基于 Web 的无指导译文消歧词模型与 *N*-gram 模型及对比研究

刘鹏远[①] 赵铁军[②]
[①](北京大学计算语言学研究所 北京 100871)
[②](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要：该文提出了基于 Web 的无指导译文消歧的词模型及 *N*-gram 模型方法，并在尽可能相同的条件下进行了比较。两种方法均利用搜索引擎统计不同搜索片段在 Web 上的 Page Count 作为主要消歧信息。词模型定义了汉语词汇与英语词汇之间的双语词汇 Web 相关度，根据汉语上下文词汇与英语译文之间的相关度进行消歧；*N*-gram 模型首先假设不同语义下的多义词 *N*-gram 序列行为模式不同，从而可对多义词不同语义类下词汇在实例中的 *N*-gram 序列进行统计与分析以进行消歧。两个模型的性能均超过了在国际语义评测 SemEval2007 的 task#5 上可比较的最好无指导系统。对这两个模型进行试验对比可发现 *N*-gram 模型性能优于词模型，也表明组合两类模型的结果有进一步提升消歧性能的潜力。

关键词：计算语言学；无指导译文消歧；词模型；*N*-gram 模型；Page Count；双语词汇 Web 相关度
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2009)12-2969-06

Comparison of Web-Based Unsupervised Translation Disambiguation Word Model and *N*-gram Model

Liu Peng-yuan[①] Zhao Tie-jun[②]
[①](*Institute of Computational Linguistics*, *Peking University. Beijing* 100871, *China*)
[②](*Department of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *China*)

Abstract: This paper describes and compares web-based unsupervised translation disambiguation word model and *N*-gram model. For acquiring knowledge of disambiguation, both two models put differents queries to search engine and statistic page counts which it returned. Word model defines Web Bilingual Relatedness(WBR) between Chinese words and English words and disambiguates word sense by maxmizing Web Bilingual Relatedness between contexts and the translations of target word. Based on the hypothesis that the pattern of a polysemant is different while different sense of it is being used, *N*-gram model makes disambiguation by statisticing and analyzing *N*-grams of words in different semantic class of that polysemant. Both of the two models are evaluated on the SemEval2007 task#5, achieving the top performance against the state-of-the-art comparable unsupervised systems. Furthmore, *N*-gram model outperforms word model and the performence has potential for promotion when combine the results of that two class model.

Key words: Computational linguistics; Unsupervised translation disambiguation; Word model; *N*-gram model; Page Count; Web Bilingual Relatedness(WBR)

1 引言

确定歧义词在特定上下文中的特定词义(WSD)或者确定歧义词的目标语译文(WTD)是为机器翻译、信息检索以及生物医学文本索引等相关任务提供服务的中间任务。词义消歧的研究一直是计算语言学研究领域中的热点和难点问题。目前主流的研究方法是利用各种机器学习技术统计各种语言学相关资源，特别是语料库和语义词典，从中获取各种语义知识来进行消歧。

根据是否需要人工标注的语料，词义消歧的研究方法可分为有指导和无指导的方法。国际语义评测SemEval-2007 [1)]的结果表明，有指导的方法均明显优于无指导的方法。但事实上，由于有指导的方法所能处理的词必须存在相对应的大量高质量的手工标注语料，因此存在着所谓知识获取瓶颈问题。不幸的是，迄今为止，没有一种语言存在已标注所有多义词的大规模语料库。针对词义消歧缺乏足够的已标注语料及相应的语义知识这一问题，除传统

2008-12-05 收到，2009-05-07 改回
国家重点基础研究发展计划(2004CB318102)资助课题

[1)] http://nlp.cs.swarthmore.edu/semeval/，本文公共标准测试语料、公共评测数据以及评测工具皆来源于此。

的基于词典的各类方法外，主要的基于统计的研究路线有三：

(1)利用种子语料以及各种半无指导方法进行词义消歧[1−3]。

(2)通过自动获取语义标注实例来进行无指导消歧的方法。利用平行语料的方法[4,5]。利用单语语料以及语义词典的方法[6−11]。

(3)本文所利用的根据Web搜索计数(Web Count)的消歧方法。

Mihalcea等[12]提出了利用Web搜索计数的词义消歧方法。该方法首先利用WordNet语义知识得到歧义词的Synset，然后利用搜索引擎得到对应不同语义的Synset下词语与上下文词语的Web搜索计数，选择该计数为大的Synset作为该上下文对应歧义词的词义。Turney[13]利用点式互信息技术的结合在Web上进行了同义词的挖掘。Rosso等[14]利用WordNet以及搜索引擎得到全名词上下文以及形容词-名词对的Web搜索计数，也就得到了不同语义与上下文的同现，然后根据同现次数对名词歧义词进行消歧。Yang[15]利用WordNet以及搜索引擎的Web搜索计数得到WordNet各个Synset之间的相关度，并由此出发对歧义词进行词义消歧，该方法取得了不错的结果。Liu等[16]将该类方法扩展到双语范畴并进行了初步的尝试。

本文提出的两种模型方法充分利用了Web这个公共海量资源，力图缓解消歧知识获取瓶颈的问题，无指导的自动获取可用于译文消歧或词义消歧的知识或词汇间联系。两种方法均以搜索引擎统计不同搜索片段在Web上的Page Count作为主要消歧信息。在整个消歧过程中，仅利用测试语料及Web中挖掘到的知识，没有利用任何已标注语料，是无指导的方法，这样就减轻了人工标注语料负担及知识获取的困难。两个模型的性能(其P_{mar}值分别为44.4%及55.9%)均超过了在国际语义评测SemEval-2007中Task#5：Multilingual Chinese English Lexical Sample Task测试集上的可比较的最好无指导系统TorMd(P_{mar}值为43.1%)1.3%及12.8%。本文还对这两个模型进行了进一步比较。

2 基于Web的无指导消歧词模型与*N*-gram模型方法

2.1 词模型方法

通过对互联网Web页面进行观察发现，当“苹果”和“香蕉”出现在同一个Web页面上，“苹果”通常会表示一种水果而非一种计算机的牌子。这说明通过Web，可以知道或者说能够找到“苹果”和“香蕉”这两个词汇之间存在一种语义相关的联系，因此“苹果”的水果的语义才会与“香蕉”相关。计算“苹果”和“香蕉”这两个词汇之间相关程度最简单的方法就是，将查询内容(Query)“苹果 香蕉”放入搜索引擎，由其返回的Web页面计数Page Counts确定。通过观察，双语词汇间也存在这样的类似关系。Page Counts可作为一种双语词汇间Web同现或者语义相关的测度。而以上建立在Web页面的Page Counts之上双语词汇之间的相关测度可被定义为Web双语相关度(Web Bilingual Relatedness,WBR)。这里首先建立如下假设：

假设 1 同现在同一个Web页面上的任意两个词汇存在一定语义联系，这种同现与语言无关。

假设 2 同现在同一个Web页面上的任意两个词汇间语义关系的必然性及系统性较偶然性与不确定性突出。

基于这两个基本假设，我们就可以直接利用特定搜索返回的双语混合Web页面的Page Counts以及词汇之间关联度的方法来考察其上双语词汇之间的语义相关强弱，也就是相关度。在对相关度的具体计算中，我们利用了点式互信息(Point-wise Mutual Information，PMI)，如公式(1)与公式(2)所示，但是将公式中的 a，b，c，d 及 N 进行基于Web的改造，用以计算任意双语词汇对(ep,ce)间的Web相关度WBR。

$$\begin{aligned}\text{WBR}_{\text{PMI}}(ep,cp)&=\lg\frac{N\times\text{freq}(ep,cp)}{\text{freq}(ep)\times\text{freq}(cp)}\\&=\lg\frac{N\times a}{(a+b)\times(a+c)}\end{aligned}\tag{1}$$

$$a=\text{freq}_{\text{web}}(cp,ep),\ b=\text{freq}_{\text{web}}(cp)-\text{freq}_{\text{web}}(cp,ep),$$
$$c=\text{freq}_{\text{web}}(ep)-\text{freq}_{\text{web}}(cp,ep),\ d=N-a-b-c\tag{2}$$

a，b，c，d，N分别表示：同时含英语词 ep 和汉语词 cp 的Web双语混合页面总数、包含汉语词 cp 但不包含英语词 ep 的Web页面总数，含英语词 ep 但不包含汉语词 cp 的Web页面总数，不含英语词 ep 和汉语词 cp 的Web页面总数及互联网上所有Web页面总数。

在得到双语词汇间的相关度WBR以后，就可选取与目标歧义词的上下文词平均相关度最大的译文作为正确译文。消歧决策可形式化如下：给定一个上下文窗口内的词汇 $c_1,c_2,\cdots,c_n$，其中 $c_k(1\leqslant k\leqslant n)$，是需要指定译文的源语言目标歧义词。假设 c_k 有 i 个可能的译文为 $t_1,t_2,\cdots,t_m$。则译文消歧的任务就是去从译文集合$\{t_1,t_2,\cdots,t_m\}$中为源语言歧义词 c_k 选出最合适的译文 t。这个过程可由式(3)表示。

$$t_i=\arg\max_{t_i}\sum_{j=1}^{n}\text{WBR}(c_j,t_i)/n,\ c_j\in C,\ t_i\in T\tag{3}$$

其中 WBR(c_j,t_i)即表示集合 C 中的任意词汇 c_j 与译文 t_i 的 WBR，可利用式(1)及式(2)计算。

2.2 *N*-gram 模型方法

来看一个具体的例子。中医在知网中的英文译文分别是 traditional Chinese medical science 和 practitioner of Chinese medicine，对应中文的含义一个是表示医学的中医，另一个就是表示医生的中医。当听到这样一个句子片段如:“是中医现代化的一项成果”,我们很容易知道这句话里的中医表示第一种含义。一种假设是你听到了“现代化”和“成果”,脑中就会反映出这两个词汇与医学之间的相关度比与医生之间的更大。但另一种可能是，你可能会经常听到 s_1:“是医学(西医/外科)现代化的一项成果”，而很少或基本听不到 s_2:“是医生(大夫/老中医)现代化的一项成果”这样的句子片段，因此会很自然知道，这里的中医是医学方面的中医含义。在经过对知网以及语料库进行初步考察之后，我们做如下假设:

假设 3　含有歧义词的语言序列在该歧义词语义不同时具有不同的模式。

假设 4　相同语义词汇的语言序列模式较之不同语义下的更容易相同。

设中文目标歧义词 w 有 n 个译文，分别对应 w 的 n 个语义。令所有含 w 的汉语句子序列为集合 S,根据假设 3，我们可以将 S 根据不同的语义分为 S_1, S_2,S_i,…,S_n，分别为 w 的 n 个语义/译文所对应的所有汉语句子序列，对应不同的模式。令 C_i={c_{i1}, c_{i2},…,c_{im_i}}，为 w 的第 i 个语义对应的汉语同义词集合，SC_i 为含有 C_i 中任意词汇的所有汉语句子序列。由假设 4 可知，SC_i 与 S_i 相对应并更容易同属于一个模式。

定义　给定一个含 w 的待消歧实例 s，可由词汇序列 $w_1w_2\cdots w\cdots w_k$ 表示,则词汇序列 $w_1w_2\cdots c_{ij}\cdots w_k$ 为一个符合 s 词汇序列的模式，以 s_c_{ij} 表示，该词汇序列出现的概率用 $P_s(c_{ij})$ 表示。

如果可以确定 s 在哪一个汉语词汇序列 S_i 中，则自然可以知道 w 的含义以及正确译文 ep。但是在无指导的方法中，由于没有标注语料，S_i 的初始划分无法确定，因此需通过比较在 SC_i 中出现符合 s 词汇序列模式的概率来对 w 进行歧义消解:

$$S_i / SC_i / C_i = \arg\max_i P_s(c_{ij}) \tag{4}$$

其中 $c_{ij}\in C_i$，是 w 的第 i 个语义所对应的汉语同义词集合 C_i 内的第 j 个词汇，对 $P_s(c_{ij})$，可进行标准 *N*-gram 模型的简化，其值可由式(5)和式(6)来进行计算，由于利用 Web 进行 *N*-gram 的统计，则 C 就是语言单位 x 在 Web 中利用搜索引擎得到的页面计数(Page Counts)。式(4)的左侧指，一旦确定了符合词汇序列模式的 SC_i，也就得到了 S_i 的模式，同时就得到对应的汉语同义词集合 C_i，他们是一一对应的，因此用 $S_i / SC_i / C_i$ 表示。式(4)要在利用同义词集合内的词汇形成的符合 s 词汇序列的模式中，找到概率最大的模式，符合这个模式同义词所对应同义词集的语义即为消歧结果。

$$P(S) = \prod_{i=1}^{n} p(w_i \mid w_{i-m+1}\cdots w_{i-1}) \tag{5}$$

$$p(w_i \mid w_{i-m+1}\cdots w_{i-1}) = \frac{C(w_{i-m+1}\cdots w_i)}{C(w_{i-m+1}\cdots w_{i-1})} \tag{6}$$

3　试验及讨论

3.1 试验设置

利用 ACL2007 评测的一个组成部分 SemEval2007 国际语义评测的中英文词汇任务(Task#5 Multilingual Chinese_English Lexical Sample Task)中的测试语料利用标准评测工具对本文方法进行评测。采用该项评测规定的评价方法 P_{mir} 与 P_{mar}(Micro Average Accuracy 与 Macro Average Accuracy):

$$P_{\mathrm{mir}} = \sum_{i=1}^{N} m_i \Big/ \sum_{i=1}^{N} n_i,\ P_{\mathrm{mar}} = \sum_{i=1}^{N} p_i / N \tag{7}$$

其中 N 为所有的目标词(all target word-types)，m_i 是对每一个特定的词所标注正确的例句数，n_i 是对该特定词所有的测试例句数，$p_i = m_i / n_i$。

根据 Liu 等人[16]对百度与谷歌(www.google.cn)的比较结果，这里采用了百度作为搜索引擎。词模型选取了以目标歧义词为中心的词袋窗口(±1，±3，±5，±7，±9，)作为上下文词汇，基于 *N*-gram 模型的方法选取以目标歧义词为中心的所有 2-gram 及 3-gram 序列,为选取语义类下的同义词还利用了知网，根据任务所提供的翻译词表半自动建立了译文与 DEF 的映射关系，并选取对应 DEF 中的所有单义词作为同义词集。实验的 baseline 为在 SemEval2007 评测中 task5 任务表现最好的无指导消歧系统 TorMd[17]、利用 Web 的无指导系统 HIT[16] 及采用最常用词义的结果 MFS。

3.2 试验结果与讨论

试验结果如表 1 所示。词模型最好结果的上下文窗口是±7，*N*-gram 模型最好的结果是 3-gram，且词序列为(−1, 0, 1)。两个模型最好的结果均超过了该项任务评测上的最好系统 TorMd 以及 MFS(该项评测的所有无指导系统均没有超过 MFS)，而 *N*-gram 模型的所有结果均超过了词模型、TorMd 以及 MFS。这说明基于 Web 的这两种方法是可行

表 1 各系统试验结果

系统	词模型					*N*-gram 模型(WBR_{PMI})					HIT	TorMd	MFS
窗口	±1	±3	±5	±7	±9	−1，0	0，1	−2，−1，0	−1，0，1	0，1，2			
P_{mir}	0.331	0.367	0.370	**0.391**	0.388	0.451	0.404	0.454	**0.494**	0.423	0.337	**0.375**	0.405
P_{mar}	0.385	0.424	0.429	**0.464**	0.455	0.506	0.467	0.502	**0.559**	0.481	0.396	**0.431**	0.462

的，且有效的。基于 Web 的无指导消歧方法中的 *N*-gram 模型方法结果的性能优于词模型的性能。但是应该注意到词模型的方法仅利用搜索引擎及 Web，没有利用任何其他资源，是一种完全无指导的方法(fully unsupervised)，而 *N*-gram 模型利用了知网这个语义资源。当然，本文对各个方法的性能比较，只要是无指导的方法，就仅从消歧效果上做比较而没有关心其所利用的资源，如 TorMd 系统也利用了词典、大量双语平行语料库的资源并有部分的人工语义映射工作。

词模型性能随着窗口的变化而变化符合对消歧任务的一般认知，之所以窗口继续扩大到±9 而模型性能没有继续提升的原因是由于引入噪音的负面影响超过了引入更多上下文的正面影响。

4 词模型与 *N*-gram 模型对比研究

4.1 *N*-gram 对比词模型

为能在尽量相同的情况下对词模型与 *N*-gram 模型进行对比研究，本小节设计了词与 *N*-gram 对比模型，该模型可做到使两个模型选取目标中心词相同，词窗口及位置与 *N*-gram 序列位置相同。该方法首先找到在 *N*-gram 模型消歧过程中已经确定的对应源语言不同语义的单义可替换词(Subtitution)，这个可替换词所在词汇序列就是形成 *N*-gram 源语言模型方法中最符合目标歧义词当前语义的词汇模式序列。然后，计算不同语义对应的可替换词与测试实例上下文中词汇的 Web 相关度，取 Web 相关度最大的那个可替换词对应的语义为消歧结果。该模型如算法 1 所示。

算法 1(输入：含 w 的测试实例；输出：正确译文)：

对含歧义词 w(m 个语义，$s_1,\cdots,s_k,\cdots,s_m$)的测试实例做：

步骤1 利用基于 *N*-gram 语言模型的消歧模型，得到分别可替换 w 的 m 个语义 $s_1,\cdots,s_k,\cdots,s_m$ 概率最大的可替换单义词 $c_1,\cdots,c_k,\cdots,c_m$。

步骤2 对所有 $c_1,\cdots,c_k,\cdots,c_m$：

(1) 计算 c_k 与 w 的上下文窗口所有词汇之间的 Web 相关度 $WR(c_{ki},c_k)$；

(2) 取 $MAXWR_k=MAX(WR(c_{ki},c_k))$，即所有 $WR(c_{ki},c_k)$中的最大值；

步骤3 取 $MAX(MAXWR_k)$，k=1，m。则 s_k 即该实例所对应语义。

这里的 WR 与前面介绍的 WBR 概念与计算方法均类似，不同的是 WR 进行的是单语词汇间 Web 相关度的计算。

4.2 试验与讨论

采用与第 4 节一样的评测语料及方法，*N*-gram 模型设置与前相同。对比模型中对相关度的计算方法不仅仅利用 PMI 模型，还利用了 Dice 系数(DICE)，ϕ^2 平方系数 (PHI)和对数似然比 (Log Likelihood Ratio, LLR)等模型，窗口设置进行了调整并与 *N*-gram 序列选取对比均列在表 2，同时仅仅选取实词，其余设置与前相同。

表 2 窗口词选取与 *N*-gram 选取对比

模型	2-gram		3-gram		
N-gram 模型序列位置	−1,0	0,1	−2,−1，0	−1,0,1	0,1,2
对比模型的词窗口词选取位置	−1	1	−2,−1	−1 ,1	1,2

试验结果如表 3 所示，可见在目标中心词以及窗口词位置与 *N*-gram 序列位置相同的情况下，对比词模型无论采用何种相关度计算方法其结果均较 *N*-gram 模型的消歧性能低。因此基本上可以说，2-gram 及 3-gram 所提供的消歧信息较相同窗口内的词汇所提供更为有效。

为考察词模型与 *N*-gram 模型在基本相同的条件下是否获得了不同的消歧知识，定义一致率 P_{cc}=两模型消歧结果相同的实例数/总实例数、一致正确率 P_{rc}=两模型消歧结果相同的实例中正确的实例数/总实例数及在一致正确比 $P_{rcc}=P_{rc}/P_c$，考察结果见表 4。

两个模型在消歧性能上的差异在 10%左右，一致率在 60%以下。模型之间的一致率不高，分类结果具有一定的多样性。特别在利用 PMI 的词模型与 *N*-gram 模型消歧结果之间，一致率在 45%以下。可以说是由于在相同条件下，词模型与 *N*-gram 模型所得利用的消歧信息有着本质不同，故在很大程度

表 3　对比词模型与 *N*-gram 模型性能比较

词位置	$DICE_{mir}$	LLR_{mir}	PHI_{mir}	PMI_{mir}	$N\text{-gram}_{mir}$	$DICE_{mar}$	LLR_{mar}	PHI_{mar}	PMI_{mar}	$N\text{-gram}_{mar}$
−1	0.3647	0.3636	0.3690	0.3701	0.4513	0.4330	0.4325	0.4392	0.4491	0.5055
1	0.3711	0.3775	0.3647	0.3540	0.4043	0.4312	0.4279	0.4141	0.4142	0.4673
−2,−1	0.3508	0.3487	0.3508	**0.3829**	0.4535	0.4150	0.4089	0.4133	**0.4508**	0.5020
−1, 1	0.3540	0.3594	0.3701	0.3529	**0.4941**	0.4011	0.4057	0.4195	0.4145	**0.5587**
1,2	0.3604	0.3711	0.3594	0.3497	0.4225	0.4144	0.4235	0.4115	0.4185	0.4813

表 4　两个模型消歧结果的 P_c，P_{rc}，P_{rcc}

词位置	$P_{c\text{-DICE}}$	$P_{c\text{-LLR}}$	$P_{c\text{-PHI}}$	$P_{c\text{-PMI}}$	$P_{rc\text{-DICE}}$	$P_{rc\text{-LLR}}$	$P_{rc\text{-PHI}}$	$P_{rc\text{-PMI}}$	$P_{rcc\text{-DICE}}$	$P_{rcc\text{-LLR}}$	$P_{rcc\text{-PHI}}$	$P_{rcc\text{-PMI}}$
−1，0	0.5837	**0.5911**	0.5511	0.4007	0.2631	0.2663	0.2588	0.2107	0.4507	0.4505	0.4697	**0.5258**
0，1	0.5749	0.5653	0.5338	0.3842	0.2353	0.2321	0.2160	0.1658	0.4093	0.4106	0.4047	0.4315
−2，−1，0	0.5827	0.5771	0.5588	0.4372	0.2449	0.2428	0.2385	0.2096	0.4203	0.4207	0.4268	0.4795
−1，0，1	0.5756	0.5695	0.5611	0.4441	0.2749	**0.2759**	0.2749	0.2096	0.4776	0.4845	0.4898	0.4720
0，1，2	0.5573	0.5435	0.5147	0.3808	0.2289	0.2299	0.2128	0.1615	0.4107	0.4231	0.4135	0.4241

上影响了消歧决策。两个模型一致正确率并不高，最高只有 27.59%。两个模型一致正确比基本都在 50%以下(除 PMI,[−1,0])。这也从另一个侧面说明，很难利用简单的投票技术来对这两类模型进行性能的提高。

4.3 模型组合的性能上限

词模型与 *N*-gram 模型的的分类结果具有一定的多样性，这是分类器组合后性能提高的基础[18]。可以将不同设置的这两类模型作为基本分类器，那么模型融合就具有进一步性能提升的潜力。表 5 给出将这两类模型的分类结果最优组合起来能够达到的性能上限(Upperbound，即利用 Oracle Model 取得的消歧结果)。

表 5 中阴影部分的每一个数值的含义是，在相同设置情况下，对比词模型利用不同相关度计算方法的结果与 *N*-gram 模型的结果组合可能达到的上限，最后一列的 ALL 标签下方表示，在相同窗口范围内，所有相关度计算方法的结果与基于 *N*-gram 模型得到的结果组合所能达到的上限，最后一行标签 ALL 的右侧表示，不同窗口下 *N*-gram 模型的结果以及利用对比词模型利用当前相关度的消歧结果组合所能达到的上限。由于要对消歧结果做整体一致性的比较，因此这里对性能的评价采用 P_{mir} 值。最终模型组合可能的性能上限是 88.98%，这与 Pedersen [19]2002 年考察 Senseval-2 ELS 上所有参与评测的有指导系统分类结果所得出的上限规律并不相符。原因可能为：针对语言不同、分别针对有指导与无指导方法及两个任务所考察对象的词义数不同，前者所有词汇平均词义约为 4.93，后者约为 3.04。

表 5　两种模型结果 P_{mir} 性能上限

窗口	DICE	LLR	PHI	PMI	ALL
−1，0	0.5529	0.5487	0.5615	0.6107	0.6406
0，1	0.5401	0.5497	0.5529	0.5925	0.6332
−2，−1，0	0.5594	0.5594	0.5658	0.6267	0.6802
−1，0，1	0.5733	0.5775	0.5893	0.6374	0.6834
0，1，2	0.5540	0.5636	0.5690	0.6107	0.6706
ALL	0.7968	0.8043	0.8160	0.8695	**0.8898**

5　结束语

基于 Web 的无指导消歧词模型及 *N*-gram 模型方法简单且性能良好，在 SemEval-2007 的 task5 上的测试表明，两种模型方法均超过了目前已知最好的无指导消歧系统，且 *N*-gram 模型方法性能尤佳。对两个模型的比较表明，2-gram 及 3-gram 较相应的词所提供的消歧信息更为有效。对两个模型一致性的考察也得出了模型组合的结果有达到很高消歧性能的潜力。

虽然初步解决了无指导消歧知识获取并极大缓解了数据稀疏问题，本文基于 Web 的方法尚存在一些问题，主要是在利用搜索引擎进行 Page Counts 的统计时，难免会将一些语义无关的网页统计进来，对消歧造成很强的噪音影响。另外就是在进行更高阶的 *N*-gram 序列统计统计时，数据稀疏现象扔对性能有一定困扰。进一步工作可围绕以下两方面展开：

(1)基于 Web 的词模型的特征优化选择及网页噪音过滤；

(2)基于 Web 的词模型与 *N*-gram 模型的模型融合。

参 考 文 献

[1] Li Hang and Li Cong. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 2004, 30(1): 1–22.

[2] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 1994: 88–95.

[3] Niu Zheng-yu, Ji Dong-hong, Tan Chew lim, and Pakhomov S. Word sense disambiguation using label propagation based semi-supervised larning. Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Morristown, NJ, USA July 2005: 395–402.

[4] Gale W A, Church K W, and Yarowsky D. Using bilingual materials to develop word sense disambiguation methods. Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada, 1992: 101–112.

[5] Hwee Tou Ng, BinWang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: an empirical study. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003: 455–462.

[6] Chodorow L M and Miller G A. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 1998, 24(1): 147–165.

[7] Mihalcea R. Bootstrapping large sense tagged corpora. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas, Spain. 2002: 1407–1411.

[8] Agirre E and Martínez D. Unsupervised WSD based on automatically retrieved examples: The importance of bias. Proceedings of the Conference on Empirical Methods in NLP. Barcelona, Spain, 2004: 25–32.

[9] 刘鹏远, 赵铁军, 杨沐昀, 李壮. 基于等价伪译词的无指导译文消歧模型研究. 电子与信息学报, 2008, 30(7): 1690–1695.
Liu Peng-yuan, Zhao Tie-jun, Yang Mu-yun, and Li Zhuang. Unsupervised translation disambiguation based on equivalent pseudotranslation model. *Journal of Electronics & Information Technology*, 2008, 30(7): 1690–1695.

[10] Kilgarriff A and Grefenstette G. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 2003, 29(3): 333–348.

[11] Martinez D, Agirre E and Wang Xing-long. Word relatives in context for word sense disambiguation. Proceedings of the 2006 Australasian Language Technology Workshop (ALTW 2006), Sydney, Australia, 2006: 42–50.

[12] Mihalcea R and Moldovan D I. Word sense disambiguation based on Semantic Density. Proceedings of COLING-ACL Wordshop on Usage of WordNet in Natural Language Processing, Montreal, Canada, July 1998: 16–22.

[13] Turney P D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the Twelfth European Conference on Machine Learning, Berlin: Springer-Verlag, 2001: 491–502.

[14] Paolo Rosso, Manuel Montes-y-G´omez, Davide Buscaldi, Aar´on Pancardo-Rodr´guez, and Luis Villase˜nor Pineda. Two Web-based approaches for noun sense disambiguation. Int. Conf. on Compute. Linguistics and Intelligent Text Processing. CICLing-2005, Springer Verlag, LNCS (3406), Mexico D. F., Mexico, 2005: 261–273.

[15] Yang Che-yu. Word sense disambiguation using semantic relatedness measurement. *Journal of Zhejiang University SCIENCE A*, 2006, 7(10): 1609–1625.

[16] Liu Peng-yuan, Zhao Tie-jun, and Yang Mu-yun. HIT-WSD: Using search engine for multilingual Chinese-English lexical sample task. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, June 2007: 169–172.

[17] Mohammad S, Hirst G, and Resnik P. TOR, TORMD: Distribtional profiles of concepts for unsupervised word sense disambiguation. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). Prague, June, zech Republic. Association for Computational Linguistics Conference. 2007: 326–333.

[18] Gavin B, Wyatt J, Harris R, and Yao Xin. Diversity creation methods: A survey and categorization. *Information Fusion Journal*, 2004, (6): 5–20.

[19] Pedersen T. A baseline methodology for word sense disambiguation. Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City. February, 2002: 17–23.

刘鹏远： 男，1974 年生，博士，研究方向为词义消歧、信息检索、机器翻译与自然语言处理.

赵铁军： 男，1962 年生，教授，博士生导师，研究方向为自然语言处理、机器翻译及人工智能.