高技术通讯 2010 年 第 20 卷 第 4 期:349 ~ 354

doi:10.3772/j.issn.1002-0470.2010.04.004

基于挖掘 Web 双语词汇关联度的无指导译文消歧①

刘鹏远② 赵铁军*

(北京大学信息科学与技术学院计算语言学研究所 北京 100871)

(*哈尔滨工业大学计算机科学与技术学院 哈尔滨 150090)

摘 要 为缓解译文消歧任务中消歧知识获取困难及数据稀疏问题,提出了一种基于 Web 的挖掘双语词汇相关关系的无指导译文消歧方法。该方法将双语词汇在语料库中的间接相关拓展到 Web,提出了基于 Web 的双语词汇间接相关模型,在此基础上又提出了一种基于 Web 的双语词汇相关度的消歧方法,通过构造不同 queries 并利用搜索引擎抽取返回页面的 page counts,最后利用点式互信息来计算词汇间的相关度并用于消歧决策。该方法最好性能($P_{mar}=0.464$)超过了国际语义评测 Semeval-2007 的 Task #5 上可比较的最好无指导系统 TorMd。

关键词 无指导译文消歧,双语词汇相关,页面计数,间接相关,基于 Web

0 引 言

确定歧义词在特定上下文中的特定词义(word sense disambiguation, WSD)或者确定歧义词的目标语译文(word translation disambiguation, 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自然语言处理及计算语言学研究领域中的热点及难点。为解决消歧缺乏足够的已标注语料及相应的语义知识这一问题,除用传统的基于词典的各类方法外,还有三种主要的基于统计的研究路线。一种是利用种子语料以及各种半无指导方法进行词义消歧[1-3],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的引入越来越强的噪音。另一种是通过自动获取语义标注实例来进行无指导消歧的方法如平行语料法[4,5]和单语语料以及语义词典的方法[6-9]。平行语料法的问题在于大规模精确对齐平行语料的获取与加工以及两种语言不同语义译文对应的问题。单语语料及词典的方法主要问题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[10]。第三种就是本文所利用的根据 Web 页面计数(page count)的消歧方法[11-14]。本文方法与以往该类方法主要区别在于:(1)是面向跨语言的译文消歧而非单语范畴内的词义消歧;(2)在进行基于 Web 的双语词汇相关的计算中,利用了双语混合 Web 页面(mix-language web Page, MLP),并取得了很好的结果;(3)没有利用其他任何语义资源,是一种完全无指导的(fully unsupervised)方法。

本文直接利用 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,计算歧义词的源语言上下文词汇与译文词汇之间的相关度,选择具有最大相关度的译文作为消歧结果,这样就减轻了人工标注语料负担及知识获取的困难。在国际语义评测 SemEval-2007 中的 Task # 5 测试集上的测试表明,该方法超过了该评测任务无指导方法的最好系统。

1 基于双语词汇相关的消歧模型

1.1 间接相关

间接相关(indirect association, IA)最早是由 Melamed[15]在研究自动构建翻译词典时提出的。如图 1所示,图中显示出了一个双语平行句对,行(1)与行(2)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

① 973 计划(2004CB318102),国家自然科学基金(60903063)和中国博士后科学基金(20090450007)资助项目。

② 男,1974 年生,博士,讲师;研究方向:词义消歧,机器翻译,计算语言学;联系人,E-mail:liupengyuan@pku.edu.cn

(收稿日期:2008-12-12)

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为直接关联或直接相关(direct association, 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，如存在搭配、复合词等各种原因。现假设 c_{k+1} 经常出现在 c_k 的上下文中，则统计模型很容易将 (e_k, c_{k+1}) 误认为是翻译候选，两者之间的相关性随着它们分别与 c_k 之间的相关性的增加而增加。相对于直接共现，这种现象被称为间接共现也就是间接相关。

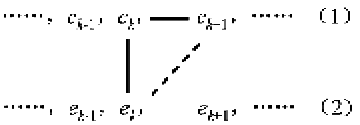

图 1 直接相关与间接相关

双语词汇的间接相关可定义如下：

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 1 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 利用间接相关的消歧模型

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)。

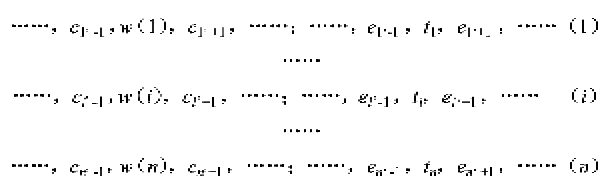

图 2 双语平行句对与译文消歧

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 $w(i)$ 表示源语言目标词且其语义为第 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。由直接相关的概念可知 $w(i)$ 与 t_i 直接相关。

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。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，直接相关。因此，由间接相关的定义可知 C_i 中的词与 t_i 间接相关。

将汉语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 $w(i)$ 所在的上下文 $C_i + E_i$ 中，同时 C_i 中的词与 t_i 是间接相关的关系。由于一般实际译文消歧任务中不存在 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相关度决定。

假设 1 将译文消歧问题视为译文与源语言目标词上下文间接相关度的问题。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)与其上下文词汇间接相关，利用这种间接相关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 基于 Web 的双语词汇间接相关

基于间接相关的消歧模型需要大规模双语平行语料来统计双语词汇的间接相关度，这样就不可避免地面临利用双语平行语料消歧方法的各种问题。若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，而从考察如何直接利用含有中间词 w 的文档空间及双语混合 Web 页面空间出发，来统计双语词汇的同现关系及间接相关度计算，则首先作假设如下：

假设 2：相对于其它源语言词汇，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：相对于其它目标语词汇，源语言词汇 w 与其目标语译文 t 更容易在同一个双语混合 Web 页面(MLP)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大。虽然在特定含有 w 的 Web 页面上常常含有其他与 w 无关的句子以及段落,因而会增大 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[16] 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,因而噪音并不会对假设 2 造成太大的影响。假设 3 的情况与之类似。

将双语词汇在双语平行句对中的同现转换为在含有中间词 w 的 MLP 中的同现,以此为基础,就可在无大规模平行语料库的情况下进行双语词汇间的间接相关度计算并进行译文消歧。

1.4 基于 Web 的双语词汇页面相关

双语词汇之间存在以含有中间词 w 双语混合 Web 页面为中介的间接联系,以此来计算双语词汇之间的间接相关,已经基本减弱了对双语平行语料库等知识的过分依赖。但是由于其思想起源仍来源于对双语平行语语料库的分析,因此该方法在对消歧知识的获取及利用上,不可避免地仍带有一定局限性。由假设 2 及假设 3 为基础的基于 Web 的双语词汇相关,需要在歧义词 w 出现的情况下计算源语言上下文词汇与目标语译文之间的同现。那么,是不是在没有歧义词 w 出现的情况下,MLP 中源语言上下文词汇与目标语译文之间的同现就是偶然的、少数的呢?

换一个角度,以 Web 页面为中心来看待词义/译文消歧问题,若在同一个 Web 页面上词汇(无论是单语还是双语)的同现不仅仅与中间词 w 相关,而是与这个 Web 页面本身的内容有关,或者说是由于特定 Web 页面的内容或者语义才决定了在该页面上词汇的同现。这样,可以利用所有 MLP,摆脱了该页面上歧义词 w 必须出现的限制,扩大了可用知识源的范围。

给定一个巨大的网页空间(至少 10^{10}个网页,即 Google 以及 Baidu 均自称已索引的网页数目),确实存在任意两个双语词汇偶然出现在同一个 MLP 上的可能,但是我们更应该考虑是否这两个词汇之间因为存在着某种语义联系而出现在同一个 MLP 上的可能性。做假设如下:

假设 4::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 5::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 以及词汇之间相关度的方法来考察其上双语词汇之间的语义相关强弱,也就是基于 Web 的双语词汇相关度来进行译文消歧。

2 基于 Web 的双语词汇相关度计算及译文消歧决策

2.1 相关度计算

采用点式互信息[17](point-wise mutual information, PMI)来计算基于 Web 的双语词汇之间的相关程度(Web-based bilingual relatedness, WBR)(初步试验的结果表明,本文方法采用点式互信息最优,限于篇幅,没有采用 Dice 系数等相关度方法如做对比)如式所示。

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

其中 a,b,c 分别表示:同时含英语词 ep 和汉语词 cp 的 Web 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。

由于基于 Web 的双语词汇间接相关及页面相关的方法均利用 Web 为知识源,则式(1)中的各个参数的意义以及计算方法就需要做出相应的改变。

可对式(1)中的 a,b,c 及 N 进行基于 Web 页面的适应性改造,如式

$$\begin{aligned} a &= freq_w(cp,ep), b = freq_w(cp) - freq_w(cp,ep), c = freq_w(ep) - freq_w(cp,ep)\\ a &= freq(cp,ep), b = freq(cp) - freq(cp,ep), c = freq(ep) - freq(cp,ep) \end{aligned} \quad (3)$$

所示。改造之后,式(1)-(3)可以计算任意双语词汇对(ep, cp)间基于 Web 的相关关系。

式(1)与式(2)计算基于 Web 的双语词汇间接相关,式(1)与式(3)计算基于 Web 的双语词汇页面相关,其区别在于:式(2)中各系数的 Web 页面空间限

于搜索引擎索引的所有含歧义词 w 的 Web 页面 P_w，总数 $N = freq(P_w)$，而式(3)中各系数的 Web 页面空间是搜索引擎索引的所有页面，总数 N 约为 10^{10}。

可根据不同搜索引擎的搜索语法来构建相应的 queries，这些页面总数通常可以由任意一个搜索引擎返回的页面中提取相应 page counts 而获得。

2.2 译文消歧决策

得到基于 Web 的双语词汇间的相关度(WBR)以后，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 $c_1, c_2, \ldots, c_n$，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 $t_1, t_2, \ldots, t_m$。则译文消歧的任务就是去从译文集合 $\{t_1, t_2, \ldots, t_m\}$ 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

$$t_i = \text{argmax}_{t_i} \sum_{j=1}^{n} WBR(c_j, t_i)/n, c_j \in C, t_i \in T \quad (4)$$

表示。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 基于 Web 的双语词汇相关度，可利用式(1)－(3)计算。

3 实验及分析

3.1 实验设置与 baseline 系统

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)对本文方法进行评测。该任务共含 40 个歧义词，语料由训练语料以及测试语料 2 个部分组成。利用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)：

$$p_{mir} = \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i, p_{mar} = \sum_{i=1}^{N} p_i / N \quad (5)$$

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，$p_i = m_i/n_i$。由于是无指导的方法，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。

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。Liu，等[18]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。Keller 及 Lapata[19]在 2003 年比较了 Google 以及 Altavista 这两个搜索引擎上的 2-gram 的 page counts，发现他们之间的区别基本可以忽略。Rosso 等[13]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于是本文利用百度作为实验搜索引擎。词模型选取了以目标歧义词为中心的词袋窗口($\pm1, \pm3, \pm5, \pm7, \pm9$)作为上下文。

作为对比，实验选取了 3 个 baseline 系统，分别为：

(1) TorMd[20]，该系统为多伦多大学参加 SemEval2007 评测的无指导系统，获得了 Task #5 Multilingual Chinese _ English Lexical Sample Task 评测第一名。该方法提出跨语言概念-词汇分布关联(Cross-lingual Distributional Profiles of a Concepts，CL-DPCs)的概念。TorMD 利用了大规模英汉双语平行语料库，一部汉英翻译词典，此外还进行了利用人工将译文映射到英语语义类的工作。

(2)利用 Web 的无指导系统 HIT[18]，该系统是哈尔滨工业大学参加 SemEval2007 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，但由于汉语上下文与英语上下文在 Web 的同现数据较为稀疏，因此效果并不理想。

(3)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。

3.2 试验结果及分析

试验结果如表 1 所示，其中 WBR _ I 与 WBR _ P 分别代表利用基于 Web 的双语词汇间接相关及页面相关的方法。可以看出，无论是间接相关还是扩展后的利用双语词汇页面相关进行译文消歧方法的最优性能均超过了该项评测任务上最好的无指导系统 TorMd，且 WBR _ P 方法的 P_{mar} 性能还超过了最常用词义结果 MFS 的性能。必须提出的一点是，在消歧过程中，本文方法仅仅利用搜索引擎以及源语言上下文词汇，不需要任何其他先验知识，甚至不需要词性标注，是一种完全无指导的方法(fully unsupervised)。

表 1 各系统性能结果比较

方法	WBR_I					WBR_P					HIT	TorMd	MFS
位置	±1	±3	±5	±7	±9	±1	±3	±5	±7	±9			
P_{mir}	0.372	0.355	**0.375**	0.359	0.364	0.331	0.367	0.370	**0.391**	0.388	0.337	**0.375**	0.405
P_{mar}	0.437	0.414	**0.438**	0.423	0.429	0.385	0.424	0.429	**0.464**	0.455	0.396	**0.431**	0.462

图 3 是 WBR_I 与 WBR_P 两种方法在不同窗口设置时的性能对比图,横坐标代表窗口大小(1~5 分别对应 ±1~ ±9),纵坐标代表召回率。可以看出两种方法性能随词袋窗口大小变化而体现不同的特点。WBR_I 方法的性能随词袋窗口的增大上下波动,而 WBR_P 方法的性能随着词袋窗口的增大大体是上升的趋势,到词袋窗口为 ±7 时到达最高点然后略微下降。

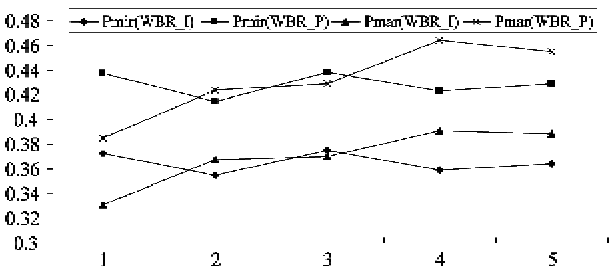

图 3 WBR_I 与 WBR_P 不同窗口设置时的性能对比

这种波动是在窗口逐步扩大的情况下,网页噪音与词汇相关信息对消歧的作用相互影响的结果。由于这两种影响的相互关联性很强,通过错误分析也很难找到一般的结论和规律,但是基本可以说,WBR_P 方法在窗口扩大的情况下得到更多有效的能够帮助消歧的知识。

4 结 论

挖掘 Web 双语词汇相关的无指导译文消歧模型由双语平行语料库出发,在以 Web 页面为中心的前提下规避了双语资源获取与加工的矛盾,直接在 Web 上获取双语词汇消歧知识。该方法实现简单且性能超过了目前可比较的最好无指导系统,证明了译文消歧向 Web 扩展的可行性,也因而缓解了消歧知识获取及数据稀疏问题,具有一定的理论和应用价值。

但该方法也存在如含有偶然同现词汇网页的噪音、无法对同一页面上同现词汇次数的统计及难以分析各个页面内的词汇关系等一系列问题。下一步研究工作可围绕以下两点进行:

(1) 研究如何与已有语义资源结合,进一步提高模型性能;

(2) 研究如何引入更多的不同消歧特征如 N-gram 等,进行更多种特征的消歧决策。

参考文献

[1] Li H, Li C. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 2004, 20(4): 563-596

[2] Yarowsky D. Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA, 1994. 88-95

[3] Niu Z Y, Ji D H, Tan C L, et al. Word sense disambiguation using label propagation based semi-Supervised Learning. In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, USA, 2005. 395-402

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

[5] Ng H T, Wang B, Chan Y S. Exploiting parallel texts for word sense disambiguation: an empirical study. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003. 455-462

[6] Leacock M C, Miller G A. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 1998, 24: 147-165

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

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

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

[10] Martinez D, Agirre E, Wang X L. Word relatives in context for word sense disambiguation. In: Proceedings of the 2006 Australasian Language Technology Workshop (ALTW 2006), Sydney, Australia, 2006. 42-50

[11] Mihalcea R, Moldovan D I. Word sense disambiguation based on semantic density. In: Proceedings of COLING-ACL Wordshop on Usage of WordNet in Natural Language Processing, Montreal, Canada, 1998. 16-22

[12] Turney P D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the 12th European Conference on Machine Learning. Berlin: Springer-Verlag, 2001. 491-502

[13] Rosso P, Montes-y-G'omez M, Buscaldi D, et al. Two web-based approaches for Noun Sense Disambiguation. In: Proceedings of the International Conference on Compute Linguistics and Intelligent Text Processing. Mexico: Springer Verlag, 2005. 261-273

[14] Yang C Y. Word sense disambiguation using semantic relatedness measurement. *Journal of Zhejiang University SCIENCE*, 2006, 7(100): 1609-1625

[15] Melamed I D. Automatic construction of clean broad-coverage translation lexicons. In: Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada, 1996. 125-134

[16] Yarowsky D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proceedings of the International Conference on Computational Linguistics (COLING), Nantes, France, 1992. 454-460

[17] Church K W, Hanks P. Word association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, Vancouver, Canoda, 1989. 76-83

[18] Liu P Y, Zhao T J, Yang M Y. HIT-WSD: using search engine for multilingual chinese-english lexical sample task. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 2007. 169-172

[19] Keller F, Lapata M. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 2003, 29(3): 459-484

[20] Mohammad S, Hirst G, Resnik P. TOR, TORMD: distribtional profiles of concepts for unsupervised word sense disambiguation. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguis, Prague, Czech Republic, 2007. 326-333

Unsupervised translation disambiguation based on mining Web relatedness of bilingual words

Liu Pengyuan, Zhao Tiejun*

(Institute of Computational Linguistics (ICL), Peking University, Beijing 100871)

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

Abstract

This paper presents an unsupervised method by mining Web relatedness of bilingual words. It intends to solve the problem of knowledge acquisition and data sparse in translation disambiguation. By introducing an indirect association model of bilingual words first, this paper expands it to bilingual web page. It goes a step further to a bilingual Web relatedness which centers around Web pages. It computes point-wise mutual information between words as relatedness and makes disambiguation by constructing different queries and extracting Web page counts through search engine. This method achieves the best performance. It outperforms the best unsupervised system TorMd on Semeval-2007 Task #5 and gets the state-of-the-art results ($P_{mar}=0.464$).

Key words: unsupervised translation disambiguation, bilingual word relatedness, page count, indirect association, web based