# 贝叶斯分类在词义消歧中的分析

汤小娜 苏劲松

(厦门大学计算机科学系,福建 厦门 361005)

**摘 要**:词义消歧是自然语言处理中的一个核心问题,尝试了基于单纯贝叶斯概率模型的消歧方法,取得了好的效果。由于该方法在抽取上下文特征时没有进行合理的选择,致使一些无用的信息混入其中降低了贝叶斯分类器的分类准确率。利用词根词性提高了上下文特征抽取的有效性,并且尝试寻找上下文中的指示词这种特征进行消歧。

**关键词**:单纯贝叶斯分类;词根;词义消歧;指示词

引言

词义消歧问题在自然语言处理的各个方面都有非常重要的作用。对于机器翻译问题,如何理解自然语言的结构与歧义是提高机器翻译质量的关键。针对于信息检索问题,如何能够很好的对词语进行消歧,才能够了解用户的检索需求,为用户检索出需要的信息。由此可见,词义消歧问题是自然语言处理中,重要的核心问题。贝叶斯模型是一个简单的模型,不考虑上下文的结构和词语顺序,并且上下文的单词彼此独立。但是实验证明在有一定训练数据的情况下消歧效果很好。

1 实验思想与算法

1.1 贝叶斯概率模型。贝叶斯分类是通过歧义词上下文窗口来判断歧义词的词义,上下文中的每一个词语提供了确定词义的潜在的有用信息。贝叶斯规则如下 $P(s_k|c)=\frac{P(c|s_k)}{P(c)}P(s_k)$

要得到正确的分类,可以消除 P(c)。若歧义词为 w,上下文为 c,上下文中的词语为 $v_j$。则为 w 指定一个语义 $s:s=\mathrm{argmax}_{S_k} P(S_k|c)=\mathrm{argmax}_{S_k} P(c|S_k)P(S_k)=\mathrm{argmax}_{S_k}[\sum_{v_j in c}\log P(v_j|S_k)+\log P(S_k)]$

其中 $P(S_k)$和 $P(v_j|S_k)$可以使用最大似然估计从标注的训练语料中计算

$$P(v_j|s_k)=\frac{c(v_j,s_k)}{\sum_t C(v_t,s_k)} \qquad P(s_k)=\frac{c(s_k)}{c(w)}$$

其中公式中 $C(v_j,S_k)$是训练语料中 $v_j$ 在语义 $S_k$ 的上下文中出现的次数,$C(S_k)$是 $S_k$ 在训练语料中出现的次数,C(w)是歧义词 w 出现的总次数。

1.2 模型参数的平滑。每一个词都可能是另外一个词的各个语义的上下文,高质量的标注训练语料是很难大量获得的,造成了训练参数的数据稀疏,用平滑来估计那些没有被统计到的参数。在这里使用 Lidstone 法则

$$P(v_j|s_k)=\frac{c(v_j,s_k)+\lambda}{\sum_t C(v_t,s_k)+\lambda N} \qquad \lambda\in(0,1)$$

选择小的 λ 是为了避免太多的概率空间转移到无知事件上,实验中 λ 值是随机取的,λ 取(0.4,0.6)时效果最好。

1.3 算法的缺点。上下文通常取自以目标词为中心的单词窗口,并且不考虑与目标词的距离及语法关系等。因为目标词往往不仅仅和其周围的词有语义联系,而且与更广的上下文有关。因此难以得到较好的结果。

1.4 改进思想。上下文单词考虑词根,词性,同时考虑更大窗口的指示词。

步骤:通过语料库找到每一义项中概率最大的几个词作为初始指示词集;然后在含有初始指示词集的训练语料中搜索,找出那些在待消歧多义词前后 6 个词语中重复出现 2 次或 2 次以上的实词作为第一批指示词集,以后的每一步循环都是在含有上一层指示词集的训练语料中进行搜索,找出那些在上一层指示词集前后 6 个词语中重复出现 2 次或 2 次以上的单词作为下一批指示词集,以此类推从而逐步扩增新的指示词,当找不到新的指示词时循环停止。

2 实验及其结果分析

2.1 实验数据。训练语料与测试语料

2.1.1 brown 语料库。该语料库(选择 Brown1)几乎所有单词都标有词性,词根和义项,义项是根据 WordNet 标注的。

根据实验统计单词表单词数是 34639.经过词根化处理的词数是 23254。对于选择测试的单词,要使得在训练语料和测试语料都达到一定的出现次数,同时要避免某个义项所占比例太大。其中预测的三个单词为 have,was,和 that。

单词义项个数分别是 30,18,8,训练句子的个数分别是 715,2366,2092 和测试句子的个数分别是 656,1577,1672

2.1.2 Senseval2 中具有大量标注语料的三个单词 ard line serve。其中每个单词义项个数 3,6,4,训练语料中标注句子的个数分别是 4224,4042,4268,测试语料中句子的个数分别是 107,110,108。其中训练语料和测试语料中义项的比例是一致的。

2.2 实验结果及其分析(见表 1,2,3)

2.2.1 把上下文的单词转化为词根后,效果更好,也更稳定。在 Brown 语料库中考虑了词性,测试率提高 2~7%个百分比。在单词 have 中,考虑词性正确率反而下降了,由于 have 的义项分歧主要在于作为动词的不同,前面后面单词词性影响并不大,这样增加前后单词的词性,反而会增加了特征的稀疏性。总体上,Brown 语料库实验效果不好,我们认为是语料不够大。在 Senseval 中有大量的训练语料,正确率提高很大。由实验可知,词性,词根和足够的标注语料都可以提高实验的结果。

2.2.2 通过寻找指示词来消歧,与改进前相比,正确率没有很大提高,有可能是上一代指示词集合中多个指示词确定下一代指示词集合中的同一个指示词。但实验证明是可行的。

2.2.3 训练语料所取上下文的个数会对识别率和正确率产生影响,当个数多时,指示词集合增大,单词的识别率也提高了,但是过多的上下文会造成噪音,使得正确率下降,同时,取相同的上下文个数,不同单词之间的识别率有较大差异,例如 hard 和 serve。这是由于不同的单词的指示词出现的位置是不同的。serve 在歧义词周围取 7 个单词时效果最好,训练文本中第 7 个词多为 prepare,cook ,food,vegetable 等性质的词,这些词是很好的指示词。对于 hard 反而是数目越多,效果越差,它的指示词多集中在上下文 3 个单词中。

2.2.4 所采用的词根表来自于 brown 语料库,还有很多单词没有收集到,对实验结果有一定的影响。

Brown 语料的实验结果

| | | 不考虑词性 | | 考虑词性 | |
|---|---|---|---|---|---|
| that | count | 单词原型 | 单词词根 | 单词原型 | 单词词根 |
| | 1 | 74.4450 | 74.7010 | 76.5383 | 75.4952 |
| | 2 | 78.1459 | 76.3589 | 79.3421 | 80.5383 |
| | 3 | 81.4785 | 79.8373 | 81.6579 | 83.4187 |
| | 4 | 82.3828 | 84.7416 | 84.5957 | 86.2708 |
| have | 1 | 55.8129 | 55.4878 | 52.5914 | 51.5244 |
| | 2 | 57.3171 | 57.7744 | 54.4207 | 53.9634 |
| | 3 | 59.8435 | 60.8232 | 55.7927 | 53.9634 |
| | 4 | 62.0427 | 63.2622 | 56.0926 | 55.7927 |
| was | 1 | 44.0076 | 44.8320 | 48.5098 | 49.7146 |
| | 2 | 46.6614 | 48.5637 | 49.7781 | 52.1097 |
| | 3 | 47.1833 | 48.8320 | 51.2365 | 52.4413 |
| | 4 | 47.5739 | 48.9201 | 52.3779 | 52.8662 |

Senseval 语料的实验结果

| 上下文左右的个数 | hard | | line | | serve | |
|---|---|---|---|---|---|---|
| | 原型 | 词根 | 原型 | 词根 | 原型 | 词根 |
| 3(*2) | 90.65~92.5% | 90.65~94.4% | 72.54% | 75.49~78.43% | 90.74~91.66% | 91.66~93.51% |
| 4(*2) | 88.78~91.59% | 88.78~91.59% | 72.54~74.50% | 75.49~76.47% | 87.03~88.88% | 90.74~93.51% |
| 5(*2) | 87.85~90.65% | 88.78~90.65% | 77.45~78.43% | 74.5~83.33% | 88.88% | 90.74~91.66% |
| 6(*2) | 82.24~84.11% | 84.11~88.78% | 82.35~87.25% | 82.35~85.29% | 91.66~92.59% | 92.59~94.44% |
| 7(*2) | 82.24~85.04% | 85.04~85.98% | 82.35~86.7% | 82.35~85.29% | 92.59~94.44% | 94.44~95.37% |
| 8(*2) | 81.30~85.04% | 84.11~8504% | 84.31~89.21% | 88.23~89.21% | 91.66~92.59% | 91.66~92.59% |
| 9(*2) | 84.11~85.04% | 85.04~85.98% | 80.39~83.3% | 85.29~86.27% | 91.66~92.59% | 91.66~92.59% |
| 10(*2) | 82.24~84.11% | 84.11~87.85% | 80.39~82.35% | 85.29~86.27% | 83.33~86.11% | 87.96~91.66% |
| max | 92.6% | 94.4% | 89.21% | 89.21% | 94.44% | 95.37% |

利用指示词的实验结果

| 上下文大小为6(*2) | hard | | line | | serve | |
|---|---|---|---|---|---|---|
| | 识别率 | 正确率 | 识别率 | 正确率 | 识别率 | 正确率 |
| 原型 | 79.11% | 82.35% | 80.39% | 75.61% | 73.15% | 82.28% |
| 词根 | 80.37% | 86.05% | 82.35% | 78.57% | 71.07% | 88.75% |

3 总结

文本使用单纯贝叶斯模型进行词义消歧,在有大量标注训练语料中,消歧效果不错。由于该方法在抽取上下文特征时没有进行合理的选择,致使一些无用的信息混入其中降低了贝叶斯分类器的分类准确率。利用词根词性提高了上下文特征抽取的有效性,并且尝试寻找上下文中的指示词这种特征进行消歧,思想是可行的,但仍需要改进。

**参考文献**

[1]卢志茂,刘挺.基于依

(下转 8 页)

# 有机化工生产中工艺条件的选择

王红军

（江苏省淮安职业技师学院，江苏 淮安 223001）

摘 要：有机化工生产中主要工艺条件的选择方法上，有很多相似之处，教师在教学过程中应及时引导学生进行归纳总结，并促进学生在学习中不断探讨，不断完善。

关键词：工艺条件；选择；归纳总结

《有机化工生产技术》是职业学校化工工艺专业的主要专业课程之一，这门课着重讲述烃类热裂解、碳一、碳二、碳三、碳四及芳烃系列有机产品的生产技术，对产品的性能和应用、工业生产方法、反应原理、工艺条件选择、工艺流程组织等进行了简明阐述，以能力培养为中心，培养学生分析问题、解决问题的综合能力，增强学生自主学习以获得知识的能力。根据这一特点，教师在教学过程中应对学生提出具体的要求，引导学生及时地进行总结归纳，有助于学生尽快地掌握这门课程的内容。

在基本有机化工生产中，虽然产品的种类和数量很多，但在生产技术的讨论方法上却有许多相似之处。比如，在工艺条件的选择这个方面，首先是可以分为气—固非均相催化反应、液相均相催化反应、非催化反应等主要类型；其次是各个产品生产技术中工艺条件的讨论方法有相似之处，可以及时总结出来。以下是本人在多年的教学工作中，充分运用归纳总结法，对有机化工生产中各方面工艺条件的选择进行总结所得出的结论，对学生理解掌握工艺条件的选择很有帮助。因为有机化工生产中绝大多数是气-固非均相催化反应，故在此主要讨论这类反应的工艺条件的选择方法。

在气—固非均相催化反应中，工艺条件主要有反应温度、反应压力、空间速度、原料配比、催化剂活性、原料纯度等方面。现就这几个重点工艺条件的选择方法的相似之处进行探讨。

1 反应温度

所有有机化工产品的生产过程都要讨论反应温度的选择，可以从热力学分析、动力学分析、热量控制、催化剂的活性四个方面着手。例如，乙烯环氧化生产环氧乙烷，主反应与深度氧化副反应之间存在着激烈竞争问题，解决这一问题的技术关键是反应温度。热力学分析，主、副反应都是强放热反应，必须采用低温操作。动力学分析，主反应的活化能比副反应的活化能低，反应温度升高，副反应的反应速率比主反应增加更快，乙烯转化率提高，选择性下降，放热量增大，低温对主反应有利。如不能及时有效控制反应热，会产生“飞温”现象，影响生产正常进行。适宜的反应温度还与催化剂的活性有关，在催化剂使用初期，活性较高，为防止催化剂过热，延长其使用时间，宜选择温度范围的下限；随着使用时间增长，活性逐渐下降，为保持生产稳定，宜相应提高反应温度。在银催化剂作用下，乙烯在373K时环氧化，产物几乎全部是环氧乙烷，但反应速率很低。工业生产中，综合考虑各方面因素，一般选择反应温度为493~573K。

2 反应压力

首先是热力学分析，从化学平衡角度来看，如主反应是气体分子数减少的反应，则理论上采用加压操作对于主反应有利（如环氧乙烷生产）；若主反应是气体分子数增多的反应，则低压操作有利（如烃类热裂解）。其次是动力学分析，压力提高，能加快反应速率，这对所有反应都是一样的。如反应平衡常数很大，压力的影响基本可以忽略，一般采用常压操作（如甲醇氧化制甲醛生产技术）。再就是要考虑加压操作所受的限制，因提高操作压力就增加了对反应器的材质、反应热的导出以及催化剂的活性和使用寿命等的要求。

3 空间速度

空间速度是影响反应转化率和选择性的重要因素之一，它对所有气-固非均相催化反应的影响都是相同的。空间速度增大，原料气与催化剂的接触时间缩短，使转化率降低，同时副反应减少，选择性提高。空间速度减小，接触时间加长，转化率提高，副反应加剧，选择性下降。空间速度的确定取决于许多因素，当其它条件确定后，主要取决于催化剂的性能，催化剂活性高可采用高空速，催化剂活性低则采用低空速。提高空间速度既有利于反应器的传热，又能提高反应器生产能力。如环氧乙烷生产中空间速度的操作范围一般为4000~8000$h^{-1}$，而采用铜基催化剂的低压法合成甲醇，工业生产上空速可达10000~20000$h^{-1}$。

4 原料配比

原料配比的选择要先看反应种类。如果是氧化反应，则首要考虑爆炸极限的制约，比如甲醇氧化制甲醛生产技术，为避免爆炸极限，银法生产中采用甲醇过量，原料混合气中甲醇的操作浓度高于爆炸极限上限(>36%)；铁钼法生产时甲醇-空气混合气中，甲醇浓度低于爆炸极限下限(<6.7%)，是在空气过量的情况下操作的。如果是非氧化反应，则需考虑在多种原料中选择哪一种或哪几种过量，比如合成气制甲醇生产技术，原料气中$H_2$与CO的理论配比是2比1，但CO过量不好，不仅对温度控制不利，而且会引起羰基铁的积聚，使催化剂失去活性，故一般采用$H_2$过量；$H_2$过量可以抑制高级醇、高级烃和还原性物质的生成，提高粗甲醇的浓度和纯度，过量的氢还可以起到稀释作用，且因氢的导热性能好，有利于防止局部过热和控制整个催化剂床层的温度，工业上采用铜基催化剂的低压法合成甲醇，一般控制$H_2$与CO的摩尔比为(2.2~3.0)比1。

5 催化剂的活性

催化剂的活性高低对转化率影响很大，而催化剂的活性又与催化剂的组成有关，通常固体催化剂由活性组分、助催化剂、载体三部分组成。比如，乙烯与醋酸气相催化氧化偶联合成醋酸乙烯酯，活性组分钯的含量愈高，催化剂的活性愈高，但催化剂的活性太高会加剧副反应的发生，而钯是贵金属，其含量高会增加成本，钯在载体表面上的分散度也应适宜，一般控制钯含量约为3.0kg/$m^3$催化剂；金的存在可防止钯的凝聚，使钯在载体上有良好的分散度，从而提高催化剂的活性，延长催化剂的寿命，金的含量一般为1.4kg/$m^3$催化剂左右；助催化剂醋酸钾，不仅可提高催化剂的活性，而且能提高反应的选择性，醋酸钾的含量通常为钯含量的10倍；载体是影响催化剂活性的重要因素，它是整个催化剂的支架，醋酸乙烯酯生产中广泛采用硅胶为载体。其它气—固催化反应的产品生产技术中，催化剂的活性的讨论方法也与此类似。

6 原料纯度

这方面比较次要，一般都是考虑原料中的杂质成分对生产过程的影响。多数产品的生产中都会提到硫和硫化物、磷化物、砷化物以及卤化物等会使催化剂中毒的杂质，一般要求它们的含量要控制得很低，如丙烯腈生产中要求原料中硫含量<0.005%，电石乙炔法制氯乙烯生产中要求原料气不含硫、磷、砷。同时，还会提到其它一些有机物或无机物杂质，通常会对反应过程造成不良影响，降低目的产物的产量和纯度，使收率下降，分离困难。

以上阐述了有机化工生产中几个主要工艺条件的选择讨论的方法，也是具有共性的一面。要对工艺条件的选择进行深入细致的探讨，还有许多问题需解决，毕竟各种产品的生产过程还有许多个性和不同之处。在教师带领学生学习的基础上，更重要的是培养学生自主学习的能力。总之，教师在教学过程中要注重养成学生的良好学习习惯，使学生积极主动地追求知识，充分运用归纳总结等各种学习方法，最大限度地促进他们在长期的学习过程中能够触类旁通，不断总结，不断探讨，不断完善。

**参考文献**

[1] 曾之平，王扶明.化工工艺学 [M]．北京：化学工业出版社,1997.

[2] 梁凤凯，舒均杰.有机化工生产技术 [M]．化学工业出版社,2003.

（上接38页） 存分析和贝叶斯网络的无指导汉语词义消歧[J].高技术通讯,2004,2.

[2]全昌勤,何婷婷,姬东鸿等.从搭配知识获取最优种子的词义消歧方法[J].中文信息学报,2005,1.

[3]David Yarowsky.Word-sense disambiguation using statistical models of Roget's categories trained on large corpora [A].In COLING 14[C].Nantes,1992.545。