文章编号:1001 - 9081(2010)06 - 1671 - 02

基于改进的 VSM 的词义排歧策略

赵晨光[1,2], 蔡东风[2]
(1. 沈阳航空工业学院 电子信息工程学院,沈阳 110136; 2. 沈阳航空工业学院 自然语言处理实验室,沈阳 110136)
(zcg0975@ tom. com)

摘 要:为了提高词义排歧的准确率,提出了一种基于改进的向量空间模型(VSM)的词义排歧策略,该模型在提取特征向量的基础上,考虑了语法、词形、语义等因素,计算语境相似度,并引入搭配约束,改进了算法的效果,在开放测试环境下,词义标注正确率可达到 80% 以上。实验结果表明,该方法对语境信息的描述更加全面,有利于进一步的语义分析。

关键词:向量空间模型;词义排歧;语境相似度;特征向量;词语搭配
中图分类号: TP39 **文献标志码**:A

Word sense disambiguation based on improved vector space model

ZHAO Chen-guang[1,2], CAI Dong-feng[2]
(1. *Institute of Electronic Information Engineering, Shenyang Institute of Aeronautic Engineering, Shenyang Liaoning* 110136, *China*;
2. *Natural Language Processing Laboratory, Shenyang Institute of Aeronautic Engineering, Shenyang Liaoning* 110136, *China*)

Abstract: To increase the word disambiguation accuracy, a word disambiguation solution based on improved Vector Space Model (VSM) was presented. Since the algorithm takes account of grammar, morphology and semantic and calculates the context similarity requiring the character vector abstraction, the algorithm is able to achieve better results by using collocation constraint. The open test precision can reach 80%. The result shows that the method can fully describe the features of context, and is beneficial to further semantic parsing.

Key words: Vector Space Model (VSM); word disambiguation; context similarity; character vector; word collocation

0 引言

一直以来,词义排歧是困扰自然语言处理的难题之一,歧义的自动消解对于自然语言理解是至关重要的。

Senseval[2] 竞赛作为语义排歧领域中最为权威的国际会议,在最近的报告中指出基于语料的方法比基于知识的方法效果要好,在参与 English All Words 单元的 20 个测试系统当中,以 Wordnet 作为语义标准,标注的准确率最高达 69%[1]。目前,国内很多学者就语义自动标注做了大量研究和探索,吴光远等人对真实文本进行语义标注的正确率稳定在 80%[2]。

通常来说,语义排歧是基于其出现的特定语境,即词语上下文。上下文在很大程度上体现了词语的基本语义特征。同时,词与词之间的搭配关系对于消歧也有不可忽视的作用。

考虑到上述原因,本文提出了一种基于向量空间模型计算语境相似度,并辅以搭配约束来进行语义标注的方法。该方法综合利用了多种知识实现词义消歧。这些知识包括:上下文词的词性、位置、语义和歧义词的词法等,在一定程度上改进了词义排歧的效果。

1 词义标注算法

1.1 词义标注流程

词义排歧的过程实际上是在特定语境下给出多义词的确切义项。而语境即上下文往往是确定词语意义的基本依据[3],找到词语在语境中出现的规律和内在联系是非常关键的。

首先选取常用的歧义词,并从已经过分词和词性标注语料中抽取得到若干个关于该歧义词的实例[4],即样本。然后,利用《同义词词林》对其中的单义词进行自动词义标注(单义词占总词数的60% 左右),对多义词采用机助手工标注。由此,构造了初始的训练集。训练集中的每个词 W_i,都有其唯一的词性标记 P_i,如果为《同义词词林》收录的单词还应有词义标记 S_i,否则为空。这样,可以充分利用当前待排歧词位置的上下文信息,将每一个样本 V 用若干个特征来表示:

$$V = (W_{-10}, P_{-10}, S_{-10}, \cdots, W_0, P_0, S_0, \cdots, W_{10}, P_{10}, S_{10}) \tag{1}$$

以上,构造了向量空间模型(Vector Space Model, VSM)中的特征向量。其中,每个特征词即 $\{W_{-10}, W_{-9}, \cdots, W_9, W_{10}\}$ 对于目标词,即待排歧词都有一定的表征程度,则定义每个特征词(除副词、连词、介词、等虚词外)的特征权值如下:

$$Weight(W_{ij}) = \frac{C(W_i, S_j)}{\sqrt{\sum_j C^2(W_i, S_j)}} \times \frac{1}{dist(W_i, S_j)} \tag{2}$$

其中:$C(W_i, S_j)$ 为特征词 W_i 与歧义词候选义类 S_j 的共现次数,分母为归一化因子;$dist(W_i, S_j)$ 为特征词 W_i 与候选义类 S_j 的距离。若 $C(W_i, S_j)$ 越大,$dist(W_i, S_j)$ 越小,则特征权值也就越大。

因此,计算语境相似度的问题就归结为计算两个向量之间的相似度。图 1 为利用 VSM 完成词义标注的过程,其中,计算语境相似度是基于语境的语义排歧的重要前提,这包括根据上下文语法、语义等信息来体现歧义词的不同语义的应用环境。

1.2 语法信息计算

为了完成上述词义标注,需要计算语境相似度,进而实现

收稿日期:2009 - 12 - 03;**修回日期**:2010 - 02 - 05。
作者简介:赵晨光(1975 -),女,辽宁沈阳人,讲师,硕士,主要研究方向:自然语言处理、信息检索。

语义排歧的目的。语境相似度是指两个语境的匹配符合程度。这不仅依赖于语境中出现的词汇，而且还依赖词汇之间的关系[5]。而词汇间关系涉及到对词形、词性等信息的统计分析。需进行以下处理：

1）初始化。在分别对训练语料和测试语料预处理后，得到训练语料向量矩阵 V 和测试语料向量矩阵 V'：

$$\begin{cases} V = \{W, P, S\} \\ V' = \{W', P', S'\} \end{cases} \tag{3}$$

2）计算分量相似度。设任意两个语境向量 V 和 V'，处理过程如下。

① 词性信息处理。

个别词性标记为虚词，包括数词、助词、量词等词语与歧义词意义联系不大[6]，但如果在排歧过程中同等对待，由此可能带来不必要的错误。因此根据词性信息可将与排歧关联不大的词语滤掉。

② 词形相关度计算。

$$sim(W, W') = MI(W, W') = \log \frac{P(W, W')}{P(W) \times P(W')} \tag{4}$$

其中：$P(W, W')$ 代表特征词 W, W' 在语境向量中出现的概率，$MI(W, W')$ 为两者的互信息。

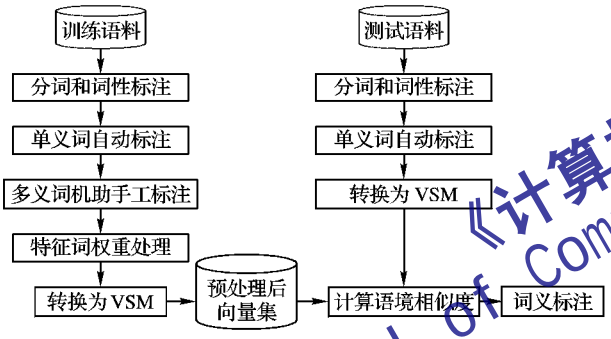


图 1 利用 *VSM* 进行词义标注流程

1.3 语义相似度计算

基于知网的概念相似度计算被应用到语义排歧，其关键是如何计算语义的义原相似度[7]。

本文提出语义相似度计算方法，具体步骤如下：

1）利用知网提供的树状义原层次体系计算路径距离的方法，计算各义原相似度，并分类得到语义大类、中类、小类匹配的词的个数。

2）利用式(5)计算语义相似度。

$$sim(S, S') = \frac{\sum_{i=1}^{3} Match(S_i, S_i') \times Val_i}{Len(S) + Len(S')} \tag{5}$$

其中：$Match(S_1, S_1')$、$Match(S_2, S_2')$、$Match(S_3, S_3')$ 分别为语义大类、中类、小类匹配的词的个数；Val_i 为义类相同时的权重，由于语义小类相同时词义已非常接近，中类相同词义稍有差异，而大类相同词义间的差异更大，因此设 $Val_1 > Val_2 > Val_3$；$Len(S)$，$Len(S')$ 为两个语义向量的长度。

1.4 修正

句子的内部结构和词语之间的深层语义联系是体现语境相似度的重要因素[8]。折中的方法就是依据上述两种方法得到词形、语义相似度，然后对每种方法赋予不同的权重，并求和。

$$sim(V, V') = sim(W, W') \times Weight_1 + sim(S, S') \times Weight_2 \tag{6}$$

其中：$sim(W, W')$ 为词形相似度，$sim(S, S')$ 为词义相似度，$Weight_1$ 和 $Weight_2$ 分别为词形和词义权重，两语境向量相似度为各分量相似度加权之和，权值根据经验值设定。

1）确定义项似然度。

在计算了测试集各句子与训练集中句子间的语境相似度后，由此可得到测试句的义项似然度：

$$SIM_{S_i}(V') = \sum_{i=0}^{NUM(V_{s_i})} \frac{sim(V_{s_i}, V')}{NUM(V_{s_i})} \times Weight(V_{s_i}) \tag{7}$$

其中：$SIM_{S_i}(V')$ 为测试句取义项 S_i 的可能性，$sim(V_{s_i}, V')$ 为测试句与义项为 S_i 的训练集中句子的相似度，$NUM(V_{s_i})$ 是义项为 S_i 的训练集中句子的数目，$Weight(V_{s_i})$ 为上文介绍的训练语境的权重。

2）利用搭配知识校正结果。

如果，测试句 V' 中待排歧词 W 的前后实词 W_x 存在于搭配词库中该词的词对形式中，则

$$SIM_{S_i}(V') = SIM_{si}(V') + \frac{1}{dist(W, W_x)} \tag{8}$$

其中 $dist(W, W_x)$ 为歧义词 W 与实词 W_x 的距离。

2 实验结果

本文选择 1998 年 1 月的《人民日报》作为训练语料，主要以歧义词作为研究对象，在进行语义标注前在语料中抽取了大量词语搭配，以《同义词词林》和《知网》分别对其进行语义标注，初步构建了一个词语搭配库。

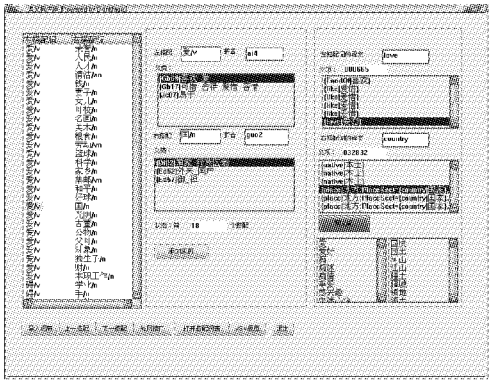


图 2 词语搭配库的结构

在网上选取了涵盖 100 个歧义词的 10 万条例句作为测试集，按窗口大小[-10,10]对例句进行截取，按《词林》标准对其中单义词自动标注后，依据以上方法对歧义词进行语义标注。测试结果如表 1 所示。

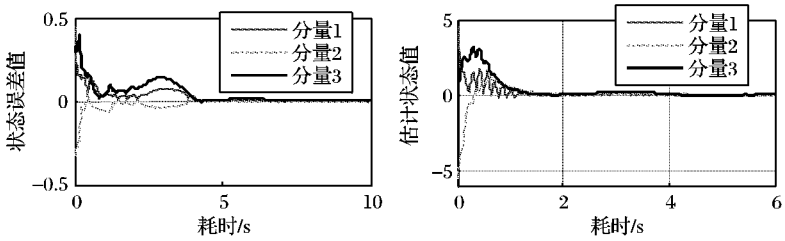
表 1 排歧测试结果

歧义词词性	词数	测试句数	基于 VSM 标注准确率/%	基于 VSM 并引入搭配信息标注准确率/%
动词	62	80000	67.2	78.3
名词	27	20000	82.7	86.5
形容词	11	10000	85.9	89.2

从对上述实验结果的分析表明，从词性角度来看，动词的歧义性是最为突出的，在语义现象相对复杂的情况下（如“发动机吃油”与“少吃油是减肥的关键”），标注时往往会发生混淆。

（下转第 1693 页）

动,$\tilde{e}$在上步已解出。由于A_1、A_2均不稳定,设$w_1 = 0.48$,$w_2 = 0.52$,而凸组合$A_0 = \sum w_i * A_i$可以稳定,对其采用切换率,延时$\eta_1 = \eta_2 = 0.05$,阈值$v = 0.000\,01$时,耗时 6 s,切换 78 次,误差范数 0.003 7,状态曲线如图 2(b),2 s 后误差趋近 0,$\hat{g}$即为x的状态估计。



(a) 切换率为3.3的估计误差收敛曲线　　(b) 切换率为3.3的状态估计收敛曲线

图 2　扰动下误差及状态估计收敛曲线

5　结语

寻找相应凸组合而使A_0稳定的前提是切换系统有可收敛的特征子空间[12],通过上述 3 类状态及组合延时切换算法可使系统稳定,利用 Matlab 工具亦可方便进行设计和实现。通过寻找较好的w和L等系数,确保系统快速稳定收敛。3 类切换率收敛算法互有优劣,工程实践中可以组合运用。比较而言,同样时间内,反馈切换次数大于阈值修正切换次数,而状态延时切换次数最少,收敛速度则相反,对于扰动抖动系统则需要进行阈值约束减少切换次数和收敛误差。

参考文献:

[1] LIBERZON D, MORSE S. Basic problems in stability and design of switched systems[J]. IEEE Control System, 1999, 19(5): 59 – 70.

[2] SUN Z D, GE S S. Switched linear systems control and design [M]. London: Springer, 2005.

[3] SUN X M, WANG W, LIU G P, *et al*. Stability analysis for linear switched systems with time varying delay[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008, 38(2): 528 – 533.

[4] JOHANSSON M, RANTZER A. Computation of piecewise quadratic Lyapunov functions for hybrid systems [J]. IEEE Transactions on Automatic Control, 1998, 43(4): 555 – 559.

[5] ZHANG LIGUO, LI HONGFENG, CHEN YANGZHOU. Robust stability and L2-gain analysis for uncertain discrete time switched systems with time-delay [C]// Proceedings of the 7th World Congress on Intelligent Control and Automation. Chongqing: [s. n.], 2008: 845 – 850.

[6] SUN Z, GE S S. Dynamic output feedback stabilization of a class of switched linear systems [J]. IEEE Transactions on Circuits and Systems: Fundamental Theory and Applications, 2003, 50(8): 1111 – 1115.

[7] XU XUPING, ZHAI GUISHENG, HE SHOULING. Stabilizability and practical stability of continuous time switched systems: A unified view [C]// Proceedings of the 2007 American Control Conference. Washington, DC: IEEE, 2007: 663 – 669.

[8] SPINELLI W, BOLZERN P, COLANERI P. A note on optimal control of autonomous switched systems on a finite time interval [C]// Proceedings of the 2006 American Control Conference. Washington, DC: IEEE, 2006: 5947 – 5952.

[9] HETEL L, DAFOUZ J, IUNG C. Stability analysis for discrete time switched systems with temporary uncertain switching signal [C]// Proceedings of the 46th IEEE Conference on Decision and Control. Washington, DC: IEEE, 2007: 5623 – 5628.

[10] SUN Z, GE S S. On stability of switched linear systems with perturbed switching paths [J]. Journal of Control Theory and Applications, 2007, 4(1): 18 – 25.

[11] MAHMOUDI A, MOMENI A. On observer design for a class of impulsive switched systems [C]// American Control Conference. Washington, DC: IEEE, 2008: 4633 – 4639.

[12] SUN Z, GE S S. Switched stabilization of higher-order switched linear systems [C]// Proceedings of the 44th IEEE Conference on Decision and Control and the European Control Conference. Washington, DC: IEEE, 2005: 4873 – 4878.

(上接第 1672 页)

值得注意的是,同一个词在不同语境下带有明显的语境色彩和倾向性[9]。而且,对输入文本切分不正确也会影响标注结果。同时,还存在着严重的数据稀疏问题。但总的来说,本文提出的这种基于改进 VSM 的词义排歧方法充分利用了现有的语义、语法甚至语用知识;并利用了现有的词语搭配库,在一定程度上改善了词义标注结果。

3　结语

本文结合实例介绍了基于向量空间模型进行词义标注的策略,正确率在 80% 以上,表明这一算法对词义排歧的效果有一定改善。未来工作的重点是扩充训练语料以及词语搭配库,并在特征词选取等环节上做细致地调整,以期得到更为理想的排歧效果。由于中文语义、语料资源的限制,词语语义自动标注仍将是一个长期而艰巨的任务,本文仅仅在语义研究中得到了初步的结果,距离实现真实文本语义标注还有一定的差距。

参考文献:

[1] YING DING. IR and AI: Using co-occurrence theory to generate lightweight ontologies [C]// Proceedings of 12th International Workshop on Database and Expert Systems Applications. Washington, DC: IEEE, 2007, 9: 961 – 965.

[2] 吴光远,何丕廉,曹桂宏,等. 基于向量空间模型的词共现研究及其在文本分类中的应用[J]. 计算机应用,2003,23(6):138 – 145.

[3] ATLAM E-S. A new method for construction field association terms using co-occurrence words and declinable words information [C]// Proceedings of 2002 IEEE International Conference on Systems, Man and Cybernetics. Washington, DC: IEEE, 2002, 4: 95 – 100.

[4] 郭池,陈家俊. 一种基于语料库的词义消歧策略[J]. 计算机工程与应用,2003,39(6):121 – 125.

[5] 周舫. 汉语句子相似度计算方法及其应用的研究[D]. 郑州:河南大学,2005.

[6] 王荣波,池哲儒. 基于词类串的汉语句子结构相似度计算方法[J]. 中文信息学报,2005,19(1):25 – 32.

[7] 刘群,李素建. 基于《知网》的词汇相似度计算[EB/OL]. [2002 – 10 – 10]. http://www.keenage.com.

[8] CHE WANXIANG, LIU TING, LI SHENG. A new Chinese nature language understanding architecture based on multilayer search mechanism [C]// The Third SIGHAN Workshop on Chinese Language Processing. Barcelona, Spain: [s. n.], 2004: 134 – 400.

[9] TURNEY P D, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315 – 346.