”中的“角色”,弧的权值为 1。

语义规则表是参照《知网》中的《中文信息结构库》中的语义描述信息归纳出的语义序列,语义规则包括词性序列和语义序列(如表 1 所示)。

词性序列:描述实词可组成的合法的词性序列。

语义序列:描述实词间应有的语义序列。

表 1 语义规则表

规则标志号	词性序列	语义序列
...
21	n + n	{部件} + {实体}
22	v + n	{方向性自移} + {地方}
...

2 算法核心思想

基于语义计算的中文歧义字段消歧算法的基本思想是通过歧义字段前后及本身语义进行计算,得到语义关联度,根据关联度计算切分是否与上下文一致,选择最优语义序列,从而起到消除歧义的作用。算法步骤如下:

首先,利用上面的所讲的歧义字段识别方法,识别出歧义。

其次,针对每种歧义切分字段以及上下文切分出来的词语,根据上面提到了词图创建方法,创建词图。

最后,利用语义规则表,依次两两计算词语相关度。并根据相关度的平均值得到最佳切分方案。

以下是词图关联度的计算:

定义 1 备选语义规则集:语义规则表中可以用来计算词语相关度的所有语义规则序列的集合,称为备选语义规则集。

定义 2 最短语义距离:在义原树中两个义原间的最短距离。

(1) 计算义原之间的语义距离。

ele1 为某一词图 G1 中义原结点所代表的义原, ele2 为另一个词图 G2 中某义原结点所代表的义原, a 为调节参数。计算公式(1)如下:

$$\text{distance}(\text{ele1}, \text{ele2}) = \begin{cases} 0, & \text{(如果 ele2 与 ele1 同是中心义原结点)} \\ 1, & \text{(如果 ele2 是 ele1 的父辈义原)} \\ \text{epth}(\text{ele2}) * a + 1, & \text{(义原树中 ele2 为根结点的树的高度,} \\ & \text{ele1, ele2 在同一棵义原树上, 且 ele2} \\ & \text{不是 ele1 的父辈义原)} \\ \text{Max}, & \text{(如果 ele2 与 ele1 不在同一棵义原树中)} \end{cases} \quad (1)$$

(2) 计算词语相似度。

在计算词语关联度之前,还需要了解词语相似度的计算,文中采用的计算方法是对已有词语相似度计算方法的改进^[5],其中 a, b 是两个调节参数。a + b = 1, a > 0.5。semhead1 是 G1 的中心义原结点, semhead2 是 G2 的中心义原结点。semother1 是 G1 中非中心义原结点, semother2 是 G2 中非中心义原结点。similar(G1, G2)是由两部分组成。sim(semhead1, semhead2)是中心义原结点的相似度, sim(semother1, semother2)是非中心义原结点的相似度。具体公式如下:

$$\text{similar}(G1, G2) = a * \text{sim}(\text{semhead1}, \text{semhead2}) + b * 1/n \sum_{i=1}^n \text{sim}(\text{semother1}_i, \text{semother2}_i) \quad (2)$$

其中 $\text{sim}(\text{semhead1}, \text{semhead2}) = 1 / (1 + \text{Distance}(\text{semhead1}, \text{semhead2}))$

sim(semother1, semother2)非中心义原结点的计算

如下:

$\text{sim}(\text{semother1}, \text{semother2}) =$

$$\begin{cases} 1 & (\text{semother1}, \text{semother2} \text{ 指向中心义原结点弧的类型相同}) \\ 1/(1 + \text{Distance}(\text{semother1}, \text{semother2})), & (\text{义原树中 ele2 为} \\ & \text{根节点的树高度, ele1, ele2 在同一棵义原树上, 且 ele2 不} \\ & \text{是 ele1 的父辈义原}) \\ 1/(1 + \text{Max}), & (\text{semother1}, \text{semother2} \text{ 指向中心义原结点弧中} \\ & \text{没有类型匹配的弧}) \end{cases} \quad (4)$$

因此非中心义原结点计算公式为:

$$\frac{1}{n} \sum_{i=1}^n \text{sim}(\text{semother1}_i, \text{semother2}_i) \quad (5)$$

(3) 计算词语关联度。

在计算合适的词语切分组合时,是通过词语两两间的关联度的平均值进行优选工作的,这就需要计算词图关联度:即首先依据词性序列找到备选的语义规则集,然后根据语义规则计算词语的关联度,选择最大值为词图关联度。 w_1 和 w_2 词图关联度如算法 1 所示:

算法 1:

1) $\text{relation} = 0$

2) if 备用语义规则为空 then 算法结束 else 做(3)

3) if 备选语义规则已经检测为空 then 算法结束 else 从备选语义规则中取出一条规则

4) 构造 w_1 和 w_2 的词图 g_1 和 g_2

5) 根据语义序列“+”左端的概念构造词图 s_1 , 根据右端的概念构造词图 s_2

6) 计算 g_1 与 s_1 的相似度 $\text{similar}(g_1, s_1)$, if $\text{similar}(g_1, s_1) < \theta$ then 返回到 3)

7) 计算 g_2 与 s_2 的相似度 $\text{similar}(g_2, s_2)$, if $\text{similar}(g_2, s_2) < \theta$ then 返回到 3)

8) if $\text{similar}(g_1, s_1) + \text{similar}(g_2, s_2) > \text{relation}$ then $\text{relation} = \text{similar}(g_1, s_1) + \text{similar}(g_2, s_2)$

9) 返回 3)

其中,如果 θ 足够小就一定能够得到匹配的语义规则。

(4) 计算最优切分方案算法。

定义 3 歧义字段切分方案集:歧义字段所有歧异切分方法的集合。

1) $\text{sign} = 0, \text{maxrelation} = 0, \text{relation} = 0$

2) 获取歧义字段在句子中前后相邻的实词 preword 、 lastword (它们允许为空,当歧义字段发生在句首或句尾时, preword 和 lastword 为空)。

3) 从歧义字段切分方案集中取一个切分方案,如: word1 、 word2 , if 歧义字段切分方案集为空 then 退出程序。

4) 把 2) 和 3) 中的结果合并成词串队列 preword 、 word1 、 word2 、 lastword 。

5) 上述词串的队列的队首词语出队,记作 tempword , 计算 tempword 与先在队首词语关联度,记作 rela 。

6) $\text{relation} = \text{relation} + \text{rela}$

7) if 词串的队列首词语出队后队列不为空 then 返回 5)

8) $\text{relation} = 1/n * \text{relation}$ (n = 词串队列的长度 - 1)

9) if $\text{relation} > \text{maxrelation}$ then $\text{maxrelation} = \text{relation}$ $\text{sign} =$ 记录歧义字段切分方案集中方案序列

10) if 歧义字段切分方案集为空 then 程序结束退出 else 返回 3)

其中 sign 则记录正确的切分方案的序列号。

3 实验分析

本算法是对已有的歧义字段歧义的消除,所以为了对算法的效果进行评估,使用最常用的歧义消除评估方法——正确率进行评估,具体计算方法如下:

正确率 = 正确消歧的词个数 / 测试文本中歧义出现的词个数 $\times 100\%$ 。

使用小规模的人工标注过歧义的语言资料,对算法进行检验。最后手工将检验结果进行验证,研究算法歧义消除的效果,将数据总结如下:

(1) 小量数据。

找到 10 个含有歧义,且歧义字段已被标注过的句子,借鉴文献[4]中歧义句子,进行手工词性标注,对算法进行测试,运算结果见表 1。

在上面数据中,句子 5 的错误是因为在知网中常用词语的检索失败造成的,随着知网词库的不断扩大与完善将改善这一问题。而 6,9 的错误则是有上下文信息模糊不清造成的;对于本算法来说,上下文是给出正确切分的关键,如果能给出充足的上下文关系便可得到更高的正确率。

以上数据与中科院 ICTCLAS 系统进行比较:中科院 ICTCLAS 系统切分正确的是(4)、(5)、(6)、(7)、(8)、(9),其正确率只有 60%;而本算法可达到 70% 的正确率。所以对本算法研究还是很有价值的。

(2) 批量数据。

为了更进一步测试算法的正确率,扩大了测试语料的数量,在语料库中随机抽取两篇文章,随机地选取具有歧义字段的句子 56 个,因为验证过程需要人工校验,所以选取的测试数据规模不大。为了对此数据进行对比,采用同样的数据对中科院 ICTCLAS 系统进行

测试。实验结果见表 2。

表 1 运算结果

例句:	结果	分析
1 (1)这/n场/n火/n把/v他/pron全部/adv的/stru希望/n化为/v了/stru灰烬/n (2)这/n场/n火把/n他/pron全部/adv的/stru希望/n化为/v了/stru灰烬/n	(1)	正确
2 (1)张/class三/num就是/adv在/v往后/n看/v的/stru时候/n被/prep人/n打/v了/stru一/num棍子/n (2)张/class三/num就是/adv在/v往后/adj看/v的/stru时候/n被/prep人/n打/v了/stru一/num棍子/n	(2)	正确
3 (1)市长/n将/adv来/v我校/n视察/v (2)市长/n将来/n我校/n视察/v	(1)	正确
4 (1)这/n人/n才/adv不怕/v呢/n (2)这/n人才/n不怕/v呢/n	(1)	正确
5 (1)这种/adj饺子/n好/adv主要/adj是v/面n的/stru质量/n好/adj (2)这种/adj饺子/n好/adv主要/adj是v/面的/n质量/n好/adj	False	错误
6 (1)三/num个/adj人/n选/v你/pron会/adj选/v谁/pron (2)三/num个/adj人选/v你/pron会/adj选/v谁/pron	(1)	错误
7 (1)他们/pron干/v的/stru确实/adj是/v一/num起/n罕见/adj的/stru高/adj科技/n犯罪/v (2)他们/pron干/v的/stru确实/adj是/v一起/adv罕见/adj的/stru高/adj科技/n犯罪/v	(1)	正确
8 (1)破产/n公司/n可/v以/adj实物/n抵偿/v债款/n (2)破产/n公司/n可以/adj实物/n抵偿/v债款/n	(1)	正确
9 (1)所有/adj裁判/n给/v了/adj她/pron十分/num分/n的/stru最高/adj成绩/n (2)所有/adj裁判/n给/v了/adj她/pron十分/num分的/stru最高/adj成绩/n	(2)	错误
10 (1)她/pron的/stru病/n因/v我/pron而/adj起/v (2)她/pron的/stru病因/n我/pron而/adj起/v	(1)	正确

表 2 实验结果

	正确切分个数	错误切分个数	正确率
中科院 ICTCLAS 系统	41	15	73.2%
文中算法	46	10	82.1%

通过实验表明:本算法歧义除的效果稍好于中

院 ICTCLAS 系统,因为测试数据的不同将在某种程度上影响算法验证的结果,因此不同算法的比较也存在某种程度上的不准确,以上数据仅供参考。

4 结束语

词语歧义消歧是自然语言处理的一大难题,它直接影响自然语言处理的其它任务,到目前为止,许多歧义消除算法都在探索状态,还要经过很长一段时间的继续深入研究。文中阐述的算法,是基于上下文理解与词语关联度计算的消歧算法,能够在一定程度上提高歧义字段消歧的正确率,但是算法中依然存在不足的地方。

首先,知网词库的局限性,使得部分生词不能够检索到,从而不能获得其词图信息,所以无法进行词语间关联度的计算,但是随着知网词库规模的不断扩大,本算法将会有更好表现。其次,对于上下文表述含糊不清的句子,算法很难保证自己的准确性。这是算法本身需要改进的地方。

最后,在词语关联度计算的时候,将义原树中所有关联的权值都当作 1 来处理,这一点有些不合常理,也是算法今后需要改进的地方。因此对本算法的性能的提高,还需要进一步的探索和研究。

参考文献:

- [1] Zhou Jingye. On ambiguity, polysemy and fuzziness of natural language[R]. Shanghai:[s. n.],1984.
- [2] Hindle D, Rooth M. Structural ambiguity and lexical relations[J]. Computational Linguistics,1993,19(1):229-236.
- [3] 孙茂松,邹嘉彦. 汉语自动分子研究中的若干理论问题[J]. 语言文学应用,1995,40(4):40-60.
- [4] 刘禹孜. 汉语自动分词中排除歧义字段算法的研究[D]. 重庆:重庆大学,2005.
- [5] 张瑞霞. 语义的汉语句法分析系统的研究与实现[D]. 西安:西北大学,2005.

(上接第 106 页)

4 结束语

文中提出一种基于边缘检测的 Contourlet 变换图像去噪方法。实验结果显示,该方法在更好地去除高斯白噪声的同时,能更有效地保留图像的边缘信息,提高了去噪图像的 PSNR 值,去噪图像边缘鲜亮。

参考文献:

- [1] Do M N, Vetterli M. Contourlets: A directional multiresolution image representation[C]//Proc of IEEE Interna-

tional Conference on Image Processing. Rochester, NY: [s. n.],2002:357-360.

- [2] Donoho D. De-noising by softthresholding[J]. IEEE Transon IT,1995,41(5):613-627.
- [3] 刘英霞,王欣. 最佳软门限去噪[J]. 电子学报,2006,31(1):167-169.
- [4] 梁栋,沈敏,高清维,等. 一种基于 Contourlet 递归 Cycle Spinning 的图像去噪方法[J]. 电子学报,2005,33(11):2044-2046.
- [5] Clark J J. Authenticating edges produced by zero-crossing algorithm[J]. IEEE-PAMI,1998,11(1):43-57.