基于多分类器集成的古代汉语词义消歧*

于丽丽[1] 丁德鑫[2,3] 曲维光[2,3] 陈小荷[1] 石民[1]

1. 南京师范大学文学院 南京 210097; 2. 南京师范大学计算机科学与技术学院 南京 210097

3. 江苏省信息安全保密技术工程研究中心 南京 210097

E-mail: ytsr0070030749@126.com

摘 要: 本文首先分析了古代汉语词义义项特点，考察了词义消歧的难点，确定出面向汉语信息处理的词语义项区分遵循的原则和方法。然后在现有的词义消歧理论基础上，采用机器学习的方法，选择合适的特征，使用高效率的 NaiveBayes、RFR_SUM、最大熵以及 CRF 等分类模型，对“將”、“如”、“我”、“信”、“聞”等高频词进行了词义消歧实验。最后采用分类集成的方法，研究了乘法法则、均值法则、最大值法则三种集成法则在古汉语词义消歧中的应用。集成实验最好平均 F 值达到了 84.10%，实验结果表明，分类器的集成对古汉语词义消歧效果良好。

关键词: 中文信息处理；古代汉语；词义消歧；分类器集成

The Ancient Chinese Word Sense Disambiguation Based on Ensembles of Classifiers

YU Lili[1] DING Dexin[2,3] QU Weiguang[2,3] CHEN Xiaohe[1] SHI Min[1]

1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097;

2. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097;

3. The Research Center of Information Security and Confidentiality Technology of Jiangsu Province, Nanjing 210097

E-mail: ytsr0070030749@126.com

Abstract: This paper firstly analyzes the ancient Chinese word sense and characteristic, inspects the difficulty of the ancient Chinese word sense disambiguation (WSD), and definites the principles and methods that should be followed by sense discrimination for Chinese language processing. Then basing on the existing theory and methods, we make use of methods of machine learning, choose the appropriate characteristic, use the high efficiency NaiveBayes,RFR_SUM model, the Condition Random Field as well as the Maximum Entropy model etc, then we test 5 Chinese high frequency words like “將”,“如”,“我”,“信”,“聞” etc. At last, we use 3 combining strategies of ensembles of classifiers and study the application of product, average, max in the ancient Chinese WSD. Of the experiment, the best average F-score achieved 84.10%, which indicates the method of ensembles of classifiers is effective to the ancient Chinese word sense disambiguation.

Keywords: Chinese information processing, the ancient Chinese, word sense disambiguation, ensemble of classifiers.

引言

传统训诂学是一门以研究词义为出发点和落脚点的具有实用意义的学问[1]，历史源远流长。在信息处理迅猛发展的今天，训诂学的发展也应与时俱进，来改变传统以手工为主的研究方式。现有的中文信息处理技术成果主要应用于现代汉语中，在古代汉语处理领域还有待进一步探索，以实现快速的检索和校对、考证研究、文白自动翻译等工作。我们的工作旨在充分挖掘信息处理技术在古代汉语中的应用价值，期望能对古代汉语的研究起到一定的推动作用。

*基金项目：国家自然科学基金项目（60773173）；国家 973 项目（2004CB318102）；国家社科基金项目(07BYY050)；江苏省社科基金项目（06JSBYY001）.

1 古代汉语词义消歧的难点

“单音节词占优势”是古代汉语词汇最突出的特点，句子使用的字少，但信息量大，短小精炼，一字多义现象普遍。其中古代汉语词义消歧的难点主要有：

(1)古代汉语中一形多义现象普遍。在现代汉语中用几个不同的词来表达的意义，古代汉语中却常用一个词来表达，加大了词形承载的内涵。而且大量的古今字、通假字等的存在，更使得词形与词义的关系复杂。(2)深层语义丰富，往往很难从句子的表层结构明确词的某个义项。古代汉语词汇往往是在比单句更大的整个语言环境中才表现出某种特定的意义，因此在消歧的过程中，简单的上下文窗口不能完全解决掉这种歧义，处理起来更为困难。(3)词的词汇意义和语法意义相互依存。词汇意义总是和一定的语法意义关联着，语法意义又总是依附在一定的词汇意义上。在古代汉语中，名词、形容词的使动、意动用法或用作一般动词；动词、形容词用作名词等都是比较常见的现象，这也直接涉及到我们在确定面向信息处理的义项区分问题。(4)现代汉语中作为主流的基于统计的词义消歧所关注的是如何从训练语料中尽可能多地学习语言知识再对同质文本进行消歧。而古代汉语语料属于封闭性资源，语料有限规模小，训练语料往往不足。

此外，目前适合于机器阅读的现汉词义消歧资源丰富，如知网、中文概念词典等，而几乎没有适合于古代汉语词义消歧的资源。总之古代汉语的词义消歧研究相对比较滞后，需要我们从基础入手，探究理论，根据古代汉语特点，寻找在语料库支持下的适合古代汉语词义消歧的方法。

2 面向词义消歧的义项区分原则和颗粒度的确定

现汉词典中绝大多数为单义词，多义词仅占 14.8%[2]。我们统计了《春秋左传》中的词汇及其词频，依据陈克炯[3]的《春秋左传详解词典》(以下简称“《详解词典》”)的释义，分析了词频表中前 150 个高频词的义项个数(详见表 1)。其中，单义词仅占 10%，多是一些专有名词和虚词，如“晋”、“楚”，“吴”、“矣”、“皆”等。多义词占了 90%之多，而且义项个数繁多，其中居于 3-8 个之间的词语占到了 52%，这就对古代汉语词义消歧提出了更高的要求。

义项个数	1	2-7	8-10	11-16
词个数	15	98	27	10
比率	10%	65.33%	18%	6.67%

表 1　义项个数分布表

高频也就意味着具有较高处理价值和必要性，所以我们选取了“將”、“我”、“如”、“信”、“聞”五个高频多义词作为实验对象。面向语言信息处理来辨析义项，确定合适的词义颗粒度成为我们的首要任务。

我们知道，义项个数的划分主观性比较强。每个词有多少个义项，并无统一标准，根据概括抽象程度的不同可以有各种不同的结果。实际上，每个词的义项个数的多少并不是最本质的问题，关键是与特定应用紧密相连，为特定应用服务[4]。面向人的义类体系对于计算机信息处理要么过于粗糙，要么过于细微，很不完备。鉴于此，我们确定多义词的义项的方法是根据词典资源提供的词义信息，在面向人和机器的比较中抽取、概括适于信息处理的义项区分，重点把握面向计算机的词义区分的颗粒度[4]。我们以《详解词典》和《汉语大词典》[5]为主要参照，依据语料中的实际出现情况和信息处理的实用需求，对目标词的义项进行了适当处理，或删掉、或合并、或细

分。主要遵循如下的原则：(1)可行性：根据相应语境,操作者(计算机或人)可以顺利地对语料中的目标词标注出义项,即义项区分对所标注的语料具有“完备性”[4]。(2)充分性：为面向自然语言处理服务，各个义项要满足古代汉语各种检索或分析工作的需要，适应文本检索和分析工作。(3)区别性：义项之间要具有明晰性，即各个分类之间没有重叠，保证义项之间的离散和不相交。(4)兼容性：尽量使义项的分合与已建立的各种资源的表示相一致，兼容性好，有利于资源共享。(5)针对性：针对本文的最终统计结果是为计算机处理规范的古代汉语服务的，对极少数具有很浓的方言、口语色彩的义项另单独处理。如上的原则，重在保证义项标注时的规范性、可操作性以及内部一致性，我们对目标词的各个义项进行了再定义，处理后的义项个数统计如表2：

词	如	將	我	信	聞
义项个数	8	8	3	8	4

表2　多义词义项个数统计

3 分类器的选择

目前，词义消歧已有的主要方法有：(1)基于规则的方法：制定歧义消解知识库进行消歧。(2)基于语料库的统计学方法：如决策树、最大熵、支持向量机等，且逐渐成为当前的主流技术。一般而言,单从一个侧面描述词义消歧知识存在一定局限性,尤其当某一模型完善到一定程度后,仅依靠学习算法的改进，面对古代汉语有限的语料资源，在词义消歧性能上很难有质的提高。本文采用分类器集成学习的方法，在古代汉语语料上进行了集成测试。目的是通过分类器的集成性研究，使各有所长的几个分类器取长补短，充分发挥各自优势，提高古代汉语词义消歧的效果。

3.1 分类器的选择原则

对于进行集成的单分类器的选择我们遵循了如下主要原则：(1)单分类器的互补程度要高，其分类结果应具有多样性。因此我们尽量选用不同类型的分类器进行集成，避免同类分类器犯同样的错误。(2)单分类器的准确率要高。单分类器的学习是集成分类器学习的一部分，对于词义消歧任务而言，更要尽量要求单分类器的准确率要高，否则势必会影响到集成的效果。本文选用的分类器都是经过测试了的高效率模型，已广泛应用于自然语言处理的各项任务中，从性能上满足选择分类器的原则;而且各个模型根据选用的不同特征,从不同侧面充分挖掘词义消歧的知识,互补程度高，适合于集成性的消歧实验。

3.2 选择的分类器及其简介

(1)朴素贝叶斯(NaiveBayes)分类器：NaiveBayes 算法是基于贝叶斯全概率公式的一种分类算法，它以贝叶斯定理为理论基础，是一种在已知先验概率和条件概率的情况下计算后验概率的分类方法。国际语义评测 SemEval-2007 的中英文对译选择词消歧任务(SE_CE)中，6 个参赛系统有 2 个使用了 NaiveBayes 分类器[6]，在词义消歧方面表现出了很好的性能。

(2)RFR_SUM 分类模型：曲维光等提出相对词频比[7]的概念，据此建立语境计算模型。该模型不仅考虑了语境中存在的词频，而且考虑了语境中出现词语相对于整个语料词频的相对比率，用相对词频比来模拟人们判断语境中出现词语对消歧的重要程度；同时又区分了语境的位置，将语境分为前、后语境，提高了语境信息计算的准确性，具体分类算法详见文献[8]。

(3)最大熵模型:最大熵提供了一种将各种上下文特征组合在一起的概率模型,基本思想是对于未知事件不作任何假设,以此得到的分布与样本的实际分布最一致[9],我们采用 Zhang Le 博士写的最大熵工具包,下载地址是 http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html。

(4)CRF 分类模型:条件随机场[10](Conditional Random Fields)是一个在给定输入节点(观察值)条件下计算输出节点(标记)条件概率的无向图模型,特别擅长处理序列标记问题,已经在各类序列标注问题,分词歧义消解[11]、明喻计算[12]等研究中显示出了很好的处理效果。本文的实验采用了 TakuKudo 编写的工具包“CRF++0.50”进行训练和测试,下载地址:http://crfpp.sourceforge.net。

3.3 集成方法研究

在确定的单分类器条件下,关键是选择一个让每个分类器都能最大发挥作用,可以显著提高预测性能的集成方法。现已有一些集成的方法,文献[13]将 9 种集成法则,如乘法法则、均值、最大值、最小值、多数投票、序列投票等,用于现代汉语词义消歧集成研究中,效果理想,我们参考了文献中比较结果,选取了三种较好的集成法则:均值(Av)、乘法法则(Pr)、最大值(Max)进行集成实验。在词义消歧的过程中,集成法则如下:设词语 w 有 m 个词义($W_1,\cdots,W_j,\cdots,W_m$),存在 R 个不同的分类器 f_i(i=1,...,R),$P(W_j|f_i)$表示分类器对词义类别的输出概率,$P(W_j)$表示词义类别的先验概率。W 应当赋予的词义类别是:$\hat{W}=\arg\max_j P(w_j|f_1,...,f_R)$;假设各分类器彼此条件独立,根据贝叶斯公式等推导可得:

Pr: $\hat{W}=\arg\max_j[P(w_j)]^{-(R-1)}\prod_{i=1}^{R}P(w_j|f_i)$;

Max: $\hat{W}=\arg\max_j[\max_{i=1}^{R}P(w_j|f_i)]$;

Av: $\hat{W}=\arg\max_j[\frac{1}{R}\sum_{i=1}^{R}P(w_j|f_i)]$;

4 实验及分析

4.1 实验数据

本文使用《春秋左传》语料作为实验语料,约 20 万字,该语料已经过分词和词性标注。先利用 NaiveBayes 分类器、RFR_SUM 模型、最大熵模型、CRF 模型分别测试,然后再按照均值法则、乘法法则、最大值法则三种集成方法对这些模型进行集成实验。实验时,划出 70%的例句作为训练,余下的 30%作为开放测试,具体见表 3(S1-S8 代表词的各个义项的句子数):

词	S1	S2	S3	S4	S5	S6	S7	S8	总句数	Baseline(%)
如	13	179	22	8	6	5	334	56	623	53.19
將	34	367	57	4	10	36	11	9	528	69.62
我	162	101	157						420	38.89
信	11	3	4	24	3	38	87	17	187	47.27
聞	8	311	4	21					344	91.18

表 3 多义词语的基本统计信息

表中 Baseline 为词全部取最高频义项时的标注正确率。这些词的平均 Baseline 为 60.10%。

本文对词义消歧结果进行评测的指标主要是正确率、召回率以及 F 值,计算公式如下:

正确率(P)=正确标注个数/标注的总数量；召回率(R)=正确标注个数/测试总数；
F 值=2*P*R/(P+R)，其中平均 F 值的计算方法均采用加权平均。

4.2 实验结果分析

(1)NaiveBayes 实验：我们以上下文窗口为 1 的词形、词性、词形与词性的共现以及目标词本身词性为特征，进行实验。在该模型中，所有的多义词均被处理，在计算正确率 P 和召回率 R 时的分母相同，因此正确率、召回率、F 值相同，故表中只列出 F 值。以下 RFR_SUM、最大熵模型同此，其测试结果如表 5。

(2)RFR_SUM 实验：在统计各个义项所在的前后语境的词频时，我们考虑到了窗口中的词与目标词的距离大小，离目标词越近的词，赋予较大的词频。我们设定窗口大小为 4，离目标词按照距离由近到远每次加的词频分别为：4、3、2、1。比单纯的未考虑距离远近时的效果提高 1.43%。

(3)最大熵实验：我们选取目标词本身词性、上下文词形、词性以及词形与词性的共现作为特征，将窗口 L 的取值从 1 依次变化到 5,迭代次数的取值从 10 依次递增 5 变化到 50，没有应用平滑,最终实验发现当窗口为 1、迭代次数为 15 时，此时平均 F 值最高。

(4)CRF 实验：我们这样定义 token：包含 3 列，分别是词、词性、标记。其中标记的定义是：对于句子中的除目标词外的其他词，标注为 X；对于目标词，标注为词义。利用目标词的上下文的词形、词性、词形和词形的共现、词性与词性的共现来作为消歧的依据特征，建立模板进行实验。当上下文窗口为 2 时，测试结果如表 4：

词	正确率	召回率	F 值
如	87.03	85.64	86.33
將	87.01	84.81	85.90
我	67.86	60.32	63.87
信	79.17	69.09	73.79
聞	96.00	94.12	95.05
平均	84.31	80.29	82.25

表 4　CRF 测试结果

(5)集成实验及分析：在上面的单分类器进行测试的基础上，进行了三种法则的集成实验，取得较理想的效果，对所有的实验结果进行了比较，见表 5：

词	NB	RFR_SUM	CRF	Maxent	Pr	Av	Max
如	84.57	85.64	86.33	90.43	90.96	92.55	91.49
將	83.54	85.44	85.90	82.91	83.54	84.18	84.81
我	55.56	64.29	63.87	63.49	63.49	64.29	64.29
信	76.36	60.00	73.79	70.91	81.82	80.00	67.27
聞	92.16	91.18	95.05	95.10	95.10	95.10	96.08
平均 F 值	79.01	79.97	82.25	82.19	83.47	84.10	82.99

表 5　集成与单分类器模型结果的比较

通过表 5 可以反映出：(1)三种集成结果中，“如”、“信”、“聞”集成后最优结果优于单个分类器，集成性能尤佳；而“將”消歧的最好结果是由 CRF 模型得到的。(2)不同的集成方法对集

成结果影响甚大，平均效果由高到低排列为：均值法则集成>乘法法则集成>最大值集成。

对单分类器而言，CRF 对于多分类问题表现出较好的效果，主要在于 CRF 模型具有表达长距离依赖和组合特征的能力，把所有特征进行全局归一化，进而求得最优值。最大熵可以任意地选择特征，由于在其每一节点都要进行归一化所以只能得到局部的最优值，同时也带来标记偏差的问题，所以 F 值略逊于 CRF 模型。RFR_SUM 模型在并未对生语料深加工的情况下，未加入词性等信息，仍取得了较好的效果。此外，现代汉语词语消歧往往需要较大的上下文窗口，需考虑更多词的搭配等信息，而古汉语实验窗口的加大往往会产生更多的噪声，导致正确率的下降，无论利用 CRF 模型还是最大熵消歧时，窗口选择 1 或 2 效果均是最好的。

由表 5 可见，集成分类器的整体表现均得到不等程度的提升，最终平均结果保持在 82%以上，均高于单个分类器的性能。多分类器的集成是能够减少单个分类器的误差，提高预测性能和分类精度的，在我们的实验中这样的优势充分显现了出来：每一个模型都试图从不同侧面来描述词义消歧的知识，进而可以尽可能多地充分利用各种有效的特征，如词频、词形、词性及其各种共现等，将这些特征一起运用于单个分类器本很困难，而通过集成，增加了信息量，更加充分利用目标词的上下文语境，减少单个分类器的误差，提高了消歧的效果。

词义消歧的工作任重而道远，本文主要是根据古代汉语句子的特点，以探索合适的词义消歧的方法，终归离词义消歧的最终解决还有很大的距离，需要我们不断的探索总结。下一步工作主要有：(1)改进集成策略，充分挖掘并利用单分类器对义项的概率估计值的信息，力求选择和设计更有效的集成法则，以期获得好的集成效果；(2)建立一个人机交互式词义半自动标注平台，以方便对更多古代汉语词语进行进消歧实验；(3)鉴于多义词的各个义项分布的不平衡，致力于解决由此带来的数据稀疏问题，这都将是我们的下一步工作。

参 考 文 献

[1] 周大璞, 黄孝德, 罗邦柱. 训诂学初稿(第三版) [M]. 武汉:武汉大学出版社, 2007.

[2] 鲁松, 白硕, 黄雄. 基于向量空间模型中义项词语的无导词义消歧[J]. 软件学报, 2002, 13(6):1082-1089.

[3] 陈克炯. 春秋左传详解词典(第 1 版) [M]. 河南:中州古籍出版社, 2004.

[4] 吴云芳, 俞士汶. 信息处理用词语义项区分的原则和方法[J]. 语言文字应用, 2006(2).

[5] 罗竹风. 汉语大词典[M]. 上海:汉语大词典出版社. 1993.

[6] Jin Peng, Wu Yunfang, Yu Shiwen.SemEval-2007 Task 05: Multilingual Chinese-English lexical sample [C/OL]//Proc of SemEval-2007.

[7] Qu Weiguang, Sui Zhifang, et al. A collocation-based WSD model: RFR-SUM. 2007, LNAI, 4570:23-32.

[8] 曲维光. 现代汉语词语级歧义自动消解研究[M]. 北京:科学出版社, 2008.

[9] Adam L.Berger, Stephen A.Della Pietra, et al.A Maximum Entropy Approach to Natural Language Processing [J]. Computational Linguistic, 1996, 22(1).

[10] John Lafferty, Andrew McCallum, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [A].Proc.International Conference on Machine Learning[C], 2001.

[11] 丁德鑫, 曲维光, 徐涛等. 基于 CRF 模型的组合型歧义消解研究[J]. 南京师范大学学报(工程技术版), 2008, 8(4):73-76.

[12] 李斌, 于丽丽, 石民等. ”像”的明喻计算[J]. 中文信息学报, 2008, 22(6):27-32.

[13] 吴云芳, 王淼, 金澎等. 多分类器集成的汉语词义消歧研究[J]. 计算机研究与发展, 2008, 45(8):1354-1361.