基于词典的半指导学习古汉语全文词义标注*

张颖杰[1]，李 斌[1,2]，陈家骏[1]，陈小荷[2]
[1]南京大学 计算机软件新技术国家重点实验室，南京 210093
[2]南京师范大学 语言信息科技研究中心，南京 210097
E-mail: zhangyj@nlp.nju.edu.cn

摘 要：词义消歧是自然语言处理中的一项基础任务。本文针对先秦古汉语这一特殊的语言材料，将 WSD 的过程分为先区分拼音后区分具体词义这两个步骤。实验过程使用了《汉语大词典 2.0》为知识来源，《左传》为语料，采用了基于支持向量机（SVM）的半指导方法。本文同时做了直接为全体词义分类的对比实验，结果证明“分两步走”的标注过程确实更充分的利用了词汇的语言学信息，达到的效果也更好。

关键词：词义消歧；古汉语；自然语言处理

The Dictionary-based All-word Word Sense Disambiguation Using Semi-supervised Learning for Ancient Chinese

Zhang Ying-jie[1], Li Bin[1,2], Chen Jia-jun[1], Chen Xiao-he[2]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093
[2] Research Center for Language Informatics, Nanjing Normal University, Nanjing 210097
E-mail: zhangyj@nlp.nju.edu.cn

Abstract: Word sense disambiguation is a foundational research of natural language processing used. In this paper we provide a knowledge-based method of WSD for ancient Chinese of pre-Qin. For each candidate we first classify their pinyin and then senses. In the experiment we use *Chinese Dictionary v2.0* as our knowledge source and *ZuoZhuan* as test corpus. An unsupervised machine learning method based on support vector machine(SVM) is used. In the experiment, it outperforms the results of the method classifying word senses directly without pinyin tagging.

Keywords: word sense disambiguation; ancient Chinese; natural language processing

1 引言

词义消歧（Word Sense Disambiguation, WSD）是一个为特定的词在特定的上下文中自动选择合适的词义的过程，故而也称为自动词义标注。在主流的词义消歧方法中，有指导的 WSD 效果最好（Pradhan et al., 2007），但需要较大的人工标注数据集，并且其结果对训练数据集有很强的依赖性，泛化能力较差。相对的，基于知识的 WSD 方法，将词语在词典中的义项数作为类别数，将词典对词语的解释和例句作为义项出现的语境信息，在一定程度上减少对训练数据的依赖性，同时也不需要额外的语料资源。虽然受限于训练数据规模，其效果通常不如有指导的方法，但义项标注的覆盖率较高。在缺乏人工标注数据集的情况下，可以提供初始的自动标注结果。

目前，古汉语的词义自动标注工作还处于起步阶段，在资源和技术上都呈稀缺状态。对于经典传世之作，虽有历代学者的大量注疏，却不是在同一个释义词典或语义体系的基础上进行的。目前较为实用的、能够服务于古汉语文献词义标注的词典是《汉语大词典》。该词典收词目 30 余万条，给出了词语的古今义项和例句，是一本质量高、释义丰富的大型语文词典。董志翘（2011）介绍了中古汉语研究型语料库，采用《汉语大词典》为主的释义词典，人工逐词标注古籍义项的工作，工作量巨大。因此，研究古汉语义项的自动标注方法，已经成为了中国古典文学和文献研

*本文承江苏省哲社重点研究基地重大项目“先秦文献词汇知识挖掘”（2010JDXM023）、211 项目“先秦汉语词汇统计与知识检索”、国家社会科学基金（10&ZD117、10CYY021、08BYY054）的资助。

究的重要需求。

对于缺乏训练数据的古汉语的词义标注来说，有指导的方法难以直接使用。在本文中，我们利用词典构建了一种基于半指导方法的全文词义标注方法，对《左传》进行了标注实验，人工抽样的统计结果显示，该方法优于最大频率词义（MFS）的标注结果，能够在古汉语全文词义标注的起步阶段提供初始结果，为人工标注词语义项提供良好的数据底本。

本文结构如下，第二节介绍了古汉语词义标注的相关研究，第三节介绍了本文使用的全文词义标注方法，第四节说明了实验的设计和结果分析，第五节给出了我们的结论及未来的研究工作。

2 相关研究

目前在古汉语的义项标注方面研究较少。于丽丽（2009）首先分析了古汉语词义义项的分布情况与特点，考察了词义消歧的难点。然后在现有的词义消歧理论和方法的基础上，基于机器自动学习的统计模型条件随机场，选择上下文的词及其词性的复合特征，并加入其他适当语言学特征，设计 6 个不同的模板，对“將”、“如”、“我”等古汉语高频词进行了词义消歧实验，平均 F 值达到了 83.04%。不过，该方法使用的词典是《春秋左传词典》，缺乏一般性；采用了有指导方法，代价太高，泛化能力有限。

对于任意语言的词义标注，最简单的基于词典的方法（Lesk, 1986）通过计算目标词的定义及其所在的上下文之间重叠的词数来确定词义。

$$score_{LeskVar}(S) = |context(w) \cap gloss(S)|$$

这种方法主要的局限在于词典中的定义通常比较简洁，未必能包含足够的能标识当前词义的词汇（Pedersen, 2003）。

随着包含了分类和语义关系的本体词典的广泛使用（如：WordNet），基于词典的 WSD 也出现了结构化的方法，主要有基于相似度计算（Pedersen, 2005）的方法和基于图（Sinha and Mihalcea, 2007; Agirre and Soroa, 2009）的方法。基于相似度计算的方法通过目标词的各个词义与文本中其他词之间的语义相似度，从中选择使得结果最高的词义来实现 WSD。

$$\hat{S} = \arg\max_{S \in Sense_D(w_i)} \sum_{w_j \in T; w_j \neq w_i} \max_{S' \in Senses_D(w_j)} score(S, S')$$

基于图的方法通常把全文表示成一个以词义为结点、语义关系为边的图结构，通过随机游走等方法确定结点的得分，从而得到最终的词义。

然而，对于古汉语这一特殊的应用领域，很难使用结构化方法。首先，古汉语的结构化词典资源缺乏，在汉语中运用频繁的 HowNet 中的概念分类仅针对现代汉语，由于古今异义等原因，无法直接应用它来计算词语间的相似度。其次，图方法通常严格的遵守一个前提，即“一段一义”，用来构成图的段落中相同的词最后将会被标注上同一个词义，此前提在很多英文的 WSD 过程中被认可。然而，古汉语词类活用现象比较频繁，同样的高频词在同一段落中表现出多种不同的词义是常见的现象，一般来讲很难完全满足这样的前提。因此需要对同一段落中的词利用更多的特征来区分词义。

考虑到以上问题，本文利用现有的古汉语词典资源，采用了一种半指导方法，根据古汉语的特性，先粗分拼音再细分释义，对大量的古汉语语料实现了全文词义标注。为古汉语的文献中的实词提供用现代的白话文表示的直观解释，方便了使用者的理解、查阅及应用。

3 Bootstrapping 的 WSD 方法

本文的半指导方法沿用了 Yarowsky(1995)提出的一种通过极少量人工标注语料来进行大量词

义标注的方法。它最大的好处在于能直接利用现有的机器学习算法，而不需要额外的处理。

3.1 Yarowsky 的 bootstrapping 方法

该方法首先对每一个需要标注词义的多义词建立上下文列表。其次，对该词的每个可能词义，手动标记一个包含典型搭配信息的可信小训练集 seed，根据“一个搭配一种含义”的先决条件给出表示搭配信息的决策表。该可信小训练集只包含了一种搭配情况。再次，在 seed 上训练决策表分类模型，并将其用于待标注集的分类，将所有概率超过既定阈值的结果增加到 seed 中，同时根据“一段一义”的约束条件扩充 seed，剩余用例仍作为待标注集用于下一次的迭代。重复此过程至结果收敛，即所有未标注用例的分类结果概率均在阈值以下。最后，为剩余用例标注结果。

3.2 改进的半指导词义标注方法

在本文中针对 WSD 需要考虑的各个方面对上述方法作出一定的修改，使之适用于古汉语这一特殊应用对象。

（1）词义粒度。本文中待标注词的词义不只两项，而是根据词典中的义项来确定。词典中凡是具有来自先秦文献的例句的义项，均被用来作为词义集合的一个元素。

（2）特征选择。这里不止采用一种搭配信息，而是选取了词形、词性的一元特征和两者搭配的二元特征，见表 1。有研究表明[3]，二元特征窗口增大反而降低词性标注结果的准确性，因此对于二元特征，仅使用前后大小为表 1 的窗口。

由于“一段一义”的约束条件并不完全适用于古汉语，尤其对一些义项较多、应用情况灵活的高频词。因此，本文降低其强制性，仅将待标注词所在的句子编号作为一个特征进行考虑。

表 1　特征选择

一元特征	词形	W-2, W-1, W0, W1, W2
	词性	P-2, P-1, P0, P1, P2
	所在段落	Position
二元特征	词形_词形	W-1_W0, W0_W1
	词性_词性	P-1_P0, P0_P1
	词形_词性	W-1 P-1, W0 P0, W1 P1

（3）可信小训练集的选取。本文中不使用手动标注的方式，而是根据词典信息自动得到。由于在古汉语的词典中其释义通常用现代汉语表示，它们的上下文在形式和内容上差别较大，不能直接使用。而词典中除了释义外通常还包含一些例句，这些例句一般都具有典型性，且能保证其与词义对应的准确性，故而我们通过这些例句得到标注之初所需的 seed。

本文为目标词的每个释义选择了一个例句，得到所需的上下文特征。因而在区分拼音时，最初的可信小训练集对于每种拼音不止有一种特征形式。

（4）方法选择。由于本文中所用的特征不再是单一的搭配信息，故而也不再使用简单的决策表，而改用了 SVM 的方法，其核函数使用了默认的线性核[11]。对于阈值，第一步拼音的区分类别较少，区分度较高，故而选用了一个较高的阈值；第二步词义的区分类别较多，区分度相对较低，因此选用了一个较低的阈值。

4 实验

4.1 数据来源

本文将人工完成了分词和词性标注的 18 万字《左传》作为实验语料（石民 2009），对其中的 4671 个实词进行了词义标注。这些实词中有 635 个多音词，占待标注词的 13.6%。

知识来源采用了《汉语大词典 2.0》，词典对词的释义中涵盖了从古至今所出现过的几乎所有词义，并给出了注有年代特征的例句。

以“忘”为例，其在词典中第一种拼音的释义表示如下图：

忘 1 [wàng ㄨㄤˋ]

[《廣韻》巫放切，去漾，微。]

1.忘记；不记得。《诗·小雅·隰桑》：“中心藏之，何日忘之。”《司马法·仁本》：“天下雖安，忘戰必危。” 宋 曾巩 《尚书都官员外郎陈君墓志铭》：“ 泉州 歲凶，君築室止窮民，飢者給食，病者給醫，人忘其窮。” 周恩来 《致柯棣华大夫家属的慰问信》：“我们受惠于他的极多，使我们永不能忘。” **2.**指健忘症。《列子·周穆王》：“ 宋 陽里華子 中年病忘。” **3.**遗弃；不顾念。《诗·秦风·晨风》：“如何，如何！忘我實多。” 马瑞辰 通释：“忘我實多，猶云棄我實甚。”《庄子·山木》：“覩一蟬，方得美蔭而忘其身；螳螂執翳而搏之，見得而忘其形。”《後汉书·宋弘传》：“貧賤之知不可忘。” **4.**玩忽，怠忽。《史记·孔子世家》：“昔 武王 克 商 ，道通九夷百蠻，使各以其方賄來貢，無使忘職業。” 唐 韩愈 《潮州祭神文》之四：“惟神之恩，夙夜不敢忘怠。” **5.**无。《史记·孟尝君列传》：“日暮之後，過市朝者掉臂而不顧。非好朝而惡暮，所期物忘其中。” 司马贞 索隐：“忘者，無也。其中，市朝之中。言日暮物盡，故掉臂不顧也。”《史记·平津侯主父列传》：“ 高皇帝 蓋悔之甚，乃使 劉敬 往結和親之約，然後天下忘干戈之事。” **6.**通“ 妄 ”。《老子》：“不知常，忘作，凶。” 朱谦之 校释：“忘、妄古通。”《韩非子·解老》：“前識者，無緣而忘意度也。” 王先慎 集解：“忘與妄通。”

图：“忘”的第一个拼音在《汉语大词典》中的释义表示

注：下划线的内容表示出处，根据出处就可以得到例句出现的年代。

4.2 实验步骤

（1）根据年代筛选义项。在本次实验中从词典抽取了包含先秦例句的释义，保证了用于标注的义项均有可能出现在先秦文献中，剔除了不可能出现的词义。如上图：“忘 1”的第四个释义“玩忽”和第五个释义“无”最初都是在汉代的《史记》中出现的，故这两个义项不包含在我们要分类的义项列表中。而第一个释义“忘记”的例句除了来自于先秦文献《诗经》和《司马法》以外，还有的选自宋代和现代的文章，本文中所用的上下文信息仅从前两者中提取。

（2）标注例句作为训练语料。为了得到最初的种子训练集，实验利用先秦古汉语的词性标注工具（石民 2009，该工具在左传上的分词和词性标注 F 值均超过 90%）对这些例句进行分词和词性标注，再根据词典中给出的拼音和释义信息，最终得到用于训练的上下文特征。由于这些上下文特征来自于词典中的例句，因而此种子训练集的标注结果必然是可信的，其特征也具有典型性，保证了它对标注的指示作用。

（3）先标音再标义的“两步走”标注方法。汉语中包含了很多的多音词，同一个词的不同拼音含义差别较大，甚至有时可以看作两个不同的词来处理。因此本文先为每个目标词粗分拼音，再为每个目标词的每种拼音分别建立分类器，用于细标词义，分两步来得到最终结果。对每个目标词 w_t 都执行以下步骤：

Step0：数据准备。将目标词在《左传》中的每一次出现集合起来建立上下文列表 $L=\{context_j\}$，$j=1,2,\cdots,m$，m 为目标词在《左传》出现的次数。待标注集 $S=\{occ(w_t)|context(w_t)\in L\}$。

Step1：自动标音

Step1.1：得到训练集。从该目标词根据年代筛选后的所有义项的例句中得到可信训练集（seed）。

Step1.2：自动标注。根据词典得到目标词的拼音列表{P_i}，i=1,2,…,n，n 为拼音的数目。使用（3.2.2）的上下文特征执行本文中的半指导方法，为 S 中的所有条目标上拼音 P_i。

Step2：自动标义

Step2.1：得到新训练集。根据 Step1 的结果将原待标注集 S 分块成为 S_1、S_2、…、S_n，n 表示该目标词拼音的数目，同一个分块 S_i 中的目标词都具有相同的拼音。同样原可信训练集 seed 也根据拼音分类成为 $seed_1$、$seed_2$、…、$seed_n$。

Step2.2：自动标注。针对每个 S_i，根据 $seed_i$ 再次使用（3.2.2）的上下文特征执行本文中的半指导方法，从而得到最终的词义标注结果 $sense_{ik}$。i=1,2,…,n，n 为拼音的数目；k=1,2,…,n_i，n_i 为拼音 P_i 中义项的数目。

（4）对比实验。在对比实验中我们不考虑多音词现象，将目标词的所有拼音的所有义项合并为一个词义列表，用相同的方法直接标注词义。

（5）系统基线（baseline）。我们设定了两个基线用于对比实验结果。baseline1 是假设目标词的全部 n 个义项平均分布，每个义项出现的概率为 1/n。由于《汉语大词典 2.0》中是将常用的拼音排在前面，而越靠前的释义越接近本意，使用也越频繁，因此本文目标词根据年代筛选释义后的第一个拼音的第一个词义作为标注结果的 baseline2。

4.3 实验结果及分析

自动标注完成后，我们对结果进行了抽样，人工检查其结果进行评测。样本包含了 4 个不同的词，240 个用例，涵盖了实词的四种不同的词性。

表 2 实验抽样结果的准确率

	忘	辟	從	强	均值
义项数/拼音数	6/2	40/5	18/2	14/2	19.5/2.75
Baseline1	0.1667	0.025	0.0556	0.0714	0.0797
Baseline2	**0.5**	0.2419	**0.4556**	0	0.2994
Two-step	0.4625	**0.3871**	0.2556	**0.375**	**0.37**
One-step[1]	0.4625	0.3065	0.1333	0.375	0.3193

注 1：即直接标义项的方法

从表 2 中可以看到，对“忘”的标注结果最好，准确率达到了 46.25%，但是它的 baseline2 的结果有 50%。“忘”字有两种拼音，总共有六个义项，相对而言义项的数量较少，分类结果也较好。在 baseline2 中，我们选择了第一个义项“忘记”作为其结果，这是它的本意，也是最常用的释义，因此其准确率反而少许高于通过学习得到的结果。而在得到了拼音信息后再区分词义虽然没有令结果更好，但也没有更差。

“辟”通过两步分类得到的结果准确率为 38.7%，而直接分类结果准确率为 30.65%，baseline2 仅有 24.2%。“辟”有五种拼音，共有 40 多种义项，前两种拼音更是均有 15 种以上的义项，这使得它的词义区分困难，导致了整体准确率不高。但是由于义项分布在多种拼音间比较均匀，再粗分了拼音后大幅度地降低了分类器中类别的数量，相对于直接分类准确率提高了 8 个百分点。

对于“從”的标注结果，baseline2 的结果为 45.56%，远高于机器学习得到的结果。其 baseline2 所取的“跟随”这个本义原本应该是使用最多的含义，但是由例句自动标注得到的特征集却只有 1 个，且该例句太过简洁不能涵盖大部分情况，反而是其他不常用的释义的例句更加完整，对标注结果造成了很大的干扰。然而分两步走的标注方式在一定程度上提高了标注的准确率。

“强”的结果中 baseline2 为 0，通过观察词典得知，该词的有先秦例句的首个义项为“硬弓”，这显然不是其常用义项，而常用的“健壮”这个释义却是第 3 个义项，故而此处的 baseline2 并没有取到它的常用义，使得结果为 0。除此之外，“强”共有 14 个义项需要区分，增加了 WSD 的难度，因此总体效果一般。

综合以上实验结果，可以看出，自动标注的平均正确率仅为 37%，但已高于两个基线的结果，主要的问题在于训练数据过小，词语义项数较大。我们希望本文提出的半指导方法，可以为人工标注词语义项提供良好的数据底本，通过反复地人机交互来提高标注精度。

5 结论与未来工作

本文针对先秦古汉语这一特殊的语言环境，将 WSD 的过程分为先区分拼音后区分具体词义这两个步骤。实验过程使用了《汉语大词典 2.0》为知识来源，《左传》作为测试语料，采用了基于支持向量机（SVM）的半指导方法。本文同时做了直接为全体词义分类的对比实验，结果证明“分两步走”的标注过程确实更充分的利用了词汇的语言学信息，达到的效果也更好。

本文的特征均由人工根据经验得到，并且都由词形和词性组成，在未来的工作中我们考虑（1）加入更多语言信息，如句法结构、语义角色、依存分析等，或者加入特征选择的过程，进一步提高词义标注的效果；（2）根据候选词义类数量的不同自动确定合适的阈值，提高标注的准确性；（3）利用前人的注释信息验证指导标注结果，来提高全词标注的效果。

参 考 文 献

[1] Pradhan, S. Loper, E. Dligach, D., and Palmer, M. Semeval-2007 task-17: English lexical sample srl and all words. Proceedings of SemEval-2007, 2007: 87-92.

[2] 董志翘. 为中古汉语研究夯实基础，燕山大学学报（哲学社会科学版），vol.12，no.1，2011 年 3 月.

[3] 于丽丽，丁德鑫，曲维光，陈小荷，李惠. 基于条件随机场的古汉语词义消歧研究，微电子学与计算机，2009 年 10 期.

[4] Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pinecone from an ice cream cone. SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, 1986: 24-26.

[5] Patwardhan, S. Banerjee, S, and Pedersen, T. Using measures of Semantic Relatedness for Word Sense Disambiguation. Proceedings of CICLing, 2003: 241-257.

[6] Pedersen, T., Banerjee, S. and Patwardhan, S. Maximizing semantic relatedness to perform word sense disambiguation. Minneaplis: University of Minnesota Supercomputing Institute, Res. rep: UMSI 2005/25, 2005.

[7] Sinha, R., and Mihalcea, R. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. Proceedings of the IEEE International Conference on Semantic Computing, 2007: 363-369.

[8] Agirre, E., and Soroa, A. Personalizing PageRank for word sense disambiguation, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009: 33-41.

[9] Yarowsky, D. Unsupervised Word-Sense Disambiguation Rival Supervised Methods. Proceeding of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995:189-196.

[10] 石民，李斌，陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究，中文信息学报，2010 年第 2 期.

[11] Jin, P., Li, F. Zhu, D. Wu, Y., and Yu, S. Exploiting External Knowledge Sources to Improve Kernel-based Word Sense Disambiguation, Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2008:222-227.