基于特征自动选取的汉语词义消歧[1]

何径舟，王厚峰

(北京大学信息科学技术学院，计算语言学研究所，北京 100871)

E-mail: {hejingzhou,wanghf}@pku.edu.cn

摘要：自然语言处理的许多问题都可以归结为分类问题，汉语词义消歧是一类典型的分类问题。在分类问题中，特征的选择至关重要。通常情况下，特征的选择由人工直接确定。这样的选取方式，要求选取者对于分类问题本身和机器学习模型的特点都有比较深刻的认识。本文设计了一套基于特征自动选取的Naïve Bayes模型用于汉语词义消歧问题。大量的实验测试表明，自动特征选取方法选取的特征，在相同的训练数据集上，可以改进词义消歧效果。

关键词：特征自动选取， Naïve Bayes模型，词义消歧

Chinese Word Sense Disambiguation Based on Automatic Feature Selection

He Jingzhou, Wang Houfeng

(Institute of Computational Linguistics, School of Electronic Engineering and Computer Science, Peking University, Beijing, Zip Code: 100871, China)

E-mail: {hejingzhou,wanghf}@pku.edu.cn

Abstract: Many tasks in Natural Language Processing (NLP) come down to classification problems. One typical instance of them is word sense disambiguation(WSD). In classification problems, feature selection is of great importance. Usually, feature selection is determined manually. It requires a deep understanding of the problem itself and machine learning models. In this paper, a Naïve Bayes Model based on automatic feature selection is designed for Chinese WSD. Experiments show that features selected with automatic methods can improve the model accuracy on the same corpora.

Keywords: automatic feature selection, Naïve Bayes Model, word sense disambiguation

1 问题简介

自然语言中的大量词汇都具有多义性，需要消解歧义。例如，汉语词“中医”可能表示“一类医生”，也可能表示“一种医学”；英语词 “pen”作为名词时，可以表示“笔”，也可以表示“围栏”。词义消歧就是在特定的上下文中对多义词确定正确义项的过程。如果将一个词的每个义项看成一个类，那么词义消歧实际上就是分类过程。

在SemEval2007国际评测的词义消歧任务中，几乎所有的参赛队都采用了机器学习的方法，包括全指导，半指导或者无指导的方法。使用机器学习方法，首先必须选取特征。

特征指的是将输入x与类别y联系起来的观测事实，它表明了对于输入x的分类决定中起作用的因素：

[1]本文受国家自然科学基金（No.60675035）和北京市自然科学基金（No.4072012）资助。

$$f(x,y) \equiv [\varphi(x) \wedge y = y_i] \tag{1}$$

特征权重则表明了特征在分类决策中的重要程度。基于训练语料和测试语料同分布的假设前提，机器学习模型在训练语料上估计特征权重参数，用于测试语料的分类决策。

虽然大多数机器学习模型在学习的过程中，能够区分重要的特征和不重要的特征，但是基于不同的模型和实现方式，这种区分能力在机器学习模型之间各不相同。很多研究和实验表明，一味地不加筛选地加入新特征，仅依赖模型本身的甄别能力，对于模型的最终学习效果可能起反作用。因此，如何选取合理的特征集，直接影响机器学习方法的效果。

在实际应用中，学习模型涉及到的特征数量可能数以万计。获取特征的一般方法是制定特征模板，然后根据模板在每条训练样例上分别抽取具体特征。而模板往往是根据领域知识和以往的实验经验由人工制定。这造成以下几个问题：(1) 对于模板制定者的要求很高。不仅对于问题本身要非常熟悉，对于机器学习模型的特点也必须有所了解；(2) 对于新的问题，由于缺乏实验数据和经验，初始特征的选取将显得异常困难；(3) 对于特征模板需要进行反复实验，以找出更合适的模板，这不仅耗时耗力，而且模板质量难以得到量化的评定。

因此，本文设计了一套基于特征自动选取的 Naïve Bayes 模型方法，用于解决汉语词义消歧问题。

2 系统设计

2.1 Naïve Bayes 模型

Naïve Bayes 模型是一种典型的概率分类模型。根据 Bayes 公式：

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{2}$$

对于给定的观测数据 x 确定类别 y，就是要选取后验概率$P(y|x)$取最大值的 y，即：

$$y = arg\ max_{y \in C} P(y|x) = arg\ max_{y \in C} P(x|y)P(y) \tag{3}$$

一般来说，观测数据x_i以特征向量的形式给出，即$x_i = (f_{i1}, f_{i2}, \ldots, f_{in})^T$。其中，$f_{in}$表示与$x_i$相关的第 n 个特征。假设这些特征条件独立，那么上式又可改写成：

$$y = arg\ max_{y \in C} \prod_{i=1}^{n} P(f_{ri}|y)P(y) \tag{4}$$

在训练分类模型时，可采取最大似然估计法；又为了各概率值不为 0，可做适当平滑。在本文涉及的模型中，对$P(f_{ri}|y)$采取如下平滑方式：

$$P(f_{ri}|y) = \frac{\alpha+\sum_{x_k} Num(f_{ri},x_k,y_r)}{\alpha m_r+n\sum_{x_k} Num(f_{ri},x_k,y_r)} \quad (5)$$

其中，α是平滑因子，取值在[0, 1]；$Num(f_{ri},x_k,y_r)$表示在y_r类的所有数据x_k中特征f_{ri}出现的次数；n 表示每条观测数据的特征向量长度；m_r表示与y_r相关的特征总数。

2.2 自动特征选取算法

本文用到的待选特征模板如表 1 所示。

特征模板类别	特征模板项	描述
以位置区分的单个特征	W_{-2}, W_{-1}, W_1, W_2	出现在目标词-2，-1，1，2 位置的词
	P_{-2}, P_{-1}, P_1, P_2	以上位置词的词性
语法和组合特征	$W_{-2}W_{-1}$, $W_{-1}W_0$, W_0W_1, W_1W_2	词的二元共现
	$W_{-2}W_{-1}W_0$, $W_0W_1W_2$	词的三元共现
	$P_{-2}P_{-1}$, $P_{-1}P_0$, P_0P_1, P_1P_2	词性的二元共现
	$P_{-2}P_{-1}P_0$, $P_0P_1P_2$	词性的三元共现
	WP_{-2}, WP_{-1}, WP_1, WP_2	词和词性的共现
主题特征（词袋特征）	$W_{[-10,10]}$	出现在[-10, 10]窗口内且不区分位置的词
实体特征	InEntity, EType	是否出现在实体内，以及实体的类别

表 1：词义消歧的待选模板

为了从待选特征模板中选取合适的模板，本文提出了自动特征选取算法。

对于具有 N 个模板项的可选特征模板，所有选择方案共 2^N-1 种。一一尝试几乎是不可能。因此自动特征选取算法采用 Bootstrapping 的策略来进行特征选取。主要思想是：在已经选取一定数量模板项的基础上，给每个尚未选取的模板项评分，来决定是否选择该模板项。每轮评分只选取得分最高的那个模板项。

自动特征选取算法需要一个机器学习分类模型算法 M，本文使用 Naïve Bayes 模型。并且需要训练语料 T 和评价语料 D 作为特征选取的依据，最后还需要待选的特征模板 FT 作为输入。算法输出一个 FT 的子集 FS 作为输出的优化特征模板。

FS 是一个有序的模板项序列，P 是记录 FS 中每一项评分的序列，初始都为空。对于 FT 中每一条尚未纳入 FS 的模板项 F_i，算法将 FS∪{F_i}作为特征模板，利用 M 在 T 上训练出一个分类模型 C_i。在 D 上可以得出每个模型 C_i 的准确率 P_i，算法选取 Pi 最高的模型 Ci 对应的模板项 Fi 加入 FS，并在序列 P 中记录 F_i 的评分。重复这一步骤，直到 FT 中所有的模板项加入 FS。最后，找到 P 中的最高评分对应的下标 k，将 FS 中前 k 个模板项作为算法输出返回。

针对多词词义消歧任务，P_i可以选用多词词义消歧结果的宏平均 MacroAve（Macro-Average Accuracy）和微平均 MicroAve（Micro-Average Accuracy）综合考虑。通过下面的公式定义：

$$P_i = \frac{2 \times MicroAve \times MacroAve}{MicroAve + MacroAve} \tag{6}$$

$$MicroAve = \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{N} n_i} \tag{7}$$

$$MacroAve = \frac{\sum_{i=1}^{N} m_i / n_i}{N} \tag{8}$$

其中，N 表示词数，m_i表示第 i 个词正确消歧的次数，n_i表示第 i 个词一共出现的次数。

如果每次只选取一个模板项，在早期的几轮评分阶段，算法更容易选择包含复杂信息的模板项。比如在 Table1 中，“语法和组合特征”就比“以位置区分的单个特征”更容易被优先选中。但是先选中的模板项获得较高的评分，并不一定代表后续选择的模板项能够得到较高评分。

解决的办法是对于选取的前 n_0个模板项，列举所有可能的组合方式找出最合适的初始项组合。这个考察需要进行$C_N^{n_0}$次。根据在 SemEval2007 语料上进行实验获得的统计结果[2]，n_0取 3 对于词义消歧是比较合适的。

并不需要检查所有的模板项，当算法若干轮内选取的模板项评分持续下降并且已经下降到某一阈值δ，就可以中止算法。

完整的特征自动选择算法伪码表示如图 1。

```
Input: M, T, D, FT
Initialize: FS=∅ Scores=∅
for each initialFSet_i={ f_i, f_j, f_k } ⊂ FT (i!=j!=k):
    use initialFSet_i as feature template for M to learn a classifier C_i on T
    p_i = accuracy of C_i on D
initialPSet = { p_i }
i' = argmax_i initialPSet
FS = FS∪initialFSet_i'
Scores[1..3] = p_i'
for k=4 to |FT|:
    for each f_i in FT:
        use FS∪{f_i} as feature template for M to learn a classifier C_i on T
        p_i = accuracy of C_i on D
    PSet = { p_i }
    i' = argmax_i PSet
    FS = FS∪{ f_i' }
```

[2] 实验使用 SemEval2007: Task #5 的测试语料进行特征自动选取，比较了 n_0从 1 到 8 的情况下特征评分的走向。其中 n_0=3，7，8 时特征评分都达到了最高值。我们选取较小的 n_0作为以后选取的参考。

```
    Scores[k] = p_l'
    if Scores[k]<Scores[k-1]<Scores[k-2] and Scores[k]< δ then end for
k' = argmax_k Scores
FS = FS[1..k']
return FS
```

图 1：特征自动选择算法伪码

3 实验

3.1 基于 SemEval2007 语料的实验

数据来自于 SemEval2007：Task #5 Multilingual Chinese_English Lexical Sample Task，评测数据中包括在 40 个歧义词，有名词 19 个，动词 21 个。训练实例规模约为测试实例规模的 3 倍。下文中，我们简单称为 SemEval2007。

我们进行了两次实验：第一次直接用人工选择的特征模板，用 Naïve Bayes 模型进行测试；第二次将 SemEval2007 的训练语料划分为 3：1 的两部分，分别作为特征自动选择算法的训练语料和评价语料，对所有词自动选取了一组统一的特征模板，再利用 Naïve Bayes 模型进行了测试。

自动选择的特征和人工选择的特征分别如表 3。符号所表示的含义见表 1。

Feature Template with FS	Feature Template without FS
W_{-1} W_0 W_1	W_{-1} W_0 W_1
P_{-2} P_0 P_1	P_{-2} P_{-1} P_0 P_1 P_2
$P_{-1}P_0$	WP_{-1} WP_1
$W_{[-10,10]}$	$P_{-1}P_0$ P_0P_1
	$W_{[-10,10]}$
	InEntity, EType

表 3：自动选择特征和人工选择特征(SemEval2007)

评测结果如表 4。可以看到，采用特征自动选择比人工选择的特征更简洁，但是效果都至少提高了 2%，已经接近 SemEval2007 的最佳成绩。考虑到特征自动选择算法是框架性质的，如果采用 MEM 等较为复杂的模型作为基本模型，应该会获得更好的效果。

	Best in SemEval2007 (%)	NB without FS (%)	NB with FS (%)
MicroAve	71.66	70.31	71.81
MacroAve	74.92	73.64	74.27

表 4：SemEval2007 语料评测结果

3.2 基于人民日报语料的实验

为了考察在更真实分布下的测试结果，我们重新选择了一组实验数据，由人民日报 1998 年 1 月 1-10 日和 2000 年 1-3 月数据构成，需要消解的名词和动词与 SemEval2007 一致。与 SemEval2007 的语料相比，这份语料内容更多，并且词的义项分布更真实。我们随机将语料分成 3：1 训练和测试。

与之前一样，我们分别测试了人工选择的特征和自动选择的特征。如表 5。

Feature Template with FS	Feature Template without FS
W_{-1} W_0 P_{-2} P_1 P_2 $P_{-1}P_0$ $W_{[-10,10]}$	W_{-1} W_0 W_1 P_{-2} P_{-1} P_0 P_1 P_2 WP_{-1} WP_1 $P_{-1}P_0$ P_0P_1 $W_{[-10,10]}$ InEntity, EType

表 5：自动选择特征和人工选择特征(人民日报语料)

表 6 给出了评测的结果。可以看到，在词义分布更为真实的情况下，自动选择特征比相对于工选择特征的改进更为明显。特别对于义项较少、分布较均匀的名词改进尤为明显。

		NB without FS (%)	NB with FS (%)
名词	MicroAve	78.91	83.24
	MacroAve	75.54	80.94
动词	MicroAve	84.13	84.44
	MacroAve	82.07	84.08
综合	MicroAve	82.57	84.10
	MacroAve	77.98	82.17

表 6：人民日报语料的评测结果

4 结论

本文针对汉语词义消歧问题，设计了一套特征自动选取算法，并以 Naïve Bayes 模型为基础进行了实验。实验数据表明，通过算法自动选择的特征比人工选择的特征更为精简并且更为有效。在 SemEval2007 的语料上得到的实验结果已经接近比赛的最佳成绩。

特征选择算法通过对特征模板项评分，实现对于特征模板项重要程度的量化评定。相比于人工选取更为可靠。根据实际问题的不同，特征选取算法还可以适配不同机器学习模型，进一步提高实际问题的解决效果。

下一步，我们将研究特征自动选择算法与其他机器学习模型的结合，并将其应用于不同的自然语言处理问题。此外，我们还将尝试设计基于半指导的机器学习模型的特征自动选择算法。

参 考 文 献

Vincent Ng and Claire Cardie. 2003. Weakly Supervised Natural Language Learning Without Redundant Views. In *Proceedings of HLT-NAACL 2003*, page 94-101.

Jin Peng, Wu Yunfang and Yu Shiwen. SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), 2007*, page 19~23.

Kwong Oi Yee. CITYU-HIF: WSD with Human-Informed Feature Preference. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, page 109~112.

Yun Xing. SRCB-WSD: Supervised Chinese Word Sense Disambiguation with Key Features. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, page 300~303.

Pedersen, T., A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In P*roceedings of the North American Chapter of the Association for Computational Linguistics(NAACL) 2000*, page 63~69.