ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.5, May 2009, pp.1292−1300
doi: 10.3724/SP.J.1001.2009.03367

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

利用语义词典Web挖掘语言模型的无指导译文消歧*

刘鹏远+, 赵铁军

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Unsupervised Translation Disambiguation by Using Semantic Dictionary and Mining Language Model from Web

LIU Peng-Yuan+, ZHAO Tie-Jun

(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: pyliu@mtlab.hit.edu.cn

Liu PY, Zhao TJ. Unsupervised translation disambiguation by using semantic dictionary and mining language model from Web. *Journal of Software*, 2009,20(5):1292−1300. http://www.jos.org.cn/1000-9825/3367.htm

Abstract: In order to solve the problem of data sparseness and knowledge acquisition in translation disambiguation and WSD (word sense disambiguation), this paper introduces an unsupervised method, based on the *n*-gram language model and web mining. It is supposed that there exists a latent relationship between the word sense and *n*-gram language model. Based on this assumption, the mapping between the English translation of Chinese word and the DEF of Hownet is established and the word set is acquired. Then the probabilities of *n*-gram in the words set are calculated based on the query results of a searching engine. The disambiguation is performed via these probabilities. This method is evaluated on a gold standard Multilingual Chinese English Lexical Sample Task dataset. Experimental results show that the model gets the state-of-the-art results (P_{mar}=55.9%) and outperforms 12.8% on the best system in SemEval-2007.

Key words: WSD (word sense disambiguation); unsupervised translation disambiguation; language model; Web mining; knowledge acquisition

摘 要: 为了解决困扰词义及译文消歧的数据稀疏及知识获取问题,提出一种基于 Web 利用 *n*-gram 统计语言模型进行消歧的方法.在提出词汇语义与其 *n*-gram 语言模型存在对应关系假设的基础上,首先利用 Hownet 建立中文歧义词的英文译文与知网 DEF 的对应关系并得到该 DEF 下的词汇集合,然后通过搜索引擎在 Web 上搜索,并以此计算不同 DEF 中词汇 *n*-gram 出现的概率,然后进行消歧决策.在国际语义评测 SemEval-2007 中的 Multilingual Chinese English Lexical Sample Task 测试集上的测试表明,该方法的 P_{mar} 值为 55.9%,比其上该任务参评最好的无指导系统性能高出 12.8%.

关键词: 词义消歧;无指导译文消歧;语言模型;Web 挖掘;知识获取

* Supported by the National Natural Science Foundation of China under Grant No.60435020 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2006AA01Z150, 2006AA010108 (国家高技术研究发展计划(863))

Received 2007-12-12; Accepted 2008-04-15

中图法分类号: TP391 文献标识码: A

确定歧义词在特定上下文中的特定词义(word sense disambiguation,简称 WSD)或者确定歧义词的目标语译文(word translation disambiguation,简称 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.最近几届国际语义评测 Senseval-2[1],Senseval-3[2],Semeval-2007(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)[3]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.然而,迄今为止还没有一种语言存在已标注所有多义词的大规模语料库.目前,最大的独立英文语义标注语料库是 SemCor,里面含有 41 497 个词义标注的词.但是对全词消歧任务而言,大约有一半的测试语料实例无法从上述已标注语料中得到任何上下文特征信息[4].手工标注语料库的语义不但代价极为高昂,而且语义标注者之间的一致性又很难达到一个很高的标准[2,5,6],这又给利用手工标注的语料库进行训练以及评测造成一定影响.

针对有指导方法缺乏足够的已标注语料这一问题,主要研究路线有 3 条:

1) 利用种子语料以及各种半无指导方法进行词义消歧[7−9].

此类方法的问题一个是初始种子语料的选择,另一个就是随着自举过程的反复进行而不可避免地引入越来越强的噪音.

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[10,11].利用词对齐的平行语料,源语歧义词对应目标语的译文即成为该词的语义标注.此类方法的问题是[11]:首先,虽然有试图从 Web 挖掘平行语料的尝试[12],但是平行语料,特别是精确对齐的平行语料仍然非常少;第二,仅仅通过平行语料一般无法区分源语言所有的歧义词的语义,特别是当平行语料相对较小的时候.当然,这也是双语方法的一个共同问题:如果语料库较小,一种语言歧义词的部分语义就不会在这个语料库中出现,因此也就无法得到用于消歧的有用实例;另外,即使语料库规模足够大,一种语言歧义词的不同语义也常常被翻译成另一种语言的同一个词.

利用单语语料以及语义词典的方法[13−16].这类方法是利用歧义词在语义词典中的各类语义同义信息来自动获取单语语料并视为已标注语义的语料,利用这些语料以及机器学习算法来进行训练以及分类.该类方法也存在一些问题,主要是语义词典中部分目标词的某些语义没有对应的同义词,而若利用远距离关系词(distant relatives)又会引入噪音[17].

3) 本文所利用的是根据 Web 搜索计数(Web count)的消歧方法.

Mihalcea 等人[18]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,选择该计数最大的 Synset 作为该上下文对应歧义词的词义.Turney[19]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 等人[20]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,也即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[21]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 等人[22]沿着这个思路将该方法扩展到双语范畴并进行了初步的尝试.

受 Yuret[4]采用的利用词替换(substitution)进行词义消歧方法的启发,针对译文消歧任务,本文从一个新的视角对如何进行消歧知识的获取及其利用进行了研究与探讨.首先利用 Hownet[23]建立中文歧义词的英文译文与知网 DEF 的对应关系,并得到该 DEF 下的词汇集合.通过搜索引擎在 Web 上搜索,并以此计算得到不同 DEF 下词汇 *n*-gram 片段的出现概率.假设歧义词在不同语义下具有不同的语言模型模式,且同一语义下的词汇具有相

同语言模型模式的概率更高,就可以根据概率值来得到该语言模型下词汇所在的语义分类.利用已经建立的对应关系,就可以得到对应的英文译文.

本文方法与以往方法的区别在于:

1) 采用 *n*-gram 语言模型以及语义类 *n*-gram 语言模型[24]而非语义模型;
2) 搜索 Web 以得到对各种 *n*-gram 的统计而非利用语料库,这样就尽可能地避免了利用语料库方法的数据稀疏问题以及语料平衡性问题.

在整个消歧过程中,本文仅利用语义词典及 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.通过在国际语义评测 SemEval-2007 中的 Multilingual Chinese English Lexical Sample Task 测试集上的测试,结果表明,该方法取得了该任务无指导方法的最好结果,比参评的最好的系统的绝对性能提高了 11.7%.

1 利用*n*-gram统计语言模型的消歧方法

1.1 统计语言模型

统计语言模型(statistical language model)以概率论和数理统计理论为基础,用来计算自然语言序列出现的概率,使得正确序列的概率大于错误序列的概率.对一个自然语言序列 $S=w_1w_2\ldots w_n$,其概率可由公式(1)表示:

$$P(S)=\prod_{i=1}^{n}p(w_i\mid w_1w_2\ldots w_{i-1}) \tag{1}$$

由于公式(1)中条件概率的值无法从训练语料中估计出来,为了计算 $P(S)$的概率,必须作一定的独立性假设而进行计算的简化.最常用的一种简化就是标准 *n*-gram 语言模型.

标准 *n*-gram 语言模型将自然语言序列看作是一个 Markov 序列,满足 Markov 属性.具体来说,标准 *n*-gram 模型对公式(1)中条件概率 $p(w_i|w_1w_2\ldots w_{i-1})$作如下两个假设:

- 有限历史假设:当前语言单位的概率仅与前 $m-1$ 个语言单位有关,与其历史信息无关.
- 时齐性假设:当前语言单位的概率仅与该语言单位自身有关,与其在序列 S 的位置无关.

由此,公式(1)被简化成:

$$P(S)=\prod_{i=1}^{n}p(w_i\mid w_{i-m+1}\ldots w_{i-1}) \tag{2}$$

其中,条件概率 $p(w_i|w_{i-m+1}\ldots w_{i-1})$可由最大似然原理(maximum likelihood estimate,简称 MLE)从训练语料中估计得到:

$$p(w_i\mid w_{i-m+1}\ldots w_{i-1})=\frac{C(w_{i-m+1}\ldots w_i)}{C(w_{i-m+1}\ldots w_{i-1})} \tag{3}$$

其中,C 是语言单位 x 在训练语料中出现的次数.

1.2 统计语言模型与译文消歧

对于译文消歧任务,以汉英翻译为例,是在给定目标词汉语上下文 C 的情况下确定其英文译文.一般来说,这里的汉语目标词为多义词,对于确定汉语单义词在英语中不同译文的任务,其重点在于译文选择.无指导译文消歧方法最常用的信息就是由汉语目标词的上下文得到的词兜(word bag),主要利用上下文词汇与歧义词各个语义类之间的距离或者同现来进行消歧.

人类在日常交流或者思考时很少会意识或者考虑到自己所用词汇的歧义性,但是却能很明确地知道歧义词在各种表述中的正确含义.我们很难想像也很少有经历过在日常交流中要借助句子中词汇与歧义词语义类之间的任何关系才能确定对方所用歧义词的正确含义.那么怎样解释在不考虑上下文众多词汇之间语义关系的情况下,来确定歧义词汇含义的熟练性呢?让我们来看一个具体的例子:“中医”这个词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”时,我们很容易知道这句话里的中医表示第 1 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,因此会很自然地知道,这里的中医是医学方面中医的含义.在经过对知网以及语料库进行初步考察之后,我们作如下假设:

假设 1. 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 2. 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.

由假设 1 可知,s_1 与 s_2 因词汇“中医”语义不同而具有不同的模式.而根据假设 2,含有相同语义词汇“医学/西医/外科”的词汇序列模式(均为是 s_1)较之不同语义下的“医生”的词汇序列模式(s_2)更容易相同.也就是说,相同语义词间的词汇序列模式有一定概率的可替换性.当然,反义词也通常会具有这样的性质,完全可以利用反义词来进行模式的替换并进行译文消歧,但是由于反义词所涵盖的词汇有限,本文仅对同义词替换进行了探讨.同时,根据我们对知网的统计,目前尚无歧义词的两个含义是反义,因此对具体的歧义词而言,假设 1 并不与歧义词的某个含义的反义词也可替换该含义模式的性质相矛盾;同时,假设 2 也不否认相同语义可能会具有不同的模式.本文仅是力图证明这两个假设的有效性,对其适用范围与适用条件暂不作进一步探讨.

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 1,我们可以将 S 根据不同的语义分为 $S_1,S_2,S_i,\ldots,S_n$,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 $C_i=\{c_{i1},c_{i2},\ldots,c_{im_i}\}$ 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 2 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 $w_1w_2\ldots w\ldots w_k$ 表示,则词汇序列 $w_1w_2\ldots c_{ij}\ldots w_k$ 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.

如果我们可以确定 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 e.但在无指导的方法中,由于没有标注语料,S_i 的初始划分无法确定,我们只能通过比较在 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:

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

$$S_i/SC_i/C_i=\arg\max_i \frac{1}{m_i}\sum_{j=1}^{m_i}P_s(c_{ij}) \tag{5}$$

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,m_i 表示 C_i 内词汇的个数,对 $p_s(c_{ij})$,我们也进行标准 n-gram 模型的简化,其值可由公式(2)和公式(3)来进行计算.公式(4)和公式(5)的左侧是指,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,它们是一一对应的,因此用 $S_i/SC_i/C_i$ 表示.公式(4)是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,记为 Ngram 方法.公式(5)是要利用同义词集合内的词汇形成符合 s 词汇序列的所有模式来找到概率均值最大的同义词集,也就找到了对应的语义,可记为 P_Ngram 方法.公式(5)可以算是利用语义类 n-gram 模型的一个特例,其区别在于,本文仅对词汇序列中的目标歧义词进行基于类的建模.

1.3 利用Web挖掘进行*n*-gram语言模型的统计

对于公式(3),通常的办法是由最大似然原理(MLE)从一个训练语料库估计词汇序列 $C(w_{i-m+1}\ldots w_i)$以及 $C(w_{i-m+1}\ldots w_{i-1})$的个数,但是本文将整个 Web 视为一个语料库[25],利用语料库对公式(3)进行 MLE 的估计.这样,公式(3)中的 C 就是语言单位 x 在 Web 中利用搜索引擎得到的 Web 计数(Web count).

采用 Web 代替语料库对词汇序列进行估计的优点如下:

- Web 是公共海量信息资源,比语料库更容易获得.
- 对所有语言来说,其上的信息均仍有越增越多的趋势,Web 比语料库更能反映随时代发展的词汇、语义及语言的逐渐变化.
- Web 比普通语料库更能体现均衡性,其数据量也能进一步削弱统计 n-gram 时潜在的数据稀疏问题.

2 基于*n*-gram语言模型及Web挖掘的消歧模型

消歧流程如图 1 所示.

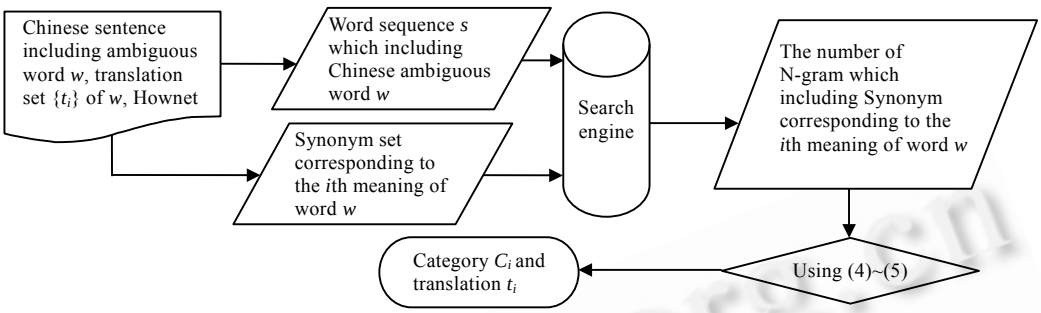

Fig.1 Unsupervised translation disambiguation model based on *n*-gram language model and Web mining

图 1 基于 *n*-gram 语言模型及 Web 挖掘无指导译文消歧模型

以汉英译文消歧为例来说明本文方法的消歧过程(对含有歧义词 w 的中文句子 s 以及该歧义词的英文译文集合 T,英汉译文消歧的任务就是要确定在 s 中歧义词 w 对应的集合 T 中的正确译文 t):

1) 由中文句子直接得到包含歧义词 w 的词汇序列 s;由该词的译文集合$\{t_i\}$以及知网得到 w 的第 i 个语义对应的同义词集合 C_i;

2) 对 C_i 中所有词汇循环进行步骤 a)、步骤 b)和步骤 c)的操作:

   a) 得到符合词汇序列 s 的模式 s_c_{ij};

   b) 利用搜索引擎得到与计算公式(2)、公式(3)所需的 Web 计数;

   c) 计算各种模式出现的概率.

3) 根据公式(4)、公式(5)来确定词汇序列 s 应属的 SC_i/C_i,也就得到了正确译文 t_i.

下面以 Semeval 2007 测试集合中的一个条目为实例来说明本文方法的具体消歧过程.利用百度(www.baidu.com)作为搜索引擎,并设网页总数为 10 亿个,总词次为 10 000 亿.

例句:攻克格罗兹尼对俄军来说是剿匪行动的一个极其重要的胜利;非法武装失守格罗兹尼,则〈head〉说明〈/head〉他们在实力上、精神上已经失败.

译文:t_1:show,对应知网同义词集合(仅各取 4 个同义词为例)C_1={表达,表露,表明,表示};

t_2:explain,对应知网同义词集合 C_2={澄清,发挥,谈论,说}.

由以上输入,按照图 1 进行消歧,整个流程如下:

1) 得到 *n*-gram 词汇序列 s(取 N=3)为“则说明他们”,则有:

C_1 符合 s 词汇序列的模式为:s_表达,则表达他们;s_表露,则表露他们;s_表明,则表明他们;s_表示,则表示他们.

C_2 符合 s 词汇序列的模式为:s_澄清,则澄清他们;s_发挥,则发挥他们;s_谈论,则谈论他们;s_说,则说他们.

2) 根据公式(2)、公式(3)对所有词汇序列模式在搜索引擎上进行搜索,经过计算可得:

P_s(表达)=1.24E−12; P_s(表露)=2.6E−13; P_s(表明)=5.99E−11; P_s(表示)=1.63E−10;

P_s(澄清)=1.82E−12; P_s(发挥)=3.44E−12; P_s(谈论)=1.22E−11; P_s(说)=1.15E−10.

3) 决策:

利用公式(4),由于 P_s(表示)最大,因此取 C_1 符合 s 词汇序列的模式,也即取译文 t_1:show.

利用公式(5),由于 $\frac{1}{m_1}\sum_{j=1}^{m_1}P_s(c_{1j})$=1/4(1.24E−12+2.6E−13+5.99E−11+1.63E−10)=5.61E−11> $\frac{1}{m_2}\sum_{j=1}^{m_2}P_s(c_{2j})$=1/4(1.82E−12+3.44E−12+1.22E−11+1.15E−10)=3.31E−11,因此取 C_1 符合 s 词汇序列的模式,也即取译文 t_1:show.

3 实 验

3.1 评测语料与baseline

利用 ACL 2007 评测的一个组成部分 SemEval 2007 国际语义评测的中英文词汇任务(Task #5 Multilingual Chinese_English Lexical Sample Task)对本文方法进行评测.该任务共含 40 个歧义词(所有词在表 3 中详细列出),语料由训练语料以及测试语料两部分组成,总体情况见表 1.

本文没有利用任何训练语料,而是对其测试语料直接进行测试.实验的 baseline 为在 SemEval 2007 评测中该任务表现最好的无指导消歧系统 TorMd 以及另一个利用 Web 的无指导系统 HIT.

Table 1 Basics of gold standard dataset

表 1 标准评测语料情况

	Average meaning	Training data	Testing data
19 nouns	2.45	1 019	364
21 verbs	3.57	1 667	571

3.2 利用知网形成评测歧义词的同义词集合

SemEval 2007 的 Task #5 提供了 40 个歧义词 w 及每个词汇到英文译文的映射 $\{t\}_i w$.为使本实验顺利进行,需要找到每个英文译文对应的知网 DEF,然后得到每个歧义词各个译文所对应的知网汉语同义词集合.这个过程是半自动进行的.

对每一个评测歧义词 w,

1) 利用知网得到 w 的英文译文集合 T_w 以及 w 各语义的 DEF 集合 D_w,同时建立空集合 HD_w;对 w 的英文译文的映射 $\{t\}_i w$ 中的每一个 t,

   若 $t \in T_w$,则:

   a) 将 t 从 T_w 移出,将 t 所对应的 D_w 中的 DEF 移到 HD_w 中,并保留对应关系;

   b) 将 t 从 $\{t\}_i w$ 中移出;

2) 若 $\{t\}_i w$ 非空,则:

   a) 由汉语、英语熟练的语言学家将 $\{t\}_i w$ 中的每一个 t 找到正确对应的知网 DEF;

   b) 将 t 从 $\{t\}_i w$ 中移出;

   c) 将 t 所对应的 D_w 中的 DEF 移到 HD_w 中,并保留对应关系.

完成以上两步以后,HD_w 就是每一个歧义词的译文所对应的知网 DEF 集合,根据这个集合就能得到在知网 DEF 下 w 不同语义所对应的同义词集合 C_i.

3.3 实验及讨论

实验以歧义词为中心,进行 2-gram 以及 3-gram 语言模型的选取.以“非法武装失守格罗兹尼,则〈head〉说明〈/head〉他们在实力上、精神上已经失败.”为例,表 2 给出了实验中具体的 n-gram 选取情况.

Table 2 Selection of n-gram

表 2 n-gram 的选取

	2-gram		3-gram		
Position	–1,0	0,1	–2,–1,0	–1,0,1	0,1,2
n-gram	则说明	说明他们	,则说明	则说明他们	说明他们在

对测试语料中所有 935 个例句不作任何处理,直接按表 2 分别进行 2-gram 以及 3-gram 的选取.对歧义词的知网同义词集合 C_i(可记为 C_i-all),我们将其分为两个集合,分别是单义词同义词集合 C_i-single、多义词同义词集合 C_i-multi.可知,C_i-all=C_i-single∪C_i-multi.利用这 3 个同义词集合分别进行实验,其结果分别以-s,-m,-a 来表示.按照上一节所示过程,搜索引擎采用百度(www.baidu.com),分别利用公式(4)、公式(5)对测试例句内的目标歧

义词进行消歧,其结果分别以 Ngram 和 P_Ngram 表示.

利用 Multilingual Chinese_English Lexical Sample Task 评测任务提供的标准评测工具进行评测,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):

$$p_{mir}=\sum_{i=1}^{N}m_i\Big/\sum_{i=1}^{N}n_i,\ p_{mar}=\sum_{i=1}^{N}p_i\Big/N \tag{6}$$

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,$p_i=m_i/n_i$.

实验结果见表 3.由表 3 可知,采用 n-gram 语言模型以及 Web 数据挖掘的方法明显优于 baseline 系统 HIT.所有组合所得到的结果均超过了在 SemEval 2007 中英文词汇任务上表现最好的系统 TorMd.最好的结果为 3-gram 语言模型,词汇位置为(−1,0,1),且利用单义词同义词集合计算的结果,其 P_{mir} 值为 0.498(见公式(5)),比 TorMd 的性能绝对提高 12.3%,P_{mar} 值为 0.559,比 TorMd 的性能绝对提高 12.8%(见公式(4)).

Table 3 Experiment results of each model

表 3 各模型的实验结果

Position	2-gram		3-gram			P_2-gram		P_3-gram		
	−1,0	0,1	−2,−1,0	−1,0,1	0,1,2	−1,0	0,1	−2,−1,0	−1,0,1	0,1,2
P_{mir} (-a)	0.398	0.415	0.415	0.447	0.414	0.390	0.398	0.419	0.437	0.416
P_{mar} (-a)	0.458	0.464	0.464	0.509	0.472	0.463	0.455	0.483	0.503	0.473
P_{mir} (-m)	0.398	0.401	0.427	0.428	0.409	0.406	0.390	0.413	0.443	0.416
P_{mar} (-m)	0.459	0.456	0.480	0.491	0.469	0.477	0.454	0.474	0.508	0.475
P_{mir} (-s)	0.451	0.404	0.454	0.494	0.423	0.455	0.403	0.470	0.498	0.425
P_{mar} (-s)	0.506	0.467	0.502	0.559	0.481	0.516	0.461	0.532	0.558	0.479
P_{mir}	TorMd		0.375			HIT		0.337		
P_{mar}			0.431					0.396		

为了进一步考察各模型的性能,我们绘制了各模型性能比较图,如图 2 所示.图 2 左图共有 10 对柱列,前 5 对代表利用 P_{mir} 考察模型性能的结果,后 5 对代表利用 P_{mar} 考察模型性能的结果.5 对柱列均按照词汇序列的位置不同依次排列(−1,0;0,1;−2,−1,0;−1,0,1;0,1,2),每对柱列分别表示采用利用公式(4)的 Ngram 方法(前)与利用公式(5)的 P-Ngram 方法(后)的结果,精度数值为考虑全部同义词集合、多义词同义词集合以及单义词同义词集合结果的平均值.从图 2 左图我们可以看出:1) Ngram 方法与 P-Ngram 方法的性能基本相当.2) 3-gram 词汇序列,位置(−1,0,1)的模型性能最优;其次是 3-gram 词汇序列,位置(−2,−1,0)的模型;2-gram 词汇序列,位置(−1,0)与 3-gram 词汇序列,位置(0,1,2)这两个模型性能接近;2-gram 词汇序列,位置(0,1)模型的性能最差.

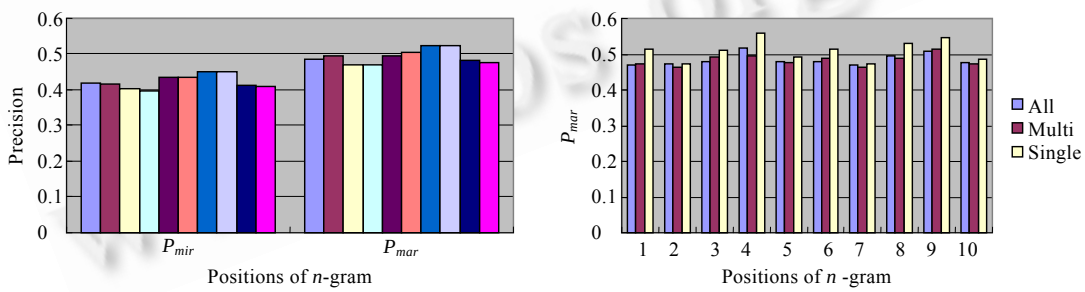

Fig.2 Comparision of models' performance

图 2 各模型性能的比较

图 2 右图共有 30 个柱列,每 3 个为一组,10 组柱列按照词汇序列位置与表 3 所列位置依次对应.每组 3 个柱列分别对应利用全部同义词集合(C_i-all)、多义词同义词集合(C_i-multi)以及单义词同义词集合(C_i-single)各模型的 P_{mar} 结果(P_{mir} 结果与 P_{mar} 结果的规律基本类似).从图 2 右图可以发现,利用单义词同义词集合的模型性能最优,其次是利用全部同义词集合,利用多义词同义词集合的模型性能最差.其原因在于,根据假设 1 可知,多义词的语言模型根据其语义不同而有不同的模式.而假设 2 中所说"模式更容易相同"系指同义词符合相同模式的

概率更大,但是并不排斥具有其他语义的词汇也会有一定的概率符合这一模式.这样,多义词在一定程度上对根据词汇序列模式进行目标歧义词消歧造成了一定的干扰.

我们将 HIT,TorMd 以及基于语言模型的最优系统(Ngram)对 40 个词消歧的 P_{mar} 结果整理在表 4 中,其中左边为 19 个名词的结果,右边为 21 个动词的结果.从表 4 中我们可以看出,基于 *n*-gram 语言模型的系统无论是对名词还是动词,都明显优于其余两个系统.与 SemEval 2007 上该任务表现最好的无指导系统 TorMd 相比,名词绝对性能提高了 7.3%,动词绝对性能提高了 17.6%.基于 *n*-gram 语言模型的方法对名词消歧的效果比对动词消歧的效果要好 5.8%,考虑到其他两个系统名词比动词消歧效果分别高 12.9%及 16.1%,且动词平均词义数比名词多 1.12 个,可以说,本文方法对动词消歧相对于其余系统更为有效.

Table 4 Detail experiment results of each system

表 4 各系统实验详细结果

Testing nouns	Meaning number	HIT	TorMd	*n*-gram	Testing verbs	Meaning Number	HIT	TorMd	*n*-gram
本	3	0.320	0.720	0.600	补	3	0.550	0.550	0.500
表面	2	0.333	0.556	0.333	成立	3	0.407	0.481	0.555
菜	2	0.632	0.474	0.579	吃	4	0.174	0.174	0.609
长城	3	0.619	0.429	0.667	出	9	0.091	0.169	0.195
单位	2	0.529	0.706	0.588	带	8	0.104	0.119	0.194
道	3	0.222	0.500	0.778	动	4	0.300	0.300	0.450
队伍	3	0.364	0.318	0.591	动摇	2	0.438	0.500	0.875
儿女	2	0.500	0.500	0.550	发	5	0.139	0.250	0.306
机组	2	0.571	0.643	0.643	赶	3	0.333	0.389	0.389
镜头	2	0.467	0.467	0.667	叫	4	0.256	0.256	0.487
面	3	0.696	0.348	0.696	进	5	0.114	0.250	0.386
牌子	2	0.529	0.353	0.412	开通	2	0.500	0.500	0.550
旗帜	3	0.111	0.500	0.389	看	4	0.294	0.294	0.412
气息	2	0.571	0.857	0.714	平息	2	0.500	0.375	0.875
气象	2	0.563	0.438	0.813	使	2	0.438	0.563	0.563
日子	3	0.344	0.281	0.375	说明	2	0.556	0.444	0.611
天地	3	0.440	0.560	0.360	挑	2	0.286	0.143	0.500
眼光	2	0.500	0.714	0.500	推翻	2	0.300	0.300	0.700
中医	2	0.500	0.438	0.938	望	2	0.462	0.462	0.538
					想	4	0.216	0.216	0.676
					震惊	2	0.571	0.714	0.786
Average P_{mar}	2.45	0.464	0.516	0.589		3.57	0.335	0.355	0.531

4 结束语

基于 *n*-gram 语言模型及 Web 挖掘的无指导译文消歧方法简单且性能良好,在 SemEval 2007 上的测试结果表明,该方法比参加该项评测的最好系统绝对性能提高了 12.8%.该方法不需要任何已标注语料,仅需要针对该语言的搜索引擎以及源语言的语义词典,一定程度上解决了消歧知识自动获取以及潜在的数据稀疏问题.但在进行大规模词义消歧以及译文消歧任务之前,需要对其在大规模词汇上进行深入实验,同时需要进一步提高精确率.进一步的研究工作可从如下几个方面入手:

1) 分析歧义词在不同语义下的 *n*-gram 词汇序列模式,以期得到不同 *n*-gram 选取对该方法性能的影响.

2) 扩大同义词集合到反义词、近义词等,同时研究本方法的适用范围.

3) 确定最优消歧决策方法.

References:

[1] Edmonds P, Cotton S. Senseval-2: Overview. In: Preiss J, Yarowsky D, eds. Proc. of the 2nd Int'l Workshop on evaluating Word Sense Disambiguation Systems. Madison: Omni Press, 2001. 1−5.

[2] Mihalcea R, Edmonds P, eds. Senseval-3. In: Mihalcea R, Edmonds P, eds. Proc. of the 3rd Int'l Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics Conf. (ACL 2004). Madison: Omni Press, 2004. 1−17.

[3] Jin P, Wu YF, Yu SW. SemEval-2007 Task 5: Multilingual Chinese-English lexical sample. In: Agirre E, Marquez L, Wicentowski

R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007). Madison: Omni Press, 2007. 19−23.
[4] Yuret D. KU: Word sense disambiguation by substitution. In: Agirre E, Marquez L, Wicentowski R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval 2007). Madison: Omni Press. 207−214.
[5] Véronis J. A study of polysemy judgements and inter-annotator agreement. In: Programme and Advanced Papers of the Senseval Workshop. Herstmonceux Castle, 1998. http://www.up.univ-mrs.fr/~veronis/pdf/1998senseval.pdf
[6] Ng HT, Lim CY, Foo SK. A case study on inter-annotator agreement for word sense disambiguation. In: Proc. of the Siglex-ACL Workshop on Standarizing Lexical Resources. 1999. 9−13. http://www.aclweb.org/anthology-new/W/W99/W99-0502.pdf
[7] Li H, Li C. Word translation disambiguation using bilingual bootstrapping. Computational Linguistics, 2004,20(4):563−596.
[8] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: Pustejovsky J, ed. Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers, 1994. 88−95.
[9] Niu ZY, Ji DH, Tan CL, Pakhomov S. Word sense disambiguation using label propagation based semi-supervised learning. In: Knight K, ed. Proc. of the 43th Annual Meeting of the Association for Computational Linguistics (ACL). Madison: Omni Press, 2005. 395−402.
[10] Gale WA, Church KW, Yarowsky D. Using bilingual materials to develop word sense disambiguation methods. In: Proc. of the Int'l Conf. on Theoretical and Methodological Issues in Machine Translation. Montreal, 1992. 101−112.
[11] Ng HT, Wang B, Chan YS. Exploiting parallel texts for word sense disambiguation: an empirical study. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, 2003. 455−462.
[12] Resnik P, Smith NA. The Web as a parallel corpus. Computational Linguistics, 2003,29(3):349−380.
[13] Chodorow LM, Miller GA. Using corpus statistics and WordNet relations for sense identification. Computational Linguistics, 1998, 24(1):147−165.
[14] Mihalcea R. Bootstrapping large sense tagged corpora. In: Proc. of the 3rd Int'l Conf. on Language Resources and Evaluation (LREC). Las Palmas, 2002. 1407−1411.
[15] Agirre E, Martínez D. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In: Lin DK, Wu DK, eds. Proc. of the Conf. on Empirical Methods in NLP. Madison: Omni Press, 2004. 25−32.
[16] Lu ZM, Liu T, Li S. Chinese word sense disambiguation based on extension theory. Journal of Harbin Institute of Technology, 2006,38(12):2026−2029 (in Chinese with English abstract).
[17] Martinez D, Agirre E, Wang XL. Word relatives in context for word sense disambiguation. In: Proc. of the 2006 Australasian Language Technology Workshop (ALTW 2006). Sydney, 2006. 42−50.
[18] Mihalcea R, Moldovan DI. Word sense disambiguation based on semantic density. In: Harabagiu S, ed. Proc. of the COLING-ACL Workshop on Usage of WordNet in Natural Language Processing. San Francisco: Morgan Kaufmann Publishers, 1998. 16−22.
[19] Turney PD. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proc. of the 20th European Conf. on Machine Learning. Berlin: Springer-Verlag, 2001. 491−502.
[20] Rosso P, Montes-y-Gómez M, Buscaldi D, Pancardo-Rodrǵuez A, Pineda LV. Two Web-based approaches for noun sense disambiguation. In: Proc. of the Int'l Conf. on Compute. Linguistics and Intelligent Text Processing, CICLing-2005. LNCS 3406, Berlin, Heidelberg: Springer-Verlag, 2005. 261−273.
[21] Yang CY. Word sense disambiguation using semantic relatedness measurement. Journal of Zhejiang University (SCIENCE A), 2006,7(100):1609−1625.
[22] Liu PY, Zhao TJ, Yang MY. HIT-WSD: Using search engine for multilingual Chinese-English lexical sample task. In: Agirre E, Marquez L, Wicentowski R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007). Madison: Omni Press, 2007. 169−172.
[23] http://www.keenage.com
[24] Brown PF, deSouza PV, Mercer RL, Pietra VJD, Lai JC. Class-Based *n*-gram models of natural language. Computational Linguistics, 1992,18(4):467−479.
[25] Kilgarriff A, Grefenstette G. Introduction to the special issue on the Web as corpus. Computational Linguistics, 2003,29(3): 333−348.

附中文参考文献:

[16] 卢志茂,刘挺,李生.基于可拓学理论的汉语词义消歧.哈尔滨工业大学学报,2006,38(12):2026−2029.

刘鹏远(1974－),男,黑龙江哈尔滨人,博士,讲师,主要研究领域为自然语言处理,词义消歧.

赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,人工智能.