DOI:10.19789/j.1004-9398.2016.03.003

一种实例库与义原关系相结合的概念消歧算法

高 璐[1] 赵小兵[1,2]

(1. 中央民族大学信息工程学院,北京 100081; 2. 国家语言资源监测与研究中心 少数民族语言分中心,北京 100081)

摘 要

本文提出了一种利用知网的实例库与知网关系进行词义消歧算法. 该方法首先利用知网提供的实例库进行初步的匹配; 若在实例库中没有完全匹配,则利用上下文搭配关键词与知网中的实例搭配词进行相似度计算,若相似度大于给定阈值,则消歧结束. 否则,我们再判断歧义词的义原与关键词的义原是否具有某种关系,根据义原权值调节算法调整义原权值. 调整后的义原权值大小不一,按照事先的约定,我们选取综合权值最大的义项. 我们发现,该方法能够弥补仅依靠实例库的覆盖率低的问题,又能减少仅依靠统计方法产生的噪音,从而提高词义消歧的正确率.

关键词: HowNet,实例库,关系网络,权重调节.

中图分类号: TP391.1

1 引 言

词义消歧是一个老生常谈的问题,它并不是单独的一个系统,而是其他系统的中间环节,比如问答系统、检索系统等等. 词义消歧经过近几十年的发展,经历了从基于AI的词义消歧,到基于知识库的消歧,再到如今的基于大规模语料库的模式,取得了长足的进步. 本文的概念模型是基于HowNet的. 知网中提供了两种可满足处理不同类型词义消歧的需求资源,一种是字典记录的实例,另一个是二次资源概念相似度衡量(CSM),以尽可能的用来进行词义消歧. 将字典记录的实例与义原的关系结合起来,不仅能够弥补仅依靠实例库的覆盖率低的问题,又能减少仅依靠统计方法产生的噪音,从而提高词义消歧的正确率.

2 知网简介

知网是由董振东先生花费十余年构建而成,目前已经收录中文词12万条,英文词10万条,共包括3万多条概念和20多万条义项. 知网的表示是基于概念的,概念的定义又是基于义原的,并描述为计算机特别容易计算的结构化标记语言形式. HowNet中还定义了各种语义关系,并用语义角色等方式体现出来. 这些关系是可完全计算化的、可直接被计算机处理. 因此知网的诞生是一个不可多得的语言资源. 词语"cure"的关系网如图1所示.

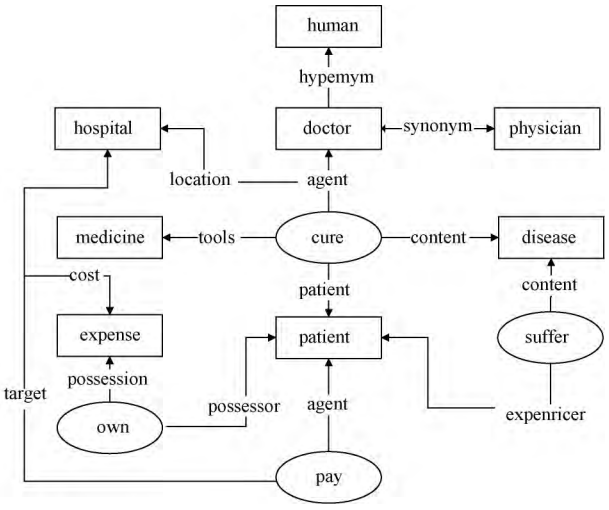


图1 cure的动态关系网

3 词义消歧模型的构建

3.1 实例匹配

HowNet最新版的系统对每一个多义词的每一个义项尽可能的提供了相应的实例,这为词义消歧

收稿日期: 2015-06-07

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

提供了方便. 如“林”字,它其中的 2 个义项在知网中的解释为:

义项一:

NO. =103108

W_C = 林

G_C = noun [in2]

S_C =

E_C = 姓~,~姓,~先生,~太太,~小姐,~女士,老~,小~,~大哥,~叔,~公,~爷爷,~伯伯,~伯母,~老爷,~家老二,~氏家族,~经理,~主任,~校长,~院长,~处长,~总理,~书记,~教授,~老师,~大夫,~护士长,~队,~总,~工

DEF = {surname|姓}

义项二:

NO. =103109

W_C = 林

G_C = noun [in2]

S_C =

E_C = 树~,森~,~木,竹~,防风~,防沙~,针叶~,阔叶~,桃~,密~深处,山~,~间小路,~中木屋,~地,在~中走失了,在~中迷失方向,承包了一片~地,这片~子,~带,~产,红~尽染,绿~

W_E = forest

DEF = {tree|树: quantity = {mass|众}}

其中,E_C 就是义项的实例.

我们可以看出,知网中的实例存储比较规范,而且每个义项的实例没有任何歧义,具有很好的区分性. 很多歧义词通过与知网的实例库进行匹配就能消除歧义. 假定歧义词为 W,在知网中它有 S1,S2,S3,……,SN 个义项,使用 ICTCLAS 提取歧义词上下文窗口为 3 的关键词 W_{-3},W_{-2},W_{-1},W_1,W_2,W_3,并且每一个关键词 W_i 在知网中的义项分别为 $S_{i,1}$,$S_{i,2}$,……,$S_{i,m}$. ($1 <= i <= 3$) 因此,基于实例库的消歧的具体步骤如下:

1) 首先将歧义词的每一个义项 Si 的实例进行初步还原. 因为知网中的实例都是以“;”来分隔的,并且词均以~来代替. 我们把歧义词还原以后,将每个歧义词的搭配拆分放到一个三维数组中. 设数组为 Array [b] [p] [q],则上面关于“林”的第 2 个义项(即 b = 2) 在二维数组 Array [p] [q] 中的表示为:

实例词	义项1	义项2	义项3	义项4
树	tree|树	cultivate|培养	establish|建立	image|图像
枫叶	plant|植物}	physiology|生理学		
山	land|陆地	shape|物形	surname|姓	
小路	route|道路			
绿	green|绿			

图 2 Array 数组存储举例

2) 实例还原以后我们就需要将关键词 Wi 与实例库进行匹配,如果 Wi 与 Array [b] [p] [q] 中的任何一个实例完全匹配,我们就称实例库完全命中,此时消歧结束;但在很多时候,在实例库中我们找不到完全匹配的实例,这时我们使用知网提供的二次资源工具概念相似度衡量(CSM) 来进行下一步的计算.

概念相似度的计算公式如下:

$$Sim(p_1,p_2) = \frac{\alpha(depth(p_1) + depth(p_2))}{\alpha(depth(p_1) + depth(p_2)) + dist(p_1,p_2) + |depth(p_1) - depth(p_2)|} \quad (1)$$

具体伪代码如下:

```
For Wi in (W-3,W-2,W-1,W1,W2,W3)
  For i from 1 to b
    For j from 1 to p
      If Wi match with Array [b] [p] [] break;
    END;
  END;
END;
For Wi in (W-3,W-2,W-1,W1,W2,W3)
  For Si,k in (Si,1,Si,2,……,Si,m(1 <=i <=3))
    If Sim(Si,k,Si) >MIN break;
    Else Next;
END;
```

3.2 义原消歧

若 $Sim(S_{i,k},S_i) < MIN$ 时,实例消歧失败. 这时我们进行下一步消歧. 我们知道,义原以网状的结构存储在数据库中,并且以一种或多种形式发生联系. 为了利用义原进行消歧,我们将义原之间不同的关系赋予不同的权值,具体的权值表如下.

当某两个义原发生如上所示的任何一种关系,则义原的权重值加上相应的权值.

具体的公式如下:

$$wt(R_i) = weight * \frac{weight(K_i)}{\sum_j weight(K_j)} \quad (2)$$

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

义原关系	权值
相同义原	3
上下位关系	2
相关关系/ 材料成品关系	1
其他关系	1

图 3　权值分配表

$$K = \max(wt(R_i)) \tag{3}$$

义原权值调节伪代码如下:

```
For Wi in (W_-3, W_-2, W_-1, W_1, W_2, W_3)
  For S_i,k in (S_i,1, S_i,2, ……, S_i,m)
    For Ki in (K1, K2, ……, Kj)
      For Ri in (R1, R2, ……, Rj)
        If(Relationship(Ki, Ri)) wt(Ri)
      END;
    END;
  END;
END;
```

因此,还是基于上面的假设,假定歧义词为 W,在知网中它有 S1,S2,S3,……,SN 个义项,其中每个义项 Si 分解为 R1,R2,……,Rj 个义原; 使用 ICTCLAS 提取歧义词上下文窗口为 3 的关键词 $W_{-3}, W_{-2}, W_{-1}, W_1, W_2, W_3$,并且每一个关键词 W_i 在知网中的义项分别为 $S_{i,1}, S_{i,2}, \cdots\cdots, S_{i,m}$,并且每个 $S_{i,k}$ 分解为 K1,K2,……,Kj 个义原. 设每一个义原的初始权值为 1.

对于歧义词 W 的每一个义项所分解的每一个义原 Ri,我们都要与关键词 Wi 的每一个义项所分解的每一个义原进行关系梳理,查看两个义原是否发生以上的关系. 若发生任意一种关系,则相应的义原加上对应权重. 具体流程图如下.

最后,我们取权值最大的义原所在的义项为我们所求的义项.

4　实验结果

实验中我们采用现代汉语语料库中常用的十个高频词进行测试. 对于每个歧义词我们都提取了50 ~ 100 个句子作为测试对象,并从知网中查询每个歧义词的义项和义原数. 具体的实验结果如下.

本文算法所采用的测试数据的平均义项数是 5,平均义原数为 7,平均正确率为 81. 1%. 我们发现,当歧义词的义项数越少,则正确率越高. 这是因为义项数越少,义项间的实例可区分度越大,消歧正确率越高. 同时我们发现,歧义词的义原数越多,相应的正确率并没有提高. 这是因为虽然义原数的增多会在一定程度上增加义原的关系,利于消歧的计算,但是义项数也会相应增加,从而不利于义项的区分. 同时我们也研究了消歧正确率在各词性之间的比例,具体如下.

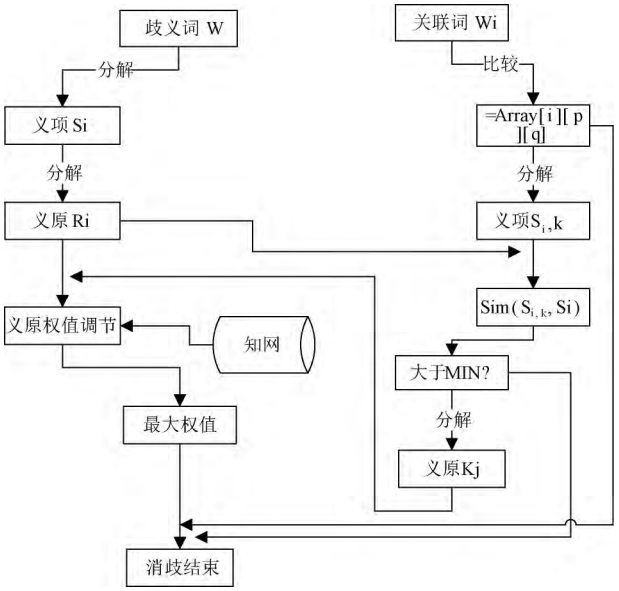


图 4　消歧流程图

歧义词	义项数	义原数	句子个数	正确率
材料	3	4	60	0.84
保守	2	2	50	0.82
穿	4	4	50	0.81
发表	2	2	60	0.84
要	8	11	90	0.77
吃	6	8	80	0.81
漏洞	2	7	70	0.83
病毒	2	12	50	0.84
起	13	20	100	0.75
改	2	2	50	0.80
平均值	5	7	66	0.811

图 5　实验结果

分类	总资源	歧义资源	正确数	正确率 (%)
名词	5893	1517	1141	75.21
动词	4431	1201	590	49.13
形容词	1551	284	210	73.94
副词	1479	299	171	57.19

图 6　各词性实验结果对比

实验结果表明,名词和形容词的正确率分别为 75. 21% 和 74. 94%,相对来说正确率比较高; 而动词和副词的正确率仅为 49. 13% 和 57. 19%,正确率较低一点. 这是因为在知网中名词和形容词的关系网络更为健全一点,从而为义原权值调节带来更多

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的节点信息. 而动词和副词则稍微差一点. 目前知网每一年都会进行添加新词和剔除旧词,并不断完善各种词语搭配. 相信随着义原的敏感性逐渐增加,消歧的结果一定会更加准确.

参 考 文 献

[1] 董振东,董强(1999),"知网"[DL]. 1999. http://www.keenage.com.

[2] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[J]. 中文计算语言学,2002,7: 59-76.

[3] 朱嫣岚,闵锦,周雅倩,等. 基于 Hownet 的词汇语义倾向计算[J]. 中文信息学报,2006,20(1):14-20.

[4] 王广正,王喜凤. 基于知网语义相关度计算的词义消歧方法[J]. 安徽工业大学学报(自然科学版),2008,25(1):71-75.

[5] 于林林. 基于知网的汉语词义消歧方法的研究[D]. 沈阳:沈阳航空工业学院,2008.

[6] 詹靓雯. 基于知网的中文词义消歧研究[D]. 杭州:中国计量学院,2012.

[7] 但汉辉. 一种汉语词义消歧方法的研究[D]. 重庆:重庆大学,2009.

[8] 董振东,董强,郝长伶. 知网的理论发现[J]. 中文信息学报,2007,21(4):3-9.

[9] 闫蓉. 基于语义的汉语词义消歧方法研究[D]. 西安:西北大学,2006.

[10] Lin Dekang. An Information2 Theoretic Definitionof Similarity Semantic distance in WordNet [C]. San Francisco: Proceedings of the Fifteenth International Conference on Machine Learning,1998.

[11] 李峰,李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007,21(3):99-105.

A Concept Disambiguation Algorithm Based on the Combination of Instance Base and Sememe Relations

Gao Lu[1] Zhao Xiaobing[1,2]

(1. Department of Information Engineering, Minzu University of China, Beijing 100081;
2. National language resources monitoring and Research Center Minority Language Branch, Beijing 100081)

Abstract

This paper proposes a WSD algorithm based on HowNet. The method firstly uses case base that HowNet provides for preliminary matching; if the case library do not exactly match, then we compute the similarity between the context matching keywords and HowNet examples collocations. If the similarity is greater than a given threshold, then the end of disambiguation. Otherwise, we judge that if there has some relations between sememes of keywords and sememes of case library and adjust the original weight value according to the original weight adjustment algorithm. After the adjustment, the weights of sememes are not the same size. In accordance with the prior agreement, we selected the largest comprehensive weight items. We find that this method can make up for the problem that only depend on the low coverage of the instance library and can also reduce the noise generated by the statistical method, thus improving the correct rate of word sense disambiguation.

Key words: HowNet, case library, Network of relations, adjustment of weights.

作者简介 高璐(1989-),中央民族大学信息工程学院硕士研究生,专业为计算机软件与理论,研究方向是自然语言处理.
赵小兵(1967-),现为中央民族大学信息工程学院研究生导师,教授. 国家语言资源监测与研究中心少数民族语言分中心主任. 研究方向是自然语言处理,计算语言学.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net