ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.4, April 2010, pp.575−585
doi: 10.3724/SP.J.1001.2010.03574

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

基于双语词汇 Web 间接关联的无指导译文消歧*

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,黑龙江 哈尔滨 150001)

Unsupervised Translation Disambiguation Based on Web Indirect Association of Bilingual Word

LIU Peng-Yuan[1+], ZHAO Tie-Jun[2]

[1](Institute of Computational Linguistics, Peking University, Beijing 100871, China)

[2](College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: liupengyuan@pku.edu.cn

Liu PY, Zhao TJ. Unsupervised translation disambiguation based on Web indirect association of bilingual word. *Journal of Software*, 2010,21(4):575−585. http://www.jos.org.cn/1000-9825/3574.htm

Abstract: To solve the problems of data sparseness and knowledge acquisition in translation disambiguation and WSD (word sense disambiguation), this paper introduces a fully unsupervised method, which is based on Web mining and Web indirect association of bilingual words. It provides new knowledge of translation disambiguation. It assumes that word sense can be determined by indirect association of bilingual words. Based on Web, this paper revises four common methods of indirect association, and designs three decision methods. These methods are evaluated on a gold standard Multilingual Chinese English Lexical Sample Task dataset of SemEval- 2007. The experimental results show that the model gets the state-of-the-art results (P_{mar}=44.4%) and outperforms the best system in SemEval-2007.

Key words: WSD; unsupervised translation disambiguation; Web indirect association; knowledge acquisition

摘 要: 为解决困扰词义消歧及译文消歧任务中存在的数据稀疏及知识获取问题,提出一种利用双语词汇 Web 间接关联的完全无指导消歧方法.首先做出词汇歧义可由双语词汇的间接关联度决定的假设,为译文消歧提供了一种新的知识.在此基础上,对 4 种常用计算间接关联的方法进了改造并定义了双语词汇 Web 间接关联.随后进行基于 Web 的词汇消歧知识获取并设计了 3 种消歧决策方法.最后,在国际语义评测 SemEval-2007 中的 Multilingual Chinese English Lexical Sample Task 测试集进行了测试.该方法的 P_{mar} 值为 44.4%,超过了该评测上最好的无指导系统的结果.

关键词: 词义消歧;无指导译文消歧;Web 间接关联;知识获取

中图法分类号: TP391 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60903063 (国家自然科学基金); the National Basic Research Program of China under Grant No.2004CB318102 (国家重点基础研究发展计划(973))

Received 2008-11-03; Accepted 2009-01-15

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的方法和无指导的方法两种.最近几届国际语义评测 Senseval-2[1],Senseval-3[2],Semeval-2007(http://nlp.cs.swarthmore.edu/semeval/,本文中公共标准测试语料、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一般需要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.但是,迄今为止,没有任何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,对于全词消歧任务而言,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及评测带来一定影响.

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,除了传统的基于词典的各类方法以外,主要的基于统计的研究路线有如下 3 种:

1) 利用种子语料等各种半有指导方法进行词义消歧[7−9]

该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,则源语歧义词对应目标语的译文即成为该词的语义标注.该类方法的问题是[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,特别是在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并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,但是一种语言歧义词的不同语义却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[13−17].该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)时又会引入噪音[18].

3) 本文所利用的根据 Web 搜索计数(Web count)的消歧方法

Mihalcea 等人[19]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[20]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[21]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[22]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,并由此对歧义词进行词义消歧,该方法得到了不错的结果.Liu 等人[23]沿着这条路线将该方法扩展到双语范畴,并进行了初步的尝试.

本文方法与以往该类方法主要区别在于:

(1) 本文首次利用这种新的消歧知识——双语词汇间的间接关联关系进行译文消歧,取得了很好的效果;

(2) 在进行间接关联的计算中,首次利用了 Web 双语混合文档页面(mix-language Web page);

(3) 本文利用 Web 直接建立双语词汇间的关联,是一种完全无指导的方法(fully unsupervised).

本文研究的出发点是:充分利用 Web 这个公共海量资源,无指导地自动获得可用于译文消歧(translation disambiguation)的双语词汇间的知识或联系,同时,力图缓解双语词汇间译文消歧知识获取瓶颈的问题.这里所关注的译文消歧,就是要确定源语言歧义词在目标语中的译文,是一个双语范畴内的词义消歧过程.该方法直接利用 Web 及搜索引擎建立并量化源语言词汇与目标语译文之间的联系,通过计算歧义词源语言上下文词汇与译文词汇之间的间接关联度来选择具有最大关联的译文作为消歧结果.由于不需要任何标注语料,仅仅需要能够应用于两种语言的搜索引擎以及相应的 Web 文本,这样就减轻了人工标注语料负担及知识获取的困难.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及词典,是一种无指导的方法.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,该方法的性能超过了参与该项任务比赛的最好的无指导系统.

本文第 1 节介绍基于间接关联的译文消歧思想及计算.第 2 节是消歧决策方法.第 3 节介绍实验、评价标准、实验结果及讨论.第 4 节是结束语.

1 基于间接关联的译文消歧思想及计算

1.1 双语词汇的间接关联

间接关联(indirect association,简称 IA)最早由 Melamed[24]在研究自动构建翻译词典时提出.如图 1 所示,图中显示为一个双语平行句对,第(1)行与第(2)行分别代表源语言以及目标语句子.c_k 与 e_k 是互译词对,一般具有明显的共现特征,所以任何计算关联度的统计模型都可以将两者联系到一起.这种真正互译词对产生的共现也被称做直接相关或者直接关联(direct association,简称 DA),反映出两词之间分布的相互依赖性.而在实际语言中,c_k 与 e_k 都可能存在一个单语中经常随之出现的词,如搭配、复合词等等各种原因产生.现在假设 c_{k+1} 经常出现在 c_k 的上下文中,则统计模型很容易就将(e_k,c_{k+1})误认为是翻译候选,两者之间的关联性随着它们分别与 c_k 之间关联性的增加而有所增加.相对于直接共现,这种现象被称为间接共现,也就是间接关联.

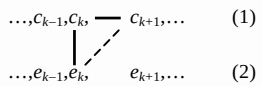

Fig.1 Direct and indirect association

图 1 直接关联与间接关联

双语词汇的间接关联可如下定义:

定义 1. 对任意源-目标语词对(c,e),如果 c 和 e 与源语言词集合 W(非空)内的词分别直接相关,则称(c,e)通过中间集 W 间接相关,用$(c,e)_W$表示.

考虑最简单的情况,当且仅当集合 W 内只有一个词 w 时,则可称词对(c,e)通过中间词 w 间接相关,以$(c,e)_w$表示.

1.2 利用间接关联进行译文消歧

以图 2 所示为例,借助双语平行句对来考察译文消歧任务.对于最一般的情况,假设源语言歧义词 w 有 n 个词义,分别对应 n 个目标语译文,译文消歧就是要利用双语平行句对(i)中源语言句子内的各种上下文信息来找到 w 在目标语句子内最适合的译文 t_i(注意,一般在译文消歧任务中,目标语句子实际上并不存在).

$$\begin{aligned}
&\ldots,c_{1,-1},w(1),c_{1,+1},\ldots;\ \ldots,e_{1,-1},t_1,e_{1,+1},\ldots \quad (1)\\
&\ldots\\
&\ldots,c_{i,-1},w(i),c_{i,+1},\ldots;\ \ldots,e_{i,-1},t_i,e_{i,+1},\ldots \quad (i)\\
&\ldots\\
&\ldots,c_{n,-1},w(n),c_{n,+1},\ldots;\ \ldots,e_{n,-1},t_n,e_{n,+1},\ldots \quad (n)
\end{aligned}$$

Fig.2 Parallel sentences pairs and translation disambiguation

图 2 双语平行句对与译文消歧

图 2 中每一行均代表英汉双语平行语料库中的一个句对,$w(i)$与 t_i 是互译词对,$c_{i,-1}$,$c_{i,+1}$ 与 $e_{i,-1}$,$e_{i,+1}$ 分别代表源语言与目标语句子的上下文.将每一个句对用元组$(w(i),C_i;t_i,E_i)$表示,其中:$w(i)$表示源语言歧义词,且其语义为第 i 个;C_i 表示 $w(i)$对应的上下文集合;$t_i \in$ 译文集合 T,表示 $w(i)$对应的目标语译文;E_i 表示译文 t_i 对应的上下文集合.由直接关联的概念可知,$w(i)$与 t_i 直接关联.

词义消歧任务中最常用到的一个基本假设就是:一段上下文决定一个词义[25].此基本假设进一步推广可表述成:类似的上下文决定类似的词义.因此,C_i 中的词会与 $w(i)$经常出现在类似的上下文中,直接关联.因此,由间接关联的定义可知,C_i 中的词与 t_i 间接关联.

将汉语以及英语所有上下文词汇合并在一起来分析,观察双语上下文,基本假设可扩展到双语范畴.由于 t_i 与 $w(i)$是互译词对,同时,类似的上下文决定类似的词义,因此,t_i 就更容易出现在 $w(i)$所在的双语上下文 C_i+E_i 中.同时,C_i 中的词与 t_i 是间接关联的关系.而在实际译文消歧任务中一般不存在 E_i,则$(w,C;T,E)$可简化成$(w,C;T)$.作如下假设:

假设 1. 源语言歧义词的译文可由该译文与该歧义词的上下文词汇通过源语言歧义词的间接关联度决定.

由一段上下文决定一个词义的假设很自然地可以得出一段源语言上下文决定歧义词一个译词的推论假设.假设 1 将译文消歧问题视为译文与源语言歧义词上下文间接关联度的问题.源语言歧义词的不同译文通过中间词(也就是源语言歧义词本身)与其上下文词汇间接关联.利用这种间接关联的不同或者程度强弱,就可以确定当前上下文情况下源语言歧义词的正确目标语译文.

1.3 间接关联的计算方法

1.3.1 现有的常用词汇关联的计算方法

4 种最常用的计算候选单元之间翻译概率的统计同现模型为点式互信息(point-wise mutual information,简称 PMI)[26]、Dice 系数(DICE)[27]、Phi 平方系数(ϕ^2)[28]和对数似然比(log likelihood ratio,简称 LLR)[29].假设英语候选单元为 ep,汉语候选单元为 cp,首先为每个候选翻译对(ep,cp)引入以下联列关系,以便计算英语单元 ep 和汉语单元 cp 之间的关联程度:

$a=freq(ep,cp)$: 同时包含英语单元 ep 和汉语单元 cp 的句对数;

$b=freq(ep)-freq(ep,cp)$: 仅包含英语单元 ep,不含有汉语单元 cp 的句对数;

$c=freq(cp)-freq(ep,cp)$: 仅包含汉语单元 cp,不含有汉语单元 ep 的句对数;

$d=N-a-b-c$: 不包含英语单元 ep 和汉语单元 cp 的句对数;

其中,N 表示语料的规模,即双语句对的总数.

(1) 点式互信息(PMI)

$$PMI(ep,cp)=\log\frac{N\times freq(ep,cp)}{freq(ep)\times freq(cp)}=\frac{N\times a}{(a+b)\times(a+c)} \tag{1}$$

(2) Dice 系数(DICE)

$$Dice(ep,cp)=\frac{2\,freq(ep,cp)}{freq(ep)+freq(cp)}=\frac{2a}{(a+b)+(a+c)}=\frac{2a}{2a+b+c} \tag{2}$$

(3) Phi 平方系数(PHI)

$$PHI(ep,cp)=\frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{3}$$

(4) 对数似然比(LLR)

$$LLR(ep,cp)=2[\log L(p_1,a,a+b)+\log L(p_2,c,c+d)-\log L(p,a,a+b)-\log L(p,c,c+d)] \tag{4}$$

其中,$\log L(p,k,n)=k\log(p)+(n-k)\log(1-p)$,$p_1=a/(a+b)$,$p_2=c/(c+d)$,$p=(a+c)/(a+b+c+d)$,$\log(0)=0$.

1.3.2 基于 Web 的间接关联的计算方法

理论上,现有统计模型无法区分直接关联还是间接关联,只要是具有明显统计共现特征的词对,利用已分词的双语平行语料库,第 1.3.1 节的各种关联度计算方法均可计算出它们之间的关联程度.考虑第 1.1 节及第 1.2

节中对间接关联的定义及其与译文消歧任务之间的关系,对于任意$(c_j,t_j)_w$,由于该双语词对是通过中间词(源语言歧义词)w 间接关联的,因此不同于直接关联,译文消歧所关心的(c_j,t_j)词对的每次出现,都应有源语言歧义词 w 的同现.也就是说,译文消歧任务所关心的译文 t_j 与源语言歧义词 w 上下文词汇 c_j 间的间接关联程度,需要考察在源语言歧义词 w 出现的情况下词对(c_j,t_j)之间的关联程度.

完全可以利用双语平行语料库来统计含有词 w 的上下文或句对中(c_j,t_j)的同现,并利用公式(1)~公式(4)来计算间接关联度.但是,由于:1) 大规模精确对齐的双语平行语料库仍是稀缺资源,且为所有语言构建大规模平行语料库的代价非常高;2) 即使建立了这样大规模的双语平行语料库,根据 zipf 定律,仍存在着一部分低频译文很少或者根本不出现的情况.也就是说,相比目前大规模的单语语料库,利用双语平行语料库的方法存在的数据稀疏问题更为严重.因此,这里采用另一种多语词汇知识源——Web 来进行间接关联度的考察、统计与计算.利用 Web 作为知识源[17]的优点如下:

1) Web 是单语、多语公共海量信息资源,比传统单语、多语语料库更容易获得;

2) 基本上对所有语种来说,Web 上的信息均有愈增愈多的趋势且相当迅猛,同时 Web 比传统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;

3) Web 比传统语料库更能体现语料的平衡性,其上不断高速增长的海量数据也能进一步削弱各种利用统计方法进行自然语言处理所必须面临的数据稀疏问题.

可以考虑将整个 Web 视作一个多语语料库,在此基础上探讨如何利用 Web 来考察双语词汇之间的间接关联度.这里,我们没有像 Resnik[12]那样试图判断所挖掘到的双语文档中是否含有并确定出双语句对.毕竟即使在目前如此海量数据的 Web 上,存在质量较高的双语平行句对仍相对较少,且仍然难以满足大规模自然语言处理任务的需求[12].

本文试图摆脱以双语平行语料为知识源的惯有对词汇间同现进行统计的思考方式,而是从考察如何直接利用文档空间或者 Web 双语页混合面(mix-language Web page)空间出发,首先作假设如下:

假设 2. 相对于其他源语言词汇,源语言词汇 w 与其上下文词汇 c 更容易在同一个 Web 页面内同现.

假设 3. 相对于其他目标语词汇,源语言词汇 w 与其目标语译文 t 更容易在同一个 Web 双(多)语混合页面内同现.

让我们对任意源语言词汇 w 以及 w_1 出现在同一个 Web 页面上的概率进行估计.在没有任何可用知识以前,根据最大信息熵原理,我们总可以假设任意这样的词对(w,w_1)出现在同一个 Web 页面上的概率是相同的.现在考虑词对(w,c),由于 c 是 w 的上下文词汇,因此 c 会经常在 w 的上下文中出现,也就会经常在含有 w 的 Web 页面中共现.这样,(w,c)出现在同一个 Web 页面上的概率就会相对其与其他源语言词汇的概率要大.虽然在特定的含有 w 的 Web 页面上常常含有其他与 w 无关的句子以及段落,增大了 w 与非其上下文词汇同现的概率而形成一定的噪音,但是与 Yarowsky[30]所讨论的情况类似,对每一个 w_1,其与 w 的同现是分散的,而 w 与 c 的同现是集中的,因而噪音并不会给假设 2 带来太大的影响.假设 3 的情况与之类似.

仿照在第 1.1 节中对词汇间直接关联以及间接关联的定义,我们可作如下定义:

定义 2. 对双语词对(c,t),在 Web 页面中 c 和 t 的统计同现称为词汇间的 Web 直接关联.其中,c 为 w 的任意上下文词汇,t 为 w 的一个译文.对单语词对也可作类似定义.在定义 2 的基础上,对任意源语言词汇 c 以及任意目标语词汇 e,给出定义 3.

定义 3. 对词对(c,e),如果 c 和 e 分别与源语言词汇集合 W(非空)内的词存在 Web 直接关联关系,则称(c,e)通过中间集 W,Web 间接关联,用 $(c,e)_W^{\mathrm{Web}}$ 表示.考虑最简单的情况,当且仅当集合 W 内只有一个词 w 时,则可称(c,e)通过中间词 w,Web 间接关联,以 $(c,e)_w^{\mathrm{Web}}$ 表示.

下面以在第 1.3.1 节中介绍的现有常用词汇间相关度的计算方法为基础,计算任意两个双语词汇之间的 Web 间接关联度.假设汉语歧义词为 w,对应的一个英语译词为 e,对应的一个汉语上下文词为 c,我们定义 IA 为双语词汇间的间接关联度(indirect association),则 $(c,e)_w^{\mathrm{Web}}$ 的间接关联度可记为 $IA_w^{\mathrm{Web}}(c,e)$.为计算双语词汇间的 Web 间接关联度(可记为 Web_*IA*)我们对 a,b,c,d 及 N 进行重新定义如下:

- $a=freq_w^{Web}(c,e)$:含有 w 且同时包含英语词 e 和汉语词 c 的 Web 页面总数(其中的 Web 页面总数可以通过向搜索引擎发送搜索请求(search query),然后根据搜索引擎返回的 Web 页面中抽取其 Page Counts 信息来得到);
- $b=freq_w^{Web}(e)-freq_w^{Web}(e,c)$:含有 w 且包含英语词 e 但不包含汉语词 c 的 Web 页面总数;
- $c=freq_w^{Web}(c)-freq_w^{Web}(c,e)$:含有 w 且包含汉语词 c 但不包含英语词 e 的 Web 页面总数;
- $d=N-a-b-c$:含有 w 且不包含英语词 e 和汉语词 c 的 Web 页面总数;
- N:互联网上所有含有 w 的 Web 页面总数.

在对同现变量的统计方法进行重新定义以后,可沿用第 1.3.1 节的公式(1)~公式(4)来计算双语词汇间的 Web 间接关联度,但是其中参数由上述定义来确定,且当 $a=0$ 时令各种关联度为 0.这样可以得到 Web_IA_{PMI}, Web_IA_{DICE},Web_IA_{PHI} 及 Web_IA_{LLR},分别对应 Web 间接关联的 4 种方法.

2 消歧决策

根据假设 1,在给定源语言歧义词上下文以及该词目标语译文集合的前提下,得到各个双语词汇间的相应 Web_*IA* 就实现了无指导译文消歧的知识获取,由此出发来确定正确的译文.即译文消歧就是在$(w,C;T)$已经给定的情况下计算 Web_*IA*,并由此来确定合适的译文 $t_i\in T$.可选择该词译文集合中与源语言目标词上下文通过源语言目标词间接相关度最高的译文 t_i,可用公式(5)来确定:

$$t_i=\arg\max_{t_i} IA_w^{Web}(C,t_i) \tag{5}$$

公式(5)中,$IA_w^{Web}(C,t_i)$ 定义为源语言歧义词的上下文词集合 C 与译文 t_i 通过词 w 的间接关联度.其中,$t_i\in T$, $i\in[1,m]$,m 为译文集合 T 中含有译文 t 的总数.根据对 $IA_w^{Web}(C,t_i)$ 的不同计算,就会有不同的决策方法. $IA_w^{Web}(C,t_i)$ 可以由计算 $IA_w^{Web}(c_j,t_i)$ 来得到,其中,$c_j\in C$,$j\in[1,n]$,n 为上下文集合 C 中含有上下文词 c 的总数.这里, $IA_w^{Web}(c_j,t_i)$ 即表示集合 C 中的任意词汇 c_j 与译文 t_i 通过词 w 的 Web 间接关联度.

令 C 在歧义词一定窗口内的词兜集合中进行选取,在基于上下文集合 C 中词汇间相互独立假设的基础上,设计了 3 种根据 Web_*IA* 进行译文消歧决策的方法,如公式(6)~公式(8)所示:

$$t_i=\arg\max_{t_i} IA_w^{Web}(c_j,t_i)\ ,c_j\in C,t_i\in T \tag{6}$$

$$t_i=\arg\max_{t_i} \sum_{j=1}^{n} IA_w^{Web}(c_j,t_i)/n\ ,c_j\in C,t_i\in T \tag{7}$$

$$t_i=\arg\max_{t_i} vote_{c_j}(\max(IA_w^{Web}(c_j,t_i)))\ ,c_j\in C,t_i\in T \tag{8}$$

公式(6)是单特征的决策方法,即求歧义词上下文所有词汇与不同译文词之间的关联度,由其中最大的一对关联度决定;公式(7)是多特征决策方法,即求歧义词上下文所有词汇与不同译文词之间的关联度,取所有关联度平均值最大的那个译文作为译文;公式(8)是多特征投票(voting)的决策方法,即选择所有上下文词汇与不同译文关联度投票支持最大的那个作为正确译文.具体过程如算法 1 所示.

算法 1. 输入:上下文词集合及歧义词的目标语译文集合;输出:正确译文.

Step 1. 令 $VOTE_i=0$;

Step 2. 对 t_i 上下文集合 C 内所有词汇 c_j,做:

Step 2.1. 建立空集合 B_j;

Step 2.2. 对每一个译文 t_i,做:

Step 2.2.1. 计算 Web 间接关联度 $IA_w^{Web}(c_j,t_i)$;

Step 2.2.2. 将 $IA_w^{Web}(c_j,t_i)$ 放入到集合 B_j;

Step 2.3. 找集合 B_j 中最大的 $IA_w^{Web}(c_j,t_i)$, 即求 MAX($IA_w^{Web}(c_j,t_i)$);

Step 2.4. 对 MAX($IA_w^{Web}(c_j,t_i)$)所对应的 i,做 $VOTE_i$++;

Step 3. 找最大的 $VOTE_i$,即求 MAX($VOTE_i$);

Step 4. MAX($VOTE_i$)对应的 t_i 即为正确译文.

3 实 验

3.1 评测语料、标准与baseline

利用 ACL2007 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(所有词在后面的表 3 中详细列出),语料由训练语料(由于本文方法是完全无指导的方法,因此没有利用任何训练语料,而是对其测试语料直接进行测试)以及测试语料两个部分组成,见表 1.同时,采用其提供的标准评测工具及相应评价指标 P_{mir} 与 P_{mar}(micro average accuracy 与 macro average accuracy)如公式(9)所示:

$$p_{\text{mir}} = \sum_{i=1}^{N} m_i \Big/ \sum_{i=1}^{N} n_i,\ p_{\text{mar}} = \sum_{i=1}^{N} p_i \Big/ N,\ PMI_{VOTE}^{5} \tag{9}$$

其中,N 为所有的目标词数(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.

Table 1 Basics of gold standard dataset

表 1 标准评测语料情况

	Average number of meanings	Number of training instances	Number of testing instances
19 nouns	2.45	1 019	364
20 verbs	3.57	1 667	571

实验采用 3 个 Baseline,分别为:

1) BL_MFS,即选取测试集答案内的测试实例最常用词义(most frequent sense)的结果,由标准测试集直接给出;
2) TorMD[31],TorMD 系统为多伦多大学参加 SemEval-2007 评测的无指导系统,获得了 SemEval-2007 Task #5 评测第一名(P_{mar}=43.1%);
3) HIT[23],该系统是哈尔滨工业大学参加 SemEval-2007 Task #5 评测的无指导系统.该方法考虑的是汉语上下文与英语上下文(由源语言翻译得来)之间的 Web 同现关系.

3.2 搜索引擎选取及其他参数设置

目前可用的搜索引擎有多个,如 Google(www.google.com),Yahoo(www.yahoo.com),MSN(www.msn.com),百度(www.baidu.com),Altavista(www.altavista.com)等.Keller 及 Lapata[32]在 2003 年比较了 Google 以及 Altavista 这两个搜索引擎上的 2-gram 的 Page Counts,发现它们之间的区别基本可以忽略.Rosso 等人[21]在对名词进行消歧时比较了 MSN,Google 以及 Altavista 这 3 个搜索引擎对消歧效果的影响,结果发现,其对消歧精度基本没有影响.Liu 等人[23]在 2007 年利用百度以及 Google 这两个搜索引擎进行汉英词汇消歧任务的考察,发现其对消歧最终结果影响很小,且百度略优于 Google.于是,本文利用百度作为实验搜索引擎.

对测试语料中所有 935 个例句分别进行上下文集合的选取,利用测试句对应的词兜(words bags),采用窗口大小为(–1,+1)到(–9,+9)的词兜内所有实词是特征词的一系列集合作为目标词在该例句的上下文集合系列.相关度计算方法采用第 1.3.2 节中介绍的 Web_IA_{PMI},Web_IA_{DICE},Web_IA_{PHI} 及 Web_IA_{LLR} 这 4 种.对于每一个例句,在消歧决策时利用公式(6)~公式(8),分别以 MAX,PAR 以及 VOTE 表示.这样,总共需要进行 3×4×5=60 组实验.

3.3 实验结果与讨论

实验结果见表 2.可知,基于双语 Web 词汇间接关联(Web_IA)的方法均取得了不错的效果.各个系统性能基本都超过了可比较系统 HIT,且无论是 MAX,PAR 及 VOTE 方法的最好结果均超过了之前最好的系统 TorMD,分别为 LLR_{MAX}^{3},高出 0.4%、PMI_{PAR}^{5},高出 0.7%、PMI_{VOTE}^{5},高出 1.3%,均为 P_{mar} 绝对性能提高值(各式主

体表示采用的关联计算方法,上标表示窗口大小,下标表示采用的决策方法).

Table 2 Experimental results of three methods

表 2 3 种方法的实验结果

Evaluation criterion	Window size	$DICE_{MAX/PAR/VOTE}$	$LLR_{MAX/PAR/VOTE}$	$PHI_{MAX/PAR/VOTE}$	$PMI_{MAX/PAR/VOTE}$
Micro average accuracy	(−1,+1)	0.349/0.354/0.340	0.362/0.355/0.347	0.360/0.358/0.350	0.360/0.372/0.369
	(−3,+3)	0.359/0.344/0.337	**0.380**/0.357/0.364	0.374/0.357/0.363	0.362/0.355/0.369
	(−5,+5)	0.356/0.343/0.337	0.374/0.349/0.363	0.365/0.350/0.369	0.346/**0.375**/**0.379**
	(−7,+7)	0.355/0.348/0.340	0.368/0.348/0.351	0.365/0.353/0.359	0.346/0.359/0.368
	(−9,+9)	0.369/0.349/0.339	0.367/0.349/0.350	0.364/0.355/0.363	0.352/0.364/0.367
Macro average accuracy	(−1,+1)	0.404/0.411/0.392	0.412/0.413/0.400	0.411/0.412/0.399	0.430/0.437/0.427
	(−3,+3)	0.413/0.401/0.396	**0.435**/0.410/0.418	0.430/0.411/0.415	0.426/0.414/0.426
	(−5,+5)	0.407/0.400/0.394	0.430/0.404/0.415	0.417/0.403/0.420	0.404/**0.438**/**0.444**
	(−7,+7)	0.408/0.402/0.395	0.424/0.402/0.405	0.419/0.407/0.415	0.402/0.423/0.442
	(−9,+9)	0.415/0.405/0.393	0.423/0.406/0.406	0.419/0.411/0.421	0.414/0.429/0.434
Baseline	TorMD	HIT	**BL_MFS**		
Micro average accuracy	**0.375**	0.337	0.405		
Macro average accuracy	**0.431**	0.396	0.462		

为考察 Web_*IA* 模型性能最佳的 VOTE 方法性能随上下文窗口大小变化的影响,我们绘制了图 3,同时令 TorMD 系统的值不随上下文窗口变化而变化作为参照,由于 P_{min} 的表现与 P_{mar} 类似,因此只考虑了 P_{mar} 的情况.

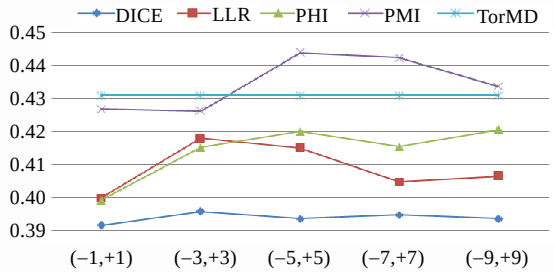

Fig.3 Comparison of Web_*IA* model-VOTE P_{mar} performance with different context windows size

图 3 Web_IA 模型的 VOTE 方法 P_{mar} 性能随上下文窗口性能变化的情况

在如图 3 所示的 VOTE 的决策方法中,各相关度计算方法的性能表现比较一致,均随上下文的扩大而增长到最高点,然后随上下文窗口的扩大性能下降并趋于平缓.其中,DICE 及 LLR 的最好性能对应为上下文窗口(−3,+3),PMI 与 PHI 计算方法的最好性能所对应的上下文窗口是(−5,+5).

由实验结果分析及错误分析可知,当所选取词兜窗口过小时,得不到足够的有效间接关联信息,因而性能得不到保障;而逐渐增大窗口虽然能够增加可为译文消歧提供支持的信息,但是这些信息基本上随着距离中心词汇的渐远而变弱,甚至在一定程度上成为干扰决策的消歧噪音,特别是在远端含有歧义的词句中.因此,在下一步工作中可以考虑过滤掉远端或者句中部分歧义词以尽量消除干扰.

将 BL_MFS,HIT,TorMD 以及基于双语词汇 Web 间接关联度表现最好的 Web_*IA*(PMI^5_{VOTE})这 4 个系统对 40 个词的消歧精度结果(P_{mar} 值,P_{mar} 值规律与之类似)整理在表 3 中.在表 3 各列各系统的 P_{mar} 值中,左侧为 19 个名词的结果,右侧为 21 个动词的结果,所有动词及相关结果均用斜体表示以示区分.最后一行是 Web_*IA*(PMI^5_{VOTE})方法对 HIT 系统以及 TorMd 系统名词和动词性能分别提升的百分比.表中,U 表示没有可用数据,粗体表示该词消歧性能的最好结果.

由表 3 可以发现,无论是对名词还是对动词,基于 Web_*IA* 模型方法消歧的性能都优于 HIT 及 TorMD 系统.在对名词的消歧上,系统的性能与 BL_MFS 基本相当,但是在对动词的消歧性能上存在一定差距.各个系统对动词消歧的性能普遍偏低,比较符合动词更难于消歧这个惯常的认识.另外,也受到该测试集合动词平均词义较名词多的客观影响.

Table 3 Detail nouns|verbs results of 3 systems (P_{mar})

表 3 各系统名词|动词实验详细结果(P_{mar})

Nouns	Number of meanings	HIT	TorMD	MFS	PMI^5_{VOTE}	Verbs	Number of meanings	HIT	TorMD	MFS	PMI^5_{VOTE}
本	3	0.320	**0.720**	0.400	0.560	补	3	**0.550**	**0.550**	0.500	0.400
表面	2	0.333	0.556	**0.611**	0.444	成立	3	0.407	**0.481**	0.370	0.296
菜	2	0.632	0.474	0.579	**0.789**	吃	4	0.174	0.174	**0.435**	0.217
长城	3	**0.619**	0.429	0.476	0.381	出	9	0.091	**0.169**	0.130	0.117
单位	2	0.529	**0.706**	0.588	0.588	带	8	0.104	0.119	**0.150**	0.060
道	3	0.222	**0.500**	**0.500**	0.278	动	4	0.300	0.300	**0.500**	0.150
队伍	3	0.364	0.318	0.455	**0.591**	动摇	2	0.438	0.500	**0.625**	**0.625**
儿女	2	0.500	0.5000	**0.500**	0.450	发	5	0.139	0.25	**0.278**	**0.278**
机组	2	0.571	0.643	0.714	**0.786**	赶	3	0.333	0.389	**0.500**	0.278
镜头	2	0.467	0.467	0.533	**0.667**	叫	4	**0.256**	**0.256**	**0.256**	**0.256**
面	3	0.696	0.348	0.435	**0.739**	进	5	0.114	0.250	**0.227**	**0.227**
牌子	2	**0.529**	0.353	0.353	0.353	开通	2	**0.500**	**0.500**	**0.500**	**0.500**
旗帜	3	0.111	0.500	**0.556**	0.444	看	4	0.294	0.294	0.294	**0.500**
气息	2	0.571	**0.857**	0.714	**0.857**	平息	2	0.500	0.375	0.500	**0.625**
气象	2	0.563	0.438	**0.625**	0.438	使	2	0.438	0.563	**0.625**	**0.625**
日子	3	**0.344**	0.281	0.313	0.219	说明	2	**0.556**	0.444	**0.556**	0.444
天地	3	0.440	**0.560**	0.400	0.320	挑	2	0.286	0.143	**0.429**	0.214
眼光	2	0.500	**0.714**	**0.714**	0.357	推翻	2	0.300	0.300	**0.600**	**0.600**
中医	2	0.500	0.438	0.625	**0.750**	望	2	0.462	0.462	**0.769**	0.538
						想	4	0.216	0.216	0.270	**0.432**
						震惊	2	0.571	**0.714**	**0.714**	0.357
Average P_{mar}		0.464	0.516	**0.528**	0.527			0.335	0.355	**0.440**	0.369
Improving performance (%)		6.3	1.1	–0.01	0			3.4	1.4	–0.71	0

虽然本文方法的性能暂没有超过 BL_MFS(所有参加评测的无指导系统均未超过),但是应该注意到,本文实验是为验证方法有效性而进行的实验,尚未包含消歧特征选择过程,也没有利用任何语言学知识及资源,是一种完全无指导的方法,因而有很大的提升空间.而与之对比的之前最好的无指导系统 TorMd 利用了多项语言学资源,如汉英翻译词典(Huang 及 Graff 开发的汉英翻译词典 3.0,http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L2.该词典含 54 000 个词条)、各种英汉双语平行语料库资源包括 LDC-distributed corpora(http://www.ldc.upenn.edu)—Chinese Treebank English Parallel News Text Version1.0 beta2,Chinese English News Magazine Parallel Text,Chinese News Translation Part 1 及 Hong Kong Parallel Text 等,且对测试语料中给定的英语译文还进行了利用人工将译文映射到英语语义类的工作.

4 结束语

由于双语词汇 Web 间接关联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,该方法性能超过了该评测任务上最好的无指导消歧系统,与绝对性能相比提高了 1.3%.该方法不需要任何已标注语料,仅需要针对两种语言的搜索引擎及相应的 Web 资源.同时,间接关联度是一种可从双语语料或 Web 中挖掘出的可用于译文消歧的新知识,因此,通过该方法所自动获得的消歧知识可以为其他各类消歧方法提供额外的决策支持,在很大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,适合大规模词义消歧以及译文消歧任务.

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数据集上进行深入实验,同时需要进一步提高精确率.今后的研究工作可从如下两个方面入手:

1) 进行特征选择以及特征优化;

2) 与语言学资源,如 WordNet 及 HowNet 等加以结合.

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. 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, 2007. 207−214.

[5] Véronis J. A study of polysemy judgements and Inter-annotator agreement. In: Proc. of the 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. College Park, 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 43rd 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. [doi: 10.1162/089120103322711578]

[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] Liu PY, Zhao TJ, Yang MY, Li Z. Unsupervised translation disambiguation based on equivalent Pseudo translation model. Journal of Electronics and Information Technology, 2008,30(7):1690−1695 (in Chinese with English abstract).

[17] Kilgarriff A, Grefenstette G. Introduction to the special issue on the Web as corpus. Computational Linguistics, 2003,29(3): 333−348. [doi: 10.1162/089120103322711569]

[18] 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.

[19] Mihalcea R, Moldovan DI. Word sense disambiguation based on semantic density. In: Harabagiu S, ed. Proc. of the COLING-ACL Wordshop on Usage of WordNet in Natural Language Processing. San Francisco: Morgan Kaufmann Publishers, 1998. 16−22.

[20] 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.

[21] Rosso P, Montes-y-Gómez M, Buscaldi D, Pancardo-Rodríguez 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.

[22] Yang CY. Word sense disambiguation using semantic relatedness measurement. Journal of Zhejiang University (SCIENCE A), 2006,7(10):1609−1625. [doi: 10.1631/jzus.2006.A1609]

[23] 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.

[24] Melamed ID. Automatic construction of clean broad-coverage translation lexicons. In: Proc. of the 2nd Conf. of the Association for Machine Translation in the Americas. Montreal, 1996. 125−134. http://www.cs.cmu.edu/afs/cs.cmu.edu/user/alavie/MT/11-734/Papers/AMTA-1996-Melamed.pdf

[25] Yarowsky D. One sense per collocation. In: Proc. of the ARPA Human Language Technology Workshop. Princeton, 1993. 266−271. http://acl.ldc.upenn.edu/H/H93/H93-1052.pdf

[26] Church KW, Hanks P. Word association norms, mutual information and lexicography. In: Proc. of the 27th Annual Conf. of the Association of Computational Linguistics. 1989. 76−83. http://acl.ldc.upenn.edu/P/P89/P89-1010.pdf

[27] Smadja F, McKeown KR, Hatzivassiloglou V. Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 1996,22(1):1−38.

[28] Gale WA, Church KW. Identifying word correspondences in parallel texts. In: Proc. of the 4th DARPA Workshop on Speech and Natual Language. 1991. 152−157. http://acl.ldc.upenn.edu/H/H91/H91-1026.pdf

[29] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993,19(1):61−74.

[30] Yarowsky D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proc. of the Int'l Conf. on Computational Linguistics (COLING). 1992. 454−460. http://acl.ldc.upenn.edu/C/C92/C92-2070.pdf

[31] Mohammad S, Hirst G, Resnik P. TOR, TORMD: Distributional profiles of concepts for unsupervised word sense disambiguation. In: Agirre E, Marquez L, Wicentowski R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval-2007). Madison: Omni Press, 2007. 326−333.

[32] Keller F, Lapata M. Using the Web to obtain frequencies for unseen bigrams. Computational Linguistics, 2003,29(3):459−484. [doi: 10.1162/089120103322711604]

附中文参考文献:

[16] 刘鹏远,赵铁军,杨沐昀,李壮.基于等价伪译词模型的无指导译文消歧技术研究.电子与信息学报,2008,30(7):1690−1695.

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.