北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

基于话题分布相似度的无监督评论词消歧方法

郭瑛媚　史晓东†　陈毅东　高燕

厦门大学信息科学与技术学院，厦门　361005；† 通信作者, E-mail: mandel@xmu.edu.cn

摘要　基于话题信息、词的位置关系和互信息等特征，提出一种无监督的跨语言词义消歧算法。该算法仅利用在线词典和 web 搜索引擎，通过上下文信息选择评论句中多义评论词的词义。实验结果表明，所提出的词义消歧算法具有较高准确率，对于具有较多候选词义的评论词仍能表现出较好的性能。

关键词　评论词消歧；话题模型；无监督

中图分类号　TP391.1

Unsupervised Opinion Word Disambiguation Based on Topic Distribution Similarity

GUO Yingmei, SHI Xiaodong†, CHEN Yidong, GAO Yan

School of Information Science and Engineering, Xiamen University, Xiamen 361005
† Corresponding author, E-mail: mandel@xmu.edu.cn

Abstract　The authors present an automatic method for choosing the correct sense of a polysemous word by using topic information, distance and mutual information of words. The only resources used in the method are an online dictionary and a Web Search Engine. The sense of ambiguous opinion word can be broadly described from words in the context. Experiments show that new approach could achieve high accuracy, and especially keep superior performance for opinion words with more alternative senses.

Key words　opinion word disambiguation; topic model; unsupervised

互联网已成为人们发布和获取信息的重要平台。在购买商品、选择服务之前，网络上的评论文章对使用者做出决策有重要的指导作用。然而，若评论文章不是用母语所写，则需要对其中重要的评论词进行跨语言词义消歧(选择)。评论词经常具有较多词义，虽然目前的机器翻译软件可以处理整句，甚至整个段落，但对多义词的处理却不够准确，影响了使用者对关键信息的获取。

针对评论文章中评论词的特点，本文提出一种算法，可以利用极少量的资源，仅使用在线词典和搜索引擎，采用无监督的方式自动获取多义评论词在特定上下文中的词义。此算法避免了很多成熟的翻译系统在整句翻译时对多义词处理不当而带来的困扰，使用者能通过本方法迅速获得主要信息。此外，这种方法无需人工参与，极大的降低了人力开销，并且与领域、语言无关，适应性很强。该算法选取了话题信息、词的位置关系和互信息作为基本特征，并加以改进。实验结果表明，选用的特征在评论词词义选择算法中均获得不错的效果。

值得一提的是，在 Klapaftis 等[1]和 Yang[2]的研究中也使用 web 搜索引擎作为消歧语料的来源，对于上下文中的词，考虑距离差异来赋予权值。然而在研究中发现，与待消歧词语义相似的词对其影响很大，但该词与其距离并非最近，在消歧过程中有必要添加语义信息。此外 Cai 等[3]和 Jordan 等[4]均使用话题模型处理过词义消歧问题，前者把消歧看

国家自然科学基金(60573189, 61005052)、国家支撑计划(2012BAH14F03)和福建省自然科学基金(2006J0043)资助
收稿日期: 2012−05−31; 修回日期: 2012−08−28; 网络出版时间: 2012-10-26 17:49
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1749.013.html

成是分类问题，利用话题信息作为朴素贝叶斯模型的一个特征，后者将话题信息直接添加到 WordNet 中。

1 基本思路

评论句子中评论词的翻译有多种方法，本文采用的方式基于一个简单的思路：评论词的词义能被上下文中的词所决定；上下文中不同的词在选择词义时的重要程度不同；按照一定准则为多义评论词的词义项排序，排在最前面的为最终翻译结果。本研究中假设评论词已明确获取。

判断上下文中的词与关键词之间关系的密切程度主要考虑如下 3 个方面的因素：1) 词对间的互信息；2) 词对间的位置关系；3) 词对在话题分布上的相似程度。

为进一步以可量化的形式表达评判准则，可引入具体的特征函数。上面的叙述可形式化地表示如下：

$$\mathrm{SC}(\mathrm{os}) = \sum_{j=1}^{n}(W_j \cdot \prod_{i=1}^{m} F_i(\mathrm{os}, ns_j))。 \quad (1)$$

SC(score 或者 sence choosing)表示多义评论词量化公式；os 与 ns_j 分别代表目标评论词的义项与上下文中词的义项(os: opionion word Sence；ns: neighboring words' Sense)，函数 $F_i(\mathrm{os},\mathrm{ns}_j)$提供了第 i 个特征的分值，计算两个义项间的相关程度，目前都具有 $F_i(\mathrm{os},\mathrm{ns}_j)=\lambda_i \mathrm{e}^{\mu_i f_i(\mathrm{os},\mathrm{ns}_j)}$ 的形式，其中 λ 和 μ 为权重，调节不同特征对模型的影响；W_i 为平衡因子，调节由于上下文中词的义项数不同对模型产生的影响；m 为特征个数；n 为上下文词义项总数。通过上下文中词的各义项的影响，得到评论词的每个翻译项的分值。

系统的总体流程如图 1 所示。首先，将含有多义评论词的句子输入，经分词、去停用词，通过在线词典获得每个词的所有词义项；构建两个集合 OS-set 与 NS-set，分别对应评论词义项集与上下文词义项集；利用两个集合中的元素及 Google 搜索引擎获取计算特征函数所需的语料资源；由式(1)计算关键词每个词义项的值，确定最终翻译结果。

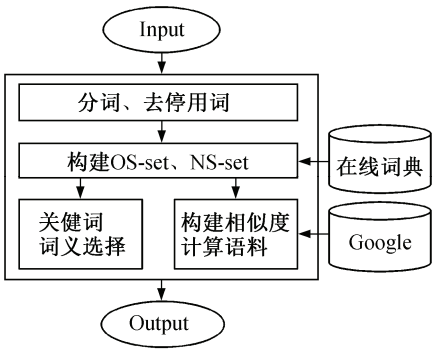

图 1 词义选择算法流程图

Fig. 1 Flow chart of the sence choosing algorithm

2 语料库构建与特征函数设计

2.1 语料库的构建

本研究利用在线词典获取词义项，通过 Google 搜索引擎得到计算特征函数所需的语料资源。首先，对带有评论词的句子进行分词，得到评论词及其上下文单词(neighbors)。然后，去除上下文中停用词，这些噪声信息对评论词词义项选择没有帮助，甚至可能影响关联程度的判断。最后，使用在线词典获得评论词与每个“neighbor”的全部词义项。建立两个集合，包含评论词所有词义项的集合，称为 OS-set；包含“neighbors”全部词义项的集合，称为 NS-set。

下面的例子可更直观地描述上述过程：“お部屋もバスルームも*綺麗*に掃除がされてあり”。

评论词“綺麗”用斜体并带有下划线标出，该词具有 3 个词义项，构成集合 OS-set={漂亮，乾淨，清楚}，os_i 代表其中一个元素。

经分词、去停用词后，得到除评论词外如下一组词：“部屋 / バスルーム / 掃除”，那么 neighbors={部屋，バスルーム，掃除}，$\mathrm{neighbor}_j$ 代表其中一个元素。通过在线词典获取每个元素的全部翻译项：“部屋→房間，屋子”；“バスルーム→浴室”；“掃除→打掃，掃除”。则 NS-set 每个元素是由 neighbors 的翻译项组成的集合，表示为{{房間，屋子}，{浴室}，{打掃，掃除}}，是集合的集合，ns_{jk} 代表 j^{th} 项的 k^{th} 个元素。

使用搜索引擎为计算特征函数的值构建语料资源。首先，提取欲输入搜索引擎的词对。词对中的词分别取自集合 OS-set 与集合 NS-set，并需保证 OS-set 中所有元素与 NS-set 的元素完全两两配对。将每个词对$(OS_i,\ ns_{jk})$输入搜索引擎获得前 200 篇文本片段加入语料集合中。

图 2 清晰地表示了两个词集的构建和词集间词的配对过程。

图 2 描述了例句的 NS-set 与 OS-set 的构建过程，以及两个词集中词之间的完全配对关系，可表示成一个二元组的集合 RS={(漂亮，房間)，(漂亮，

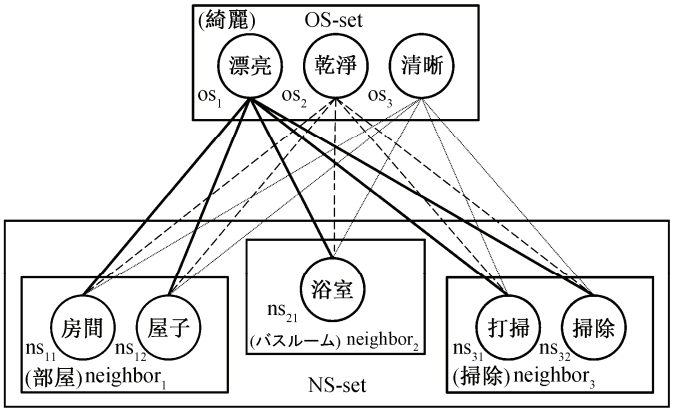

图 2　词集的构建与词集间词的配对过程

Fig. 2　Process of Building the word set and matching two words

屋子), …,(清楚, 满意)}。共可获得 15 个词对, 添加约 3000 个文本片段到语料库中。

2.2　特征函数与权重

在目前的工作中引入 3 个特征: 互信息、词的位置关系以及话题关系。根据本研究的特点, 对互信息公式进行修正, 并添加依存信息构成词的相对位置。

2.2.1　平衡因子

在分析特征前, 首先介绍平衡因子。平衡因子的作用是均衡上下文中的词在词义选择模型中的影响。从式(1)中可以看到, 上下文词的每个词义项都会成为公式中的一个加和项, 具有较多词义项的上下文对公式的影响较大, 可以通过平衡因子来调节。形式化的表示如下:

$$W_{jk} = p(\mathrm{ns}_{jk} / \mathrm{neighbor}_j), \tag{2}$$

表示上下文词选择其自身某个义项的可能性, 其值为上下文中词具有的翻译项个数分之一, 其中 $\mathrm{ns}_{jk} \in$NS-set。比如, 某上下文中的词 ns_j 有 3 个词义项 ns_{j1}, ns_{j2}, ns_{j3}, 那么定义 $p(n_{jk} / \mathrm{neighbor}_j) = 1/3$。

需要注意的是, 在式(1)以及下面的介绍中, 仅把 NS-set 看成是简单元素的集合, 并使用 j 对所有上下文义项计数。

2.2.2　话题信息

话题信息在词义选择上具有重要的作用, 多义词在给定话题下的词义几乎被明确下来。在本研究中使用的 latent dirichlet allocation(LDA)模型[5]是三层贝叶斯生成模型, 通过该模型得到两组条件概率分布: 话题与文档间、词与话题间的条件概率分布。后面一组反应词 w_i 在话题集合上的分布状况。我们认为两个具有相似话题分布的词之间的关系更为密切, 希望可以利用话题分布计算词与词之间的相似程度。

目前有很多衡量两概率分布间差异程度的方法, 本文选取 KL 散度(Kullback-Leibler divergence) [6], 也叫做相对熵(relative entropy):

$$\mathrm{topic\_distribution\_score}(\mathrm{os}, \mathrm{ns}_j) = -\sum_{k=1}^{K} p(z_k \mid \mathrm{os}) \ln\left(\frac{p(z_k \mid os)}{p(z_k \mid ns_j)}\right)。 \tag{3}$$

使用式(3)求得词 os 与 ns_j 在主题上的概率分布 $p(z_{1,\ldots,n} \mid \mathrm{os})$ 和 $p(z_{1\ldots n} \mid \mathrm{ns}_j)$ 的差异程度, 进而判别两个词之间的关联性。构造话题关系特征函数如下:

$$\mathrm{T\_KL}(\mathrm{os}, \mathrm{ns}_j) = \lambda \mathrm{e}^{\mu \cdot \mathrm{topic\_distribution\_score}(\mathrm{os}, \mathrm{ns}_j)}, \tag{4}$$

这里的 topic_distribution_score 是通过 KL 散度计算的两个词之间话题条件分布的差异程度, 值越大, 表明两个词之间的话题分布差异越大, 词之间的关系越不密切。

2.2.3　带有依存分析的词的相对位置信息

在考虑词与词之间影响时, 距离是较常用的因素, Beeferman 等[7]详细讨论了不同位置关系的两个词在关联程度上的差异, 并通过若干组实验的分析表明词与词间的相互影响程度随着距离的增大而指数下降。

$$\mathrm{Dis}(\mathrm{os}, \mathrm{ns}_j) = \lambda \mathrm{e}^{\mu \cdot \mathrm{distance}(\mathrm{os}, \mathrm{ns}_j)} \tag{5}$$

利用极大似然估计得到

$$\lambda = \mu = -\frac{1}{E_{\tilde{p}}[k]} = -\frac{1}{\sum_{k \geqslant 0} k\tilde{p}(k)}, \tag{6}$$

$\tilde{p}(k)$ 为当两个词距离为 k 时发生的概率。

本文的做法与其有较大区别: 本文使用的是进行依存关系的分析后所产生的距离, 是一种相对距

离。

本研究采用 CaboCha 的分析器对句子做依存分析。下面通过一个简单的例子来观察 CaboCha 的分析结果，以及本文对结果的使用方法。

例句："価格について質問てす"，"关于价格，有个问题"。

使用 CaboCha 分析器对以上例句进行依存分析，XML 结果如下所示，是一个三层分析树，根节点为<sentence>，其下为语块<chunk>，叶子节点是词<tok>。在这个分析结果图中，每一个语块有一个主词 head，func 标记非核心词，依附于主词存在。标签 link 表示主词与非核心词之间的关系，如果 link 值为 1 说明两词关系密切；为–1 时表明关系不强。根据这一特点，在分析依存距离时，如果两个词在同一语块中且 link 值为 1，则两词之间边的权重为基础权重 1；如果两个词位于相同语块但 link 值为–1，那么两个词之间边的权重为 1.5；如果两个词位于不同语块，那么两个词之间边的权重为 2。带有依存关系的相对位置特征函数表示如下:

$$D\_\mathrm{Dep}(\mathrm{os},\mathrm{ns}_j)=\lambda e^{\mu\cdot k\cdot \mathrm{distance}(\mathrm{os},\mathrm{ns}_j)}, \tag{7}$$

k 值表示依存信息，取值如上分析。

CaboCha 分析器的 XML 结果:

<sentence>
<chunk id="0" link="1" rel="D" score="0" head="0" func="1">
<tok id="0" read="かかく" base="価格" pos="名詞−一般" ctype="" cform="" ne="O">国債</tok>
<tok id="1" read="ニツイテ" base="について" pos="助詞−格助詞−連語" ctype="" cform="" ne="O">について</tok>
</chunk>
<chunk id="1" link="-1" rel="O" score="0" head="2" func="3">
<tok id="2" read="シツモン" base="質問" pos="名詞−サ変接続" ctype=""cform="" ne="O">質問</tok>
<tok id="3" read="デス" base="です" pos="助動詞" ctype="特殊・デス" cform="基本形" ne="O">です</tok>
<tok id="4" read="。" base="。" pos="記号−句点" ctype="" cform="" ne="O">。</tok>
</chunk>

2.2.4 互信息

目前有很多方法用来计算两个词的关联强度，我们对这些方法进行总结，通过实验和比较，发现尽管在计算语言学中被广泛运用的互信息在直接使用时结果并不理想，但经过调整在本研究中取得了很好的效果。式(8)给出的是逐点互信息计算公式:

$$MI(w_1,w_2)=\log\frac{p(w_1,w_2)}{p(w_1)p(w_2)}, \tag{8}$$

$p(w_1)$ 和 $p(w_2)$ 代表词 w_1 和 w_2 出现的概率。在我们的研究中以句子为一个基本划分单位来计算两个词共同出现的次数，$p(w_1,w_2)$ 表示两词出现在相同句子中的概率。

Smadja 等 [8]指出互信息在计算文本相似关系时存在重要缺陷，既在共现率相同的情况下，出现频率较低的词可以获得更高的互信息量。而本研究恰好避免这个弱点。计算词对的关联强度所使用的数据集是以每个评论词与其上下文词构成词对，输入到 web 搜索引擎中抓取的文本片段的集合。由于计算相似性差异的语料构成方式特殊，使得每个词在数据集中出现的频率比较均匀。

然而，在直接利用互信息计算时并没有取得预想的效果。经分析发现，词对由 OS-set 与 NS-set 中的所有词的两两组合产生，有些词对间没有明显关系。使用关联性过差的词对抽取文本，很容易造成词对共现句子的稀疏，甚至为零，既 $p(\mathrm{os},\mathrm{ns}_j)\to 0$ 时 $MI(\mathrm{os},\mathrm{ns}_j)\to-\infty$。这种负无穷大的结果，给下一步计算翻译项的分数带来了困扰。

另外，在上面的分析中可以看到，调节因子 W 与特征函数 T_KL 和 D_Dep 总为正值，所以当互信息的值出现负数时就会产生问题。考虑最简单的情况，当两个词的 MI 值相同并且都为负值时，原本较大的值在乘以一个负数时变的反而较小。但事实上在 MI 值相同的情况下，其余项的值较大时结果应该较大。

为了解决上述问题，把 MI 映射到指数函数上，得到修正的互信息特征函数:

$$\mathrm{MI\_E}(\mathrm{os},\mathrm{ns}_j)=\lambda \mathrm{e}^{\mu\log\frac{p(\mathrm{os},\mathrm{ns}_j)}{p(\mathrm{os})p(\mathrm{ns}_j)}}。 \tag{9}$$

进行指数修正过的互信息能有效地避免趋近于 0 的概率带来的计算困难，并且避免了互信息值为负时产生的错误。实验结果表明，这种修正能够在很大程度上提升词义选择算法的准确率。与其它的计算关联程度的方法相比，得到的结果也要更好一些。

2.2.5 权重设置

确定了特征后，要为特征分配权重，我们通过

表 1　采用的参数值

Table 1　Parameters

对应特征	λ	μ
话题关系	1	-10
位置特征	$-\frac{1}{\sum_{k\geqslant 0}k\tilde{p}(k)}$	$-\frac{1}{\sum_{k\geqslant 0}k\tilde{p}(k)}$
互信息	1	1

实验得到了一组经验值。表 1 中列出了每个特征函数的参数值, 其中位置特征参数的设定参考 Kullback 等[6]的做法。

3　实验与分析

本实验数据分中文和日文两部分, 均来源于互联网。在实验中逐一添加特征以测试不同特征函数的效果, 并设计与现有的成熟机器翻译系统的对比试验, 以测试本算法的有效性。

3.1　实验数据

中文部分, 将词对输入搜索引擎后返回的前 200 个文本片段, 用以计算互信息以及话题相关性。日文部分, 来自日本乐天旅游网站, 共 956892 篇评论文章, 4341266 句。选择 10 个最常见的多义词进行实验, 并随机抽取 1200 句含有评论词的句子做标记用作测试。

表 2 包含了多义评论词的词义项数、用于测试的实例数、实例的最大、最小以及平均长度。

表 2　标注后的测试集

Table 2　Tagged text set

多以评论词	词义项数	实例数	平均长度	最小长度	最大长度
明るい	2	992	36.5	7	135
甘い	2	808	40.3	7	145
暖かい	2	979	41.6	6	147
丁寧	2	1057	37.9	11	125
冷たい	2	957	44.0	6	174
薄い	2	1041	38.3	6	141
綺麗	3	736	35.2	9	113
きつい	3	755	41.9	10	136
寂しい	3	794	38.7	8	120
厳しい	4	506	45.0	7	141
平均	2.5	862.5	39.9	7.7	137.7

3.2　评估指标

本实验采用出现频率最大的词义作为基准:

$$\mathrm{MFS}=\frac{\#\text{Most Frequent Sense}}{\#\text{test sentence}}。\tag{10}$$

使用处理正确的句子数占总句数的比值作为准确率:

$$\mathrm{Accuracy}=\frac{\#\text{right sense}}{\#\text{test sentence}}。\tag{11}$$

用 10 个测试词正确率的平均值评价系统性能。由于不同多义词标注一致的数据量差别较大, 故采用两种取均值的方法: 宏平均(macro-average)和微平均(micro-average), 来保证评估的客观性。

3.3　实验设计与结果分析

设计两组实验, 一组是逐步添加特征, 来观察不同特征函数对词义选择算法的影响; 另一组是与广泛使用的 Google 在线翻译系统以及权威的 Excite

表 3　添加不同特征的实验结果

Table 3　Result with different features

多义评论词	词义项数	实例数	MFS/%	MI/%	MI_E/%	MI_E+*D*/%	MI_E+*T*/%	MI_E+*D*+*T*/%
明るい	2	992	72.0	86.56	89.42	91.98	92.05	92.74
甘い	2	808	77.6	80.62	84.16	86.08	87.93	88.46
暖かい	2	979	53.2	66.85	79.98	81.83	82.16	86.59
丁寧	2	1057	81.6	50.45	83.92	86.79	87.02	88.62
冷たい	2	957	92.0	93.82	87.88	92.10	91.84	93.31
薄い	2	1041	89.6	87.34	92.22	92.93	93.26	96.15
綺麗	3	736	49.6	73.41	77.04	78.75	79.52	80.20
きつい	3	755	56.8	65.63	72.72	76.49	77.33	78.84
寂しい	3	794	48.8	62.63	61.59	63.52	64.47	65.03
厳しい	4	506	50.8	55.71	54.74	60.26	59.78	63.27
微平均	—	—	67.2	73.27	80.23	82.80	83.26	85.04
宏平均	—	—	72.0	72.30	78.37	81.07	81.54	83.32

表 4　对比实验结果

Table 4　Result of comparison experiment

多义评论词	词义项数	MFS/%	Excite/%	Google/%	MI_E/%	MI_E+*D*+*T*/%
明るい	2	72.0	72.0	70.4	89.42	92.74
甘い	2	77.6	77.2	70.0	84.16	88.46
暖かい	2	53.2	46.8	71.6	79.98	86.59
丁寧	2	81.6	13.2	80.8	83.92	88.62
冷たい	2	92.0	92.0	89.6	87.88	93.31
薄い	2	89.6	89.6	83.2	92.22	96.15
綺麗	3	49.6	47.2	54.8	77.04	80.20
きつい	3	56.8	0	4.8	72.72	78.84
寂しい	3	48.8	15.6	8.0	61.59	65.03
厳しい	4	50.8	50.6	28.0	54.74	63.27
平均	—	67.2	50.4	56.1	80.23	85.04

在线翻译系统的对比实验，用来评估本算法的有效性。实验结果如表 3 和 4 所示。

表 3 中，MI_E 表示改进的 MI，MI_E+*T* 表示加入话题分布相似度，MI_E+*D* 表示加入依存距离，MI_E+*D*+*T* 表示两者共同作用于模型。可以看到，互信息的改进明显改善了词义选择效果，话题信息与依存距离的添加也有较好的表现。

在对照实验中清晰地看到，针对多义评论词的翻译本算法比现有机器翻译系统更好。当词义项较多时，已有系统性能明显下降，几乎全部低于 baseline，而我们的系统虽有下降，但准确率仍稳定在较高水平。该结果表明，此算法稳定、有效，在词义项较多时优势尤其明显。

4　结论

本文提出一种无监督的方法，自动对评论句子中评论词的词义进行选择。有监督方式需要大量人力成本，限制语料的使用规模，而对统计方法来讲，语料规模是影响模型训练效果的重要因素之一。此种方法直接利用庞大的互联网资源，仅用在线词典及搜索引擎便可较好的确定评论句子中多义关键词的词义，为跨语言信息获取提供可靠参考。此方法不受语言种类及语料领域范围限制，资源极易获取，适应性强，适用范围广泛。模型考虑了改进的互信息、带有依存信息的词的位置关系、词的话题关系等因素，实验结果表明该词义选择算法效果较好。本模型还有较大的改进空间。首先，实验语料领域单一，主题范围比较狭窄，话题模型的语义区分表现的还不明显，可进一步扩大实验语料的选取范围。其次，词与词之间的关系信息挖掘比较有限，例如还可找出与评论词具有直接关系的被评论词，通过为其分配较高权重来调整词义选择时的概率偏向。另外，目前模型的权重仅简单的按经验设定，可以考虑采用更加客观的方式训练各特征的权重。

参考文献

[1] Klapaftis I P, Suresh M. Google & wordnet based word sense disambiguation // Proceedings of the 22nd ICML Workshop on Learning & Extenging Ontologies. Bonn, Germany, 2005: 25–27

[2] Yang C Y. Word sense disambiguation using semantic relatedness measurement. Journal of Zhejiang University SCIENCE A, 2006, 7(10): 1609–1625

[3] Cai J F, Lee W S, Teh Y W. NUS-ML: improving word sense disambiguation using topic features // Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic, 2007: 249–252

[4] Jordan B G, Blei D, Zhu Xiaojin. A topic model for word sense disambiguation // Proceedings of the EMNLP-CoNLL. Prague, Czech Republic, 2007: 1024–1033

[5] Blei D, NG A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022

[6] Kullback S, Burnham K P, Laubscher N F, et al. Letter to the editor: the Kullback–Leibler distance. The American Statistician, 1987, 41(4): 340–341

[7] Beeferman D, Bergen A, Lafferty J, et al. A model of lexical attraction and repulsion // Proc 35th Annu Meeting Association for Computational Linguistics. Stroudsburg, PA, USA, 1997: 373 –380

[8] Smadja F, McKeown K. Translating collocations for use in bilingual lexicons // Proceedings of the workshop on Human Language Technology. Plainsboro, NJ, 1994: 152–156