上下文边界可变的贝叶斯分类器词义消歧方法

吴崇斌[1,2] 张全[2]
[1]中国科学院研究生院 北京 100049
[2]中国科学院声学研究所 北京 100190
E-mail：bearwcb007@163.com

摘 要：词义消歧在自然语言处理中具有重要作用，而基于贝叶斯分类器的的方法是一种较为常见的词义消歧方法。本文针对基于贝叶斯分类器词义消歧方法所需的特征的上下文边界进行探究，提出在训练过程中对特定多义词给出在训练中使该词的词义消歧效果最优的上下文边界，即上下文边界可变的贝叶斯分类器词义消歧方法，通过实验证明该种方法对词义消歧正确率有一定的提高。
关键词： 词义消歧 贝叶斯分类器 上下文边界可变

Word Sense Disambiguation Based on Bayesian Classifier with Variable Context Window

Wu Chongbin[1,2] Zhang Quan[2]
[1]Graduate University of Chinese Academy of Sciences, Beijing 100049
[2]Insitute of Acoustics, Chinese Academy of Sciences, Beijing 100190
E-mail：bearwcb007@163.com

Abstract: Word sense disambiguation (WSD) takes an important place in the Natural Language Processing, and the Bayesian Classifier is a usual model to realize WSD. To improve the performance of the Bayesian Classifier-based method in WSD, this paper proposes a solution that can choose an optimal context window in train data for each word which has several senses. Also, by carrying out experiments, this paper proves that the solution proposed can really improve the accuracy of WSD.
Keywords: Word sense disambiguation, Bayesian Classifier, variable context window

1 引言

一词多义的在多数自然语言中都是常见的现象。在人际交流过程中，这种现象往往可以被利用来进行模棱两可的表示或起到一语双关的作用。然而，对于机器而言，要正确分析语句结构和进行语义分析，或者要正确的实现机器翻译，却需要对多义词进行词义消歧。

当前，词义消歧的方法主要分为有监督的消歧方法和无监督的消歧方法两大类。在有监督的消歧方法中，训练数据是已知的，即每个词的语义分类是被标注了的；而在无监督的词义消歧方法中，训练数据是未被标注的。另外还有基于词典信息的消歧方法，也被人们作为一种专门的词义消歧方法加以研究。[1]

那么，在有监督的词义消歧方法中，目前主要有基于互信息的消歧方法和基于贝叶斯分类器的消歧方法这两种。本文的研究是针对后者进行的，希望可以找到一些改进的方法。

2 理论基础及相关研究

2.1 基于贝叶斯分类器的词义消歧方法

基于贝叶斯分类器的消歧方法由 W. A. Gale 等人于 1992 年提出，其基本思想是：多义词的

语义取决于该词所处的上下文语境 c，如果某个多义词 w 有多个语义 s_i(i≥2)，那么可以通过下列式子计算得到 w 的词义 s：

$$s = arc\max_{s_i} P(s_i \mid c) = arc\max_{s_i} \frac{P(c \mid s_i)P(s_i)}{P(c)}$$

在计算过程中，可以忽略分母 $P(c)$，并运用如下独立性假设：

$$P(c \mid s_i) = \prod_{v_k \in c} P(v_k \mid s_i)$$

因此实际的词义计算式子是：

$$\hat{s} = arc\max_{s_i}[P(s_i)\prod_{v_k \in c} P(v_k \mid s_i)]$$

尽管在实际文本中，上下文中每个词并非互相独立，因此独立性假设似乎不太合理。但在很多情况下这种假设作为一种简化的方法却很有效，使得贝叶斯分类器在词义消歧中的应用得到推广。

2.2 相关研究

在基于贝叶斯分类器的消歧方法中，窗口的大小是影响处理结果的因素之一。文献[2]采用交叉验证[1]的方法，针对SemEval-2007 的 40 个中文多义词——包括 19 个名词和 21 个动词——数据进行处理后得出该数据集的上下文最优边界为[-2, +2]。然而，这种方法得出的结果仅限于对特定数据集有效，毕竟实验对象只有 40 个词。如果换成SemEval-2010 的中文数据，其中包括 27 个多义动词，且只有 2 个动词出现在SemEval-2007 的数据中，那么[-2, +2]这个上下文边界就未必是最优的。[2]

本文认为，文献[2]的问题在于，它对不同的多义动词都套用同一个上下文边界。当前针对基于贝叶斯分类器的消歧方法改进有诸多方向，比如文献[3]提出基于信息增益的特征选择方法，挖掘上下文中词语的位置信息，提高贝叶斯模型知识获取的效率，从而改善词义分类效果；又如文献[4]利用依存句法分析，从句子的内部结构，寻找词语之间支配与被支配的关系，借以确定能够对词语语义构成内在限制的上下文，克服单纯贝叶斯分类器中无关上下文造成的噪声影响。然而，通过上下文边界的改进来优化贝叶斯分类器的方法则比较鲜见，现有的此类研究多数也如文献[2]一样将同一个上下文边界套用到不同的多义词上。

以常识推理，在贝叶斯分类模型中，每个多义词应该各自有其最佳上下文边界，而不是套用统一的上下文边界。因此，本文的主要研究目的就在于探究一个为不同多义词设置不同上下文边界的贝叶斯分类器消歧方法，即上下文边界可变的贝叶斯分类器词义消歧方法，是否有助于提高消歧效果。

3 上下文边界可变的贝叶斯分类器消歧方法

前面提到，每个多义词都有一个适合自己的上下文边界，这仅仅是从常识上进行推理。那

[1] 交叉验证（Cross-validation）是一种统计学上将数据样本切割成较小子集的估计方法，有多种具体实现方法，本文涉及的是 K-折交叉验证（K-fold cross-validation），即把训练数据切分成 K 份，取一份作为测试验证数据，其余 K-1 份作为训练数据，重复 K 次，每一份都验证一次，平均 K 次的结果得到单一估值。

[2] SemEval 是由 ACL-SIGLEX 主办的国际语义评测活动，2007 年为第四届，共 19 个任务，2010 年为第五届，共 18 个任务。

么，本文对 SemEval-2010 的部分中文多义动词采用朴素贝叶斯方法消歧，上下文边界统一地从[- 1, + 1]到[- 5, + 5]渐变，得到表 1 的数据，可以为这一推理提供实验证据。

表 1

动词	最大准确率（%）	最小准确率（%）	最大准确率 上下文边界	最小准确率 上下文边界
带	75.00	37.50	[-1, +2] [-1, +3] [-2, +3]	[-2, +1] [-3, +1] [-4, +1] [-5, +1]
来	50.00	12.50	[-3, +1] [-3, +2]	[-1, +2] [-1, +3]
死	90.00	50.00	[-3, +4]	[-2, +2]
有	76.92	67.69	[-3, +5]	[-1, +5] [-4, +3] [-5, +2] [-5, +3] [-5, +4]

由表 1 不难看出，不同的多义动词都有适合自身的上下文边界，而且所取边界的不同对消歧效果有显著的影响。假如能够在训练过程中，给每个多义动词分别设置合适的上下文边界，消歧的效果应当能有所提升。

为了实现给每个多义动词分别设置合适上下文边界的目的，本文的基本思路是：在训练过程中，将上下文边界从[- 1, + 1]到[- 7, + 7]或者从[- 1, + 1]到[- 5, + 5]渐变，对训练语料进行反复测试，记录取得最高正确率的上下文边界，将其作为该多义动词的合适上下文边界。而如表 1 所示，有些动词可能在若干上下文边界的条件下都能取得最高正确率，因此本文采取的处理方法是：先将这若干边界作为候选集，然后在候选集中取中间的一个作为合适边界，例如候选集有 2n 个或 2n-1 个选项就取第 n 个，选项顺序由先前上下文边界渐变的方式决定。当然，这只是权宜之计，更科学更有效的选取方法仍需进一步研究。至于边界选择范围的上限 5 或 7 的选取，首先是考虑到文献[5]指出[- 8, + 9]是一个通用的上下文边界，而文献[2]又得出 SemEval-2007 中文多义动词数据的最优上下文边界是[-2, +2]，那么加上对 SemEval-2010 中文多义动词数据的初步研究并对上述两篇文献研究结果进行折中，从而选取了 5 和 7，在二者之间也能做一些比较，看看选取范围对消歧准确率的影响如何。

而从利用训练语料进行测试的方法上，实验对采用两种方式并做了比较：第一种是在初步训练后，将各个动词的全部训练语料作为测试语料进行测试，将测试结果作为被测边界对应的正确率；第二种是将每个动词的训练语料均分为五份，进行交叉验证，即每次利用其中四份进行初步训练，然后对余下的一份进行测试，一共进行五次，将五次测试结果的均值作为被测边界对应的正确率。

4 实验结果与分析

4.1 实验使用数据

本文实验的数据来自 SemEval-2010 的中文动词词义消歧任务的训练数据和测试数据，包含多义动词 27 个，每个词的训练数据数量和测试数据数量如表 2 所示：

表 2

词形	发生	发展	给	获得	没有	属于	有	吃	穿
训练数量	41	127	41	40	41	40	271	65	41
测试数量	9	31	9	10	9	10	65	16	9

词形	带	分	干	管	见	交	结束	进行	开展
训练数量	42	41	41	42	47	41	41	78	40
测试数量	8	9	9	8	10	9	9	19	10

词形	来	弄	去	送	算	跳	写	做	死
训练数量	42	41	41	41	40	42	44	41	40
测试数量	8	9	9	9	10	8	6	9	10

4.2 实验结果数据及初步分析

实验中选取词语的词性标注符号作为统计特征，标点符号也被当做词对待。用 SemEval-2010 的训练数据完成训练，再用其测试数据进行测试检验，而测试数据也有答案。实验得到的结果如表 3 所示：

表 3（a）：Micro-P＝SUM(动词 i 词义消歧准确率)/27

左\右	1	2	3	4	5	6	7
-1	72.10	74.01	72.34	73.14	74.95	74.74	72.92
-2	73.93	73.64	73.72	73.25	75.55	74.38	71.60
-3	72.74	73.96	74.84	75.12	77.09	73.79	72.54
-4	71.88	77.21	74.32	74.03	76.04	75.29	73.35
-5	72.97	74.91	73.13	74.10	75.65	74.04	74.31
-6	71.97	75.13	75.16	73.76	75.63	75.39	75.41
-7	73.71	76.66	75.19	75.18	75.85	74.99	75.35

表 3（b）：Macro-P＝SUM（动词 i 词义消歧准确率*动词 i 测试数量）/27 个动词测试数量总和

左\右	1	2	3	4	5	6	7
-1	74.78	75.07	74.18	74.18	74.78	74.18	73.00
-2	75.67	74.48	74.78	74.48	75.96	74.48	72.40
-3	73.59	74.18	75.37	76.56	78.34	75.37	73.89
-4	73.29	76.85	74.78	75.07	76.56	75.67	73.89
-5	73.89	74.78	73.89	75.07	76.56	74.48	74.48
-6	71.51	74.78	74.78	73.89	75.96	75.37	75.37
-7	73.29	76.85	75.37	75.37	75.96	75.37	75.37

表 3（c）：综合数据对比

	Micro-P（5）	Macro-P（5）	Micro-P（7）	Macro-P（7）
最大值	77.21	78.34	77.21	78.34
最小值	71.88	73.29	71.60	71.51
均值	74.19	75.09	74.31	74.86
第一种训练	75.59	76.26	73.24	73.89
第二种训练	73.68	78.04	72.86	76.85

表 3（a）和（b）是在统一上下文边界情况下，对不同的上下文边界取值得到的 Macro-P 值

和 Micro-P 值，这两个数值的计算公式已在表头给出。表 3（c）展示的是上下文边界在[-5, +5]和在[-7, +7]范围中选取时分别得到的上下文边界固定与可变两类方法的结果；最大值、最小值和均值是在表 3（a）和（b）的数据中统计得到，作为统一边界数据的代表与第一种训练和第二种训练的结果进行比较，其中第一种训练无交叉验证，第二种训练有交叉验证；列标题中的 5 和 7 是指上下文边界的取值范围是[-5, +5]或是[-7, +7]。

从表 3（c）初步来看，不管是否采用交叉验证的方法来训练，由本文思路得出的实验结果没有明显的优势。不过，如果将本文思路得到的结果和统一边界得到的结果均值进行细致分析，则可以得到一些有意义的发现。这里采用均值作为比较对象，原因在于：

首先，最大值尽管是最理想的结果，是本文力图达到甚至超越的对象，但这是在得到正确答案后才能证实为最佳的数据，因此将其作为比较对象不合理；而与最小值比较则没有意义。

其次，由于统一边界的取值因人而异，可能是取某个经验值，抑或是通过某种算法计算得到，因此不妨将统一边界的取值视为某种随机事件。而在其它条件（包括测试数据和算法等）都不变的情况下，可以视最终处理准确率为边界的函数，则函数的值也将满足某种随机分布。而均值作为这种随机分布的一个基本统计量，将其作为与本文思路得到的结果的比较对象是合理并且有意义的。

通过仔细研究和比较会发现，对于不采用交叉验证的训练方法，将上下文边界限制在[-5, +5]之内选定所得到的测试结果比统一边界的均值略高，而将限制扩大到[-7, +7]时得到的结果却比统一边界的均值略低，而且 Micro-P 值和 Macro-P 值相对接近，差距都在 0.7 个百分点以内；而对于采用交叉验证的训练方法，Micro-P 值和 Macro-P 值的差距明显拉大，达到 4 个百分点，由此也使得这种训练方法的 Macro-P 值超过统一边界的 Macro-P 均值高出 2 个百分点，尤其是在[-5, +5]的选择范围内，其 Macro-P 值与统一边界的 Macro-P 最大值相差无几，而反观 Micro-P 值，却要比统一边界的均值略低。

4.3 进一步分析

那么，从以上的初步分析可知，对于上下文边界选取范围的限定，并非越宽松越好。以下将选取测试效果较好的将范围限定在[-5, +5]的数据做进一步分析。 对于通过采用交叉验证得到的训练结果，之所以测得的的 Macro-P 值比 Micro-P 值高出 4 个百分点以上，而从这两个值的均值来看又和未采用交叉验证得到的对应数值相差无几，原因在于采用交叉验证后对测试数量多的动词测得的准确率比未采用交叉验证的准确率要高，而对测试数量少的动词测得的准确率却可能反而比未采用交叉验证的准确率要低。这一点可以从表 2 和表 4 的数据得到印证。

表 4：未采用交叉验证与采用交叉验证的准确率（%）

词形	发生	发展	给	获得	没有	属于	有	吃	穿
无交叉验证	100.00	77.42	33.33	90.00	88.89	70.00	70.77	93.75	77.78
交叉验证	100.00	80.65	44.44	90.00	88.89	60.00	75.38	87.50	77.78

词形	带	分	干	管	见	交	结束	进行	开展
无交叉验证	62.50	44.44	88.89	87.50	50.00	88.89	88.89	100.00	80.00
交叉验证	37.50	66.67	88.89	87.50	50.00	88.89	88.89	100.00	80.00

词形	来	弄	去	送	算	跳	写	做	死
无交叉验证	25.00	100.00	88.89	55.56	100.00	75.00	66.67	66.67	70.00
交叉验证	37.50	100.00	100.00	44.44	100.00	50.00	50.00	44.44	70.00

结合表 2 和表 4 数据可知，测试数量最多的三个动词是：发展、有、进行。这三个词采用交叉验证后，除了“进行”维持在 100%（不可能再提高）外，“发展”和“有”都又提高，而且这两个词的测试数量之和为 96 个，占全部 337 个测试数据的 28%，因此这两个词的测试准确率对 Macro-p 值产生影响。再看看余下 24 个动词的测试准确率，与无交叉验证相比，有 4 个提高，13 个持平，7 个下降，除了“吃”具有 16 个测试数据外，其他动词都只有 8 到 10 个测试数据。因此出现了 Macro-P 值升高而 Micro-P 值降低的结果。再从训练数据来看，只有“发展”和“有”的训练数据超过 100 个，第三、四位的“进行”和“吃”的训练数据分别只有 78 个和 65 个，其余的都只有 40 个左右。那么，在进行交叉验证时，原本不多的训练数据却又要将其中五分之一作为测试数据而不能参与训练，尽管在均分数据是充分考虑到词义的平衡问题，但仍可能使训练结果发生偏移，导致交叉验证的置信度下降，影响最终对合适上下文边界的选取。

5 总结

本文针对基于贝叶斯分类器的词义消歧提出了在训练过程中为不同多义词训练得出不同的、对该多义词较合适的上下文边界，从而提高各多义词的词义消歧准确率，最终提高词义消歧整体性能的方案。通过实验，本文发现该方案确实对词义消歧的准确度有所提高，尽管实验得到的数据体现出来的提高效果并不明显，但本文认为这主要是由于训练数据量太小所导致，同时，在从候选集中选取合适上下文边界的环节上也还有较大的改进空间，因此本文认为，在训练数据量较为充足的情况下，上下文边界可变的贝叶斯分类器消歧方法的改进是有效的。

参 考 文 献

[1] 宗成庆. 统计自然语言处理. 清华大学出版社, 2008

[2] 李纲, 寇广增, 夏晨曦, 全吉, 张东赫. 中文词义消歧上下文最优边界问题研究. 知识组织与知识管理, 2009 第 7/8 期

[3] 王达, 张坤. 贝叶斯模型在词义消歧中的应用. 计算机时代, 2009 年 第 7 期

[4] 范冬梅, 卢志茂, 张汝波, 潘树粲. 基于信息增益改进贝叶斯模型的汉语词义消歧. 电子与信息学报, 2008 年 第 12 期

[5] 鲁松, 白硕. 自然语言处理中词语上下文有效范围的定量描述. 计算机学报, 2001, 24 (7) : 742 - 747.

[6] 卢志茂, 刘挺, 张刚, 李生. 基于依存分析改进贝叶斯模型的词义消歧