doi: 103969/j. issn. 0490-6756. 2016. 11. 014

基于主题模型的中文词义归纳

高章敏[1,2], 何 祥[1], 刘嘉勇[1,2], 汤殿华[2]
(1. 四川大学电子信息学院, 成都 610065; 2. 保密通信重点实验室, 成都 610041)

摘 要: 词义归纳是在给定包含多义词语料的条件下,识别出多义词词义的过程,通常是采用聚类的方法. 本文提出了基于主题模型的方法来解决中文词义归纳问题,基于主题模型的词义归纳方法关键之处在于使用文档的主题概率分布来推断多义词的词义分布. 实验结果表明,本文方法在测试数据上获得了 77.58% FScore 值.

关键词: 词义归纳; 主题模型; 隐含狄利克雷分布; 词义消歧
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0490-6756(2016)06-1269-04

Chinese word sense induction based on topic model

GAO Zhang-Min[1,2], *HE Xiang*[1], *LIU Jia-Yong*[1,2], *TANG Dian-Hua*[2]
(1. College of Electronics and Information, Sichuan Universiry, Chengdu 610065, China;
2. Science and Technology on communication Security Laboratory, Chengdu, China)

Abstract: Sense Induction is the process of identifying the word sense given its context, often treated as a clustering task. In this paper, the authors present a approach to Chinese word sense induction which is based on topic modeling. Key to the methodology in this paper is the use of probabilistic assignment of topics distributions to documents to estimate sense distributions. Experimental results show that the method in the paper could achieve 77.58% scores of Fscore on the development data set.

Keywords: Word sense induction; Topic model; Latent Dirichlet Allocation; Word sense disambiguation

1 引 言

词义归纳(Word sense induction,简称 WSI)的任务就是要自动识别出未标注语料中多义词的词义,也就是给定多义词以及包含该词语的上下文集合,根据上下文所包含的信息,使用无监督的学习方法自动获取该多义词在语料中的词义. 和词义消歧(Word Sense Disambiguation,简称 WSD)不同,词义消歧是有监督的学习,它需要标注语料来进行训练,然而标注语料又较难获得. 词

义归纳克服了词义消歧的这个缺点,它自动从未标注语料中获取多义词的词义,是无监督学习,无需利用已标注的语料来进行训练. 词义归纳自动从未标注语料中获取正确的词义,这能改善很多自然语言处理任务的表现,例如信息检索[1],信息提取[2]及机器翻译[3]等.

大多数词义归纳算法基于分布式假设[4],该假设认为具有相近词义的词语总是出现在相似的上下文中[5]. 目前词义归纳的方法主要是基于特征向量的聚类方法. 基于特征向量聚类是通过特征选择算法来构建特征向量空间,然后使用各种聚类算法来进行词义归纳. 聚类算法主要涉及到 K-means, Agglomerative Clustering, Information Bottleneck 等[6-8]. Pantel 和 Lin 根据共现词词向量的相似度

收稿日期: 2015-12-25
基金项目: 保密通信重点实验室基金 (9140C110401140C11053)
作者简介: 高章敏(1991-), 男, 硕士生, 研究方向为自然语言处理. E-mail: 2014222050138@stu. scu. edu. cn.
通讯作者: 刘嘉勇. E-mail: ljy@scu. edu. cn

对多义词进行聚类,提出了 Clustering By Committee(简称 CBC)算法[9]. Agirre 和 Klapaftis 对基于相似度矩阵的词义归纳方法进行了研究[10,11],取目标多义词的上下文作为结点,上下文之间的相似度来度量边权值,取得了较好结果. 最后基于统计模型的方法在词义归纳中也有应用,Brody 等人首次将 LDA 主题模型运用到英文词义归纳中[12],取得了较好的成绩.

相对于国外语言,中文词义归纳的研究起步较晚. 中文词义归纳的研究中,使用最多的是基于特征向量的方法[13]. 特征选择主要是使用词性、上下文等作为特征,并对两者进行组合. 聚类算法主要涉及到 K-means、The Sequential Information Bottlenech(简称 sIB)、Expectation Maximization(简称 EM)等. Zhang 等人使用单个汉字、词语、二元组作为特征,并对其进行组合,构建特征向量,使用 K-means、EM 等算法进行词义归纳,实验取得了 64.67%的 FScore 值[14]. Xu 等人使用除停用词外的所有词语来构建特征向量,基于该特征空间构造相似度矩阵,并使用 k-means、Chinese Whispers 聚类算法进行词义归纳,最后的实验结果 FScore 分别取得了 74.93%和 63.93%[15].

本文使用潜在狄利克雷分布(Latent Dirichlet Allocation,简称 LDA)主题模型[16]推导中文多义词词义. 在自然语言处理当中,主题(Topics)可以看成是一系列与主题相关词项的概率分布,主题模型是一种无监督学习算法,它自动推导语料的主题,并为文档计算主题概率分布. 其中,语料主题是相关词项的多项式分布,主题概率分布在语料主题上的一个多项式概率分布. 基于主题模型的词义归纳方法将多义词的上下文词语看成是多项式词义分布的采样,将词义归纳放在概率模型中. 本文使用 LDA 来对多义词上下文词语进行采样. 同时提取了多种词汇搭配作为特征,如全局词汇搭配,局部词汇搭配.

2 主题模型词义归纳

在主题模型中,假设语料包括多个主题,主题模型为语料库中每篇文档计算基于这些主题的概率分布,文档中的词汇都是通过"以一定概率从文档主题概率分布中选择了某一个主题,并从这个主题中以一定概率选择某个词语"这样一个过程得到的.

2.1 特征提取

由于中文文本没有类似英文空格之类的显示标志来标示词的边界,所以在提取文本特征之前需要进行分词处理. 待文本分词后,去掉文本中的停用词,提取目标多义词上下文的词语搭配,依据词袋模型建立特征向量. 词袋(bag of words)假设是主题模型中的一个重要假设,即一篇文档内的单词可以交换次序而不影响模型的训练结果[17].

本文提取了两种词汇搭配作为特征:全局词汇搭配和局部词汇搭配.

多义词某种词义的使用与该多义词所在的上下文有关,在抽取全局词汇搭配时,将多义词上下文的所有词语抽取出来,作为全局词汇搭配.

考虑到多义词附近的词与多义词词义联系更大,更能表达多义词的词义,将多义词局部上下文的词语搭配抽取出作为局部词汇搭配,局部词汇搭配指的是与多义词共现的有实际意义的词汇. 例如,包含多义词"暗淡"的例句:"经济增长暗淡也前所未有地激起欧盟内部对发展模式之争."例句中用到了"经济"、"增长"等一些具有实际意义的词汇. 本文在多义词左右开了一个大小各为 3 个词的窗口,来抽取多义词左右各 3 个具有实际意义的词汇作为局部的词汇搭配.

表 1 给出了从包含多义词"暗淡"文本中提取出的全局和局部词汇搭配.

表 1 词汇搭配

Tab. 1 Global collections and local collections

多义词汇	暗淡
句子	经济增长暗淡也前所未有的激起欧盟内部发展模式之争.
分词结果	经济 增长 暗淡 前所未有 激起 欧盟 内部 发展 模式 之争.
全局词汇搭配	经济 增长 暗淡 前所未有 激起 欧盟 内部 发展 模式 之争
窗口±3 的局部词汇搭配	经济#-2 增长#-1 前所未有#1 激起#2 欧盟#3

最后,合并从文本中提取出全局和局部词汇搭配,根据词袋计数模型建立特征向量.

2.2 主题模型

本文使用主题模型对文档的生成过程进行模拟,再通过参数估计得到各个主题. 参数估计可以看成是文档生成过程的逆过程. 通过参数估计,主题模型计算语料主题. 然后基于语料主题,计算语料库中每篇文档的主题概率分布. 对于文档中每个单词,主题模型定义了如下的生成过程.

1) 对于文档 d,从其隐含主题的分布中抽取一个主题 z,数学表示如式(1)所示.

$$P(t = z|d) \tag{1}$$

2）从抽取的主题 z 所对应的单词分布中抽取一个单词 w，数学表示如式(2)所示.

$$P(w|t=z) \tag{2}$$

3）重复上述过程，直至遍历文档中的每一个单词.

这样可以得到文档 d 中每个词 w 的生成概率. 这个过程的数学表示如式(3)所示.

$$P(w|d)=\sum_{z=1}^{T}P(w|t=z)P(t=z|d) \tag{3}$$

其中，t 是关于主题的一个变量；T 是主题的数量，主题数量 T 是主题模型的一个重要输入，T 的大小需要在模型训练前指定，而且存在一定的经验性. 实验中将每个目标多义词汇语料的主题数量 T 设置等于多义词词义数 量. 在基于主题模型的词义归纳方法中，首先对训练语料应用主题模型推导主题，语料主题即为多义词的词义列表. 接下来，将训练完成的 LDA 模型应用到计算测试语料中文档主题概率分布中，计算文档主题概率分布的过程即从词义列表中选择正确词义的过程.

为了选择正确的词义，本文从文档的主题概率分布中选择概率最大的那项主题作为多义词的正确词义，数学公式表示如式(4)所示.

$$\arg\max_z P(t=z|d) \tag{4}$$

3　实验结果及其分析

3.1　实验数据与评价标准

实验数据包含两部分，训练语料和测试语料. 训练语料包括 50 个多义词汇，每个多义词汇包括 50 个实例，一共包括 2500 个实例，训练语料中所有句子的正确意思未知，但给出了词汇包含的词义数量，词义数量从 2 到 21 不等. 和训练语料一样，测试语料同样包括 2500 个实例，并且人工标注了每个实例中多义词的词义. 所有实验数据从互联网上获得.

训练数据中的每个多义词汇的语料规模仅包含 50 个实例，语料规模太小，主题模型难以从中推导出有意义的主题. 为此，我们从互联网上为每个多义词提取了 5000～10,000 条包含多义词汇的中文句子来扩展训练数据语料规模，每条语句长度在 20～30 个词之间.

中文词义归纳评价指标一般采用 FScore[18]，相关定义如下.

本文把具有相同词义标签的测试数据作为一个类，设 C_r 是测试数据中的一个类，对应的实例个数为 n_r. S_i 是实验模型输出的一个类，对应的实例个数为 n_i. n_{ir} 表示模型正确预测的实例个数.

召回率：实验模型输出类中正确预测实例个数占测试数据对应类中的实例比例，如下式所示.

$$R(C_r,S_i)=\frac{n_{ir}}{n_r} \tag{5}$$

准确率：实验模型输出类中正确预测实例个数占模型输出类中实例总数比例，如下式所示.

$$P(C_r,S_i)=\frac{n_{ir}}{n_i} \tag{6}$$

FScore 指标：FScore 是在考虑了召回率和准确率的基础上，实验模型正确识别结果的可能性，其值越大说明实验模型的效果越好，如下式所示.

$$F(C_r,S_i)=\frac{2\cdot R(C_r,S_i)\cdot p(C_r,S_i)}{R(C_r,S_i)+p(C_r,S_i)} \tag{7}$$

对于某个类别 的 F 值定义如下式所示.

$$F(C_r)=\mathrm{Max}_i F(C_r,S_i) \tag{8}$$

对于整体的 FScore 定义如下式所示.

$$F\mathrm{Score}=\sum_{r=1}^{c}\frac{n_r}{n}F(C_r) \tag{9}$$

其中，c 是测试数据中包含类别总数；n 是测试数据实例总数.

3.2　实验及其结果分析

实验中首先将语料主题数量设为词义个数作为先验，然后使用 LDA 模型推导多义词汇训练语料的主题，模型训练完成后，使用该模型对测试语料中包含的多义词词义进行推断：LDA 模型计算测试语料中每篇文档的主题概率分布，多义词的词义即文档的主题分布中概率最大的那项主题. 为了描述实验中 LDA 推导出的多义词词义，表 2 列出了从包含多义词“暗淡”语料中推导出的主题，从主题词汇中可以得到多义词“暗淡”包含的两种词义：词义 1 用于形容光线等昏黑不明，词义 2 形容没有前途等没有希望.

表 2　词汇“暗淡”语料的主题词

Tab. 2　The topic words of '暗淡'

主题	主题词汇
1	光线 色彩 灯光 颜色 变得 明亮 显得 眼睛 色泽 光泽
2	前景 经济 中国 美国 市场 企业 前途 发展 世界 日本

表 3 是在测试数据上进行不同窗口大小实验后的结果. 窗口大小表示抽取的局部词汇数量，例如窗口大小等于 0，主题模型特征中不使用多义词的局部词汇，仅包含全局词汇. 窗口大小等于 3 表示抽取多义词上下文 3 个词汇作为额外特征.

从表 3 不同窗口大小的实验结果可以看出，局部词汇窗口大小为 2 时，其 FScore 取得了最佳的

性能. 当局部词汇窗口大小为 0 时,即不使用局部词汇作为额外特征,FScore 表现最差. 实验结果说明多义词附近的词汇包含更丰富语义信息,有助于正确判断多义词词义.

表 3 不同窗口大小实验结果

Tab. 3 Experimental results for different window sizes

窗口大小	准确率	召回率	FScore
0	0.7416	0.8040	0.7405
2	0.7680	0.8408	0.7758

4 结 论

本文使用主题模型 LDA 解决中文词义归纳问题,实验中设置语料主题数量等于多义词的词义数量作为先验. 鉴于单个多义词的训练语料规模太小,LDA 难以从中推导出有实际意义的主题,实验从互联网上提取包含多义词的语料若干,扩展了训练语料的规模. 最后的实验结果表明,本文基于主题模型的词义归纳方法在测试数据上获得了最高 77.58%的 FScore 得分.

参考文献:

[1] Uzuner O, Katz B, Yuret D. Word sense disambiguation for information retrieval[C]// Proceedings of the 16th National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence. Orlando: American Association for Artificial Intelligence, 1999.

[2] Chai J Y, Biermann A W. The use of word sense disambiguation in an information extraction system[C]// Proceedings of the 16th National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence. Orlando: American Association for Artificial Intelligence, 1999.

[3] Vickrey D, Biewald L, Teyssier M, *et al*. Word-sense disambiguation for machine translation[C]// Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver: Association for Computational Linguistics, 2005.

[4] Harris Z S. Distributional structure[J]. Synth Lang Libr, 1954, 10(2-3): 146.

[5] Denkowski M. A survey of techniques for unsupervised word sense induction[J]. Lang & Stat II Lit Rev, 2009(1): 1.

[6] Purandare A, Pedersen T. Word sense discrimination by clustering contexts in vector and similarity spaces[C]//Proceedings of the Conference on Computational Natural Language Learning. Boston: SIGNLL, 2004.

[7] Schütze H. Automatic word sense discrimination[J]. Comput Linguist, 1998, 24(1): 97.

[8] Niu Z Y, Ji D H, Tan C L. I2r: Three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation [C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Prague: Association for Computational Linguistics, 2007.

[9] Pantel P, Lin D. Discovering word senses from text [C]//Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton: ACM, 2002.

[10] Agirre E, Soroa A. Ubc-as: a graph based unsupervised system for induction and classification[C]// Proceedings of the 4th International Workshop on Semantic Evaluations. Prague: Association for Computational Linguistics, 2007.

[11] Klapaftis I P, Manandhar S. Word sense induction & disambiguation using hierarchical random graphs [C]//Proceedings of the 2010 conference on empirical methods in natural language processing. Massachusetts: Association for Computational Linguistics, 2010.

[12] Brody S, Lapata M. Bayesian word sense induction [C]//Proceedings of the 12th Conference of the European Chapter the ACL. Athens: Association for Computational Linguistics, 2009.

[13] 孙玉霞, 曲维光, 狄颖, 等. 词义归纳综述[J]. 计算机科学, 2014, 41(2): 23.

[14] Jin P, Sui R, Zhang Y. A knowledge based method for Chinese word sense induction[C]// Proceedings of the 2010 Fourth International Conference on Genetic and Evolutionary Computing (ICGEC). Shenzhen, china: IEEE, 2010.

[15] Liu Z, Qiu X, Huang X. Triplet-based Chinese word sense induction[C]//CIPS-SIGHAN Joint Conference on Chinese Language Processing. Beijing: CIPS, 2010.

[16] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. J Mach Learn Res, 2003(3): 993.

[17] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423.

[18] Zhao Y, Karypis G, Fayyad U. Hierarchical clustering algorithms for document datasets [J]. Data Min&Knowl Disc, 2005, 10(2): 141.