Research on Key Technologies for Cross-lingual Topic Analysis

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Tang Guoyu

Dissertation Supervisor : Professor Thomas Fang Zheng

December, 2014

摘 要

本文进行跨语言话题分析研究，致力于解决如下科学问题：给定一个涉及多种语言的文档集（限定每个文档只采用一种语言），如何识别出跨语言话题。跨语言话题分析面临的主要问题是翻译歧义。翻译歧义有两种情况：跨语言歧义和单语言歧义。为解决上述问题，本文深入研究了跨语言话题分析中的文档建模技术，针对不同问题提出了一系列基于语义的跨语言文档模型，主要包括：

- 针对跨语言歧义问题，将单语广义空间向量模型拓展为**跨语言广义向量空间模型**，同时提出了基于广义空间向量模型的特征选择算法。实验结果证明了这个方法在两个数据集上相对于向量空间模型分别提高 1.46% 和 2.18% 。
- 提出了**一种基于全局词义的跨语言文档建模方法**。分析翻译歧义的本质是词的同义性和多义性问题。针对这一问题，本文提出了基于词义的文档建模方法，采用统计方法定义局部词义和全局词义，提出了跨语言词义归纳算法获取跨语言词义，并分别在向量空间模型和潜狄利克雷模型中探讨了词义的贡献。实验结果表明，相对于跨语言广义空间向量模型，这个方法在两个数据集上性能分别提高 6.16% 和 5.74%。
- 在经典主题模型潜狄利克雷（Latent Dirichlet Allocation，LDA）模型的基础上，提出了**三种基于统计词义的 LDA 模型**（Sense-based LDA，SLDA）：第一种模型考虑词义对主题的影响（独立 SLDA）；第二种模型考虑词义和主题的相互影响（点估计协同 SLDA）；第三种模型考虑词义分布和主题的相互影响（词义混合协同 SLDA）。实验结果表明, 三种 SLDA 模型均优于 LDA 模型，而词义混合协同 SLDA 模型性能最好。概率分布分析也验证了这一结果，性能好的模型具有更加尖锐的主题分布，可以提供更强的后验信息。
- 针对跨语言话题分析问题，在词义混合词义协同 SLDA 的基础上融入词对齐信息，设计了**两种基于词义的跨语言主题模型**：第一种模型将词对看作变量，将主题看作词对的分布（词对齐词义 LDA）；第二种模型除了考虑词对齐信息，还进一步估计了词义对齐，从而构造了主题、词义和词义对齐的迭代学习过程（词义对齐词义 LDA）。实验结果表明相对于基于全局词义的跨语言文档建模方法，这个方法在两个数据集上性能分别提高 6.67% 和 3.28%，而词义对齐词义 LDA 在本文提出的所有方法中性能最好。

关键词：跨语言；话题分析；文档建模；主题模型

Abstract

Focusing on cross-lingual topic analysis, this dissertation seeks to resolve the following scientific problem of detecting cross-lingual topics from a multilingual corpus in which each document is written in only one language. The most challenging issue is translation ambiguity. There are two cases of translation ambiguity: the cross-lingual ambiguity and the monolingual ambiguity. To address the translation ambiguity issue, a series of cross-lingual document representation models are explored. The major contributions of this work are four-fold:

- *Cross-lingual Generalized Vector Space Model*. Aiming at addressing the cross-lingual ambiguity issue, the model considers word relationship in cross-lingual case and employs several word similarity calculation methods. In our work, the monolingual Generalized Vector Space Model was extended to the cross-lingual case by incorporating the cross-lingual word similarity. Meanwhile, a new feature selection algorithm is designed for the Generalized Vector Space Model. The experimental results on cross-lingual topic analysis task show that this model outperforms Vector Space Model by 1.46% and 2.18% respectively in two test datasets.
- *A cross-lingual document representation model based on global senses*. Analysis on the translation ambiguity issue indicates that the underlying reasons are synonymy and polysemy phenomena. It is thus argued in this work that word senses can be promising to represent documents to reduce the translation ambiguity. In the proposed word sense based document modeling methods, we defined local word senses and global word senses based on word co-occurrence statistics. We proposed the cross-lingual word sense induction algorithm to obtain statistical word senses and used word senses in two document presentation models: Vector Space Model and Latent Dirichlet Allocation(LDA). The experimental results on cross-lingual topic analysis task shows that this model outperforms the Cross-lingual Generalized Vector Space Model by 6.16% and 5.74% respectively in two test datasets.
- *Three topic models based on statistical senses*. Topic models rely on the co-occurrences of surface words to capture their semantic relations, but a surface word is likely to presents different word senses. We argued that disambiguating word senses for topic models can enhance their discrimination capabilities. We incorpo-

rated the statistic word senses into topic models and designed three revised LDA models based on word senses: Standalone Sense based LDA(SA-SLDA), Point estimate Collaborative Sense based LDA(PCo-SLDA) and Sense mixture Collaborative Sense LDA(SCo-SLDA). The experimental results shows that SCo-SLDA performs the best and the other two models also perform better than the LDA model. Distribution analysis also shows the same results: the model with better performance presents a sharper distribution of topics, thus provides more confidence on the posterior estimation. We also evaluated SCo-SLDA in WSI and verified that using proper topics as pseudo feedback will induce more accurate word senses.

- *Two cross-lingual topic models based on statistical senses*. In order to address the cross-lingual ambiguity issue, we incorporated word alignment information into SCo-SLDA and designed two cross-lingual topic models based on senses: Word Aligned Sense based LDA(WA-SLDA) and Word Sense Aligned Sense based LDA(WSA-SLDA). In WA-SLDA model, aligned word pairs are taken as variables and topics are taken as distributions on aligned word pairs. In WSA-SLDA model, besides aligned word pairs, the alignments of word senses are also estimated and the generative story of word senses, topics and word sense alignment is identified as iteratively interchangeable steps. The experimental results on cross-lingual topic analysis show that WSA-SLDA outperforms the Cross-lingual document representation models based on global senses than 6.67% and 3.28% respectively in two test datasets.

Key words: cross-lingual; topic analysis; document representation; topic model

目 录

第 1 章 引言 1
1.1 研究背景和意义 1
1.1.1 话题分析 1
1.1.2 话题分析的跨语言问题 2
1.2 相关研究工作 3
1.2.1 单语言话题分析 3
1.2.2 跨语言话题分析 9
1.3 论文的主要内容 12
1.3.1 研究目标 12
1.3.2 研究难点 12
1.3.3 研究思路 13
1.3.4 论文的贡献 16
1.3.5 论文的结构 17
第 2 章 跨语言广义向量空间模型 19
2.1 本章引论 19
2.1.1 研究问题 19
2.1.2 问题分析 19
2.1.3 解决思路 20
2.2 跨语言的词相似度计算 20
2.2.1 词相似度的相关研究工作 20
2.2.2 跨语言的词相似度计算 21
2.3 跨语言广义向量空间模型 26
2.3.1 广义向量空间模型 26
2.3.2 跨语言广义向量空间模型 26
2.4 广义向量空间模型的特征选择 26
2.4.1 基于 GVSM 的词语重要性 26
2.4.2 特征选择算法 27
2.5 实验评测 28
2.5.1 基于跨语言广义空间模型的跨语言话题分析 28
2.5.2 实验数据 29
2.5.3 评测指标 29
2.5.4 实验 #1：不同的词相似度计算 30
2.5.5 实验 #2：不同的特征选择 32

2.5.6 实验 #3：不同的文档表示模型 33
2.6 本章总结 34
第 3 章 基于全局词义的跨语言文档建模方法 35
3.1 本章引论 35
3.1.1 研究问题 35
3.1.2 问题分析 35
3.1.3 解决思路 36
3.2 跨语言词义 37
3.2.1 词义的相关研究 37
3.2.2 跨语言词义的定义 39
3.2.3 跨语言词义的生成 41
3.3 基于词义的文档建模 45
3.3.1 跨语言词义消歧 45
3.3.2 基于词义的 VSM 46
3.3.3 基于词义的 LDA 47
3.4 基于词义的文档建模总结以及在跨语言话题分析的应用 49
3.5 实验评测 50
3.5.1 实验设置 50
3.5.2 实验 #1 ：不同的词义聚类方法 51
3.5.3 实验 #2：不同的基于词义的文档建模模型 52
3.5.4 实验 #3：不同的文档建模模型 53
3.6 本章小结 54
第 4 章 基于统计词义的主题模型 55
4.1 本章引论 55
4.1.1 研究问题 55
4.1.2 分析问题 55
4.1.3 解决思路 56
4.2 基于词义的主题模型 56
4.2.1 独立词义 LDA 57
4.2.2 点估计协同词义 LDA 59
4.2.3 词义混合协同词义 LDA 61
4.3 实验评测 64
4.3.1 话题分析 64
4.3.2 概率分布分析 67
4.3.3 词义归纳 72
4.4 本章小结 74

第 5 章 基于统计词义的跨语言主题模型 76
5.1 本章引论 76
5.1.1 研究问题 76
5.1.2 问题分析 76
5.1.3 解决思路 76
5.2 基于词义的跨语言主题模型 77
5.2.1 JointLDA 模型 78
5.2.2 词对齐词义 LDA 79
5.2.3 词义对齐词义 LDA 81
5.3 实验评测 87
5.3.1 跨语言话题分析 88
5.3.2 概率分布分析 91
5.3.3 文档建模模型性能比较 93
5.3.4 复杂度问题 95
5.4 本章小结 95
第 6 章 总结和展望 97
6.1 论文工作总结 97
6.2 下一步研究展望 98
参考文献 100
致 谢 111
声 明 112
个人简历、在学期间发表的学术论文与研究成果 113

第 1 章 引言

1.1 研究背景和意义

1.1.1 话题分析

话题分析是一项旨在帮助人们应对信息过载问题的研究，曾以新闻为处理对象，将涉及某个话题的报道组织起来以某种方式呈现给用户。它的研究起源于美国军方的话题检测与跟踪（TDT）需求。

1996 年，美国国防高级研究计划局（DARPA）根据自身需求，提出开发一种能自动判断新闻数据流主题的新技术。从 1998 年开始，美国国家标准技术研究所（NIST）在 DARPA 支持下，联合马萨诸塞大学、卡耐基 -梅隆大学和 Dragon systems 公司，每年举办一次话题检测与跟踪国际研讨会，在共同的新闻媒体流测试集上对比不同 TDT 算法的准确性[1]。根据 TDT 评测官方定义，话题（topic）是指由一个核心事件或活动以及所有与之直接相关的事件或活动[1]。话题分析（topic analysis）技术由此产生，其目标是识别给定新闻集中每个新闻的话题。早期 TDT 评测室在传统新闻媒体信息流数据（如在新闻专线和广播新闻等）上开展的。近年来，随着互联网的发展，话题分析的研究进一步关注互联网载体上的新闻，例如网页新闻[2]、博客新闻[3]、微博新闻[4] 等。更有研究将话题分析技术应用于科技论文[5]，实现科技学术话题的自动分析。这些研究的共同点是：**给定一个文档集，识别出该文档集中每个文档的话题**。

同时，话题分析技术是网络舆情分析系统的核心模块。

网络舆情是指在一定的社会空间内，通过网络围绕中介性社会事件的发生、发展和变化，民众对公共问题和社会管理者产生和持有的社会政治态度、信念和价值观① 。近年来，互联网（Internet）在我国迅猛发展，已经成为了人们生活工作必不可少的一部分。根据《第 34 次中国互联网发展状况统计报告》截至 2014 年 1 月，中国网民规模达到 6.32 亿，互联网普及率为 46.9%。其中，手机网民规模达 5.27 亿。随着规模的增加，互联网已成为影响社会稳定、国家安全和世界和平的重要因素。网络舆情分析对政府和企业都有重大作用。2013 年 8 月全国宣传思想工作会议提出，要把网上舆论工作作为宣传思想工作的重中之重来抓。舆情监测成为各级政府的重要工作之一。对于企业来说，有效地监测企业相关的舆情，

① 百度百科，网络舆情。http://baikebaiducom/view/2143779htm

第一时间了解和处理企业在网络上的相关信息，可以有效地保持企业的健康良好形象。舆情分析系统主要包括以下几个核心功能[①]：1、热点话题、敏感话题识别；2、倾向性分析；3、话题跟踪；4、自动摘要；6、突发事件分析。可见，舆情分析系统的核心模块是话题识别，进而寻找热点和敏感话题。在海量网络信息中，与同一话题相关的信息往往孤立地分散在不同的时间段、不同的网站中，因此迫切需要一种能自动汇总特定话题相关信息的话题分析算法。这是话题分析研究的应用背景。

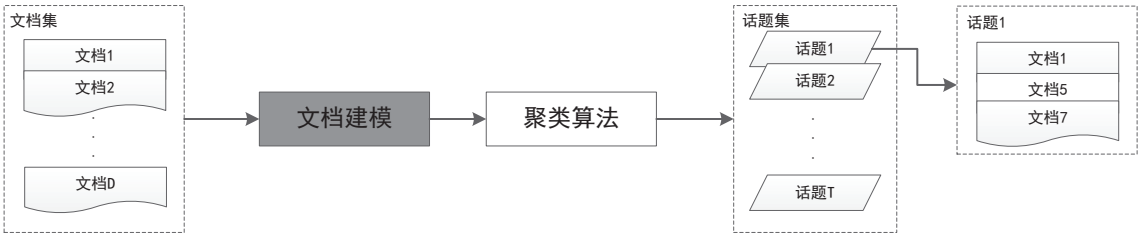


图 1.1　话题分析基本流程

从应用角度看，**话题分析系统的输入是文档集；输出是话题，并赋予每个文档一个特定话题**。图1.1是话题分析的基本流程图，对于文档集的每个文档，首先进行文档建模，然后根据文档的模型采用聚类算法进行聚类。

1.1.2　话题分析的跨语言问题

语言障碍是阻碍人们日常信息获取的瓶颈。全世界有超过 150 多种语言。由于语言障碍的存在，大多数用户无法快速获得其他语言的信息。随着全球经济的发展，互联网多语言内容已经十分普遍。英语独霸互联网的时代已经过去，汉语内容紧随英语之后并有后来居上之势，其他语种也发展很快。如何从非母语的互联网内容中获取信息，是困扰人们学习需要的问题之一。

语言障碍也是阻碍计算机信息处理的瓶颈。以话题分析为例，目前话题分析研究在单一语言上取得了快速发展，但无法满足人们对非母语信息的需求。随着互联网的进一步发展，各种语言网页的数量快速增长，人们已不再满足于仅从自己国家或地区的角度来看待事物。在事件发生或者变化时，人们还渴望了解其他国家或地区的报道或评论。我国是一个多民族国家，由 56 个民族组成，语言资源极其丰富。除了汉语外，我国还有 80 多种少数民族语言，而少数民族文字则有 19 种。通过跨语言话题分析可以更好地获取少数民族民众的观点，更有效的进行舆情分析。因此，跨越语言鸿沟、获取跨语言话题信息成为个人乃至一个国家了解

① 百度百科，互联网舆情监控系统。http://baikebaiducom/view/2143779htm# 7_1

和掌握全球信息的重要手段。因此，通过计算机自动地识别多种语言中相关话题信息，已成为信息检索和自然语言处理的研究热点之一。

综上，在话题分析中解决语言障碍问题，进行跨语言分析的研究是很有必要的。从应用角度看，**跨语言话题分析系统的输入是多语言文档集（限定每个文档仅采用一种语言）；输出是跨语言话题，并赋予每个文档一个特定的跨语言话题。**

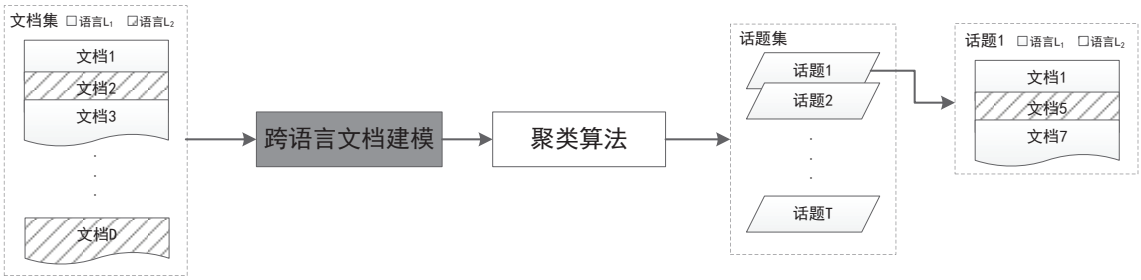


图 1.2　跨语言话题分析基本流程

图1.2分别是跨语言话题分析的基本流程图。图中不同语言的文档以不同的图案表示。从图中可以看出，单语言话题分析和跨语言话题分析的流程基本一致，区别在于跨语言话题分析需要对不同语言的文档进行建模，识别话题。跨语言话题分析中的话题可包含不同语言的文档。因此跨语言话题分析的关键问题是两种语言之间的意义映射。从自然语言理解的角度看，多种语言的同一对象的信息表示是这个对象在不同语言符号系统的字符串表示。如何对这些不同表示进行建模，获得它们表示的对象是跨语言信息处理的重要研究内容，也是跨语言话题分析的核心问题。

本文认为语义可以反映不同语言符号系统下的实际意义，深入研究了基于语义的跨语言文档建模，使不同语言相似话题的文档具有相似的文档表示，并将其应用到话题分析任务，提高跨语言话题分析的准确性。

1.2　相关研究工作

1.2.1　单语言话题分析

单语言话题分析的目标是从文档集中识别话题，并赋予每个文档特定话题。基本流程如图1.1所示。

1.2.1.1　话题检测

自 1998 年开始，NIST 召集了五次 TDT 评测，相关工作的综述已在文献 [6] 和文献 [1] 中分别给出。本文关注话题检测技术。话题检测的目标是识别新闻文

档中的话题，主流方法是文档聚类，即先以特定文档模型表示文档，然后采用聚类算法（如 K-means 或者层次聚类算法）将相似度高的文档聚在一起，构建成一个话题。因此，话题检测的关键问题是如何进行文档建模。

早期的话题检测与跟踪系统主要采用两种文档表示模型：向量空间模型（vector space model，VSM）[7] 和语言模型（language model，LM）[8]。

- 向量空间模型（VSM）

向量空间模型（VSM）是最早的文档表示模型[7]。它采用特征空间的向量来表示文档，其性能关键是特征的选择，目前主要采用了文档中的词汇作为特征。早期的一些话题检测与跟踪研究工作针对新闻这个特殊对象，着眼于如何准确地计算新闻中的词的权重。研究显示，词性对新闻表示有一定的影响。文献 [9] 评测了使用不同词性表示新闻对系统性能的影响。此研究表明，只用名词和形容词来表示新闻性能最好，当加入名词词组后，使用名词、动词、形容词以及名词词组表示新闻性能最好。文献 [10] 使用共现命名实体集（frequent itemsets）来表示话题。在文献 [11] 中，新闻表示中命名实体和普通词的权重不同，话题表示则使用卡方准则计算特征权重。文献 [12] 和文献 [13] 使用命名实体表示新闻和话题，并且不同种类的命名实体权重不同。文献 [14] 分别用命名实体，话题词以及所有词表示文档，并且采用三种表示相似度的线性组合作为文档相似度。国内也有研究采用了相似的策略，文献 [15] 将词划分为人物、时间、地点、内容 4 个组，并在这 4 个向量空间上分别进行权重和相似度计算。由于新闻的长度差异性大，短新闻的特征较少，文献 [9] 采用了新闻扩展技术，当两篇新闻的相似度高于预先定义的阈值时，将这两篇新闻相互扩展。文献 [9] 尝试根据话题将新闻划分成不同片段，计算片段之间的相似度。但是这种改进并不明显。

- 语言模型（LM）

语言模型的主要思想是构建 n 元文法模型，通过概率公式计算报道和话题的相关性，从而识别出话题[8]。文献 [16] 针对新闻建立了一元语法模型，进而计算新闻文档与话题的相似度。很多研究在语言模型中加入了新闻的其他因素。例如文献 [17] 采用了多模型方法，分别用人物、时间、地点、关键词来表示新闻和话题。由于新闻的时间以及其他内容是不同类型的特征，它们的模型也不同。针对时间，使用高斯混合模型表示，而其他内容则用一元文法表示。模型参数用最大似然方法估计。文献 [18] 构建了一个相关模型并且通过比较 Kullback-Leibler（KL）距离来判定话题，研究表明相关模型与传统语言模型相比有很大提高。文献 [19] 将词划分成不同的语义类，并为每个语义类构建了一元文法语言模型，采用每个语义类对数似然值的线性组合作为相似度。

除了文档建模方面的研究工作，还有一些研究致力于通过改进文本聚类

的方法以提高话题检测的准确性。文献 [20] 首先在在线事件监测任务中提出 Single-Pass 聚类方法。还有一些研究根据不同的任务要求改进了聚类算法[11,21–23]。国内的一些研究也对此作出了贡献[24–27]。聚类算法是机器学习领域的研究热点，而本研究的侧重点是文档建模。

尽管学术界已在话题检测研究中取得了大量研究成果，但传统话题检测技术仍然在文档建模方面存在不足。词汇是传统话题检测技术的主流特征，并假设特征之间的独立性。两个问题值得关注：首先，它将一个词看作是“死”的字符串，因此只能表示一个意思，这样就不能解决一词多义问题。比如，“苹果”这个词至少具有两个不同的意思：一个指一种水果，另一个指乔布斯创建的苹果电脑公司。其次，它的独立性假设忽视了词之间的关联性，例如同义词。比如，“电脑”和“电子计算机”互为同义词。举个极端的例子，假设一篇文档中仅有一个词“电脑”，而另一篇文档中只含有“电子计算机”，传统技术认为这两篇文档是完全不同的。

为解决上述问题，越来越多的研究将语义信息结合到话题检测算法中。有的研究加入了本体或知识库。例如文献 [28] 用词汇链表示报道，该文献只考虑名词，并且根据 WordNet 来确定候选序列的语义关系。文献 [29] 也采用了相似的表示方式。文献 [30] 将词分为四类：地名、专有名词、时间词和普通词。对于每一类，将具有相近含义的词项放在同一语义组。本体和语义知识库需要人工编辑，其覆盖面和更新速度都不能满足互联网快速发展的需要。另外一些研究考虑词的之间的共现程度。例如文献 [2] 将经常在一个句子共现的词定义为一个基本概念，然后用基本概念表示报道。文献 [31] 通过计算词的共现得到词的相关度，并将与新闻文档中的词的相关度大的词作为概念词。但是在这些仅考虑词共现的研究中，一个词只有一个概念，因此不能解决多义词问题。

1.2.1.2 主题模型

主题模型（Topic Model）近年来被广泛用于信息检索、文档分类等领域。主题模型也可以用做话题分析。主题模型的代表模型包括潜语义分析 (Latent Semantic Analysis，LSA)[32]、概率潜语义分析 (Probability Latent Semantic Analysis，PLSA)[33] 以及潜狄利克雷分布 (Latent Dirichlet Allocation，LDA)[34]。

LSA 模型的基本思想是将文档从高维词汇空间映射到低维的潜语义空间[32]。通过词—文档矩阵进行奇异值分解，消除词语间相关性，构造出一组标准正交基，即隐含的“人工概念”。但是在这个模型中，一个词只有一个映射，无法解决多义词问题。

在 LSA 的基础上，文献 [33] 提出了 PLSA 模型，假设每篇文档都是“潜在主

题”的混合。与 LSA 不同的是，PLSA 将文档表示成主题的多项分布，将主题表示为词语的多项分布，所以 PLSA 中的主题与 LSA 中的“人工概念”相比更直观，也在一词多义问题上具有潜力。但 PLSA 需要对文档层面的概率分布进行估计，因此不能预测新文档。针对这一问题，文献 [34] 提出了 LDA 模型。在 PLSA 的基础上，LDA 模型在文档 -主题概率分布中加入狄利克雷先验，从而可以对新文档进行预测。

LDA 作为第一个成熟的主题模型，对“主题”这一抽象概念进行建模，大大提升了深层描述文本语义的能力，在多个领域得到了广泛的应用。但是它仍然存在着一些问题，研究者们陆续在 LDA 的基础上提出了一些改进的主题模型。

在 LDA 模型中，各个主题是弱相关的，这与事实不否，有些主题之间可能存在层次关系。针对主题间的层次关系，文献 [35] 提出了层次化 LDA 模型（Hierarchical LDA）。该模型构造了一个高度固定的主题树，假设一个文档的所有主题都位于主题树的一条路径上。通过 Gibbs 抽样算法，自动地将文本中比较抽象的单词分配给主题树中较高层次的主题，从而描述了主题间的层次关系。但是该模型要求固定主题树的高度，当文本中隐含的主题层次关系不那么规律的时候该模型并不适用。为此文献 [36] 使用 Griffths-Engen-McCloskey（GEM）分布代替层次 LDA 中的词在主题树上的层次狄利克雷分布，打破了主题树高度必须固定的限制。针对主题间可能存在的两两相关性，文献 [37] 提出了关联主题模型（Correlated Topic Model, CTM），用多维高斯分布代替 LDA 模型中的狄利克雷分布，通过模型训练，估计高斯分布的协方差矩阵来表示主题之间的两两相关性。但是 CTM 在模型推理和参数模拟问题上存在缺陷，当维度过高的时候会很不准确。文献 [38] 提出了独立因子主题模型（Independent Factor Topic Model），借鉴因子分析的思想，假设主题是主题因子的线性混合。通过估计混合参数，获得主题和主题因子的关联性，间接获得主题间关联性。

还有一些改进模型加入了其他信息，如作者[39]、时间[40,41]、事件约束[42,43]等。还有研究将 LDA 与 n-gram 模型结合起来[44–47]。

主题模型试图同时解决“一词多义”和“多词同义”两个问题，解决效果并不理想，这是因为它们需要从文档集全局层面直接寻找主题，同时解决同义词和一词多义问题的难度较大，导致效果下降。因此，一些主题模型[48–50] 加入了知识库如 WordNet 的语义信息，但是这些方法面临着知识库覆盖率的问题。

基于主题模型的话题分析方法有两种，一种是直接利用主题模型抽取话题，将主题模型产生的主题作为文档包含的话题[5,34]。第二种是在主题空间上利用主题表示文档和话题，然后利用聚类的方法进行话题分析[51]。

1.2.1.3 其他基于语义的文档建模

近年来，语义信息不仅被用在话题分析中，还被广泛地用在文档分类、文档聚类以及信息检索的文档建模步骤。

研究者尝试采用概念对文档进行建模，概念主要来自 Wikipedia、WordNet 和 HowNet 等语义知识库。文献 [52] 提出了明确语义分析 (Explicit Semantic Analysis，ESA) 方法，将文档表示为一个维基百科相关概念序列，根据文档中概念的相关程度来衡量文档相似度。ESA 方法对文档的语义拓展比较好，但是在语义拓展过程中会加入很多外围的相关词，可能与目标不吻合。文献 [53] 则研究了维基百科对文本向量的语义扩展问题，将文档词向量中的每个词匹配到维基百科概念，利用同义词、上层概念、关联概念等实现向量语义相关性扩充。文献 [54] 首先识别出文档中的概念，然后给相关分类节点打分。分类节点的权重依赖于文档中出现该分类标题、子分类标题或者概念的次数以及分类的大小。文献 [55] 则在文档聚类中分别利用了维基百科中的类别、上下层概念关系修改文档相似度公式。文献 [56,57] 采用 WordNet 或者 HowNet 等人工编撰的本体中的概念表示文档。它们采用词义消歧的算法确定词的概念然后用概念表示文档。采用概念进行文本建模有以下两个问题，首先，这些概念需要人工定义的，构建比较困难，同时受到知识库规模和更新速度的影响，在海量文档话题分析中难以奏效。其次，人工概念的粒度不一，难以形成符合实际的概念划分。

另外一些研究者利用词的共现信息挖掘词的关系，并将这些关系用于文档建模。GVSM[58] 改进了特征之间的正交性，将文档表示在一个非正交的空间上。空间上的维度之间的关系就是词的相关度，词的相关度则通过统计词的共现得到。在 GVSM 模型基础上，文献 [59] 改进了词相关度的计算方法并在语义空间上进行了降维处理。以上这两种模型着眼于解决同义词问题，无法处理一词多义问题。文献 [60] 提出采用联合聚类的方法同时聚类文档和词。文献 [61] 利用 EM 算法得到词簇，再用词簇表示文档。这种基于词簇的方法通常假设一个词仅属于一个类簇，同样不能解决一词多义的问题。

文献 [62] 用上下文表示词义，利用无指导词义归纳算法从原始文本语料库里自动获得词义。该方法用于无指导词义消歧[63]，后来用于短文本 (如电子邮件) 分类[64]。词义归纳还可用于信息检索领域，文献[65,66] 在信息检索领域采用无指导词义归纳技术确定词义。但是这两个研究只考虑到一个词的多义现象，没有考虑到词与词之间的关系。

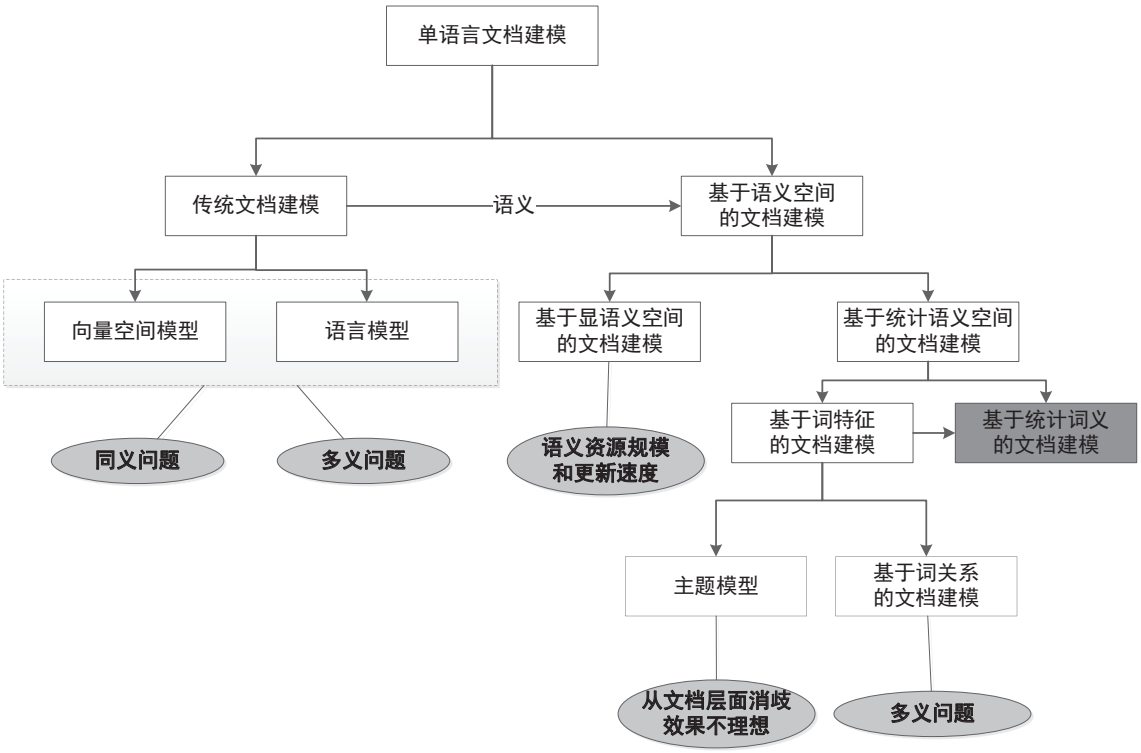


图 1.3　单语言文档建模文献总结

1.2.1.4　单语言话题分析小结

单语言话题分析的研究集中在文档建模模块上，研究的总结如图1.3所示。

早期的单语言话题分析起源于话题检测与跟踪技术，采用传统的文档建模方法，无法解决文档表示中的多义词和同义词问题。为了解决上述问题，单语言话题分析开始采取基于语义的文档建模。基于语义的文档建模有两种：

1. 基于显语义空间的文档建模。这类方法利用知识库的概念表示文档，依赖于知识库，受到知识库规模和更新速度的影响，在知识库资源缺乏的语言不适用。
2. 基于统计语义空间的文档建模。基于统计语义空间的文档建模方法有两种：第一种是主题模型。主题模型试图同时解决“一词多义”和“多词同义”两个问题，解决效果并不理想，这是因为它们需要从文档集全局层面直接寻找主题，解决同义词和一词多义问题的难度较大，导致效果下降。第二种方法考虑词之间的关系如利用词相似度或者词簇。但是这种方法无法解决多义词问题。

近年来，利用词义进行文档建模的尝试开始涌现，尤其是无指导词义归纳方法的出现，给文档建模带来新的希望。词义是介于词和概念之间并携带语义的语言元素，可有效避免词和概念所面临的问题，在文档建模中颇具潜力。

1.2.2 跨语言话题分析

随着互联网的全球化, 话题分析面临着多语言的挑战，跨语言话题分析的研究也越来越多。跨语言话题分析的任务是识别跨语言文档集中每个文档的话题，并进行汇总。跨语言话题分析的话题可能包含不同语言的文档。基本流程如图1.2所示，跨语言话题分析和单语言话题分析的区别在于它需要对不同语言的文档进行建模，识别话题。

1.2.2.1 跨语言话题检测

在 TDT 国际评测中，TDT2 开始提供英文和中文的评测语料，TDT3 还加入了阿拉伯文等其他语种的新闻语料。话题检测与跟踪研究从一开始就考虑到跨语言的问题。

传统的跨语言 TDT 研究主要有基于机器翻译和基于双语词典翻译两种策略。

- 基于机器翻译

NIST 为 TDT 国际评测提供了机器翻译的功能。因此早期的 TDT 评测系统都是用机器翻译工具将源语言翻译成目标语言，然后采用单语言的方法进行相似度计算[67]。评测结果表明, 与单语言话题检测与跟踪相比, 采用机器翻译方法会导致 50% 的性能下降。这是由于机器翻译中的错误使源语言被翻译成目标语言后，不仅没有准确的表达原新闻的意思，还会产生噪声影响文档建模的准确性。文献 [68] 验证了跨语言话题跟踪在原始语料上的性能比机器翻译后语料上的性能更好。文献 [68] 首先用机器把多语言新闻文档统一翻译成一种语言，生成话题模型。如果检测到的后续相关新闻的语言与话题模型不相同，那么生成这种语言的话题模型。这种方法依然对机器翻译依赖性很强，如果机器翻译提供了错误的训练样本，那么后续的话题模型会产生偏差。

- 基于双语词典翻译

基于双语词典翻译的跨语言话题检测与跟踪方法是把文本中的词对应成另一种语言的词。文献 [69] 将中文话题的动词和名词都翻译成英文，来关联中文话题和英文话题。文献 [69] 用中英词典翻译新闻文档中的词，同时考虑到歧义问题保留多个翻译候选词。文献 [21] 利用多语言主题词表 Eurovoc 构造跨语言文档向量。文献 [70] 从知网中获取义原, 用义原向量空间来表示文档。以上基于词典的跨语言文档聚类方法不能判断歧义词的正确翻译，难以解决翻译歧义问题。

1.2.2.2 跨语言主题模型

主题模型在单语言话题分析中取得了很好的效果。跨语言主题模型是在单语言主题模型的基础上借助词典或者跨语言语料库对应不同语言。跨语言主题模型的拓展方式有两种。

第一种是将平行或者可比语料中不同语言的对应文档看做伪文档，利用主题模型在平行/可比语料上抽取跨语言主题。文献 [71] 利用 LSA 在平行空间中构造了多语言语义空间，不同语言的单词和文档都能映射到同一个语义空间中，然后进行多语言文档聚类。一些研究在信息检索领域上采取了相同的做法[72–75]。文献 [76] 在双语平行语料库利用 PLSA 获得跨语言主题概率，然后利用这个概率计算词的翻译概率。但是这些算法简单地将不同语言的平行文档当做一个伪文档，当主题个数特别大的时候，主题中的词会很稀疏。同一语言的主题分布比较集中，不利于获得对齐的主题。针对这一问题，文献 [77,78] 提出多语言主题模型（Polylingual Topic Model，PTM），将不同语言的文档放在不同的元组中，从多语言语料库中建立多语言的对齐主题。文献 [79,80] 使用了相同的模型进行跨语言聚类。文献 [81] 提出了 Coupled Probabilistic Latent Semantic Analysis (CPLSA)，用于跨语言信息检索和跨语言文档分类中。但是这种方法需要从开发集中训练主题，而开发集中的主题与目标集中的主题不完全一样，这种偏差会影响性能。

第二种是从词典或者平行语料中获得不同语言词的对应关系，然后用这种对应关系拓展主题模型。文献 [82] 用词与词的对应关系拓展 PLSA，抽取双语主题。文献 [83] 则在有指导的 LDA 基础上加入词典拓展为多语言有指导的 LDA 模型。文献 [84] 利用了双语词对拓展 LDA。但是这个方法限制词在一个数据集中只存在一个翻译。文献 [85] 中的方法则没有这一限制，可以在一定程度上解决语言之间的对应问题和翻译歧义问题。但是它的消歧是在文档集全局层面上进行的，考虑的是文档层面的上下文，难度较大，会导致结果下降。

1.2.2.3 其他跨语言文档建模方法

近年来，跨语言文档建模不仅被用在跨语言话题分析中，还被广泛地用在跨语言文档分类、跨语言文档聚类以及跨语言信息检索中。

一些研究利用维基百科进行跨语言文档信息检索。文献 [86] 利用维基百科中的知识结构（如跨语言链接、类别等）建立不同语言文档之间的联系。文献 [87] 在 ESA 的基础上，用维基百科的概念表示文档，并且并通过跨语言链接对应不同语言的概念。但是这种方法会受到维基百科的范围和更新速度的影响，在没有维基百科资源的语言中不适用。

一些研究利用矩阵分解的方法在平行语料中进行不同语言的对应。文献 [81] 提出 Oriented Principal Components Analysis(OPCA) 的方法在信息检索和文本分类上解决跨语言问题。此外，还有研究将非负矩阵分解（Non-negative Matrix Factorization，NMF）[88–90]、典型相关分析（Canonical Correlation Analysis，CCA）[91,92] 和基于核方法的典型相关分析（Kernel Canonical Correlation Analysis，KCCA）[93–95] 应用于跨语言信息检索或者跨语言文档分类。文献 [96] 提出一种谱聚类的方法，将不同语言的文档看做加权图，在可比语料中采用传播算法获取先验信息，合并多语言空间，解决跨语言聚类问题。但是这些方法仍然面临不同文档集的偏差问题。

文献 [97] 首先对文档进行机器翻译，然后将翻译的文档作为原始文档的相关视图进行聚类。聚类过程中采用了约束机制。这个方法受到机器翻译性能和效率的限制。

文献 [98] 等在英语和西班牙语的可比语料库中识别同源命名实体，然后聚类。实验结果显示，仅使用同源命名实体进行匹配可以获得好的聚类性能。文献[99] 使用英语的命名实体识别印度语言的命名实体。通过识别同源词进行跨语言文档建模的优点是不需要双语词典等双语资源，但是这个方法只适用于同一语系，对于不同语系的语言（如英语和汉语）之间作用不大。

1.2.2.4 跨语言话题分析小结

跨语言话题分析的主要方法可分为基于直接对应的文档建模和基于语义空间的文档建模。

基于直接对应的文档建模用于跨语言话题分析早期，一般是借助机器翻译翻译文档或者词典翻译文档中的词或者关键词实现。基于机器翻译的方法的性能会受到机器翻译性能和效率的影响，而基于词对应的方法面临的主要问题是翻译歧义问题。

语义空间有显语义空间和潜语义空间之分。显语义空间需要语义资源的支持，同样受到语义资源的规模和更新速度的影响。潜语义空间方法有两种：一种是直接在平行/可比语料中构造语义空间，特征选择和提取也都是在平行/可比语料的语义空间中进行的。这种语义对齐是主题层面的，而主题通常是和文档息息相关的，不同文档集之间的主题差异性较大。目标语料中的主题可能在平行/可比语料中并未涉及，同时平行/可比语料中的冗余主题会产生噪音。因此，这种主题偏差会严重影响方法的性能。而另一种方法则是利用词对齐信息直接在文档集中构造跨语言语义空间。但是现有的跨语言主题模型在主题层面上解决翻译歧义问题，解决效果并不理想。

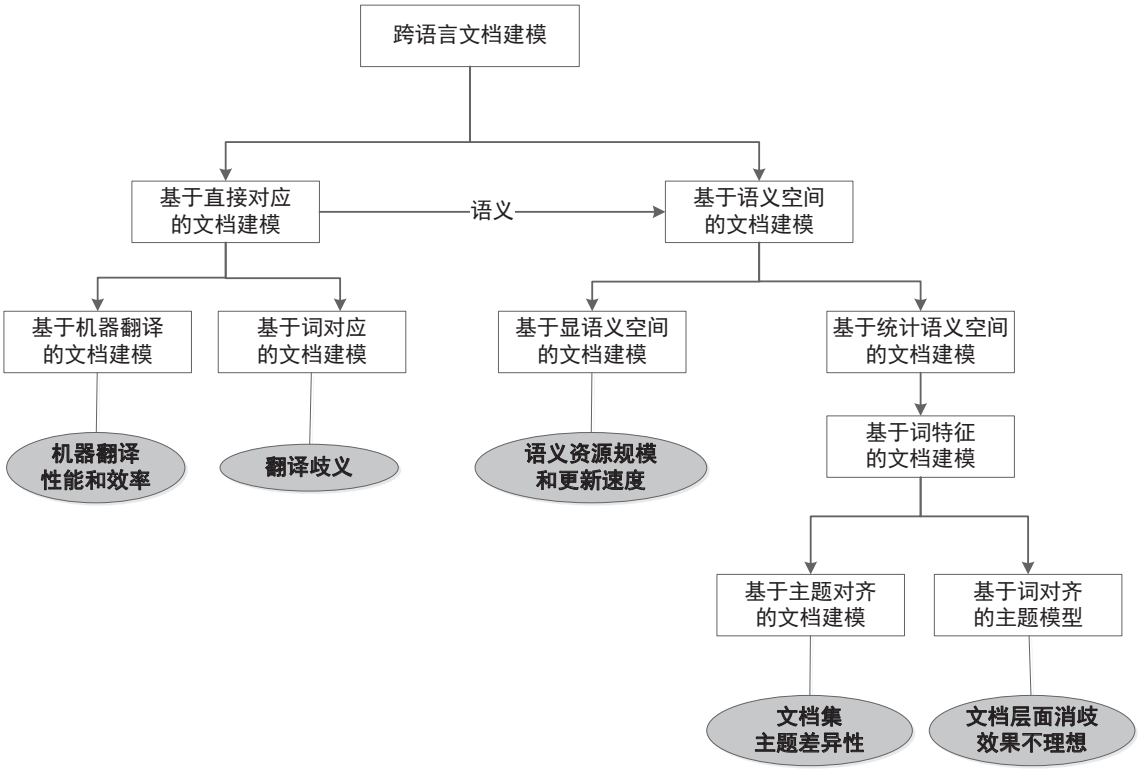


图 1.4　跨语言文档建模文献总结

1.3　论文的主要内容

1.3.1　研究目标

本文的研究目标是解决**跨语言话题分析问题**，跨语言话题分析问题的定义如下：

给定一个多语言文档集，限定单个文档仅采用一种语言。设计文档模型，将文档表示为特定模型实例，以便借助聚类算法，将文档集划分为多个话题类簇，赋予每个文档一个特定的话题。

本文围绕跨语言文本建模这一核心问题，通过更好的文档表示模型提高跨语言话题分析的性能。

1.3.2　研究难点

跨语言话题分析在文档建模中面临着翻译歧义的这一研究难点。分析翻译歧义，有两种情况：

1) 一个词可以有不同的翻译选择，2) 多义词在不同的情况下翻译不同。

分析翻译歧义产生的原因，第一种情况来源于目标语言的同义性，我们称为跨语言的歧义；第二种情况来源于源语言的多义性，我们称为单语言的歧义。本文参考单语言文档建模的解决方案，加入语义信息，在解决同义词和多义词问题

的同时解决翻译歧义问题。现有的基于语义的文档建模主要有基于知识库和基于统计语义两种。由于基于知识库的文档建模方法受到知识库的限制，本文主要考虑如何从文档中获取统计语义，然后利用统计语义表示文档。

根据上述分析，本文提出的方法需要解决以下两个科学问题：

1.3.2.1 如何在文档建模中更好地解决跨语言的歧义

跨语言的歧义性来源于语言的同义性。例如，在表示人体器官的情况下，词“arm”既可以被翻译成“胳膊”，也可以被翻译成“手臂”。因此需要对词和它的翻译有着相同或者相近的表示。本文考虑利用统计语义表示文档，但是现有基于统计语义的方法或者只能获取主题层面的对齐或者通过词对齐构造跨语言主题空间，效果不理想。因此如何充分利用语义空间获取更好的对齐信息是本文研究要解决的首要问题。

本文认为细粒度的对齐信息，如词或者词义的对齐则通常与文档无关。使用细粒度的对齐信息可以针对不同的目标集选择不同的特征，从而产生不同的话题，鲁棒性较好。

1.3.2.2 如何在文档建模中更好地解决单语言的歧义

单语言的歧义来源于语言的多义性。比如词“arm”有两个意思：a）手臂，胳膊 b）武装。因此“arm”在描写人体的上下文中需要翻译成“手臂”，在描写军事的上下文中需要翻译成“武装”。因此，需要根据不同的上下文判断词的语义，对于表达不同含义的同一个词给出不同的表示。比如说，当词“arm”在描写人体的上下文中和描写军事的上下文的表示不同。

现有的基于统计语义的方法只能从文档层面进行消歧义，不能充分利用近距离上下文信息。本文认为近距离的上下文信息比如句子级别的上下文，比文档层面的上下文即文档级别的上下文更能反映词的含义。

针对以上两个科学问题，本文将以跨语言文本建模为研究对象，探索如何有效的利用统计语义表示跨语言文档。需要强调的是，虽然本文中跨语言文档建模应用在跨语言话题分析上，它同样可以应用在跨语言信息处理的其他任务上，如跨语言文档分类，跨语言信息检索等。

1.3.3 研究思路

针对如何更好地抽取对齐语义信息，提高跨语言文档建模的科学问题，本文研究思路如图1.5 所示。围绕跨语言歧义和单语言歧义这两个研究难点，从以下三

个方面开展研究：选取合适的跨语言语义表示，在语料中提取跨语言语义以及利用跨语言语义进行文档建模。本文选取了词相似度和词义这两个语义表示，分别提出了对应的语义提取算法，然后针对词义与话题的影响，提出更好的文档建模方法。下面将从语义表示、语义提取和文档建模三个方面详细介绍文本的研究思路。

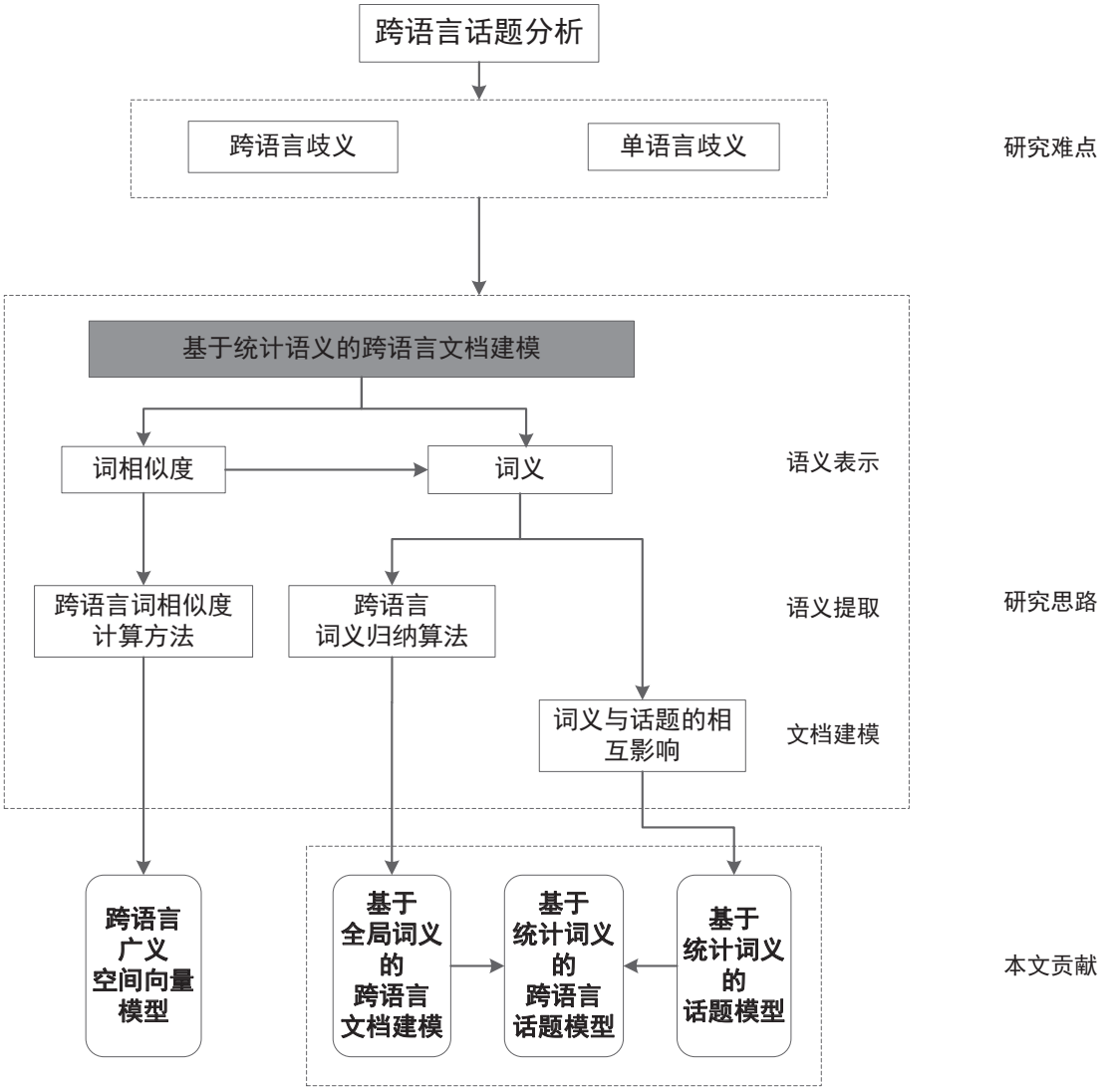


图 1.5　研究内容

1.3.3.1　选取合适的跨语言语义表示

现有的统计语义主要有两种表示：主题模型抽取的潜在语义（主题）以及词与词之间的关系（词类簇或者词的相似度）。这些统计语义表示主要是针对同义词和多义词现象的。本文除了要解决同义词和多义词问题，还需要考虑到跨语言问题。因此，需要改进现有的统计语义表示或者构造新的语义表示。本文的统计语

义表示需要满足以下两个条件：

1) 可以解决跨语言歧义问题。本文的统计语义需要将不同语言中的相同语义表现为相同或者相似的形式。因此能很好的反映不同语言之间的对齐形式是本文选择统计语义表示的一个首要标准。

2) 可以解决单语言歧义问题。本文的统计语义需要识别出词在表达不同含义时候的不同翻译，因此是否能表示出词的歧义是本文选择统计语义表示的一个重要标准。

为了满足以上两个条件，本文比较了两种语义表示：词的相似度和词义。

如图1.5所示，本文首先探索了使用词的相似度作为语义表示，将单语言词的相似度拓展为跨语言词的相似度。这是由于词的相似度相比词簇可以更准确反映词的关系。相比主题可以更灵活的在不同语料库中构造不同的语义空间。但是词相似度只能反映同义现象，无法表示多义现象，也就是说只能解决跨语言歧义问题，无法解决单语言歧义问题。

因此本文进一步选择词义作为语义表示。词义是意义的表达，有“指称说”、“因果说”、“四因素说”、“用法说”等多种定义和表示方式。考虑到计算的方便性，本文采用“用法说”，用上下文中能够表征特定语义的词语集合来定义一个词义。

举例来说，“包袱”一词至少包含如下两个词义，如例1.1和例1.2 。

例 1.1： 包袱 #T1：用布包起来的包儿。如，“把你的包袱拿好。”

例 1.2： 包袱 #T2：某种心理负担。如。“你不要有思想包袱。”

采用“用法说”，以上词义可借助其上下文词语表示为例1.3和例1.4。

例 1.3： 包袱 # E1：{ 东西, 行李，衣服, 布, 拎, 丢，整理, 收拾}

例 1.4： 包袱 # E2：{ 思想，心理, 背, 压力, 负担, 克服, 消除, 加深}

通俗地说，第一组定义是通过解释其用途和现象等意义（内涵）来完成的，而第二组是通过具体的同现词汇实例（外延）来完成的。前者能准确表达词义，但不方便计算机计算。后者能否准确表达词义依赖于同现词汇的选择，但非常适合计算。因此本文采取后者词义定义方式。同时我们注意到相同或相近的词义具有相似的上下文词语集合。因此词义具有解决同义词能力。而同一词可能有多个词义，因此词义也具有解决多义词的能力，也就是说词义可以同时反映由同义和多义现象导致的翻译歧义问题。

为了解决跨语言问题，本文构造跨语言的词义表示，同时用不同语言的上下文中能够表征特定语义的词语集合来表示一个词义。

举例来说，词义“包袱 # S2”被扩展为例1.5。

例 1.5： 包袱 #S2：{思想, 心理, 背, 压力，负担, 甩掉，消除，加深, burden, psychological, suffer, pressure, endure, eliminate, alleviate, add}

在上述跨语言词义表示中，多语言上下文相关词被收集进来。因此不同语言的相同或者相近的词义也具有相似的上下文词语集合。因此跨语言词义具有解决跨语言问题的能力。

同时与主题模型生成的主题相比，词义的粒度比较细，可以根据不同文档集的不同词以及词频，构造不同的语义空间和特征空间，鲁棒性更好。

1.3.3.2 在语料中提取跨语言语义

在确定了语义表示后，还需要解决语义提取的相关问题。现有的研究大多都是基于单语言的，跨语言语义生成算法较少。因此需要提出新的跨语言语义生成算法，解决跨语言语义生成问题。针对词的相似度，本文在平行语料上扩展了相似度计算方法，生成跨语言相似度。针对跨语言词义，本文在提出了新的跨语言词义归纳方法。

1.3.3.3 利用跨语言语义进行文档建模

利用跨语言语义表示需要考虑如何确定目标文档中词的语义，如何提取语义特征，计算基于语义的文档相似度等。针对词的相似度，参考单语言文档建模，本文采用经典的 GVSM 表示文档。针对词义，本文直接将词义作为特征，采用词义消歧的方法确定每个词的词义，利用现有的文档建模方法表示文档。本文还提出新的基于词义的文档表示模型，主要思想是考虑到词与主题的相互影响，将词的词义看做主题模型的一个隐藏变量，提出基于词义的主题模型，从文档中自动学习词义和话题。之后，本文对这个话题模型进行拓展，解决跨语言话题分析问题。

1.3.4 论文的贡献

本文的主要贡献是图1.5中所列出的四个文档建模方法，包括：

1) 针对跨语言歧义问题，借助跨语言词相似度，将单语广义向量空间模型（Generalized Vector Space Model, GVSM）拓展为跨语言广义向量空间模型。同时，

提出了适用于 GVSM 的特征选择算法。实验证明，GVSM 在跨语言话题分析的性能优于现有方法，新的特征选择算法可以进一步改进跨语言话题分析。

2) 为了同时解决跨语言歧义问题和多语言歧义问题，提出一种基于全局词义的跨语言文档建模方法。该方法首先利用词义归纳算法从平行语料中获取每个词的词义，然后利用词义聚类算法发现不同词词义的关系。最后利用词义消歧获得目标文档集中每个词的词义，并用词义表示跨语言文档。实验表明该方法在跨语言话题分析任务上优于 GVSM。

3) 为了获得更好地表示文档，现在单语言情况下进行尝试，提出三种基于统计词义的主题模型。该模型可以同时生成文档的主题和文档中词的词义。与之前的研究不同，我们并没有使用一些语义资源来表示词义，而是将词的词义看做主题模型的一个隐藏变量，从文档中自动学习词义和话题。该主题模型与上一个基于词义的方法相比，优势在于同时考虑到了词义对主题的影响和主题对词义的影响。实验结果表明，词义作为主题模型的特征可以获得更加集中的话题分布。将词义与主题的相互影响作为伪回馈，可以生成更准确的词义和主题。

4) 在跨语言情况下更好的表示文档，提出两种基于统计词义的跨语言主题模型。该方法在单语言主题模型的基础上解决了对齐词对解决跨语言的对应问题。此外，这个方法还进一步考虑了词义对齐，实现了主题、词义以及词义对齐的相互影响，实验结果表明，该主题模型在跨语言话题分析上优于第二个方法。

1.3.5 论文的结构

本文的内容共六章，具体安排如下：

第1章是绪论部分。首先介绍了话题分析的研究背景和意义，引出研究跨语言话题分析的重要性。接着综述了单语言话题分析和跨语言话题分析的国内外研究现状，然后分析了跨语言话题分析的研究任务以及面临的挑战，最后阐述了本文的研究思路和贡献。

第2章将单语广义向量空间模型拓展为跨语言广义向量空间模型。首先介绍了跨语言词相似度的计算方法，然后利用跨语言词相似度改进广义空间向量模型，进行跨语言文档建模。在此基础上还提出了适用于 GVSM 的特征选择算法。本章在跨语言话题分析的任务下比较了不同相似度的性能。

第3章提出一种基于全局词义的跨语言文档建模方法。首先给出了词义的定义，然后介绍了词义归纳算法和词义聚类算法，之后介绍了词义消歧的方法和基于词义的文档建模方法。最后在跨语言话题分析任务上评测了基于全局词义的跨语言文档建模方法的性能。

第4章提出三种基于统计词义的主题模型。首先分析了词义作为主题模型特征的优势，然后设计了三种基于词义的 LDA 模型（Sense based LDA, SLDA）：独立 SLDA（Standalone SLD, SA-SLDA）, 点估计协同 SLDA（Point estimate Collaborative SLDA，PCo-SLDA），词义混合协同 SLDA（Sense mixture Collaborative SLDA, SCo-SLDA），之后介绍了三个模型的推理求解过程，最后评测了三种 LDA 模型的性能。

第5章提出两种基于统计词义的跨语言主题模型。首先利用对齐词对，拓展了词义混合的协作 SLDA（Sense mixture Collaborative SLDA, SCo-SLDA）使其可以处理跨语言的情况。然后介绍了该拓展模型的推理求解过程，最后在跨语言话题分析的任务下评测了该模型的性能。

第6章总结了论文的工作并进行了展望。

第 2 章　跨语言广义向量空间模型

2.1　本章引论

2.1.1　研究问题

传统的话题分析采用向量空间模型（Vector Space Model, VSM）[7] 表示文本，它利用词袋（Bag of Word, BOW）模型来构建特征空间，并将每个文档转化为一个向量。词袋模型导致在特征匹配中的“硬匹配”问题。例如，当词“海岸”被选为特征时，除非“海边”也被选为特征，否则“海边”无法影响到文档表示。这是因为在“硬匹配”中，“海岸”和“海边”完全不同。

在跨语言话题分析中，一些研究[100] 利用双语词典或者机器翻译工具翻译文档或者特征。这会导致翻译歧义问题。当一个词存在多个翻译目标词时，很难选择采取哪一个翻译目标词作为翻译结果，这就是跨语言话题分析中的跨语言歧义问题。可见，一旦这个词的某一个翻译被选为特征时，这种跨语言歧义加剧了“硬匹配”问题。

因此，本章主要针对跨语言歧义和特征选择中的“硬匹配”问题开展研究。

2.1.2　问题分析

针对解决跨语言歧义和“硬匹配”问题，现有的文档建模研究做了一些尝试。

为了解决 VSM 的缺陷，很多研究提出基于统计语义的文档表示模型。一些文档表示模型如 LSA[32]，LDA[34] 等将文档映射到低维空间来获取统计潜语义。另外一些模型如单语广义向量空间模型（Generalized Vector Space Model, GVSM）[58,59] 则采用显式语义，在文档中计算词的相似度，将文档表示在一个非正交的空间上。但是上述模型都是基于单语文档集的，不能直接用到跨语言文档集中。

为了解决跨语言歧义问题，文献 [71] 在平行语料训练 LSA，获得 LSA“潜概念”，然后生成这些“潜概念”在目标文档集的权重。与单语言 LSA 不同，跨语言 LSA 在固定开发集上选择特征。但由于目标文档集通常与开发集存在内容和用词的显著不同，会导致过度适应问题。

因此，为了更好的解决跨语言歧义和“硬匹配”问题，跨语言文档建模方法在文档表示中需要使词和它的翻译表示相同或者相近，在特征选择中需要避免“硬匹配”问题。同时需要注意目标文档集和开发集的特征差异性。

2.1.3　解决思路

为了解决跨语言歧义，本章利用跨语言词相似度将 GVSM 拓展到跨语言文档建模中，即跨语言广义空间向量模型（Cross-lingual GVSM，CLGVSM）。本章在平行语料中统计词的相似度，比较了几种相似度的性能，同时利用词典和翻译概率进一步改进了跨语言词相似度。考虑词相似度后，词和它的不同翻译有较大的相似度，这样可以使词的不同翻译都能获得对应，有利于解决跨语言歧义问题。

为了避免特征选择中的“硬匹配”，本章提出了基于“软匹配”的特征选择方法。在新的特征选择方法中，最有代表性的词是根据软词频和软文档频在语义空间中选择出来的。这样，一个非特征的词可以影响它的近义词的权重，从而为跨语言文档建模做出贡献。词的跨语言歧义具有相近的表示，有利于解决跨语言歧义问题。

同时本章提出的方法直接在目标文档集上选择特征，避免了文档集之间的差异性对性能造成的影响。

本章后续部分按如下方式组织：2.2节介绍跨语言词相似度计算方法；2.3 节介绍跨语言广义向量空间模型；2.4节介绍基于广义向量空间模型的特征选择方法；2.5节评测了基于跨语言广义向量空间模型在跨语言话题分析任务下的性能；2.6节对本章进行了总结。

2.2　跨语言的词相似度计算

词相似度计算是一个自然语言处理研究热点，已经应用在机器翻译和词义排歧等研究中。词相似度在 CLGVSM 的构建中起到重要的作用，本节先总结了词相似度计算的研究现状，然后介绍本章的词相似度计算方法。

2.2.1　词相似度的相关研究工作

在传统的 GVSM[58] 和改进的 GVSM[59] 中，词的相似度在目标文档集中进行计算。性能最好的词相似度计算方法如下：假设词是高斯分布的随机变量，计算词协方差向量的内积。本章则用词协方差向量的余弦相似度来计算词的相似度。本章将这种方法称为基于 COV 的方法。但是这种相似度计算方法是在目标文档集中进行计算，无法计算不同语言词的相似度。

近年来提出的的词相似度的计算算法或基于语义网络，或基于统计技术。

文献 [101,102] 提出基于 WordNet 的英文语义相似度计算方法。文献 [103] 在机器翻译中用知网计算词相似度。文献 [104] 提出用知网的概念定义计算跨语言相似度。

基于统计技术的词相似度计算方法更为广泛。最经典的方法是点互信息（Pointwise Mutual Information, PMI）[105]。PMI 值越大，说明词汇越有可能出现在同一语境下。很多研究都是在 PMI 的基础上进行改进。文献 [106] 提出了基于 PMI-IR 的同义词获取方法，利用 Alta Vista Adanvced 搜索引擎计算单词之间的概率。SOCPMI 方法[107] 利用 PMI 将两个目标词的相邻词按重要性排序，并通过计算相邻词的 PMI 实现目标词之间的相似度计算。

这两种方法各有特点，基于语义网络的方法简单有效，无需训练语料库，但是这种方法得到的结果受到人的主观意识影响大，有时并不能准确反映客观事实。另外，这种方法比较准确地反映了词之间语义上的相似性和差异，但是对于词之间的语法和语用考虑的较少，基于统计技术的方法比较客观，综合反映了词在语法、语义、语用等因素的相似性和差异。但是这种方法比较依赖于训练语料，计算量大，计算也比较复杂。

2.2.2　跨语言的词相似度计算

跨语言词汇相似度（Cross-lingual Word Similarity，CLWS）在 CLGVSM 的构建中起到重要的作用。本章实现了三种代表性的词相似度计算方法：基于知网的词相似度[104] 、基于 COV 的词相似度方法[59] 和基于 SOCPMI 的词相似度[107]。第一种是基于语义网络的相似度计算方法，后两种则都是基于统计的计算方法。下面将分别进行介绍。

2.2.2.1　基于知网的词相似度

我们采用基于知网的跨语言词相似度计算方法[104]，基本思想是利用知网中词汇的语义定义。

按照知网的创造者董振东先生的说法：“知网是一个以汉语和英语的词语为代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。”[108]。知网中有两个基本概念：概念与义原。概念是对词汇语义的描述。一个词可以表达多个概念。义原是用于描述概念的最小意义单元。与一般的语义网络（如 WordNet）不同，知网并不是简单地将所有概念归结到一个树状的概念层次体系，而是用义原对概念进行描述。[109] 中文词“海岸”和英文词“coast”在 HowNet 的概念如表2.1 所示。从概念定义（表2.1 的 DEF）得出，海岸是一块“陆地”并且“靠近”“水域”。同时，表2.1显示 DEF 的结构包括两部分：（1）基本义元，例如“land| 陆地”; (2) 关系义元，例如 BeNear| 靠近: existent= , partner=waters| 水域。关系义元会包含几个二级义元，进一步定义了概

念。知网中的词涵盖了中文和英文，义元也具有双语结构。因此，HowNet 可以用来计算双语词相似度。

表 2.1　中文词“海岸”和英文词“coast”在 HowNet 中的概念定义

No. =048973	//概念标识
W_C= 海岸	//中文词
G_C=N[hai3 an4]	//中文词的词性和拼音
W_E=coast	//英文词
G_E =N	//英文词的词性
DEF={land| 陆地: {BeNear| 靠近: existent={~}, partner={waters| 水域}}}	// 概念定义

基于知网的中文词相似度和英文词相似度都有研究。文献 [103] 提出基于知网的中文词相似度计算，文献 [110] 提出基于知网的英文词相似度计算。本章采用文献 [104] 提出的跨语言词相似度计算，可以计算中文词和英文词之间的相似度。给一个中文词 w^{CN} 和英文词 w^{EN}，跨语言词相似度计算如下：

1) 首先在 HowNet 搜索出 w^{CN} 的相关概念集 $C^{\mathrm{CN}} = \{c_i^{\mathrm{CN}}\}(i = 1...N^{\mathrm{CN}})$，相关义元集 $P^{\mathrm{CN}} = \{p_i^{\mathrm{CN}}\}$ 和相关二级义元集 $S^{\mathrm{CN}} = \bigcup_{i=1}^{N_{\mathrm{CN}}} S^{\mathrm{CN_i}} = \{s_{i,k}^{\mathrm{CN}}\}(k = 1...M^{\mathrm{CN}})$。我们也可以得到英文词 w^{CN} 的相关概念集 $C^{\mathrm{EN}} = \{c_i^{\mathrm{EN}}\}(i = 1...N^{\mathrm{EN}})$，相关义元集 $P^{\mathrm{EN}} = \{p_i^{\mathrm{EN}}\}$ 和相关二级义元集 $S^{\mathrm{EN}} = \bigcup_{i=1}^{N_{\mathrm{EN}}} S^{\mathrm{EN_i}} = \{s_{i,k}^{\mathrm{EN}}\}(k = 1...M^{\mathrm{EN}})$。

2) 计算中文词的每个义元 $s^{\mathrm{CN}} \in (P^{\mathrm{CN}} \cup S^{\mathrm{CN}})$ 和英文词的每个义元 $s^{\mathrm{EN}} \in (P^{\mathrm{EN}} \cup S^{\mathrm{EN}})$ 的义元相似度 Sim^{SEM}。义元相似度是根据 HowNet 中义元的距离和深度差计算的。计算公式如公式 (2-1)。

$$\begin{aligned} &Sim^{\mathrm{SEM}}(s^{\mathrm{CN}}, s^{\mathrm{EN}}) = \\ &\max\left\{0, 1 - \lambda_{\mathrm{dis}} \cdot V^{\mathrm{dis}}(s^{\mathrm{CN}}, s^{\mathrm{EN}})\right\} \cdot \min\left\{\lambda_{\mathrm{dep}} \cdot V^{\mathrm{dep}}(s^{\mathrm{CN}}, s^{\mathrm{EN}}), 1\right\} \end{aligned} \tag{2-1}$$

其中 $V^{\mathrm{dis}}(s_1, s_2)$ 和 $V^{\mathrm{dep}}(s_1, s_2)$ 从 HowNet 中得到的两个义元的距离和深度差；λ_{dis} 和 λ_{dep} 分别是距离和深度差的归一化系数。根据 [110] 的实验，我们设置 $\lambda_{\mathrm{dis}} = 40$，$\lambda_{\mathrm{dep}} = 4$。

3) 根据义元相似度 Sim^{SEM}，计算中文词的每个概念 c_i^{CN} 和英文词的每个概念 c_i^{EN}，计算公式如公式 (2-2) 和公式 (2-3)。

$$Sim^{\mathrm{CON}}(c^{\mathrm{CN}}c^{\mathrm{EN}}) = Sim^{\mathrm{SEM}}(p_i^{\mathrm{CN}}, p_i^{\mathrm{EN}}) \cdot Sim^{\mathrm{SEM}}(S_i^{\mathrm{CN}}, S_i^{\mathrm{EN}}) \tag{2-2}$$

$$Sim^{\mathrm{SEM}}(S^{\mathrm{CN}}, S^{\mathrm{EN}}) = \frac{\sum_{m=1}^{M^{\mathrm{CN}}} \sum_{n=1}^{M^{\mathrm{EN}}} Sim^{\mathrm{SEM}}(s_m^{\mathrm{CN}}, s_n^{\mathrm{EN}})}{M^{\mathrm{CN}} \times M^{\mathrm{EN}}} \tag{2-3}$$

4) 根据概念相似度计算词 w^{CN} 和词 w^{EN} 的词相似度 Sim^{HN}，计算公式如公式 (2-4)。

$$Sim^{\mathrm{HN}}(w^{\mathrm{CN}}, w^{\mathrm{EN}}) = \max_{i=1}^{N^{\mathrm{CN}}} \max_{j=1}^{N^{\mathrm{EN}}} Sim^{\mathrm{CON}}(c_i^{\mathrm{CN}}, c_j^{\mathrm{EN}}) \tag{2-4}$$

2.2.2.2　基于 COV 的词相似度

在单语言 GVSM[58,59] 中，词的关系是在目标文档集中进行计算的。

假设 $D = \{d_j; j = 1, \cdots, N\}$ 表示包含 N 篇文档，M 个词的文档集。X 表示一个 $m \times n$ 的矩阵，它的元素 x_{ij} 表示词 w_i 在文档 d_j 的权重。G 表示词与词的关系矩阵。例如，3 个词的关系矩阵如下：

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.2 & 1 \end{pmatrix}_{3\times3} \end{array} \tag{2-5}$$

其中每一行和每一列都代表一个词。

传统的 GVSM[58] 中，词表示为文档的对偶空间中的向量。词和词的相似度是通过计算向量的余弦相似度得出的，G 的计算公式如公式 (2-6)。

$$G = L^{-1/2} X X^T L^{-1/2} \tag{2-6}$$

其中 L 是一个对角阵，每个元素对应 XX^T 的对角元素。

在改进的 GVSM 中，性能最好的 G 为词向量的协方差矩阵。首先假设文档的词是高斯分布的随机变量，然后计算词的协方差矩阵。协方差矩阵计算如公式 (2-7)。

$$G_{\mathrm{COV}} = \frac{1}{N-1} X H X^T; \text{其中} H = 1 - \frac{1}{n} e e^T \tag{2-7}$$

其中 e 是一个元素都为 1，长度为 n 的矩阵。

G_{COV} 将不相关的词映射到正交的方向，将负相关的词映射到相反的方向。而传统的 G 则将不相关和负相关的词均映射到正交方向，因此 G_{COV} 的性能好于传统的方法。本章选用余弦相似度来代替内积进行计算词相似度。

2.2.2.3　基于 SOCPMI 的词相似度

SOCPMI[107] 主要是利用 PMI 选取相邻词。两个词的相似度与他们的共现词是否相同有关。它的好处是可以计算不经常共现的词的相似度。

首先介绍单语的 SOCPMI 计算流程[107]。

假设 $C = \{w_j; j = 1, \cdots, N\}$ 表示包含 N 个词的语料集。词集合 $T = \{t_i; i = 1, \cdots, M\}$ 是所有词形的集合。值得注意的是，C 是一个有序的列表，同一个词有可能多次出现，而 T 则是所有词形的集合，不包含重复的词。词 w_g，w_h 的相似度计算流程如下：

1) 分别获得 w_g ，w_h 的共现词集合。共现词指的是在一个共现窗口中同时出现的词。文献 [107] 中，取前后 α 个词作为窗口。共现窗口包括目标词本身，则该窗口一共有 $2\alpha + 1$ 个词。定义词频（type frequency）如公式 (2-8).

$$f^t(t_i) = |\{k : c_k = t_i\}|; \text{其中 } i = 1, 2, \cdots, N \tag{2-8}$$

定义词 w_g 的二元词频（bigram frequency）如公式 (2-9)。

$$f^b(t_i, w_g) = |\left\{k : c_k = w_g; c_{k\pm j} = t_i\right\}|; \text{其中} i = 1, 2, ..., N;\ -\alpha \le j \le \alpha \tag{2-9}$$

二元词频 $f^b(t_i, w_g)$ 反映了词 t_i 和词 w_g 在共现窗口的共现次数。从而可以得到当 $f^b(t_i, w_g) > 0$ 时，词 w_g 和 t_i 的 PMI 的计算如公式 (2-10)。

$$f^{\text{pmi}}(t_i, w_g) = \log_2 \frac{f^b(t_i, w_g) \times M}{f^t(t_i) f^t(w_g)} \tag{2-10}$$

这样，可以得到 w_g 的共现词集合。共现词集合 X 是按照 PMI 值降序排列的前 β_g 个 $f^{\text{pmi}}(t_i, w_g) > 0$ 的词。w_g 的共现词集合的定义如公式 (2-11)。

$$\begin{aligned} X = \{x_i\}; \text{其中} f^{\text{pmi}}(t_1, w_g) \ge f^{\text{pmi}}(t_2, w_g) \ge \cdots \\ f^{\text{pmi}}(t_{\beta_g - 1}, w_g) \ge f^{\text{pmi}}(t_{\beta_g}, w_g); i = 1, 2, \cdots, \beta_g \end{aligned} \tag{2-11}$$

同样我们可以获得 w_g，w_h 的共现词集合如公式 (2-12)。

$$\begin{aligned} Y = \{y_i\}; \text{其中} f^{\text{pmi}}(t_1, w_h) \ge f^{\text{pmi}}(t_2, w_h) \ge \cdots \\ f^{\text{pmi}}(t_{\beta_h - 1}, w_h) \ge f^{\text{pmi}}(t_{\beta_h}, w_h); i = 1, 2, \cdots, \beta_h \end{aligned} \tag{2-12}$$

2) 分别计算 w_g ，w_h 与共现词集合的 β – PMI 总和（β – PMI summation）。对于 w_g ，β – PMI 总和按照公式 (2-13) 计算。

$$f^\beta(w_g) = \sum_{i=1}^{\beta_g} \left(f^{\text{pmi}}(x_i, w_g)\right)^\gamma; \text{其中} f^{\text{pmi}}(x_i, w_g) > 0;\ f^{\text{pmi}}(x_i, w_h) > 0 \tag{2-13}$$

同样可以得到 w_h 的 β – PMI 总和如公式 (2-14)。

$$f^\beta(w_h) = \sum_{i=1}^{\beta_h} \left(f^{\text{pmi}}(y_i, w_h)\right)^\gamma; \text{其中} f^{\text{pmi}}(y_i, w_g) > 0;\ f^{\text{pmi}}(y_i, w_h) > 0 \tag{2-14}$$

3) w_g ， w_h 的相似度按照公式 (2-15) 计算。

$$Sim(w_g, w_h) = \frac{f^{\beta}(w_g)}{\beta_g} + \frac{f^{\beta}(w_h)}{\beta_h} \tag{2-15}$$

SOCPMI 算法有两个参数 β 和 γ 。在文献 [107] 中，β 按照公式 (2-16) 计算。

$$\beta_i = (log\,(f^t(w_i)))^2 \frac{\log_2(n)}{\delta}; \text{其中} i = g, h \tag{2-16}$$

其中 δ 是一个常数。本章的参数与文献 [107] 相同：$\delta = 6.5$，$\gamma = 3$ 。

原始的 SOCPMI 算法只针对单语，不能解决跨语言问题。为此，我们对 SOCPMI 进行拓展，使其可以计算跨语言词相似度。拓展的 SOCPMI 方法是针对平行语料的，拓展的策略是：共现窗口设定为一个句子，同时平行语料中不同语言的对应句子都算作共现窗口。假设给一个中文词 w^{CN} 和一个英文词 w^{EN} ，首先获得 w^{CN} 的中文共现词集合以及 w^{EN} 的英文共现词集合，然后分别计算 w^{CN} 与 w^{EN} 的共现词集合中的 β – PMI 值以及 w^{EN} 与 w^{CN} 的共现词集合的 β – PMI 值。

2.2.2.4　结合词典或翻译概率的统计相似度

前面提到的两种基于统计的跨语言词相似度都是在开发集（平行语料）中计算的。但是我们认为目标集的词共现信息对文档建模同样有作用。因此，我们改进跨语言词相似度的计算方法。改进策略主要是在目标集中计算单语言词相似度，然后利用词典或者翻译概率获得跨语言词相似度。下面将分别介绍如何利用词典和翻译概率来获得跨语言词相似度。

使用词典时，假设词 w^{CN} 在词典的中翻译为 $T^{\mathrm{CN}} = \{t_k^{\mathrm{CN}}; k = 1, \cdots, K^{\mathrm{CN}}\}$ 同时词 w^{EN} 在词典中的翻译为 $T^{\mathrm{EN}} = \{t_k^{\mathrm{EN}}; k = 1, \cdots, K^{\mathrm{EN}}\}$ 。我们选择 w^{CN} 与 T^{EN} 中的每个词以及 w^{EN} 与 T^{CN} 中每个词的相似度最大值作为 w^{CN} 与词 w^{EN} 的相似度。

使用翻译概率时，假设词 w^{CN} 在平行语料中的翻译概率为 $P^{\mathrm{CN}} = \{t_k^{\mathrm{CN}} : p_k^{\mathrm{CN}}; k = 1, \cdots, K^{\mathrm{CN}}\}$ ，词 w^{EN} 在平行语料中的翻译概率为 $P^{\mathrm{EN}} = \{t_k^{\mathrm{EN}} : p_k^{EN}; k = 1, \cdots, K^{\mathrm{EN}}\}$，$w^{\mathrm{CN}}$ 与词 w^{EN} 的相似度的计算公式如公式 (2-17)

$$\begin{aligned} Sim^{\mathrm{PR}}(w^{\mathrm{CN}} \rightarrow w^{\mathrm{EN}}) &= \sum_k p_k^{\mathrm{CN}} \times Sim^{\mathrm{Mono}}(w^{EN}, t_k^{\mathrm{CN}}) \\ Sim^{\mathrm{PR}}(w^{\mathrm{EN}} \rightarrow w^{\mathrm{CN}}) &= \sum_k p_k^{\mathrm{EN}} \times Sim^{\mathrm{Mono}}(w^{CN}, t_k^{\mathrm{EN}}) \end{aligned} \tag{2-17}$$

$$Sim^{\mathrm{PR}}(w^{\mathrm{CN}}, w^{\mathrm{EN}}) = \max\{Sim^{\mathrm{PR}}(w^{\mathrm{CN}} \rightarrow w^{\mathrm{EN}}), Sim^{\mathrm{PR}}(w^{\mathrm{CN}} \leftarrow w^{\mathrm{EN}})\}$$

其中 Sim^{Mono} 表示单语言相似度。

2.3　跨语言广义向量空间模型

为了便于描述，我们首先介绍传统的广义向量空间模型。

2.3.1　广义向量空间模型

假设 $D = \{d_j; j = 1, \cdots, N\}$ 表示包含 N 篇文档，M 个词的文档集。GVSM[58] 将文档表示在一个非正交空间中，文档的相似度计算公式如公式 (2-18)

$$Sim^{\mathrm{GVSM}}(d_1, d_2) = \frac{d_1^T G d_2}{\sqrt{d_1^T G d_1}\sqrt{d_2^T G d_2}} \tag{2-18}$$

其中 G 的定义和在单语言 GVSM[58,59] 的计算方法已经在 2.2.2.2节介绍过，这里就不再赘述了。

2.3.2　跨语言广义向量空间模型

前人的工作 [59] 中，词的关系矩阵 G 反映了词向量的内积，词向量的长度反映了词在文档中的重要程度。因此他们用到的 G_{COV} 实际上是词的相关度，与词相似度并不完全等同。同时 G_{COV} 很难处理双语情况。词向量是在目标文档中计算出来的，缺乏双语信息。如果直接在双语语料中计算 G_{COV}，不同文档集中的词的重要性不同，会产生噪音。因此，本章选择词的相似度代替相关度构造词的关系矩阵。选择词的相似度的另一个好处是可以忽略噪音和贡献度小的相似度。这样，关系矩阵 G 变得比较稀疏，有利于节省运算时间。因此本章设置了相似度的阈值。本章还比较了2.2.2节中介绍的几种词相似度在 CLGVSM 的性能。

2.4　广义向量空间模型的特征选择

2.4.1　基于 GVSM 的词语重要性

在 VSM 中，词对文档的重要性与词在文档中的词频（Term Frequency，TF）成正比，与词在文档集中的倒文档频（Document Frequency，DF）成反比。我们认为 TF 和 DF 在 GVSM 中是可以改进的。

考虑一个文档出现“criminal”3 次、“imprisonment”10 次。虽然词“criminal”的词频比较低，但是它对该文档仍然非常重要。这是由于“imprisonment”与“criminal”是语义相似的。如果将 VSM 中按照字符串进行特征匹配的方法叫做“硬匹配”，而我们将提出基于 CLGVSM 模型“软匹配”方法。

在 GVSM 模型中，词的重要性由两种指标衡量: 软词频和软文档频。

2.4.1.1　软词频

软词频（Soft Term Frequency，TF^{s}）考虑的是词与语义相近的词。给定文档 d 和词 w，我们首先在文档 d 中查找 w 的相似词 $S = \{w_i; i = 1, \cdots, M^S\}$。软词频 TF^{s} 的定义如公式 (2-19) 所示。

$$TF^{\mathrm{s}}(t,d) = \sum_{i=1}^{M^S} TF_i \times Sim(w, w_i) \tag{2-19}$$

其中 TF_i 代表词 w_i 在文档 d 中的词频。软词频是在一篇文档中计算的。

2.4.1.2　软文档频

一个词的软文档频（Soft Document Frequency，DF^{s}）不仅需要考虑含有该词的文档还需要考虑含有相似词的文档。给定词 w 和文档集 $D = \{d_j; j = 1, \cdots, N\}$，假设 $d_j = \{w_{i,j}; i = 1, \cdots, M^D\}$ 代表文档 d_j 中的词，软文档频定义如公式 (2-20)。

$$DF^s(w) = \sum\nolimits_{d_j \in D} \max\nolimits_i \{Sim(w, w_{i,j})\} \tag{2-20}$$

软词频是在文档集中计算的，选取的是文档中的所有词的最大值，这是为了减少文档长度和相似度低的词的影响。

2.4.2　特征选择算法

得到基于 GVSM 的词重要性后，可以进行特征选择。参考 TF-IDF 思想，我们定义到软文档频如公式 (2-21)。

$$IDF^s(w) = log\Big(\frac{N}{DF^s(w)}\Big) \tag{2-21}$$

其中 N 表示文档的个数。

因此词 w 在文档 d 的软权重计算如公式 (2-22)。

$$r^s(w,d) = TF^s(w,d) \cdot IDF^s(w) \tag{2-22}$$

如果我们单纯依靠权重进行特征选择，相似度比较高的单词会同时被选为特征。这是因为相似度比较高的单词含有相近的权重，会造成特征集的冗余。为此，我们提出了一个改进的特征选择算法，只给相似词集中的一个词比较高的软词频，其余词汇则降低权重。具体做法是按照初始软词频的从大到小更新软词频，删除相似度所造成的冗余。

特征选择前，我们按照如下流程更新软词频 TF^{s}:

1) 创建一个空的候选词列表。

2) 将文档的词按照它们的 TF^{s} 的大小降序排列，作为文档词列表

3) 将第一个词 w_0 移动到候选列表中。

4) 对于文档词列表中的每个词，按照公式 (2-23) 更新它们的 TF^{s}。

$$TF^{s}_{g+1}(w,d) = TF^{s}_{g}(w,d) - \sum_{w_k \in d} Sim(w,w_k) Sim(w_0,w_k) TF^{s}(w_k,d)) \tag{2-23}$$

其中 t_k 代表文档 d 的词，g 代表迭代次数。

5) 删除文档词列表中权重小于 0 的词。

6) 重复步骤 2)~5) 直到文档词列表变为空。

获得候选列表后，我们使用公式 (2-22) 计算词的权重，并按照特征权重的大小选择每个文档的特征，然后将文档集中的每个文档的特征合并为一个特征集。

2.4.2.1　GVSM 的文档表示

获得文档集后，下面我们考虑如何使用特征集表示文档。假定 $F = \{f_i\}_{i=1...M^{F}}$ 表示特征集，$d = \{w_j\}_{j=1...M^{d}}$ 表示文档 d 中的词。我们现在尝试将 d 映射到文档集中。

在映射过程中，不仅要考虑特征集的词，还需要考虑特征集之外的词对文档表示的影响。但是为了避免信息冗余，我们只允许一个特征映射一个词。

因此，给定按照文档 d 的软词频 TF^{s} 排序的特征集 F，文档 d 中的特征 f_i 的特征权重计算如下：

1) 创建一个空的候选词列表 T^{d}。

2) 按照公式 (2-24) 找出文档中候选词列表 T^{d} 外的最相似的词 w^{*}。

$$w^{*} = argmax_{w_i \in d, w_i \notin T_M} Sim(w, f_i) \tag{2-24}$$

3) 特征 f_i 的权重 r^{a} 重新计算如公式 (2-25)

$$r^{a}(f_i,d) = TF(w^{*}) IDF^{s}(f_i) \tag{2-25}$$

4) 将 w^{*} 加入到候选词列表 T^{d} 中。

5) 重复步骤 2)~ 4) 直到所有特征都被匹配过。

2.5　实验评测

2.5.1　基于跨语言广义空间模型的跨语言话题分析

跨语言话题分析系统的任务是识别出给定文档集中每个文档的话题。本章采用跨语言广义空间模型表示文档，获得文档的相似度后，采用聚类算法进行

话题聚类。本文采用 Bisecting K-Means[111] 算法进行聚类。Bisecting K-Means 是 K-Means 的拓展方法，研究证明它的性能优于标准的 K-Means 算法和层次聚类算法。它的基本思想是，假定要得到 k 个簇，将所有点的集合用 k-means 算法分裂成两个簇，从这些簇中选取一个继续分裂，如此下去，直到产生 k 个簇。

2.5.2　实验数据

开发集

我们从 LDC 数据集 (LDC2004E12, LDC2004T08, LDC2005T10, LDC2003E14, LDC2002E18m LDC2005T06, LDC2003E07 和 LDC2004T07)① 中随机抽取 1M 的语料作为开发集。开发集主要是用来训练跨语言词相似度和获得翻译概率。

词典

我们从 HowNet 中抽取翻译词对作为词典。

翻译概率

我们在开发集中使用 Giza++[112] 获取翻译概率。

测试集

我们采用 TDT4 语料的两个数据集进行测试。两个评测集的信息如表2.2所示

表 2.2　测试集的统计数据

语料	TDT41 (2002)	TDT42 (2003)
英文 (话题数/文档数)	38/1270	33/617
中文 (话题数/文档数)	37/657	32/560
总计 (话题数/文档数)	40/1927	37/1177

TDT4 语料共有 98452 篇新闻文档，包含三个语言：英文，中文和阿拉伯语。新闻时间为 2000 年 10 月 1 日至 2001 年 1 月 31 日之间。语料基本是原始的新闻文本或语音识别得到的电视或广播新闻文本。本实验分别采用 2002 年的评测语料（TDT41）和 2003 年的评测语料（TDT42）。[113]。

2.5.3　评测指标

国际 TDT 评测的评测指标是由漏检率和错检率组成的评测公式（detection cost function）C_{Det}[1] 以及它的归一化形式 $(C_{\mathrm{Det}})_{\mathrm{Norm}}$。

① https://www.ldc.upenn.edu/

而近期的话题分析[4,79,114] 使用文档聚类的评测方法。文档聚类通常采用 F-Measure 来表示性能。F-Measure 的值越大，聚类的性能就越高。本文采用 F-Measure 作为评测指标。

F-Measure 的计算方法[111] 如下：

首先，假设 A_i 代表系统生成的类簇 c_i 的文档，代表 A_j 人工标注的类簇 c_j 的文档，则准确率和召回率的计算如公式 (2-27) 和公式 (2-26)。

$$p_{i,j} = \frac{|A_i \cap A_j|}{|A_i|} \tag{2-26}$$

$$r_{i,j} = \frac{|A_i \cap A_j|}{|A_j|} \tag{2-27}$$

对于簇 c_i 和类 c_j 的 F-Measure 可由公式 (2-28) 计算得到。

$$f_{i,j} = \frac{2 \cdot p_{i,j} \cdot r_{i,j}}{p_{i,j} + r_{i,j}} \tag{2-28}$$

其次，每一类 c_i 的 F-Measure 值都被定义为对于所有簇计算 $f_{i,j}$ 的最大值，召回率和准确率也按这种方法进行计算，将他们记作 f_i，r_i，p_i（公式 (2-29)）。

$$\begin{aligned} p_i &= \max_j\{p_{i,j}\} \\ r_i &= \max_j\{r_{i,j}\} \\ f_i &= \max_j\{f_{i,j}\} \end{aligned} \tag{2-29}$$

最后对于整体的召回率、准确率和 F-Measure 值按照他们每一类的文档比重加权平均得到。假定 N 为总文档数。计算公式如公式 (2-30) 所示。

$$\begin{aligned} r &= \sum_i \frac{|A_i|}{N} r_i \\ p &= \sum_i \frac{|A_i|}{N} p_i \\ f &= \sum_i \frac{|A_i|}{N} f_i \end{aligned} \tag{2-30}$$

2.5.4　实验 #1：不同的词相似度计算

本实验旨在比较不同跨语言相似度在跨语言话题分析任务下的性能。我们实现了七个跨语言相似度的计算方法：

- **HN**：基于 HowNet 的跨语言词相似度。
- **SOCPMI^DEV**：在开发集中计算基于 SOCPMI 的跨语言词相似度。
- **SOCPMI&DIC**：在测试集中采用 SOCPMI 计算单语言相似度，利用词典计算跨语言相似度。

- **SOCPMI&TRAN**：在测试集中采用 SOCPMI 计算单语言相似度，利用翻译概率计算跨语言相似度。
- **COVˆDEV**：在开发集中计算基于 COV 的跨语言词相似度。
- **COV&DIC**：在测试集中采用 COV 计算单语言相似度，利用词典计算跨语言相似度。
- **COV&TRAN**：在测试集中采用 COV 计算单语言相似度，利用翻译概率计算跨语言相似度。

表 2.3　不同相似度下话题分析系统的 F-measure

系统 \ 数据集	TDT41	TDT42
HN	0.683	0.697
SOCPMI ˆ DEV	0.733	0.716
SOCPMI &DIC	0.734	0.728
SOCPMI &TRAN	**0.752**	**0.738**
COV ˆDEV	0.727	0.688
COV &DIC	0.738	0.700
COV &TRAN	0.744	0.732

七种相似度的的跨语言话题分析系统在其他方面都采用相同的配置：采用 GVSM 表示文档，Bisecting K-Means 进行聚类。根据我们的实验，相似度的阈值统一设为 0.4。实验结果如表 2.3。

实验结果分析

1) 我们从表2.3中可以看出，基于 HowNet 的相似度性能远远低于其他系统。我们发现有些词在 HowNet 中语义相似度特别高。比如说，词“美联储”和“银行”的在 HowNet 的相似度为 1。错误分析显示，基于 HowNet 的跨语言相似度更加注重的是给定词的语义性质而非语义本身。因此，对于语义性质接近的两个词，即使它们的语义相关性比较低，HowNet 仍然给这两个词很高的相似度。这显然不利于话题聚类。因此，虽然基于 HownNet 的词相似度取得的很好的效果，但是由于话题分析这个任务更偏重于词的实际用法而非语义性质，因此，基于 HowNet 的跨语言词相似度在话题分析这个任务性能比较低。

2) 我们还发现当单语言采用相同的相似度方法时，基于翻译概率的系统要高于基于词典的系统。SOCPMI&TRAN 和 COV&TRAN 的性能分别优于 SOCPMI&DIC 和 COV&DIC 的性能。这是因为有两个值得注意的原因：首先，从 HowNet 抽取的词典比开发集的未登录词要多。其次，翻译概率在翻译歧义存在

的时候更具有区分性，翻译概率大的翻译对更常出现，在词相似度计算的时候权重也高。

3) 从表2.3中可以看出，SOCPMI&TRAN 在两个数据集上的结果都优于 SOCPMI^DEV 的结果，同时 COV&TRAN 的性能也要好于 COV^DEV。这些系统都是从开发集中获得跨语言信息，唯一的区别在于基于翻译概率的方法在测试集上计算单语言词相似度。因此我们可以总结出结合测试集上的词相似度信息和翻译概率在话题分析中是比较有效的。

4) 在所有的七个系统中，SOCPMI&TRAN 的性能最好。因此我们下面的实验都采用 SOCPMI&TRAN 作为我们的相似度计算方法。

2.5.5　实验 #2：不同的特征选择

这个实验旨在比较我们提出的特征选择方法与现有方法的区别。我们实现了三个话题分析系统。

- **SFS**：在 GVSM 中使用我们提出的基于软词频，软文档频和软匹配策略的特征选择方法。
- **HFS**：在 GVSM 中使用传统的 TF-IDF 以及硬匹配策略的特征选择方法。
- **NFS**：GVSM 中不使用特征选择方法，与实验一中的 SOCPMI&TRAN 系统相同。

表 2.4　采用不同特征选择的话题分析系统的 F-measure

系统 \ 数据集	TDT41	TDT42
SFS	**0.762**	**0.748**
HFS	0.755	0.740
NFS	0.752	0.738

四个测试集的实验结果如表2.4所示。

实验结果分析

从表2.4中可以看出，使用基于软匹配策略的系统比使用基于硬匹配策略的系统的性能高，不使用特征选择的系统的性能最差。这意味着特征选择对 GVSM 是有效的。使用软词频 TF^s 和软词频 DF^s 可以选择在特征集中最有效的词，经过合适的文档表示，文档可以有效地映射到特征集中。

2.5.6　实验 #3：不同的文档表示模型

这个实验旨在比较 CLGVSM 与现有文档表示的性能。

- **VSM&DIC**：采用 VSM 表示文档，余弦相似度计算文档相似度。并从知网获得词汇翻译信息。
- **VSM&TRAN**：采用 VSM 表示文档，余弦相似度计算文档相似度。并从平行语料翻译概率获取翻译信息。
- **PTM**：使用文献 [78] 的方法。用 PTM 在平行语料中获取跨语言主题，然后再测试集中推测主题。主题个数设为 1000。
- **CLGVSM**：我们提出的模型。与实验二的 SFS 系统相同。

表 2.5　采用不同文档表示模型的话题分析系统的 F-measure

系统 \ 数据集	TDT41	TDT42
VSM&DIC	0.730	0.721
VSM&TRAN	0.751	0.732
PTM	0.493	0.482
CLGVSM	**0.762**	**0.748**

四个测试集的实验结果如表2.5所示。

实验结果分析

1) 从表2.5中可以看出，在两个数据集上 CLGVSM 的性能好于 VSM&DIC 和 VSM&TRAN 。这意味着通过使用 SOCPMI&TRAN 相似度，CLGVSM 可以改进跨语言话题分析。词相似度对跨语言话题分析是有贡献的。通过使用词的二阶共现信息，SOCPMI 使得与相同词共现的词对有更好的相似度。这样符合词相似度计算的实际要求。比如说，“犯罪分子”和“imprisonment”（监禁）在 SOCPMI 算法中的相似度为 0.49。因此，当“犯罪分子”被选为特征时，即使文档中没有相同的词，含有“imprisonment”的文档与含有“犯罪分子”的文档也会有一个合理的相似度。

2) CLGVSM 比 PTM 的性能好。这说明了方法 CLGVSM 更适合跨语言文档聚类。分析原因如下：PTM 所构建的语义空间是在固定的平行语料中构建的，因此它没有考虑到目标聚类集的特征的重要性。相比之下，方法 CLGVSM 充分利用了测试集的信息构建语义空间。

2.6　本章总结

针对跨语言歧义问题，本章利用了跨语言词相似度计算拓展了单语言 GVSM，即 CLGVSM，同时比较了不同词相似度对 CLGVSM 的影响。

针对特征选择的硬匹配问题，本章提出了基于 CLGVSM 的新特征选择算法并通过实验证明了这个算法的有效性。

最后本章比较了 CLGVSM 和现有文档表示方法在跨语言话题分析的性能，实验证明，CLGVSM 的性能优于现有文档表示方法。

本章得出三个结论：

1) 基于知网的词相似度方法不利于跨语言话题分析。

2) 从开发集中获得的翻译概率比知网词典更有利于处理跨语言翻译。

3) 结合测试集中的词相似度以及开发集的翻译概率在跨语言话题分析中可以获得较好的性能。

CLGVSM 虽然可以在一定程度上解决跨语言问题和同义问题，在处理翻译歧义的时候加重了常用翻译的概率，但是它并不能根据词的不同含义提供不同的翻译，因此在处理多义问题和单语言歧义上有很大的局限性。因此，下章将采用词义作为语义表示，提出基于词义的跨语言文档建模方法。

第 3 章　基于全局词义的跨语言文档建模方法

3.1　本章引论

3.1.1　研究问题

跨语言广义向量空间模型利用了跨语言词相似度，可以在一定程度上解决跨语言问题和同义词问题。比如词“arm”既可以被翻译成“胳膊”，也可以被翻译成“手臂”。在跨语言广义向量空间模型中，词“arm”跟词“胳膊”以及词“手臂”的相似度为 1。这样，词“arm”和词“胳膊”和词“手臂”都有对应，跨语言广义空间向量模型可以通过词相似度反映跨语言歧义问题。

但是对于多义词，比如词“arm”在描写人体的上下文中需要翻译成“手臂”，在描写军事的上下文中需要翻译成“武装”。在跨语言广义向量空间模型中，词“arm”对“手臂”和“武装”的相似度在任何上下文都相同，无法区分“arm”的不同含义。因此，在文档建模的时候对于多义词的翻译，需要根据词表达的不同含义选择不同的翻译，这就是跨语言话题分析中的单语言歧义问题。而跨语言广义空间向量模型通过词相似度只反映了同义词现象，无法识别词的多种含义，也就是无法解决单语言歧义问题。

因此，本章主要同时针对跨语言歧义和单语言歧义展开研究。这就要求文档建模方法可以同时识别一个词在表达同一含义时的不同翻译选择和一个词在表达不同含义的不同翻译。

3.1.2　问题分析

跨语言文档建模需要对词的相同含义的不同翻译选择有相同或者相近的表示，解决跨语言歧义；对词的不同含义选择恰当的翻译，解决单语言歧义。现有的研究利用语义空间进行文档建模。

语义空间有显语义空间和潜语义空间两类。

基于显语义空间的文档建模方法通常是利用维基百科等语义资源的概念表示文档[52,87]。这种方法需要语义资源的支持，受到语义资源的规模和更新速度的影响。

潜语义空间方法有两种：第一种是基于主题对齐[71,74,76–78,81]。这种方法在平行/可比语料中构造主题空间，而主题通常是和文档息息相关的，不同文档集之间

的主题差异性较大。目标语料中的主题可能在平行/可比语料中并未涉及，同时平行/可比语料中的冗余主题会产生噪音。因此，这种主题偏差会严重影响方法的性能。第二种是基于利用词对齐信息直接在文档集中构造跨语言语义空间[82–85]。但是现有的跨语言主题模型考虑文档层面上的上下文，解决翻译歧义问题的效果并不理想。

因此，跨语言文档建模在获取语义的时候需要注意不同文档集的差异性；在消除歧义的时候需要考虑更具体的上下文，更好地解决单语言歧义问题。

3.1.3　解决思路

在1.3.2中提到，翻译歧义的两种情况分别来源于同义现象和多义现象。本章将这两种情况分别等同于同义问题和多义问题。同时把词和它在不同语言上的翻译也看做同义问题。为了更好的解决同义问题和多义问题，本章采用词义这一语义表示，提出基于词义的文档建模方法。之前有研究利用词义改进文档聚类[56], 文本分类[115] 以及信息检索[116] 。但是这些研究都是基于语义资源（如 WordNet 或者 Wikipedia）。本文的词义则是用词义归纳方法从语料中自动抽取出来的。但是现有的词义归纳的方法只能提取单语言的词义，无法处理跨语言情况。为此本文重新定义了词义的建模方法，并且提出了跨语言词义归纳算法。

本章采用两种文档表示模型：基于词义的 VSM 和基于词义的 LDA。优势在于：

1) 同义问题可以通过词义解决。同种语言中的同义词属于同一个词义。对于跨语言情况，不同语言相同含义的词也属于同一个词义。这样，基于词义的跨语言文档建模更加准确。

2) 多义问题也可以通过词义解决。跨语言情况下，多义词可以通过不同上下文的选择不同的翻译。因此，通过词义消歧可以更好的表示跨语言文档。同时，本章考虑了句子级别的上下文，与考虑文档层面上下文的研究[82–85] 相比更有利于消除歧义。

因此，基于词义的跨语言文档建模比特征为词的文档建模具有优越性。同时与基于主题对齐的方法相比，词义可以获得粒度更小的语义信息，与文档相关性小，具有更好的文档鲁棒性。

本章剩余部分按照如下方式组织：3.2节介绍跨语言词义的定义以及归纳算法；3.3 节介绍基于词义的文档建模方法；3.4节总结了基于词义的文档建模整体流程和在跨语言话题分析的应用。3.5节评测了基于词义的跨语言话题分析的性能。3.6节对本章进行了总结。

3.2　跨语言词义

3.2.1　词义的相关研究

计算机应用中对词义的表示通常有两种，一种是基于解释的定义，通常表示为词典资源提供的词条信息，如 HowNet 中的义元等。另一种是基于归纳的定义。利用词的上下文归纳词义，通常表示为一组意思相关的词。基于归纳的定义通常在词义归纳任务（Word Sense Induction, WSI）中使用。这种表示是基于由 Harris 提出的分布性假设[117]：“词语的意义可以从他周围的词语获知”（A word is characterized by the company it keeps）[118]。显而易见，基于解释的定义难以获得，而基于归纳的定义却十分便于计算。这是本章采取基于归纳定义词义的原因。有关词义的研究成果十分丰富，这里只阐述与本章有关的词义相关研究，包括词义归纳、词义消歧。

3.2.1.1　词义归纳

词义归纳（Word Sense Induction，WSI）是指通过归纳办法获得词义，尤其以无指导词义归纳最具吸引力。无指导词义归纳的目标是从无标注语料中自动获取多义词的不同意义，可归结为聚类问题，一个词通常根据它的不同语言上下文划分为不同的类簇，每个类簇都代表这个词的一个词义。Sense Cluster[62] 是一个基于上下文的词义区分系统，也用于词义归纳。文献 [119] 提出了一种基于词聚类的词义归纳方法。该方法选用了词在文本中的共现词作为特征，利用 CBC（Clustering By Committee）聚类算法进行词义归纳。文献 [120] 则采用共现词对代替共现词作为特征进行聚类。文献 [121] 在小范围语料中利用词的自拓展技术改进词义归纳算法，首先通过互信息计算出一个共现词表，然后用共现词表代替共现词进行聚类。文献 [122] 则直接对上下文进行聚类。文献 [123,124] 采用基于图的方法归纳词义。文献 [125] 采用贝叶斯的方法，假设目标词的上下文中的词是从一个多项分布抽样得出的，他们拓展 LDA 模型，利用多种特征的多层模型，性能要好于 semEval 2007 的评测[126] 的最好系统。文献 [127] 采用层次狄利克雷过程（Hierarchical Dirichlet Process，HDP）[128] 模型进行词义归纳，实验证明 HDP 模型在词义归纳任务的性能好于 LDA。这是因为 LDA 需要人工指定主题个数而 HDP 可以从语料中学习出合适的主题个数。文献 [129] 则利用平行语料中的词语对齐信息归纳词义。他假设具有等价翻译的词含有相同的意思，但是这个并不符合实际情况。一个词在两个不同的语境下含有不同的词义，但它的翻译可能相同。文献 [130] 则提出了一个基于层次聚类的中文词义归纳方法。

词义归纳最早用于信息检索，文献 [65,66] 在信息检索领域采用无指导词义归

纳技术确定词义。

本章着眼于如何在平行语料中进行词义归纳。参考文献 [125,127] 的思想，本章提出针对平行语料提出一个新的贝叶斯词义归纳模型（CL-HDP），在平行语料中归纳词义。

3.2.1.2　词义消歧

词义消歧的目标是确定一个词在特定上下文中的词义，是自然语言处理的一个经典研究方向，国内外相关研究非常广泛。词义消歧方法大致分为三类：基于规则的词义消歧方法、基于词典资源的词义消歧方法和基于语料库的词义消歧方法，基于语料库的方法又分有指导和无指导方法。本章涉及的词义消歧以文档建模为目标，为文档中的每个词汇确定一个预先归纳获得的词义，是限定词义消歧，因此并不需要训练资源。与本项目相关的词义消歧方法是无指导统计方法。文献 [131] 将训练语料中歧义词的上下文聚成若干个类, 每个类别代表一个抽象词义，词义的识别和判断在这些类别里进行。文献 [63] 利用 Sense Clusters 工具实现无指导词义消歧。无指导词义消歧与无指导词义归纳的任务十分类似，有时可以采取相同的方法。

早期的一些研究在信息检索中采用词义消歧的技术[132,133]，但实验结果并不理想，这是由于查询词长度较短，词义消歧精度比较低。一些研究在文本分类任务使用词义消歧技术。文献 [134] 利用词义消歧技术加强领域特征的重要性。文献 [56] 利用 WordNet 中的概念表示文档并采用了词义消歧技术确定概念。文献 [115] 则利用 SUMO/MILO 概念表示文本并采用词对齐和词义消歧技术确定概念进行文本分类。

还有一些研究通过平行语料进行跨语言词义消歧的工作。1991 年，文献 [135] 指出两种语言包含的信息比一种语言多。文献 [136] 使用第二种语言帮助词义消歧。文献 [137] 在语言数据协会（LDC）提供的汉英双语语料上，采用 Giza++ 进行词对齐，然后对译文进行人工选择，译文对源语言词的作用相当于词义标注。在该文献中，词义分类器使用贝叶斯模型构造，英文词义采用 WordNet 的标准定义。文献 [138] 将文献 [62] 的方法拓展到双语，并证明了比单语的效果好。文献 [139] 提出了一个无指导的方法，他的系统首先利用 Giza 程序自动对齐，生成多种语言的词义标注语料，为多语种的词义消歧提供了框架。文献 [140] 充分利用了 WordNet 的语义和概念体系来确定语义模型和概念模型的结构,。实验结果表明文献 [140] 建立的语义模型在词义消歧上比文献 [139] 的 SALAAM 系统的性能好，而概念模型优于语义模型。上述工作都是处理来自知识库（WordNet）的词义，与本章的词义来源不相同。

本章从语料中采用贝叶斯方法产生词义，在词义消歧的时候也使用贝叶斯的方法，选择概率大的作为词义。

3.2.2　跨语言词义的定义

3.2.2.1　局部词义

定义 3.1：　特定词 w 在特定语言 l 上的局部词义 s_w^l 可以统计地表示为在特定语言 l 中的一组上下文的词的概率分布,

$$s_w^l = \{c_i^l : p(c_i^l|s_w)\}; \quad i = 1...N \tag{3-1}$$

其中 s_w^l 表示词 w 在特定语言 l 上的局部词义，c_i^l 表示语言 l 中的上下文词，$p(c_i^l|s_w)$ 表示上下文词 c_i^l 对于词义 s_w^l 的概率，即给定词义 s_w^l ，词 c_i^l 出现在上下文中的概率。

前人的研究[52,56,87,115,116] 通常依赖于本体来获取词义。但是本体的构建需要耗费巨大的人力和时间。本章采用上下文词以及它们的概率表示词义。比如词“arm”，它的两个词义可表示为例3.1和例3.2。“arm”的词义包括它有关的上下文词和概率。上下文词的概率可以通过词义归纳算法在文档集中估算出。一个词可以有多个词义，因此通过局部词义可以解决多义问题。

例 3.1：　arm# 1={ limb: 0.159, forelimb: 0.069, sleeve: 0.019 }

例 3.2：　arm# 2={weapon: 0.116, war: 0.039, battle: 0.026 }

3.2.2.2　跨语言局部词义

为了处理跨语言情况，我们拓展局部词义的定义，使它可以处理多语言的上下文。多语言的上下文可以从平行语料抽取出来。

定义 3.2：　特定词 w 的跨语言局部词义 s_w 可以统计地表示为多种语言中的一组上下文的词的概率分布,

$$s_w = \begin{bmatrix} \{c_i^{l_1} : p(c_i^{l_1}|s_w)\}, & i = 1...N_{l_1} \\ ... & \\ \{c_j^{l_1} : p(c_i^{l_L}|s_w)\}, & j = 1...N_{l_L} \end{bmatrix} \tag{3-2}$$

其中 $c_i^{l_k}$ 是在语言 l_k 中的上下文词，$p(c_i^{l_k})$ 是语言 l_k 中的上下文词 $c_i^{l_k}$ 对于词义 s_w 的概率。

例如，词“arm”在中英文两种语言的情况下的两个跨语言局部词义可表示为例3.3和例3.4。

例 3.3：　arm# 1={limb: 0.159, forelimb: 0.069, sleeve: 0.019; 手臂: 0.137, 上肢: 0.079, 衣袖: 0.017 }

例 3.4：　arm# 2={weapon: 0.116, war: 0.039, battle: 0.026; 装备: 0.153, 武器: 0.027; 战争: 0.026 }

跨语言局部词义可以通过跨语言词义归纳算法（Cross-lingual Word Senses Induction，CL-WSI）在平行语料中获得。从上例可以看出，跨语言局部词义可以在跨语言的情况下处理多义问题。但是跨语言词义是对于每个词单独归纳的。这样会导致大量同义词义存在。为此，我们进一步定义全局词义，表示所有词的词义。

3.2.2.3　跨语言全局词义

定义 3.3：　一个全局词义 g 指的由一组同义的局部词义组成，

$$g = \{s_w^j\};\ \ j = 1...M \tag{3-3}$$

其中 s_w^j 表示布局词义。

如果局部词义是跨语言的，那么全局词义自然也可以处理跨语言情况。本章的全局词义的通过如下流程得到：将上下文的词作为特征，概率作为权重，计算局部词义的相似度，然后通过聚类算法得到全局词义。同样，包含“arm”的词义的全局词义如例3.5和例3.6所示。

例 3.5：　g#1={ arm#1, 手臂 #1 } ={
{limb: 0.159, forelimb: 0.069, sleeve: 0.019; 手臂: 0.137, 上肢: 0.079, 衣袖: 0.017},
{arm: 0.189, forelimb: 0.058, sleeve: 0.025; 胳膊: 0.159, 上肢: 0.089, 衣袖: 0.014}
}

例 3.6：　g#2={ arm#2, weapon#1, 装备 #1 } ={
{weapon: 0.116, war: 0.039, battle: 0.026; 装备: 0.153, 武器: 0.027; 战争: 0.026 },
{arm: 0.12, battle: 0.04, war: 0.016; 装备: 0.133, 武器: 0.035; 战士: 0.028 }, {arm: 0.14, weapon: 0.12, war: 0.016; 装备: 0.133, 战争: 0.035; 战士: 0.028 }
}

从例3.5和例3.6可以看出，局部词义 arm#1 和手臂 #1 都属于全局词义 g#1 。这是因为局部词义 arm#1 和手臂 #1 上下文分布是相似的。不同语言的同义词义通过这种方式聚到一个全局词义中，可以解决同义问题。下面我们将具体介绍如何从平行语料中获得跨语言词义。

3.2.3　跨语言词义的生成

跨语言词义的生成分为两步：1) 从平行语料中归纳跨语言局部词义。2) 在跨语言局部词义上生成全局词义。下面将一一介绍。

3.2.3.1　局部词义的归纳

本章采用贝叶斯方法进行归纳局部词义。具体来讲，本章拓展了 HDP 模型，使其可以处理跨语言情况。本文称这种模型为跨语言 HDP 模型（CLHDP）。为了介绍 CLHDP，先简要的介绍用传统的 HDP 模型及其在词义归纳上的应用。

HDP 模型

层次狄利克雷过程最早是由文献 [128] 提出来，由文献 [127] 应用到词义归纳中。这里主要针对词义归纳任务进行介绍。

用 HDP 进行词义归纳时，对每个需要归纳词义的词都要建立 HDP 模型。这里我们定义需要归纳词义的词为目标词。同时还定义一个上下文窗口，上下文窗口中的词影响目标词的词义。我们称目标词的上下文窗口为上下文，称目标词的上下文窗口中的词为目标词的上下文词。

HDP 是一种生成模型，可以随机生成可观测的数据。对于目标词 w 的每个上下文 v_i，上下文词 c_{ij} 的词义 s_{ij} 服从先验分布 G_i 。G_i 是从全局随机分布 G_w 中抽样出来的。词义 -上下文词分布 η_{s_w} 是从分布 H_w 中随机生成的: H_w:$\eta_{s_w} \sim H_w$。H_w 是超参数为 ϵ_w 的狄利克雷分布。一个目标词 w 的 HDP 生成过程如下所示：

1. 选择 $G_w \sim DP(\gamma_w, H)$.
2. 对每个上下文 v_i 中的词 w:
   (a) 选择 $G_i \sim DP(\rho_w, G_w)$.
   (b) 对目标词 w 的上下文词 c_{ij}:
      i. 选择 $s_{ij} \sim G_i$.
      ii. 选择 $c_{ij} \sim Mult(\eta_{s_{ij}})$.

超参数 γ_w 和 ρ_w 是狄利克雷过程（Dirichlet Process，DP）的聚集度参数，分别控制 G_w 和 G_i 的可变性。HDP 的概率图表示可见图3.1，其中可见变量，即上下文词 c_{ij} 用阴影表示。HDP 的过程可以用折棍（strick-breaking）构造和中国餐馆过程（Chinese Restaurant Franchise）构造。具体构造过程可参见文献 [128]

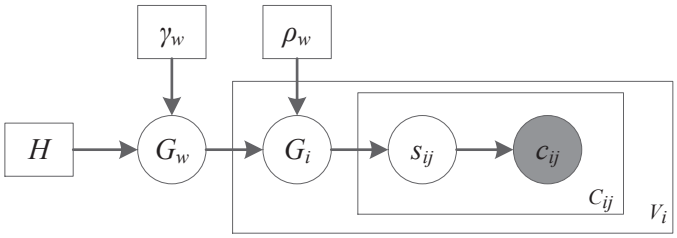


图 3.1　HDP 模型

CLHDP 模型

CLHDP 模型通过跨语言元组对词义进行建模。每个元组（tuple）代表一组含义相同语言不同的上下文。针对 CLHDP，我们做出了两个假设：首先，一个元组中的上下文共享相同的词义分布；其次，每个词义对每个元组的上下文词有一个分布。也就是说，CLHDP 并不象 HDP 那样，对每个词义 s_w 有一个上下文词分布 η_{s_w}，而是有 L 个语言相关的词义 -上下文词分布（$\eta^1_{s_w},\cdots,\eta^L_{s_w}$）。这 L 个语言相关的上下文词分布是从语言相关的先验分布 H^l_w 抽样出来的。H^l_w 是超参数为 λ^l_w 狄利克雷分布。

一个目标词 w 的 CLHDP 生成过程如下所示：

1. 选择 $G_w \sim DP(\gamma_w, H)$.
2. 对词 w 的每个上下文 v_i :
   (a) 选择 $G_i \sim DP(\rho_w, G_w)$.
   (b) 对目标词 w 的在语言 l 的上下文词 c^l_{ij}:
      i. 选择 $s^l_{ij} \sim G_i$.
      ii. 选择 $c^l_{ij} \sim Mult(\eta^l_{s_{ij}})$.

超参数 γ_w 和 ρ_w 是狄利克雷过程（Dirichlet Process，DP）的聚集度参数，分别控制 G_w 和 G_i 的可变性。CLHDP 的概率图表示可见图3.2，其中可见变量，即上下文词 c^l_{ij} 用阴影表示。

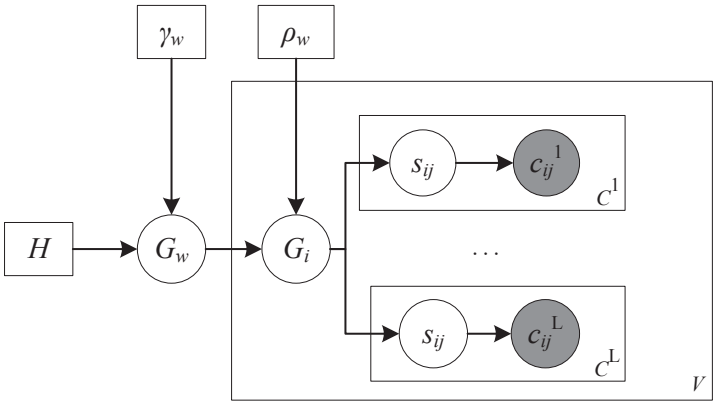


图 3.2　CLHDP 模型

CLHDP 模型的参数推导

文献 [128] 采用折叠吉布斯抽样（Collapsed Gibbs Sampling）学习 HDP 模型，介绍了三种对 HDP 的采样方法：基于中国餐馆过程（Chinese Restaurant Franchise）的后验采样方法；基于增强表示（Augumented Reresentation）的后验采样方法和直接分配后验采样方法。前两种的实现相对比较繁琐。本章主要参考后面一种采样方法，吉布斯抽样只对模型中的隐变量进行抽样。在初始化的时候随机选取隐变量，然后在每一轮迭代中，隐变量顺序地从给定该变量以外的其他变量的分布进行抽样。

本章的抽样过程分为三步：

1) 给定 $\boldsymbol{s} = \{s_{ij}^l\}$ 和词义 -上下文词分布 $\{\eta_{s_w}^l\}$，抽样 G_w。这个抽样过程和文献 [128] 类似。

由于每个词的先验概率 G_w 是一个狄利克雷过程（Dirichlet Process, DP），$G_w \sim DP(\lambda_w, H_w)$，$G_w$ 可以用折棍（strick-breaking）构造。具体构造过程如公式 (3-4)。

$$G_w = \Sigma_{s_w=1}^{\infty} \pi_w^{s_w} \delta_{(\eta_{s_w}^1, \cdots, \eta_{s_w}^L)} \tag{3-4}$$

其中 $\eta_{s_w}^1, \cdots, \eta_{s_w}^L$ 分别从 $H_w^1, \cdots, H_w^L$ 中生成，代表词义 s_w 在语言 $1, \cdots, L$ 上的词义 -上下文词分布。这个分布在步骤 1) 中是已知的。$\{\pi_w^{s_w}\}$ 是词义的混合参数。它们是从折棍（strick-breaking）构造中抽样出来的。在抽样过程中，假设目标词 w 已经有 S_w 个词义。由于词义 -上下文词分布 $\{\eta_{s_w}^l\}$ 和每个上下文词的词义 $\boldsymbol{s} = \{s_{ij}^l\}$ 已知，G_w 可以用公式 (3-5) 表示。

$$G_w = \Sigma_{s_w} \pi_w^{s_w} \delta_{(\eta_{s_w}^1, \cdots, \eta_{s_w}^L)} + \pi_w^u G_w^u \tag{3-5}$$

其中 G_w^u 是狄利克雷过程。$G_w^u \sim DP(\gamma_w, H_w)$。

因此，G_w 依赖于 $\boldsymbol{\pi_w} = \{\pi_w^{s_w}\}$。下面介绍 G_w 中的 $\{\pi_w^{s_w}\}$ 的抽样过程。根据文献 [128]，$\{\pi_w^{s_w}\}$ 可根据中国餐馆过程中餐桌数进行抽样。具体抽样公式参见公式 (3-6)。

$$(\pi_w^1, \cdots, \pi_w^{S_w}, \pi_w^u) \mid \{\eta_{s_w}^l\}, \boldsymbol{s} \sim \mathrm{Dir}(m_{\cdot 1}, \cdots, m_{\cdot S_w}, \gamma_w) \tag{3-6}$$

其中 m_{kj} 表示餐馆 k 中点了菜 j 的餐桌数，$m_{\cdot j}$ 表示所有餐厅中点了菜 j 的餐桌数。$\boldsymbol{m} = \{m_{kj}\}$ 的抽样将在步骤 3 中介绍。

2) 给定 G_w，抽样 $\boldsymbol{s} = \{s_{ij}^l\}$。假设上下文 v_i 在语言 l 上的上下文词 $c_{ij}^l = c$，s_{ij}

的抽样概率用公式 (3-7) 计算。

$$P(s_{ij}=s_w,|\boldsymbol{s_{-ij}},\boldsymbol{c_i})=\begin{cases}(n^{v_i}_{-ij,s_w}+\rho_w\pi^{s_w}_w)\frac{n^c_{-ij,s_w}+\lambda^l_w}{n_{-ij,s_w,l}+V_{l,w}\lambda^l_w} & \text{如果}s_w\text{已经出现过}\\ \rho_w\pi^u_w\frac{n^c_{-ij,s_w}+\lambda^l_w}{n_{-ij,s_w,l}+V_{l,w}\lambda^l_w} & \text{如果}s_w=s^{\text{new}}_w\end{cases} \tag{3-7}$$

其中 n^c_{-ij,s_w} 是上下文词 c 的词义是 s_w 的个数，$V_{l,w}$ 是目标词 w 在语言 l 上的上下文词个数。$n_{-ij,s_w,l}$ 是语言 l 中词义是 s_w 的上下文词的个数，$n^{v_i}_{-ij,s_w}$ 是上下文 v_i 中词义是 s_w 的词的个数，上面变量中的 $-ij$ 在本文中均表示计数时排除词 c_{ij}。

3) 给定 G_w、$\boldsymbol{s}=\{s^l_{ij}\}$，抽样 $\boldsymbol{m}=\{m_{kj}\}$。根据文献 [128]，$\boldsymbol{m}=\{m_{kj}\}$ 按照公式 (3-8) 进行抽样。

$$p(m_{kj}=m|\boldsymbol{s},\boldsymbol{m_{-kj}},\boldsymbol{\pi_w})=\frac{\Gamma(\rho_w\pi^{s_w}_w)}{\Gamma(\rho_w\pi^{s_w}_w+n_{k.j})}\mathrm{s}(n_{k.j},m)(\rho_w\pi^{s_w}_w)^m \tag{3-8}$$

其中 $n_{k.j}$ 表示餐厅 k 点了菜 j 的餐桌上的顾客数，$\mathrm{s}(n_{k.j},m)$ 为 Stirling 数。

经过一定的迭代次数后，吉布斯抽样可以估计出 $\boldsymbol{s}=\{s^l_{ij}\}$，据此可以计算出词义 -上下文词的分布 $\eta^l_{s_w}$，计算公式如公式 (3-9)。

$$\eta^l_{s_w}(c)=\frac{n^c_{s_w,l}+\lambda^l_w}{n_{s_w,l}+V_{w,l}\lambda^l_w} \tag{3-9}$$

其中 $n^c_{s_w,l}$ 是语言 l 中目标词 w 的上下文词 c 的词义是 s_w 的个数。$V_{w,l}$ 是语言 l 中目标词 w 的上下文词的个数。$n_{s_w,l}$ 是语言 l 中词义是 s_w 的上下文词的个数。

本章的上下文窗口设为一句话，在平行语料中抽取跨语言上下文。例如，当目标词出现在某句话中，我们将这句话和它在平行语料中对应的句子放入跨语言元组中。

3.2.3.2　全局词义的生成

我们将全局词义的生成看作聚类任务。它的目标是将相似度高的局部词义聚成一个全局词义。

本章用局部词义的上下文词的作为特征，然后用聚类算法合并相似的词义。对于跨语言局部词义，我们将所有语言的上下文词都合并到一个向量中。

我们采用下面两个算法聚类局部词义。

1) Bisecting K-Means。这个算法曾在2.5.1中用作了话题聚类。这里用作局部词义的聚类。Bisecting K-Means 是 K-Means 的拓展方法，研究证明它的性能优于标准的 K-Means 算法和层次聚类算法。它的基本思想是，假定要得到 k 个簇，将所有点的集合用 k-means 算法分裂成两个簇，从这些簇中选取一个继续分裂，如此下去，直到产生 k 个簇。

2) 基于图的聚类。这个算法主要是基于图分割的算法。它首先将数据在最近邻图上建模，然后将用最小切割算法将图分为 k 个簇。

3.3　基于词义的文档建模

3.3.1　跨语言词义消歧

本章采用词义归纳相同的模型 (CLHDP) 进行词义消歧。假设文档集 $D = \{d_j; j = 1, \cdots, N\}$ 表示包含 N 篇文档，M 个词的文档集。我们首先在 D 中以句子作为上下文窗口，抽取每个词的上下文集，然后在每个上下文集中分别估算词义的分布。

假设语言 l 上的词 w 在平行语料中训练出来的词义 -上下文词分布为 $\eta^l_{s_w}$。在文档集 D 的上下文集 $\{\hat{v}_i; i = 1, \cdots, \hat{V}_w\}$ 中具体的参数推导过程与3.2.3.1相同。唯一的区别是在步骤 2) 中，给定 G_w，抽样 $\boldsymbol{s} = \{s^l_{ij}\}$ 时采用公式 (3-10)。

$$P(s_{ij} = s_w, |\boldsymbol{s_{-ij}}, \boldsymbol{c_i}) = (\hat{n}^{\hat{v}_i}_{-ij,s_w} + \hat{\rho}_w \hat{\pi}^{s_w}_w)\eta^l_{s_w}(c_{ij}) \tag{3-10}$$

其中 $\hat{n}^{\hat{v}_i}_{-ij,s_w}$，$\hat{\rho}_w$，$\hat{\pi}^{s_w}_w$ 是 CLHDP 在上下文集 $\{\hat{v}_i; i = 1, \cdots, \hat{V}_w\}$ 相应的参数。

经过一定的迭代次数后，可以估计出上下文集 $\{\hat{v}_i; i = 1, \cdots, \hat{V}_w\}$ 中每个上下文的词义分布。上下文 $\hat{v}$ 的词义分布 $\theta_{\hat{v}}$ 的计算公式如公式 (3-11)。

$$\theta_{\hat{v}}(s_w) = \frac{\hat{n}^{\hat{v}}_{s_w} + \hat{\rho}_w \hat{\pi}^{s_w}_w}{\hat{n}^{\hat{v}} + \hat{\rho}_w \sum_{s'_w} \hat{\pi}^{s'_w}_w} \tag{3-11}$$

其中 $\hat{n}^{\hat{v}}_{s_w}$ 是上下文 $\hat{v}$ 中词义是 s 的上下文次的个数，$\hat{n}^{\hat{v}}$ 是上下文 $\hat{v}$ 中词的个数。

获得 $p(s_w|\hat{v})$ 即 $\theta_{\hat{v}}(s_w)$ 后，我们将概率最大的词义作为词 w 在上下文 $\hat{v}$ 中的词义。比如，例3.10，例3.11和例3.12给出了三句话。

例 3.7：　S_1: That man with one arm lost his other limb in an airplane crash.

例 3.8：　S_2: The nation must arm its soldiers for battle.

例 3.9：　S_3: 国家必须为了战争武装它的士兵。

经过分词、去停用词以及词形还原后，这三句话变成了例3.10，例3.11和例3.12的形式。

例 3.10：　$\overline{S_1}$: man arm lost limb airplane crash

例 3.11：　$\overline{S_2}$: nation arm soldier battle

例 3.12：　$\overline{S_3}$: 国家战争武装士兵

经过词义消歧，词义”arm# 1“在句子 S_1 的概率是 0.998005，而词义”arm# 2“在句子 S_2 的概率是 0.944096。所以词”arm“在句子 S_1 中的词义是”arm# 1“，在句子 S_1 中的词义是”arm# 2“。由于”arm# 1“属于全局词义”g# 1“，”arm# 2“属于全局词义”g# 2“，词”arm“在句子 S_1 中的词义是”g# 1“，在句子 S_1 中的词义是”g# 2“。对于句子 S_3，词”武装“的词义也是”g# 2“。这样，可以识别出多义词的不同含义和不同词的同义情况，翻译歧义和跨语言问题可以同时解决。

词义消歧后，我们在两种流行的文档表示模型 VSM 和 LDA 中使用词义作为特征，分别提出基于词义的 VSM（Sense based VSM，SVSM）和基于词义的 LDA（Sense based LDA，SLDA）。

3.3.2　基于词义的 VSM

在介绍基于词义的 VSM 之前，先简要的介绍一下 VSM。

3.3.2.1　VSM

VSM 是一种简单高效的文档表示模型，把文档解析成特征的集合，通过权重公式刻画这些特征，从而把文档表示成向量。在向量空间模型中，$D = \{d_i; i = 1, \cdots, N\}$ 表示包含 N 篇文档的文档集。文档集的一篇文档 d_i 定义为：$d_i = \{w_{ij} : r_{ij}\}_{j=1...M^{d_i}}$。$r_{ij}$ 表示文档 d_i 中第 j 个特征词 w_{ij} 的权重，M^{d_i} 是文档 d_i 中特征词的总数。计算权重最常用的方法是 TF*IDF 公式。TF*IDF 的 $TF(w_{ij}, d_i)$ 因子是特征词 w_{ij} 在文档 d_i 中的词频。$DF(w_{ij})$ 是文档集中含有 w_{ij} 的文档的个数。则 TF*IDF 公式可以使用公式 (3-12) 表示。

$$r_{ij} = TF(w_{ij}, d_i) \cdot IDF(w_{ij}); IDF(w_{ij}) = log\left(\frac{N}{DF(w_{ij})}\right) \tag{3-12}$$

文档之间的相似度可以用两个向量之间的距离来度量。余弦公式（cosine similarity）是应用最广泛的计算相似度的方法。两个文档之间相同的词越多同时这些词的权重越高，则相似度越高。余弦公式如公式 (3-13) 所示。

$$sim(d_1, d_2) = \frac{\sum_{j=1}^{M}(r_{1j} \times r_{2j})}{\sqrt{\sum_{j=1}^{M} r_{1j}^2} \times \sqrt{\sum_{j=1}^{M} r_{2j}^2}} \tag{3-13}$$

3.3.2.2　基于词义的 VSM

传统的 VSM 模型用词作为特征表示文档。SVSM 则使用全局词义作为特征表示文档。具体流程是：经过词义消歧，首先确定文档中每个词的词义，则文档

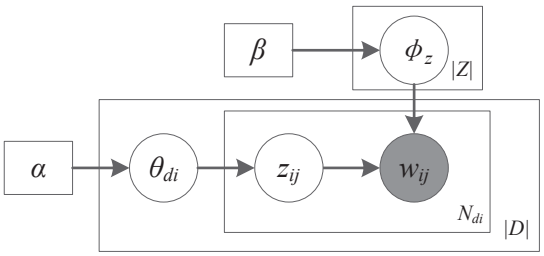


图 3.3　LDA 模型

d_i 可表示为：$d_i = \{g_{ij} : \hat{r}_{ij}\}_{j=1...M^{d_i}}$。同样可以计算词义 g_{ij} 在文档 d_i 中的词义频 $GF(g_{ij}, d_i)$ 和文档集中的文档频 $DF(g_{ij})^{\mathrm{g}}$，那么词义 g_{ij} 的权重按照公式 (3-14) 计算。

$$r_{ij} = GF(g_{ij}, d_i) \cdot IDF(g_{ij})^{\mathrm{g}}; IDF(g_{ij})^{\mathrm{g}} = log\left(\frac{N}{DF(g_{ij})^{\mathrm{g}}}\right) \tag{3-14}$$

最后得到每个文档的词义向量。采用余弦相似度可以计算出词义向量的相似度。

3.3.3　基于词义的 LDA

在介绍基于词义的 LDA 之前，先简要的介绍一下 LDA。

3.3.3.1　LDA

图3.3表示 LDA 的概率模型，其中可见变量即词 w_{ij} 用阴影表示。给定文档集 D、文档中的词集 W 以及 LDA 中的主题集为 Z，LDA 的概率生成模型如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 中的词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择词 $w_{ij} \sim Mult(\phi_{z_{ij}})$.

其中 d_i 表示文档集 D 中的第 i 个文档，w_{ij} 表示文档 d_i 的第 j 个词，z_{ij} 表示词 w_{ij} 的主题。α, β 是模型的超参数。$\phi_{z_{ij}}$ 和 θ_{d_i} 分别表示主题 -词分布和文档 -主题分布。这两个分布都有从狄利克雷分布抽样出来的。在图3.3中，N_{d_i} 表示文档 d_i 中词的个数。

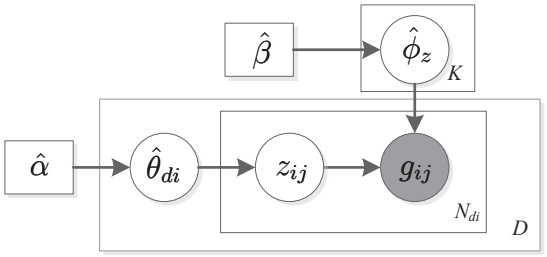


图 3.4　SLDA 模型

LDA 采用吉布斯抽样（Collapsed Gibbs Sampling）学习模型的参数[5]。在抽样过程中，词 $w_{ij}=w$ 的条件概率公式按照公式 (3-15) 计算。

$$P(z_{ij}=z|\boldsymbol{z_{-ij}},\boldsymbol{w}) \propto \frac{n_{-ij,z}^{d_i}+\alpha}{n_{-ij}^{d_i}+Z\alpha} \times \frac{n_{-ij,z}^{w}+\beta}{n_{-ij,z}+W\beta} \tag{3-15}$$

在公式 (3-15) 中，$n_{-ij,z}^{d_i}$ 表示文档 d_i 中主题为 z 的个数；$n_{-ij,z}^{w}$ 表示主题为 z 的词 w 的个数；$n_{-ij}^{d_i}$ 表示文档 d_i 的词的个数；$n_{-ij,z}$ 表示主题为 z 的词数。上面变量中的 $-ij$ 表示在计数过程中去掉词 w_{ij}。

3.3.3.2　基于词义的 LDA

我们使用词义代替词作为特征将 LDA 模型拓展为 SLDA。具体流程是：首先使用词义消歧算法得到每个词的词义。然后根据文档 -主题分布 $\hat{\theta}_{d_i}$ 生成文档的主题。对于文档中的词 w_{ij}，词义 g_{ij} 由词 w_{ij} 的主题根据主题 -词义分布 $\hat{\phi}_{z_{ij}}$ 生成，其中 $\hat{\phi}_{z_{ij}}$ 代表主题 z_{ij} 对词义的生成概率，它是从狄利克雷分布 $Dir(\hat{\beta})$ 生成的。

图3.4是 SLDA 的概率模型，其中可见变量即词义 g_{ij} 用阴影表示。SLDA 的概率生成模型如下：

1. 对于每个主题 z:
   (a) 选择 $\hat{\phi}_z \sim Dir(\hat{\beta})$.
2. 对于每个文档 d_i:
   (a) 选择 $\hat{\theta}_{d_i} \sim Dir(\hat{\alpha})$.
   (b) 对于文档 d_i 中的词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\hat{\theta}_{d_i})$.
      ii. 选择词义 $g_{ij} \sim Mult(\hat{\phi}_{z_{ij}})$.

SLDA 的参数推导过程和 LDA 的参数推导过程相同，唯一的区别是用词义代替了词作为特征。因此，在抽样过程中，词 w_{ij} 的词义 $g_{ij}=g$ 时，条件概率公式按照公式 (3-16) 计算。

$$P(z_{ij}=z|\boldsymbol{z_{-ij}},\boldsymbol{g}) \propto \frac{n^{d_i}_{-ij,z}+\hat{\alpha}}{n^{d_i}_{-ij}+Z\hat{\alpha}} \times \frac{n^{g}_{-ij,z}+\hat{\beta}}{n_{-ij,z}+G\hat{\beta}} \tag{3-16}$$

3.4　基于词义的文档建模总结以及在跨语言话题分析的应用

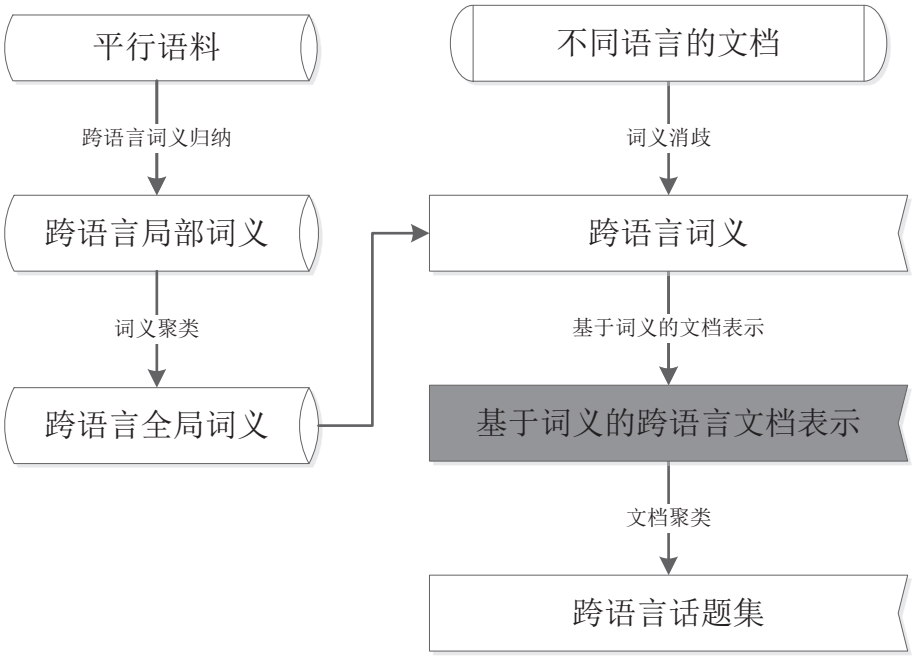


图 3.5　基于词义的跨语言话题分析流程

图3.5是基于词义的跨语言话题分析流程。首先用跨语言词义归纳算法从平行语料库中归纳出每个词的词义。这样我们可以获得局部词义（local sense）。然后使用聚类算法将相同含义的局部词义聚为一类，产生全局词义。获得了全局词义后，我们开始进行跨语言话题分析，首先通过词义消歧确定文档中每个词的词义。然后采用词义表示文档，最后采用聚类算法进行话题聚类。

聚类算法与具体的文档建模模型有关系，如果采用 SVSM 表示文档，则用 Bisecting K-Means 来进行聚类。如果采用 SLDA，则将文档的主题看成话题并将每个文档分到概率最大的主题中。

图3.6给出了本章提出的基于全局词义的文档建模模型和 VSM 模型以及其他跨语言模型的区别。VSM 模型以及其他跨语言模型采用词或者词簇来表示文档，而基于全局词义的文档建模模型采用词义来表示文档，使得基于全局词义的文档建模模型可以更准确地反映多义和同义现象，同时解决跨语言问题和翻译歧义问题。

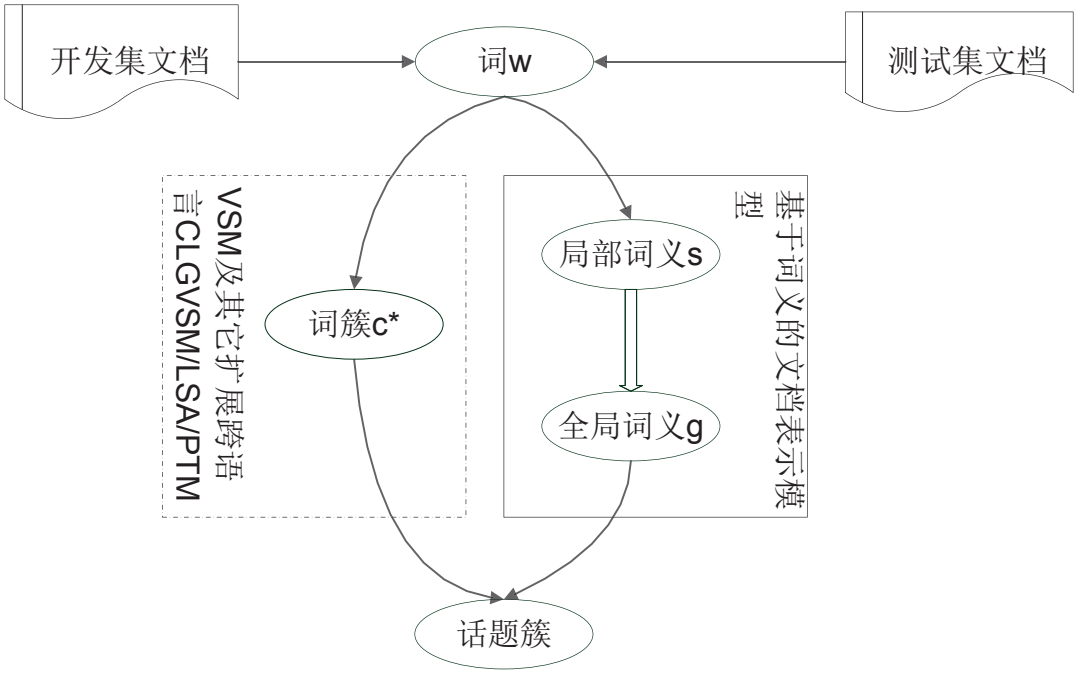


图 3.6　本章方法与 VSM 及其他模型的差别

3.5　实验评测

3.5.1　实验设置

实验数据

本章的实验数据（开发集和测试集）与2.5节相同。

评测指标

本章采用与2.5节相同的评测指标，即 F-measure。

系统参数

本章的方法参数有以下几类：

1) WSI 参数：我们对每个词每种语言的参数都设为同一个值，设置为：$\gamma \sim \text{Gamma}(1, 0.1), \rho \sim \text{Gamma}(0.01, 0.028)$，$\lambda = 0.1$。

2) SVSM 参数：我们将类的个数设置为每个文档集的话题数。

3) SLDA 参数：我们在测试集 CLTC1 中调整超参数 $\hat{\alpha}$ 和 $\hat{\beta}$ 的值，使其最优化。最后的设置为 $\hat{\alpha} = 0.1$，$\hat{\beta} = 0.1$。我们将主题的个数设置为每个文档集的话题数。

4) 全局变量的个数。我们通过实验来观察这个参数对文档聚类的影响。

在所有的实验中，我们先迭代 2000 次，然后选取在之后 200 轮迭代中，每 20 轮取一次，最后结果取平均。

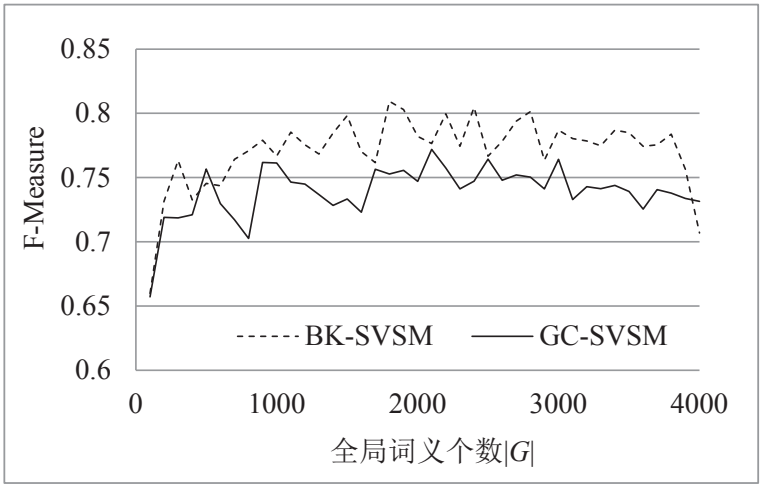


图 3.7　数据集 TDT41 上不同全局词义个数的系统性能

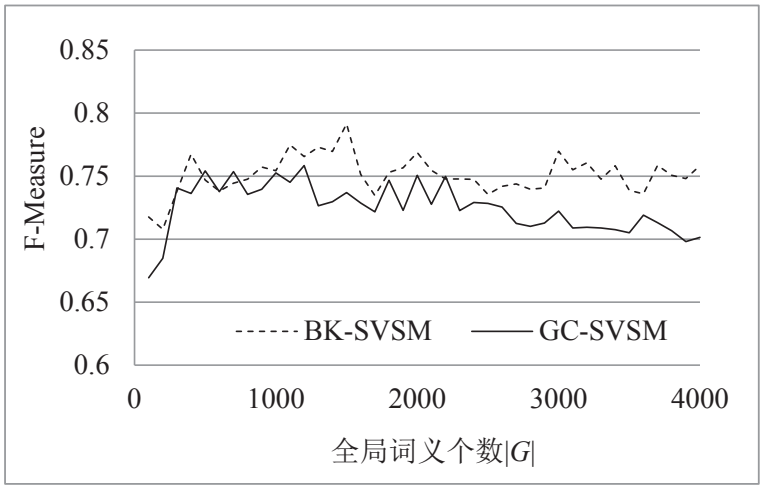


图 3.8　数据集 TDT42 上不同全局词义个数的系统性能

3.5.2　实验 #1 ：不同的词义聚类方法

在本实验中，我们旨在观察不同的词义聚类方法对系统性能的影响。我们实现了两个用不同的词义聚类方法的跨语言话题分析系统。

- **BK-sVSM**：这个系统采用 Bisecting K-Means 生成全局词义，SVSM 表示文档，同时采用 Bisecting K-Means 聚类文档。
- **GC-sVSM**：这个系统采用基于图的聚类算法生成全局词义，SVSM 表示文

表 3.1　不同词义聚类算法的最好性能

系统 \ 数据集	TDT41	TDT42
BK-SVSM	**0.809@1800**	**0.791@1500**
GC-SVSM	0.771@2100	0.752@1000

档，同时采用 Bisecting K-Means 聚类文档。

我们从 100 到 4000 逐步的增加全局词义的个数 $|G|$，并且在两个数据集里分别测试系统的性能。实验结果如图3.7和图3.8所示。每个数据集上的最好结果以及相关的全局词义个数在表3.1中列出。

实验结果分析

1) 全局词义个数的影响

我们比较了不同词义个数的性能，发现使用过高或者过低的词义个数均会造成性能的降低。原因很简单：词义个数过低时，很多不相似的局部词义划分到一个全局词义中，造成了性能的下降；当词义个数过高时，相似的局部词义不能识别，这样会导致不同语言相同含义的词不能识别，严重影响跨语言文档相似度的准确性，降低性能。比较不同的数据集，我们发现有更多词数的数据集会有很大的最优全局词义个数。比如说，系统 BK-SVSM 下，TDT42 数据集有 5435 个词，最好的 F-measure 出现在全局词义个数设为 1500 的时候，而 TDT41 数据集有 7871 个词，最优的全局词义个数为 1800。这与拥有较多词的数据集有较多的词义这个事实是一致的。

2) 词义聚类算法的影响

从图3.7和图3.8可以看出，BK-SVSM 在大多数情况下都优于 GC-SVSM。分析原因，我们发现这是由于基于图的聚类方法利用到最近邻，容易产生不平衡的类簇，而 Bisecting K-Means 利用全局的性质，可以产生更加平衡的类簇。因此，我们在之后的实验中都采用 Bisecting K-Means 作为词义聚类算法。

3.5.3　实验 #2：不同的基于词义的文档建模模型

在本实验中，我们旨在比较不同的基于词义的文档建模模型对系统性能的影响。我们实现了除了 BK-SVSM，我们还实现了 BK-SLDA 系统。

- **BK-SLDA** 使用 SLDA 表示文档的系统。

表 3.2　不同基于词义的文档建模模型的最好性能

系统 \ 数据集	TDT41	TDT42
BK-SVSM	**0.809@1800**	**0.791@1500**
BK-SLDA	0.795@3100	0.780@1900

实验结果如图3.9和图3.10所示。每个数据集上的最好结果以及相关的全局词义个数在表3.2中列出。

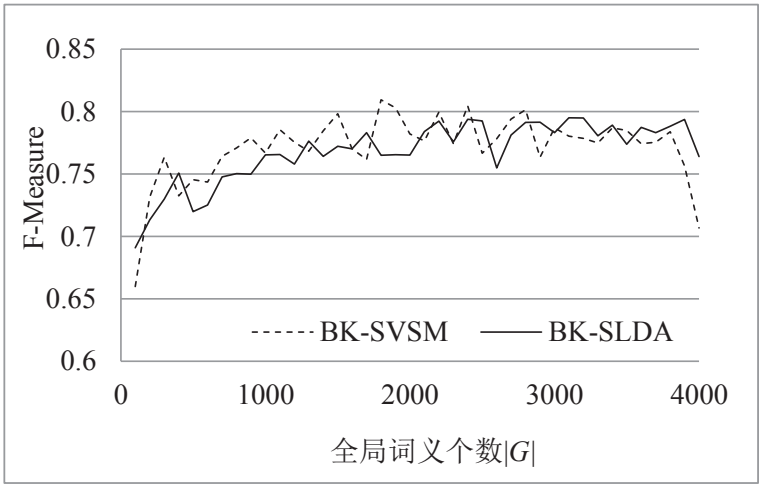


图 3.9　数据集 TDT41 上不同全局词义个数的系统性能

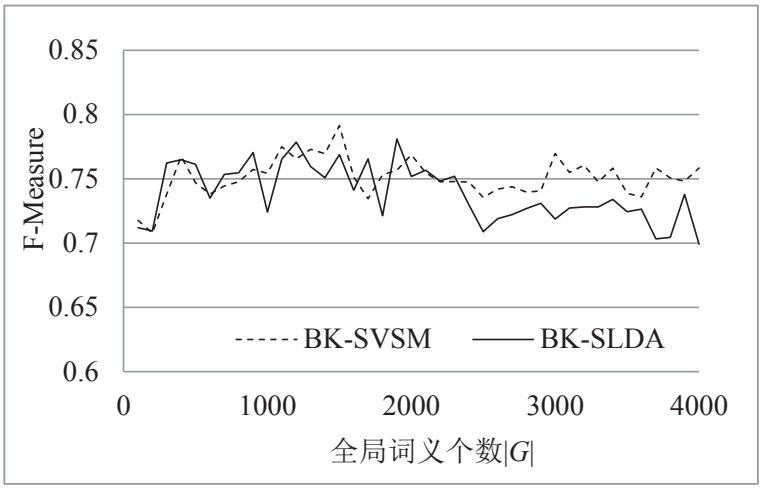


图 3.10　数据集 TDT42 上不同全局词义个数的系统性能

实验结果分析

我们比较了两个基于词义的文档建模模型的系统，发现 SVSM 的性能在所有数据集下都好于 SLDA。这是由于细粒度的特征空间在文档聚类比较有利而 LDA 推导出来的主题不能很好的解决这个问题。这种现象在文献 [141] 也有提到。

3.5.4　实验 #3：不同的文档建模模型

表 3.3　采用不同文档建模模型的话题分析系统的 F-measure

系统 \ 数据集	TDT41	TDT42
CLGVSM	0.762	0.748
BK-SVSM	**0.809**	**0.791**

本实验主要是比较我们提出的模型与2章提出的跨语言广义向量空间模型。

两个测试集的实验结果如表3.3所示。

实验结果分析

从表3.3中可以看出，BKSVSM 在所有的数据集中均好于 CLGVSM。这是由于 CLGVSM 考虑了词相似度，它只能直接反映词的同义关系，无法反映多义问题。而我们的方法可以同时反映这两种问题。

3.6　本章小结

针对跨语言歧义问题和单语言歧义问题，本章分别将这两个问题看做广义的同义问题和多义问题，提出用统计词义进行文档建模的方法。本章首先定义了局部词义和全局词义，提出了 CLHDP 模型，用于跨语言局部词义归纳，同时用聚类算法聚类局部词义，学习全局词义。然后提出了基于全局词义的两种文档建模方法：基于词义的 VSM 和基于词义的 LDA。

本章提出的方法旨在同时解决同义问题和多义问题：

1) 不同语言的同义词识别为一个词义，这样可以解决跨语言歧义问题。

2) 多义词在表示不同意思的时候识别成不同的词义，这样可以解决单语言歧义问题。在两种语言两个数据集上的实验证明了我们的方法优于其他文本表示方法。

但是本章的思想是在现有的文档建模模型中用词义替换了词的特征，没有进一步考虑到文档信息对词义生成的影响。因此下章将重点研究如何提出更好的基于词义的文档建模模型。

第 4 章　基于统计词义的主题模型

4.1　本章引论

4.1.1　研究问题

基于全局词义的文档建模方法介绍了跨语言词义的构造，在文档建模中利用统计词义代替词作为特征。本章进一步考虑如何用统计词义更好的表示文档。具体来说，利用统计词义表示文档需要解决两个问题：第一个是如何进行词义的学习，第二个是如何用学习的词义表示文档。比如说词“robot”可以表示“一种电子机械装置”，也可以表示“电影名称”①。第一个问题涉及“robot”的词义学习，第二个问题涉及如何用“robot”表示文档。

本章主要考虑如何利用词义更好的表示文档。具体来说，就是如何更好地学习词义以及如何更好地用学习到的词义表示文档。本章先考虑单语言的情况，跨语言的情况将在下章基于统计词义的跨语言主题模型介绍。

4.1.2　分析问题

主题模型目前已经被广泛的用于文档建模、话题分析中。本章采用主题模型进行文本表示和话题分析。Latent Dirichlet Allocation（LDA）是最经典的主题模型。LDA 模型依赖于词的共现来挖掘语义信息。但是，词通常与多个主题相关，并且在不同的主题中表达了不同的含义。例如给定两个主题：“T#1 电子科技”和“T#2 电影”，在 LDA 中，“robot”在不同的上下文中没有区分，通过与不同词的共现来区分这两个主题。理想情况下，如果一个模型可以根据上下文区分词义，词义与词相比对相关主题可以做出更多贡献。例如，如果识别出词“robot”的两个词义，即 S#1 (machine robot) 和 S#2 (film robot)，词义 S#1 对于主题 T#1 有比较高的概率，而对于主题 T#2 有比较低的概率；S#2 则相反。因此，在主题模型中使用词义作为附加特征可以增强主题模型的区分性。

以上的分析中只考虑到词义对主题的影响，接下来考虑主题对词义的影响。同个主题的词倾向于表达同一个含义，不同主题的词倾向于表达不同的含义。比如词“robot””分别在主题“T#1 电子科技”与主题“T#2 电影”下表示不同的词义。因此，如果在词义学习的过程加入主题，可以提高词义归纳的准确性。

① 维基百科，Robot (disambiguation)，http://enwikipediaorg/wiki/Robot (disambiguation)

4.1.3　解决思路

本章不仅在主题模型中用词义替代词作为特征，还进一步提出了改进主题模型。本章提出三个基于词义的主题模型：独立词义 LDA，点估计协同词义 LDA 以及词义混合协同词义 LDA。独立词义 LDA 模型首先用词义归纳算法获得词义，然后在 LDA 中用词义作为元素生成主题。点估计协同词义 LDA 和词义混合协同词义 LDA 则将词义看做主题模型的隐藏变量，协同迭代估计词义和主题。点估计协同词义 LDA 是通过点估计的方法确定词义而词义混合协同词义 LDA 则是考虑词义的分布。本章在话题分析、词义归纳任务中分别评测了以上三种模型。

本章的贡献主要在于两个方面：第一，将词义特征用于主题模型中并且构造了词义和主题协同推导过程。不同于上章先生成词义再用词义表示文本，本章将词义作为主题模型的隐藏变量，对词义和主题迭代进行估计。其次，前人的工作[48–50] 尝试将 WordNet 中的词义加入到主题模型中，但是这些模型需要额外的资源（WordNet）来获得给定词的词义。而我们的方法可以自动的从语料中估计词义。因此这种方法特别适用于资源缺乏的语言和一些特殊领域的文本。

本章剩余部分按照如下方式组织：4.2节介绍了基于词义的主题模型。4.3节评测了基于词义的主题模型在话题分析、词义归纳任务上的性能并且分析了这些主题模型的词义 -主题分布和文档 -主题分布。4.4节对本章进行了总结。

4.2　基于词义的主题模型

图4.1对比了传统的 LDA 模型与本章提出的基于词义的主题模型。如图4.1所示，LDA 模型认为词是文档的基本元素，并且为文档里面的每个词指定主题。基于词义的主题模型则首先获得词的词义，然后为每个词义指定主题，因此这个模型的基本元素是词义。词义是作为隐藏变量加入到主题模型中，根据上下文词归纳出来。基于词义的主题模型有两个步骤：首先确定每个词的词义，也就是词义归纳步骤；然后用词义表示文档，然后确定主题。本章设计了三种方法来实现基于词义的主题模型。

- 独立词义 LDA（Standalone SLDA，SA-SLDA）：这个方法将词义归纳的步骤看做独立的步骤，然后用归纳的词义进行 LDA。这个方法是本章的基线系统，用于验证词义对主题的影响。
- 点估计协同词义 LDA（Point estimate Collaborative SLDA，PCo-SLDA）：这个方法有两个交替迭代步骤：给定已知的主题，根据主题确定词义。给定已知的词义，确定词义的主题，其中词义是词义分布的点估计。
- 词义混合协同词义 LDA（Sense mixture Collaborative SLDA，SCo-SLDA）：这

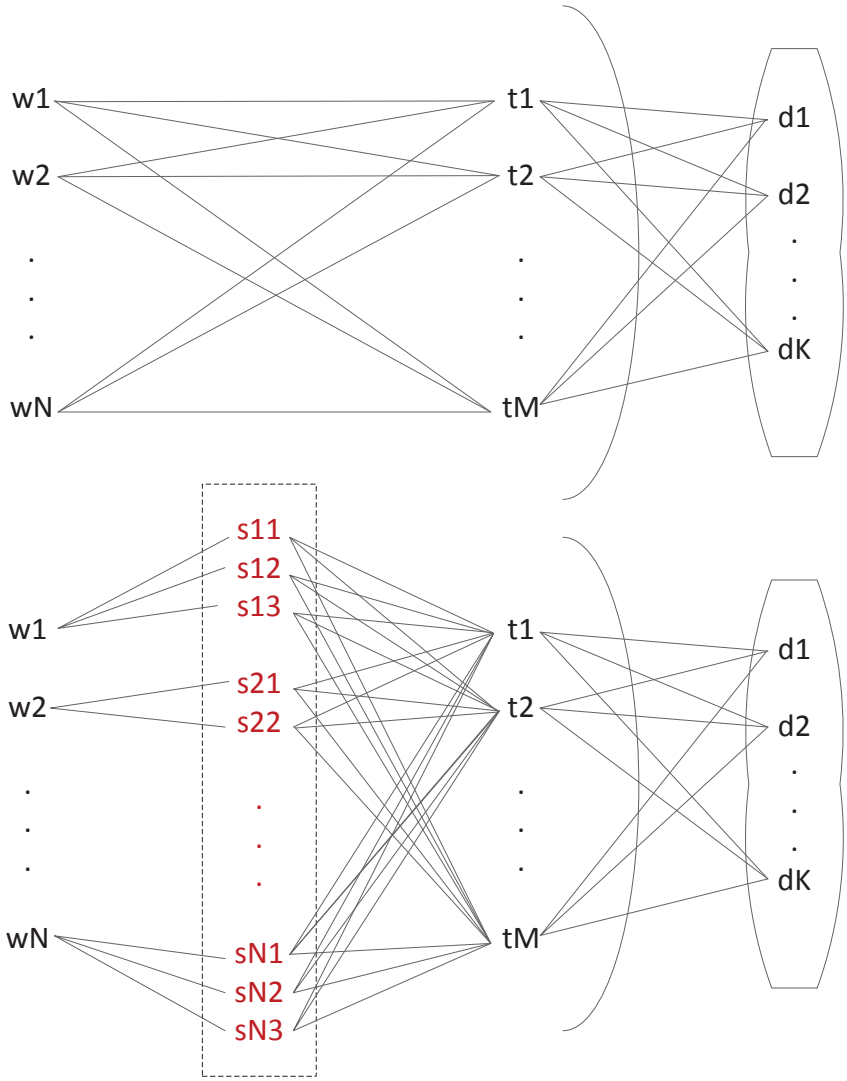


图 4.1　LDA 模型（上图）和基于词义的主题模型（下图）对比示意图。虚线框中的变量为潜变量（即词义）

个方法与点估计协同词义 LDA 的区别在于它没有用点估计确定词义而是直接从词义分布中获得主题。

下面将分别介绍提出的三个方法。

4.2.1　独立词义 LDA

在 SA-SLDA 模型，WSI 和 DR 两个步骤是独立的两个模块，DR 直接采用 WSI 作为输入（见图4.2）。

WSI 步骤中采用 HDP 模型归纳词义，具体的模型与3.2.3.1节相同，这里就不在赘述了。词义归纳步骤可以获得文档中每个词的词义分布，SA-SLDA 模型采用词义分布的众数作为该词的词义，之后在 LDA 模型中用词义表示文档。如图4.2所示，图中的阴影部分表示可见变量词义，SA-SLDA 中文本表示部分的概

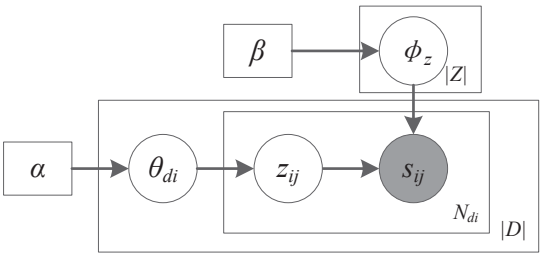


图 4.2　SA-SLDA 模型

率生成模型如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 中的词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择词义 $s_{ij} \sim Mult(\phi_{z_{ij}})$.

以词“robot”为例，从语料中可以归纳出以下两个词义：

Example #1: sense *robot#1*

film:	0.159
role:	0.069
performance:	0.019

...

Example #2: sense *robot#2*

computer:	0.116
system:	0.039
software:	0.019

...

关于词义的影响有以下分析：假设一篇有关电影“robot”的评论含有上下文“…it’s an inspired performance from Robot that keeps the film fresh”。而在一篇技术报告中，“robot”的上下文是“There may be a computer operating system designed mainly for robots”。词义归纳步骤后，“robot”归纳出两个词义“robot#1 ”和词义“robot# 2”，同时根据词义分布判断第一个上下文中的“robot”的词义为“robot#1 ”，第二个上下文中的“robot”的词义为“robot#2 ”。传统的 LDA 模型无法识别词的语义信息，只能判断主题，区分性较弱，而在 SA-SLDA 模型中，词义可以识别出语义信息，主题的区分性较强。

SA-SLDA 采用吉布斯抽样（Collapsed Gibbs Sampling）学习模型[5]，它的参数推导过程与 LDA 过程相同，唯一的区别在用词义代替了词。因此，在抽样过程

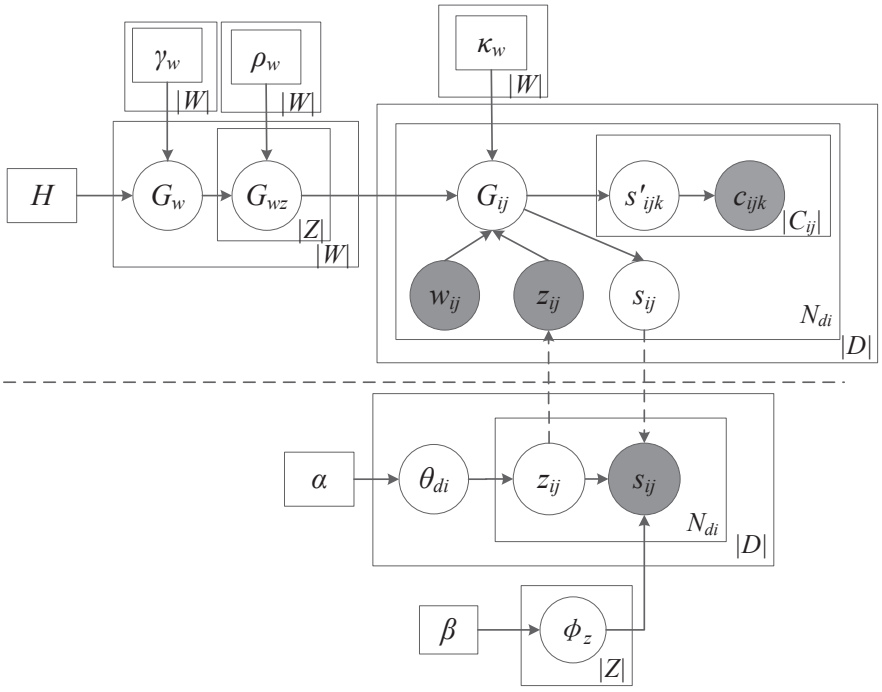


图 4.3　PCo-SLDA 模型

中，词 w_{ij} 词义为 $s_{ij} = s$ 时，主题条件概率公式按照公式 (4-1) 计算。

$$P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{s}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{s}_{-ij,z} + \beta}{n_{-ij,z} + S\beta} \tag{4-1}$$

在公式 (4-1) 中，$n^{d_i}_{-ij,z}$ 表示文档 d_i 中主题为 z 的词数；$n^{s}_{-ij,z}$ 表示主题为 z 词义为 s 的词数；$n^{d_i}_{-ij}$ 表示文档 d_i 的词数；$n_{-ij,z}$ 表示主题为 z 的词义数；S 表示词义个数。上面变量中的 $-ij$ 表示在计数过程中去掉词 w_{ij}。

4.2.2　点估计协同词义 LDA

SA-SLDA 模型中，词义归纳和文档建模是相互独立的两个过程。SA-SLDA 模型只考虑了词义对主题的作用，我们进一步考虑主题对词义的影响，提出了协同词义 LDA 模型。该模型将主题作为词义分布的伪回馈，采用点估计确定词义然后用词义估计主题分布，如此循环迭代的估计主题和词义。因为这个模型是采用点估计的方法确定词义，我们称这个方法为点估计协同词义 LDA 模型。

具体来说，PCo-SLDA 模型采用三层 HDP 模型获得目标词 w 的词义和主题的关系（见图4.3）。在词 w 的三层 HDP 模型中，每个主题都有一个概率分布 G_{wz}，这个分布是从狄利克雷过程 $DP(\rho_w, G_w)$ 中生成的，对于文档 d_i 中的每个词 w_{ij}，给定主题 $z_{ij} = z$，我们使用 w_{ij} 上下文的 G_{wz} 作为基概率分布并且根据基概率分布

从狄利克雷过程中抽取词义概率分布 G_{ij}，$G_{ij} \sim DP(\kappa_{wz}, G_{wz})$。这意味着词 w 在不同的主题下具有不同的词义分布。

图4.3表示 PCo-SLDA 模型，其中 C_{ij} 表示文档 d_i 中词 w_{ij} 的上下文窗口 v_{ij} 的词个数。虚线之上的部分为 WSI 步骤而虚线之下的部分为 DR 步骤。给定主题 $\{z_{ij}\}$ 已知，词义 $\{s_{ij}\}$ 在 WSI 步骤中进行估计；而给定已知的词义 $\{s_{ij}\}$，主题 $\{z_{ij}\}$ 在 DR 步骤中进行估计。两个过程交替进行。图4.3用虚线箭头连接 $\{s_{ij}\}$ 和 $\{z_{ij}\}$ 表示两个过程的交替时这两个变量从隐藏到已知的变化。

词义归纳过程如下：

1. 对于每个词 w:
   (a) 选择词义分布 $G_w \sim DP(\gamma_w, H_w)$.
   (b) 对于每个主题 z:
      i. 选择词义分布 $G_{wz} \sim DP(\rho_w, G_w)$.
2. 对于每个文档 d_i:
   (a) 对于词 w_{ij} 的上下文窗口 v_{ij} :
      i. 选择词义分布 $G_{ij} \sim DP(\kappa_{w_{ij}z}, G_{w_{ij}z})$.
      ii. 对于目标词 w_{ij} 的每个上下文词 c_{ijk}:
         A. 选择词义 $s_{ijk} \sim G_{ij}$.
         B. 选择上下文词 $c_{ijk} \sim Mult(\eta_{s_{ijk}})$.
      iii. 指定词义 $s_{ij} = \arg\max_s P(s|G_{ij})$.

文档表示过程如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 的每个词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择词义 $s_{ij} \sim Mult(\phi_{z_{ij}})$.

PCo-SLDA 的模型参数仍然采用吉布斯抽样（Collapsed Gibbs Sampling）进行学习[5]。该模型有两组隐藏变量需要交替估计：

1. 给定每个词的主题 z_{ij} 已知，估计词义分布 G_{ij}。这个过程与文献 [128] 相同。接着我们将概率分布 G_{ij} 中概率最大的词义指定为该词的词义。
2. 给定每个词的词义 s_{ij} 已知，估计主题 z_{ij}。这个过程与 SA-SLDA 相同。

由估计过程可以看出，PCo-SDA 模型迭代的改进主题和词义的学习，因此直观上,PCo-SLDA 要优于只有一轮变量估计的 SA-SLDA 模型。

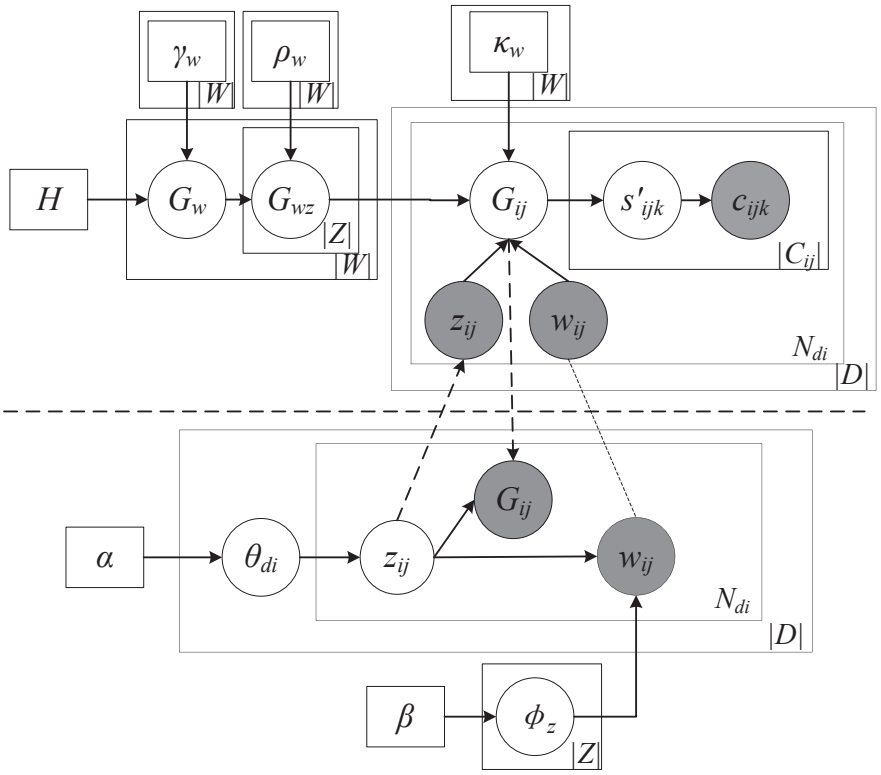


图 4.4　SCo-SLDA 模型

4.2.3　词义混合协同词义 LDA

在 PCo-SLDA 的 WSI 步骤中，词的词义设定为词义分布中概率最高的值。现在我们解除这个假设，考虑整个词义分布，提出词义混合协同词义 LDA (SCo-SLDA)。

SCo-SLDA 仍然采用三层 HDP 模型进行词义归纳。PCo-SLDA 模型和 SCo-SLDA 模型的区别在于 SCo-SLDA 模型考虑上下文的整体词义分布而不是指定词义。

图4.4表示 SCo-SLDA 模型。在图4.4中，虚线之上的部分表示 WSI 步骤而虚线之下的部分表示 DR 步骤。两个步骤与 PCo-SLDA 一样迭代交替进行。但是不同于 PCo-SLDA，主题 $\{z_{ij}\}$ 是在给定词义混合 $\{G_{ij}\}$ 已知的条件下进行估计的。同样，我们将两个过程交替时从隐藏状态到已知状态中的变量用虚箭头连接。我们还将两个过程的 w_{ij} 用虚线连接, 表示他们是相同的。

词义归纳过程如下：

1. 对于每个词 w:
   (a) 选择词义分布 $G_w \sim DP(\gamma_w, H_w)$.
   (b) 对于每个主题 z:
      i. 选择词义分布 $G_{wz} \sim DP(\rho_w, G_w)$.

2. 对于每个文档 d_i:
   (a) 对于词 w_{ij} 的上下文窗口 v_{ij} :
      i. 选择词义分布 $G_{ij} \sim DP(\kappa_{w_{ij}z}, G_{w_{ij}z})$.
      ii. 对于目标词 w_{ij} 的每个上下文词 c_{ijk}:
         A. 选择词义 $s_{ijk} \sim G_{ij}$.
         B. 选择上下文词 $c_{ijk} \sim Mult(\eta_{s_{ijk}})$.

文档表示过程如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 的每个词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择词 $w_{ij} \sim Mult(\phi_{z_{ij}})$.
      iii. 选择词义分布 $G_{ij} \sim DP(\kappa_{w_{ij}z}, G_{w_{ij}z})$.

在 SCo-SLDA 模型中，词的主题决定于三个因素：

1. 文档的主题分布，
2. 给定主题，生成词的概率，
3. 给定词和主题，生成词义的概率。

前面两个因素与 LDA 相同而最后一个因素反映了词义的影响。

SCo-SLDA 采用吉布斯抽样（Collapsed Gibbs Sampling）进行学习[5]。该模型有两组隐藏变量需要交替估计。给定每个词的主题 z_{ij}，估计词义分布 G_{ij}。这个过程与文献 [128] 相同。下面将主要介绍给定词义分布，如何估计 z_{ij}。

SCo-SLDA 模型中，主题 $\{z_{ij}\}$ 直接基于词义分布混合 $\{G_{ij}\}$ 进行估计的。因此，我们将首先介绍 $\{G_{ij}\}$。

每个目标词的先验概率 G_w 是一个参数为 λ_w，基概率分布为 H_w 的狄利克雷过程。它可以用折棍（stick-breaking）过程表示（如公式 (4-2)）。

$$G_w = \Sigma_{s_w=1}^{\infty} \pi_w^{s_w} \delta_{\eta_{s_w}} \tag{4-2}$$

其中 η_{s_w} 是基于 H_w 生成的，$\delta_{\eta_{s_w}}$ 是参数为 η_{s_w} 的概率度量。$\{\pi_w^{s_w}\}$ 是词义的混合参数。它们是从折棍构造中抽样出来的。在抽样过程中，经过几轮迭代，假设目标词 w 现在已经有 S_w 个词义。上下文词的分布 $\{\eta_{s_w}\}$ 已经生成，上下文词的词义都

已经指定。G_w 可以表示为公式 (4-3)。

$$G_w = \Sigma_{s_w} \pi_w^{s_w} \delta_{\eta_{s_w}} + \pi_w^u G_w^u \tag{4-3}$$

其中 G_w^u 从狄利克雷过程 $DP(\gamma_w, H_w)$ 抽样生成。每个主题 z 的词义分布 G_{wz} 可以类似的表示为公式 (4-4)。

$$G_{wz} = \Sigma_{s_w} \pi_{wz}^{s_w} \delta_{\eta_{s_w}} + \pi_{wz}^u G_{wz}^u, \tag{4-4}$$

对于文档 d_i 的每个上下文窗口 v_{ij}，G_{ij} 可以表示为公式 (4-5)。

$$G_{ij} = \Sigma_{s_w} \pi_{ij}^{s_w} \delta_{\eta_{s_w}} + \pi_{ij}^u G_{ij}^u \tag{4-5}$$

其中 $\{\pi_{wz}^{s_w}\}$ 和 $\{\pi_{ij}^{s_w}\}$ 都是词义混合，G_{wz}^u 和 G_{ij}^u 是从狄利克雷过程 $DP(\gamma_w, H_w)$ 抽样生成。值得注意的是，对于每个目标词 w，整体词义分布 G_w，主题词义分布 $\{G_{wz}\}$ 以及上下文词义分布 $\{G_{ij}\}$ 共享相同的上下文词分布集合 $\{\eta_{s_w}\}$。但是他们的词义混合分布不同。不同目标词的词义 -上下文词分布集合 $\{\eta_{s_w}\}$ 也不相同。

本章采用与文献 [128] 相同的抽样过程。$\{\pi_w\}, \{\pi_{wz}\}$ 是抽样出来的而 $\{\eta_{s_w}\}$、G_w、$\{G_{wz}\}$、$\{G_{ij}\}$ 以及 $\{\pi_{ij}\}$ 可以直接积分出来。为了抽样文档 d_i 中词 $w_{ij} = w$ 的主题 z_{ij}，条件概率可以按照公式 (4-6) 计算。

$$\begin{aligned} & P(z_{ij} = z | \boldsymbol{z_{-ij}}, \boldsymbol{s}, \boldsymbol{w}) \\ & \propto P(z_{ij} = z | \boldsymbol{z_{-ij}}, \boldsymbol{w}) P(\boldsymbol{s_{ij}} | \boldsymbol{z}, \boldsymbol{s_{-ij}}, \boldsymbol{w}) \end{aligned} \tag{4-6}$$

条件概率 $P(z_{ij} = z | \boldsymbol{z_{-ij}}, \boldsymbol{w})$ 的计算公式如 LDA 相同（见公式 (3-15)）。条件概率 $P(\boldsymbol{s_{ij}} | \boldsymbol{z}, \boldsymbol{s_{-ij}}, \boldsymbol{w})$ 可以通过公式 (4-7) 计算。

$$\begin{aligned} & P(\boldsymbol{s_{ij}} | \boldsymbol{z}, \boldsymbol{s_{-ij}}, \boldsymbol{w}) \\ & \propto \frac{\Gamma(\kappa_{wz})}{\Gamma(\kappa_{wz} + C_{ij})} \frac{\prod_{s \in \{s_w\}} \Gamma(\kappa_{wz} \cdot \pi_{wz}^s + n_{ij}^s)}{\prod_{s \in \{s_w\}} \Gamma(\kappa_{wz} \cdot \pi_{wz}^s)} \\ & = \frac{\prod_{s \in \{s_w\}} \prod_{g=0}^{n_{ij}^{s-1}} (\kappa_{wz} \pi_{wz}^s + g)}{\prod_{g=0}^{C_{ij}-1} (\kappa_{wz} + g)}) \end{aligned} \tag{4-7}$$

其中 $\{s_w\}$ 是目标词 w 的词义集合，C_{ij} 表示文档 d_i 中上下文窗口 v_{ij} 词数，n_{ij}^s 是上下文窗口 v_{ij} 中词义为 s 的词数。因此，条件概率 $P(z_{ij} = z | \boldsymbol{z_{-ij}}, \boldsymbol{s}, \boldsymbol{w})$ 可以按照公式 (4-8) 进行估计。

表 4.1　测试集信息

数据集	# 文档	# 主题	# 词
TDT41	1270	38	18511
TDT42	617	33	11782
Reuters20	9101	20	25748

$$
\begin{aligned}
&P(z_{ij}=z|\boldsymbol{z_{-ij}},\boldsymbol{s},\boldsymbol{w}) \\
&\propto \frac{n^{d_i}_{-ij,z}+\alpha}{n^{d_i}_{ij}+Z\alpha}\frac{n^{w}_{-ij,z}+\beta}{n_{-ij,z}+W\beta}\frac{\prod_{s\in\{s_w\}}\prod_{g=0}^{n^{s-1}_{ij}}(\kappa_{wz}\pi^{s}_{wz}+g)}{\prod_{g=0}^{C_{ij}-1}(\kappa_{wz}+g)}
\end{aligned}
\tag{4-8}
$$

4.3　实验评测

基于词义的主题模型的评测一共有三个部分：第一，评测主题；第二，分析所提出模型的词义 -主题分布和文档主题分布；第三，评测模型归纳出的词义。

4.3.1　话题分析

本章在话题分析任务上评测模型生成的主题，并且比较所提出模型与 LDA 和 K-Means 等基线系统的性能。

4.3.1.1　实验设置

数据集

实验采用三个测试集，具体的信息见表4.1。

1. **TDT4 数据集**: 本实验分别采用 TDT4 语料中的 2002 年的评测语料（TDT41）和 2003 年的评测语料（TDT42）[113]。
2. **Reuters**: 本实验采用 Reuters-21578[142] 中文档数最大的前 20 个类作为第三个数据集（Reuters20）。

实验中仅保留名词和动词作为目标词进行词义消歧和主题学习。我们采用句子作为目标词的上下文窗口。TreeTagger[143] 用来进行词性标注和词形还原。

评测指标

本实验采用与2.5节相同的评测指标，即 F-measure。

表 4.2　不同词义归纳方法（LDA 和 HDP）下的 SA-SLDA 结果

系统 \ 数据集	TDT41	TDT42	Reuters20
SA-SLDA(LDA)	0.787	0.842	0.490
SA-SLDA(HDP)	0.792	0.870	0.512

4.3.1.2　实验 #1.1：不同的词义归纳方法

实验 #1.1 旨在研究不同的词义归纳方法对主题模型的影响。我们实现了两个用不同的贝叶斯词义归纳模型的 SA-SLDA 话题分析系统。

- **SA-SLDA(LDA)**：这个系统在 SA-SLDA 模型上采用 LDA 模型归纳词义，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **SA-SLDA(HDP)**：这个系统在 SA-SLDA 模型上采用 HDP 模型归纳词义，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。

系统参数

SA-SLDA 有两个步骤：WSI 步骤和 DR 步骤。下面将分别两个步骤的参数：

1. 在 WSI 步骤，对于 HDP 模型，参考文献 [128]，γ_w 和 ρ_w 使用 gamma 先验。每个词的 γ_w , ρ_w , ϵ_w 的超参数分别为 $\gamma_w \sim Gamma(1, 0.001)$, $\rho_w \sim Gamma(0.1, 0.028)$ 和 $\epsilon_w = 0.1$ 。LDA 模型的参数则设定为：$\alpha = 0.2$,$\beta = 0.1$，每个词的词义数均设为 4。
2. 在 DR 步骤，参数设定为：$\alpha = 1.5$ and $\beta = 0.1$。

本实验中所有的超参数均是在 TDT42 数据集上进行优化的。文档的主题个数设为该数据集的类簇数。在所有实验中，为了避免吉布斯抽样的随机性，先进行 2000 轮抽样，然后在进行 200 轮，抽取其中的每 20 轮结果做平均。

实验结果如表4.2所示。

实验结果分析

从表4.2中，我们可以发现在 SA-SLDA 模型中，使用 HDP 模型进行词义归纳的性能要优于 LDA 模型。这是因为 LDA 模型是参数模型，需要人工指定词义个数，而 HDP 是非参数模型，可以自动学习每个词的词义个数。HDP 模型为词义建模提供了合理的解释并为文档建模提供了更大的灵活性。由于 HDP 的优越性，我们在模型中均采用 HDP 作为词义归纳的非参数先验。

表 4.3　所提出模型和基线系统的性能对比

系统 \ 数据集	TDT41	TDT42	Reuters20
VSM	0.727	0.843	0.501
LDA	0.744	0.867	0.496
SA-SLDA	0.792	0.870	0.512
PCo-SLDA	0.825	0.874	0.597
SCo-SLDA	0.864	0.905	0.612

4.3.1.3　实验 #1.2：不同的拓展主题模型

实验 #1.2 在话题分析任务中评测所提出的模型。除了所提出的模型（SA-SLDA，PCo-SLDA 以及 SCo-SLDA），本实验还比较了 VSM 和 LDA 等基线系统的性能。具体来说，我们实现了五个话题分析系统

- **VSM**：这个系统采用 VSM 表示文档，在 TF-IDF 特征上计算文档的余弦相似度，同时采用 Bisecting K-Means 聚类文档。
- **LDA**：这个系统用 LDA 模型对文档的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **SA-SLDA**：这个系统用 SA-SLDA 模型对文档的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。这个系统与实验 #1.1 中的 SA-LDA(HDP) 相同
- **PCo-SLDA**：这个系统用 PCo-SLDA 模型对文档的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **SCo-SLDA**：这个系统用 SCo-SLDA 模型对文档的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。

系统参数

PCo-SLDA 和 SCo-SLDA 的参数相同。

1. 在 WSI 步骤中，每个词的超参数 γ_w, ρ_w , ϵ_w 设定为：$\gamma_w \sim Gamma(8, 0.1)$, $\rho_w \sim Gamma(5, 1)$, $\kappa_w \sim Gamma(0.1, 0.028)$, $\epsilon_w = 0.1$ 。
2. 在 DR 步骤中，$\alpha = 1.5$ ，$\beta = 0.1$ 。

LDA 的参数设定为：$\alpha = 1.5$, $\beta = 0.1$。主题的个数设定为数据集中的类簇个数。K-Mean 中的 K 值也设定为数据集中的类簇个数。

实验结果如表4.3所示。

实验结果分析

从表4.3我们可以看出：

首先，所有 SLDA 模型在全部数据集上的性能均优于两个基线系统。这意味着使用词义作为特征可以改进话题分析的结果。这是因为 SLDA 模型采用更细粒度的特征，可以根据上下文的不同选用不同的特征。

其次，PCo-SLDA 和 SCo-SLDA 在全部数据集上的性能要好于 SA-SLDA。这意味着词的主题和词义的协同学习过程对结果有着正面的影响。原因可能有两个：

- 对于常用的词义，同一个词在不同的主题下可能具有不同的词义，同一个词在相同的主题下通常意思相同。PCo-SLDA 和 SCo-SLDA 模型中主题和词义相互影响，同一个主题下的词倾向于识别为同一个词义而不同主题下的相同词倾向于识别为不同的词义，词义的识别更加准确。
- 用主题作为词义的伪回馈可以生成主题相关的词义。例如，词”election“通常只有一个意思。但是，在 TDT42 数据集中，主题的粒度比较细。例如，下面两个句子是分别从不同的主题中抽取出来的：z_1：“Ilyescu Wins Romanian Elections”，z_2：“Ghana Gets New Democratically Elected President”。两个主题均是关于选举的，但是选举的国家不同。通过主题和词义的协同学习，词“election”有两个词义，“election#1 ”和“election#2 ”，分别指的是罗马尼亚和加纳的选举。有了上述两个词义，词“election”在上下文含有词“Romania”或者其他与之相关的词时被识别为词义“election#1 ”因此它的主题更倾向于 z_1，而词“election”在上下文含有词“Ghana”或者其他与之相关的词被识别为词义“election#2 ”因此它的主题更倾向于 z_2。

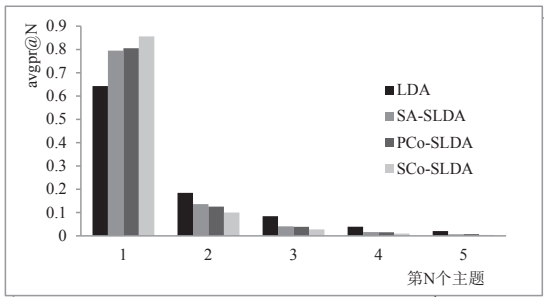
第三，SCo-SLDA 的性能在全部数据集上均优于 PCo-SLDA。这意味着直接考虑目标词的词义分布可以进一步改进性能。原因可能是考虑多个可能的词义可以降低词义消歧的风险。另外，SCo-SLDA 模型中考虑的词义混合可以为文档 -主题和主题 -词义分布的估计提供更大的灵活性。

4.3.2　概率分布分析

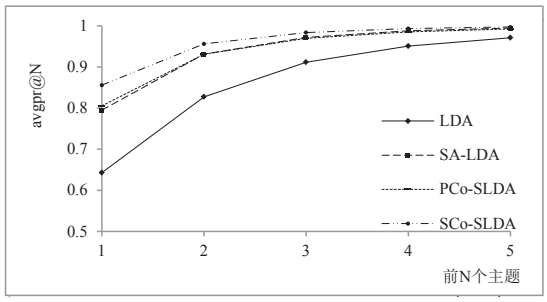
本小节旨在分析模型的词义 -主题以及文档 -主题的平均分布。除了所提出的模型（SA-SLDA，PCo-SLDA 以及 SCo-SLDA），本小节还比较了 LDA 的词义 -主题以及文档 -主题的平均分布情况。

4.3.2.1　实验 #2.1：整体概率分布分析

在图4.5中，*argpr@N* 在 LDA 模型中表示所有词的词 -主题平均分布，在 SLDA 系列模型中表示所有词义的词义 -主题平均分布。它的具体计算过程是，在

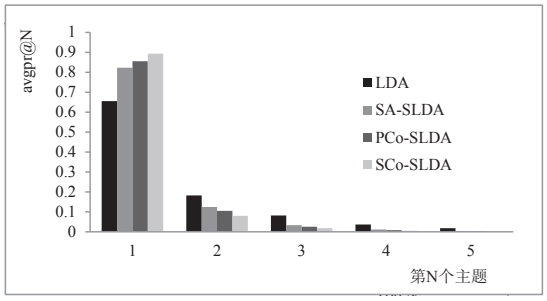


(a.1)

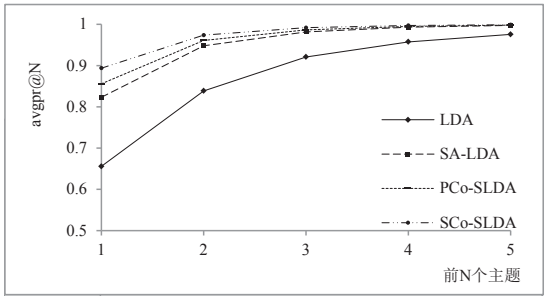


(a.2)

(a) TDT41

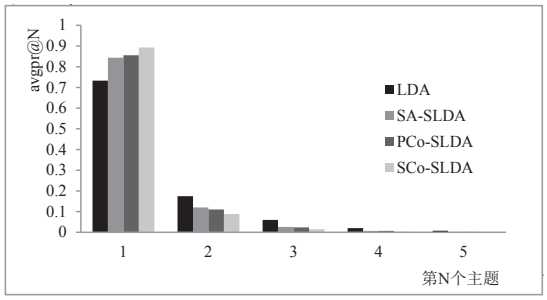


(b.1)

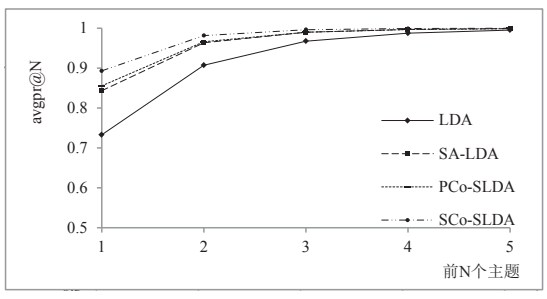


(b.2)

(b) TDT42



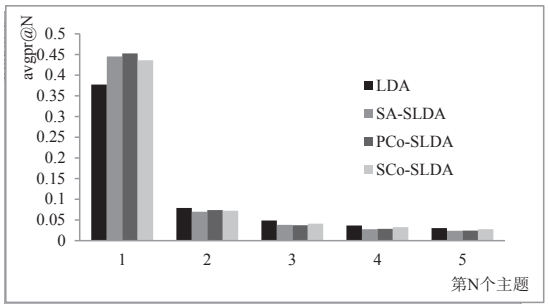
(c.1)



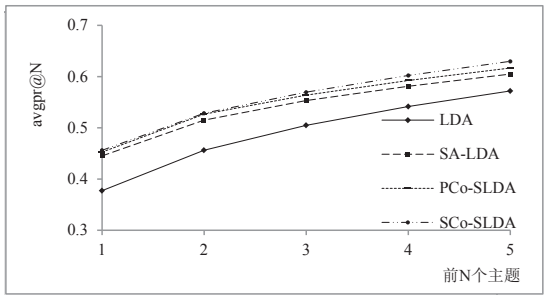
(c.2)

(c) Reuters20

图 4.5　前五个主题的词（词义）-主题平均分布 (1) 第 k 个主题的直方图 (2) 前 k 个主题的曲线图

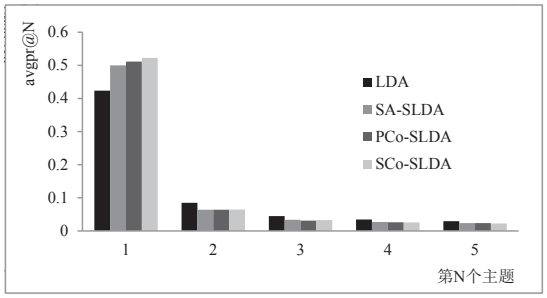


(a.1)

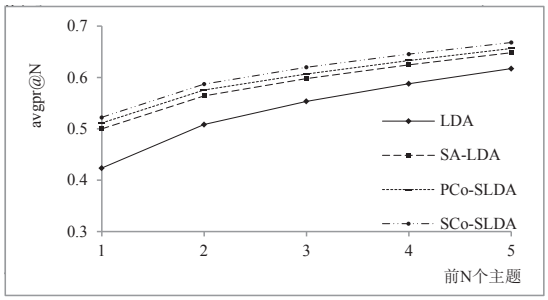


(a.2)

(a) TDT41

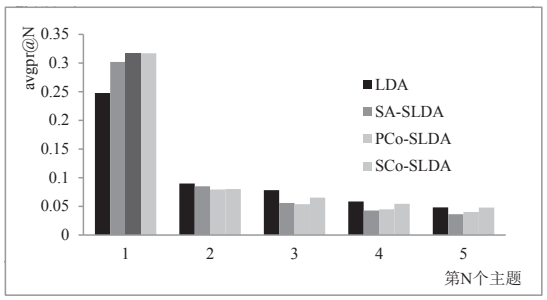


(b.1)

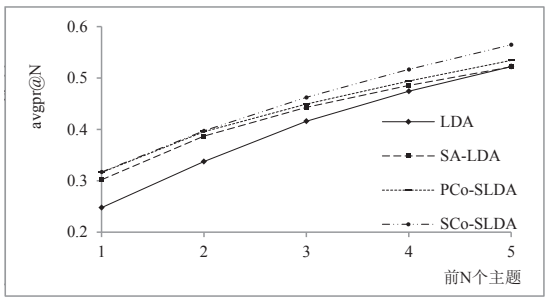


(b.2)

(b) TDT42



(c.1)



(c.2)

(c) Reuters20

图 4.6　前五个主题的文档 -主题平均分布 (1) 第 k 个主题的直方图 (2) 前 k 个主题的曲线图

LDA 模型中，对于每个词，将它所相关的主题按照概率 $p(z|w)$[①] 排序。*argpr@N* 是所有词前 *N* 个主题取平均的结果。在 SLDA 系列模型中，*argpr@N* 是对于每一个词义基于 $p(z|s)$ 进行计算的，计算过程与 LDA 类似。在 LDA 和 SLDA 系列模型中，我们均使用最后一轮迭代（即 2200 轮）生成的词义和主题。对于每一个数据集，结果通过以下的两张图展示：左边的图是第 N 个话题的直方图而右边的图则是前 N 个话题的曲线图。此外，我们还计算了文档 -话题分布的 *avgpr@N* 值，这个值是在模型中通过对每个文档的文档 -主题分布，即 $p(z|d)$ 概率值，取平均进行计算的。LDA 模型和 SLDA 模型在三个数据集上的结果如图4.6所示。

实验结果分析

1) 从图4.5中可以看出：

首先，在直方图中，SLDA 系列模型的词义 -主题平均分布与 LDA 的词 -主题平均分布相比要更加尖锐；在曲线图中，SLDA 系列模型的曲线都在 LDA 模型之上。这意味着在主题模型中，词义可以提供更好的区分性。

其次，在直方图中，PCo-SLDA 以及 SCo-SLDA 的词义 -主题平均分布与 SA-SLDA 相比要更加尖锐；而曲线图中，PCo-SLDA 和 SCo-SLDA 模型的曲线在 SA-SLDA 模型之上。这得益于使用主题作为伪回馈 PCo-SLDA 和 SCo-SLDA 可以归纳出主题相关的词义。主题相关的词义与常用词义相比更具有主题区分性。

第三，在直方图中，SCo-SDLA 的词义 -主题分布比 PCo-SLDA 更加尖锐；而曲线图中，SCo-SLDA 的曲线在 PCo-SLDA 之上。这可能是 SCo-SLDA 的性能在话题分析任务上优于 PCo-SLDA，因此 SCo-SLDA 可以获得更好的主题，也可以归纳出更好的词义。

2) 从图4.6中可以看出：

首先，SLDA 系列模型与 LDA 相比文档 -主题分布更加尖锐；在曲线图中，SLDA 系列模型的曲线均高 LDA 模型。这意味着 SLDA 模型中文档集中于更少的主题，这使得 SLDA 模型的主题更具有区分性。

其次，PCo-SLDA 与 SCo-SLDA 的文档 -主题分布更加尖锐；曲线图中，PCo-SLDA 的曲线在 SA-SLDA 之上。这意味着迭代的改进主题和词义可以为文档中的主题提供更强的后验概率估计。

① $p(z|w)$ 可以根据 $p(z|w) \propto p(w|z)\Sigma p(z|d)p(d)$ 进行计算，其中 $p(w|z)$ 和 $p(z|d)$ 是模型可以估计出来的参数而 $p(d)$ 与文档的长度成正比。

表 4.4　百分比 (%)/# 文档，其中文档满足两个条件：1) LDA 和 SLDA 模型中话题均识别正确；2）$p_{\mathrm{SLDA}}(\hat{z}|d) > p_{\mathrm{LDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42	Reuters20
SA-SLDA	87.6/774	89.6/473	75.7/2064
PCo-SLDA	90.5/736	92.7/472	78.5/2293
SCo-SLDA	80.0/767	93.2/487	79.3/2849

表 4.5　百分比 (%)/# 文档，其中文档满足两个条件：1) LDA 和 SLDA 模型中话题均识别错误；2）$p_{\mathrm{SLDA}}(\hat{z}|d) > p_{\mathrm{LDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42	Reuters20
SA-SLDA	54.2/286	62.6/83	50.7/4313
PCo-SLDA	62.0/208	51.5/68	52.3/4077
SCo-SLDA	67.7/138	55.2/67	58.3/3572

4.3.2.2　实验 #2.2：基于文档标签的概率分布分析

为了进一步分析所提出的模型，我们以 LDA 模型作为基线，比较了 SLDA 系列模型中每个文档概率最大的主题。我们将文档中概率最大的主题作为该文档的“话题”（伪标准话题）。与话题分析实验的设置相同，我们选择每个数据集中人工标注的话题作为标准话题，并且对系统生成的伪标准话题和人工标注的标准话题计算 F 值，将每个系统生成的伪标准话题根据最大 F 值对应到人工标注的标准话题上。

接着我们在三类文档上分别比较了伪标准话题的权重：

- LDA 和 SLDA 模型中话题均识别正确的文档；
- LDA 和 SLDA 模型中话题均识别错误的文档；
- 所有文档。

在每类文档中，我们都计算了 SLDA 模型中的 $p(\hat{z}|d)$ 大于 LDA 模型中的 $p(\hat{z}|d)$ 的文档百分比，还给出了该类的文档个数，其中 $\hat{z}$ 表示伪标准话题。结果如表4.4，4.5和表4.6。

实验结果分析

从表4.4中可以看出，在 LDA 和 SLDA 系列模型都识别正确的文档集中，SLDA 可以在大多数文档上提供一个更大的 $p(\hat{z}|d)$（结果百分比均大于 50%）。这说明了当文档识别正确时，SLDA 系列模型可以增强正确话题的权重。表4.5显示，

表 4.6　百分比 (%)/# 文档，其中文档满足 $p_{\mathrm{SLDA}}(\hat{z}|d) > p_{\mathrm{LDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42	Reuters20
SA-SLDA	78.2/1270	84.0/617	60.0/9101
PCo-SLDA	81.3/1270	85.9/617	55.5/9101
SCo-SLDA	79.5/1270	89.0/617	57.3/9101

即使识别错误，与 LDA 模型相比，SLDA 模型在大多数文档中仍然会给正确话题一个较高的权重。换言之，与 LDA 模型相比，SLDA 模型给错误的话题一个较低的权重。表4.6显示，在所有文档中，与 LDA 模型相比，SLDA 模型在大多数文档中正确话题的概率值更高。上面的分析证明了 SLDA 模型可以增加正确话题的概率减少错误话题的概率。因此 SLDA 模型更加适合于一些需要用话题分布的权重作为输入的复杂系统。

4.3.3　词义归纳

本小节在词义归纳任务中评测我们提出模型的词义。在这个实验中，我们仅评测了 SCo-SLDA 的词义。

4.3.3.1　实验设置

测试集

我们使用 Semval-2007 标准数据集中针对词义归纳和识别任务的部分[126]。这个数据集包含了宾州树库 II 数据集中的文本，来自 1989 年《华尔街时报》（Wall Street Journal，WSJ）的前半部分。目标词是 35 个名词，并且根据 OntoNotes[144] 人工标注了词义。平均词义个数 3.9。

训练集

参考文献 [125]，我们分别使用英国国家语料库（The British National Corpus，BNC) 和 WSJ 语料库分别作为跨领域和同领域数据集。对于 BNC 数据集，我们对每个目标词都构造一个数据集，包括含有目标词的所有文档。因此一共有 35 个目标词的 35 个数据集。对于 WSJ 语料库，我们使用 1987-1989 年的所有文档（排除宾州树库 II 部分）并且同样构造了 35 个目标词的数据集。在 BNC 和 WSJ 数据集中，我们对于每个词都分别构造单独的模型。数据集中所有词的词义都进行估计但是仅仅评测目标词的词义。另外，我们使用 Senseval-2 和 Senseval-3 的示例数据作为开发集调试模型的超参数。

评测指标

词义归纳任务有两种评测指标[126]：

1. 将系统输出与标准结果作比较，评测指标直接采用聚类任务的评测指标（纯度，熵，F 值等）。
2. 将标准结果分为测试语料和训练语料。后者主要是用来映射系统归纳的词义和标准标签，接着用得到的映射计算系统在测试集上的 F 值。

Semval-2007 的评测[126] 表明，第一种评测指标不能很好的评测实际系统。这种评测方法忽略了标签的实际意义，同时对单词义（每词仅指定一个词义）方法有所趋向。因此，我们使用第二种评测指标。

系统参数

对于 SCo-SLDA 模型，在 WSI 步骤中，每个词的超参数 γ_w, ρ_w , ϵ_w 设定为：$\gamma_w \sim Gamma(1, 0.1), \rho_w \sim Gamma(1, 0.1), \kappa_w \sim Gamma(0.1, 0.028), \epsilon_w = 0.1$；在 DR 步骤中，参数为 $\alpha = 1.5$ 和 $\beta = 0.1$ 。所有的超参数均在开发集中进行优化，在所有的实验中，为了避免吉布斯抽样的随机性，先进行 2000 轮抽样，然后在进行 200 轮，抽取其中的每 20 轮结果做平均。

4.3.3.2　实验 #3.1：不同的主题个数

实验 #3.1 旨在研究主题个数对词义归纳的影响。所有目标词的数据集均选用相同的主题个数。为了降低计算复杂度，我们随机的选取 9 个词来评测话题个数。我们在 10 到 70 的范围内尝试了若干主题个数的值。实验结果如图4.7所示。

实验结果分析

从实验结果中可以看出，当主题个数设为 30 的时候系统获得最好的性能。同时结果还显示用 WSI 语料训练的模型性能要好于 BNC 语料，这是因为测试集是从 WSJ 语料中抽取出来的。同领域的上下文信息比跨领域的上下文信息一致性更高，效果更好。

4.3.3.3　实验 #3.2：不同的 WSI 模型

实验 #3.2 旨在验证 SCo-SLDA 在 WSI 的有效性，我们采用 HDP 作为基线系统。HDP 的参数与 SCo-SLDA 模型在 WSI 部分相同。实验结果如表4.7所示。

实验结果分析

从表4.7中可以看出，SCo-SLDA 模型在 BNC 和 WSJ 数据集上均好于 HDP 模型。这证明了词的主题和词义的协同推导过程对词义归纳有着积极的影响。

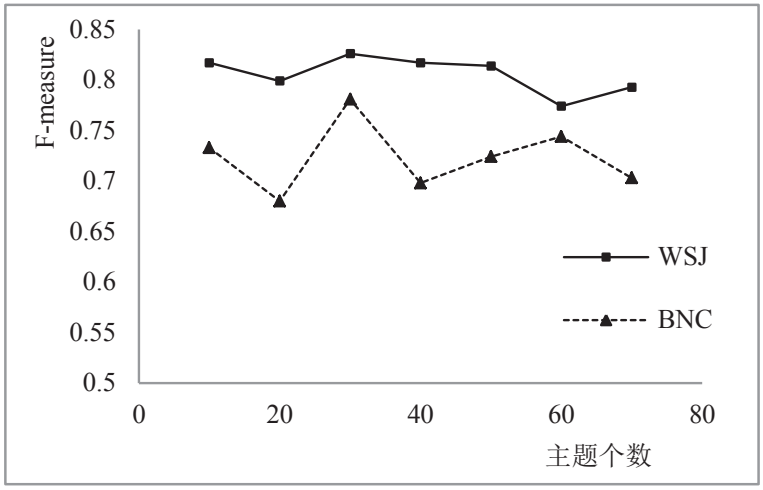


图 4.7　不同主题个数下 SCo-SLDA 模型的 WSI 结果

表 4.7　HDP 模型和 SCo-SLDA 模型在 WSI 任务中的性能

系统 \ 数据集	WSJ	BNC
HDP	0.857	0.843
SCo-SLDA	0.867	0.855

表4.8显示的是对于词“plant”，SCo-SLDA 模型分别在同领域语料和跨领域语料中学习到的词义。这里仅展示了概率最高的几个上下文词。这些词义被人工的映射到 OntoNotes 词义。表中还列出了每个词义中最高的上下文词。从表4.8中可以看出，WSJ 数据集中学习到的词义可以更好的映射到 OnteNote 上。另外，在 BNC 数据集中学习到的“plant”的词义的粒度更细。表4.9还分别列出了从 BNC 数据集和 WSJ 数据集中归纳出每个词的词义平均个数。SCO-SLDA 模型中从 BNC 数据集中学习到更多的词义。这是因为 BNC 数据集和测试集（SemEval-2007，来自 1989 年的 WSJ 数据集）有着领域的差异，BNC 数据集的领域更加广泛。

4.4　本章小结

针对如何用词义更好地表示文档这一问题，本章将词义看做主题模型的一个隐藏变量，设计了三个基于词义的主题模型：SA-SLDA，PCo-SLDA 和 SCo-SLDA。SA-SLDA 模型首先用词义归纳算法获得词义，然后在 LDA 中用词义作为元素生成主题。PCo-SLDA 和 SCo-SLDA 则将词义看做主题模型的隐藏变量，协同迭代估计词义和主题。PCo-SLDA 是通过点估计的方法确定词义而 SCo-SLDA 则是考虑词义的分布。

根据词义 -主题和文档 -主题的平均概率分布分析，基于词义的主题模型生成

表 4.8　SCo-SLDA 模型分别从 WSJ 和 BNC 语料中学习到的词“plant”的词义与 OntoNotes 词义的对比

Sense induced from WSJ	Sense induced from BNC	Sense in OntoNotes
词义 #1：{company, product, build, facility, capacity}	词义 #1：{company, build, sale, worker, production} 词义 #2：{people, work, worker, job, manager}	词义 #1：{A building for industrial activity}
词义 #2：{seed, gene, corn, produce, pollen}	词义 #3：{grow, garden, flower, seed, plot} 词义 #4：{species, animal, tree, insect, forest}	词义 #2：{Living photosynthesizing organism}
词义 #3：{power, energy, gas, operate, cogeneration}	词义 #5：{nuclear, power, chemical, waste, process}	词义 #3：{A contrivance or stratagem}

表 4.9　用 WSJ/BNC 数据集训练，SemalEval-2007 数据集测试时，SCo-SLDA 模型学习到每个词的平均词义个数

系统 \ 数据集	WSJ	BNC
SCo-SLDA	5.3	8.4

更加尖锐的主题分布，也就是说该模型可以提供更强的后验分布估计。实验结果表明基于词义的主题模型可以改进单语言话题分析的性能。

具体来说，协同学习模型（PCo-SLDA 和 SCo-SLDA）比独立模型（SA-SLDA）的性能更好，因为词义和话题的估计可以协同的改进。另外，SCo-SLDA 考虑到整个词义混合，因此性能优于采用点估计方法估计词义的模型 PCo-SLDA。

我们还评测了 SCo-SLDA 在词义归纳任务上的性能，实验结果验证了使用主题作为词义的伪回馈可以归纳和识别出更准确的词义。

但是本章的模型仅能用于单语言的情况，无法解决跨语言障碍。为了进行跨语言话题分析，下章将重点研究基于词义的跨语言主题模型。

第 5 章　基于统计词义的跨语言主题模型

5.1　本章引论

5.1.1　研究问题

基于统计词义的主题模型介绍了如何用词义改进单语言主题模型。但是这个模型不能处理文档集中含有多种语言的情况。如果一个文档集中既含有中文文档又含有英文文档，基于统计词义的主题模型无法识别中文文档和英文文档中包含的同一话题。

因此，文档建模需要识别不同语言文档中的同一话题。针对这一跨语言问题，本章将介绍基于统计词义的跨语言主题模型。跨语言主题模型与单语言主题模型相比，难点在于如何解决跨语言问题。

5.1.2　问题分析

现有的跨语言模型通常是在单语言主题模型的基础上借助词典或者平行/可比语料库进行拓展。在1.2.2.2节中提到，跨语言主题模型的拓展方式有两种：第一种是主题的对齐，在平行或者可比语料中训练跨语言主题空间[77–81]。这种方法的问题在于不同数据集上的主题偏差。第二种是根据词对齐信息直接在数据集上构造跨语言主题空间[82–85]。文献 [82] 和文献 [83] 分别拓展了 PLSA 和有指导的 LDA 模型，与我们的工作不太相关。文献 [84] 用双语词对拓展了 LDA（Muto 模型），但是这个方法有一个限制：词在一个数据集上只能有一个翻译。JointLDA[85] 则没有这一限制，可以在一定程度上解决语言的对应问题和翻译歧义问题。但是它的消歧是在文档集全局层面上进行的，难度较大，会导致结果下降。

第3章考虑了词义对齐，在平行语料中生成统计词义，提出基于跨语言全局词义的文档表示方法，但是这种方法没有考虑到主题信息对词义生成的影响。第4章将词义看做主题模型的一个隐藏变量，设计了基于词义的主题模型，但是这种方法无法解决跨语言问题。

5.1.3　解决思路

本章在第4章提出的 SCo-SLDA 模型的基础上，引入词对齐信息，提出了两个基于词义的跨语言主题模型：词对齐词义 LDA 和词义对齐词义 LDA。在词对

齐词义 LDA 模型中，将词对看作变量，加入到 SCo-SLDA 模型中，解决跨语言的对应问题。在词义对齐词义 LDA 中，除了考虑词对齐信息，还估计了词义对齐，构造了主题、词义、词义对齐的迭代估计过程。本章在跨语言话题分析任务中评测了两个模型。

本章的贡献主要在两个方面：第一，用词对齐信息拓展了 SCo-SLDA，使其可以解决跨语言问题；第二，考虑了词义，词义对齐以及话题的相互影响，构造了词义，词义对齐以及话题的协同估计过程，有利于解决翻译歧义的问题。与前人利用词对齐信息的跨语言主题模型[82–85] 相比：首先，本章提出的方法可以自动的估计词义，改进主题的性能；其次，本章提出的词义对齐词义 LDA 模型可以自动的获得词义对齐信息，与仅用词对齐的方法比，更有利于获得准确地词义分布，进一步改进主题的性能。

本章剩余部分按照如下方式组织：5.2节介绍了基于词义的跨语言主题模型。5.3节评测了基于词义的主题模型在话题分析任务上的性能并且分析了这些主题模型的词义 -主题分布和文档 -主题分布。5.4节对本章进行了总结。

5.2 基于词义的跨语言主题模型

图5.1对比了 JointLDA 模型[85] 与本章提出的基于词义的跨语言主题模型。如图5.1所示，JointLDA 模型在传统 LDA 的基础上加入词对齐，并为每个对齐词对指定一个话题，然后选择词对齐中的一个词。基于词义的跨语言主题模型同样为每个词对指定一个话题，然后选择词对的词义。同单语言的主题模型类似，这个模型的基本元素是词义。词义也是作为隐藏变量加入到主题模型中，根据上下文词归纳出来。本文设计了两种方法来实现基于词义的跨语言主题模型。

- 词对齐词义 LDA（Word aligned SLDA，WA-SLDA）：这个方法直接将词对齐加入到 SCo-SLDA 模型中，然后迭代的学习每个词的话题、词对以及词义。
- 词义对齐词义 LDA（Word sense aligned SLDA，WSA-SLDA）：这个方法在 WA-SLDA 的基础上进一步学习了词义对齐，构建了话题、词对、词义和词义对齐的迭代过程。

下面我们将分别介绍提出的两个方法。为了便于对比，我们首先介绍 JointLDA 模型[85]。

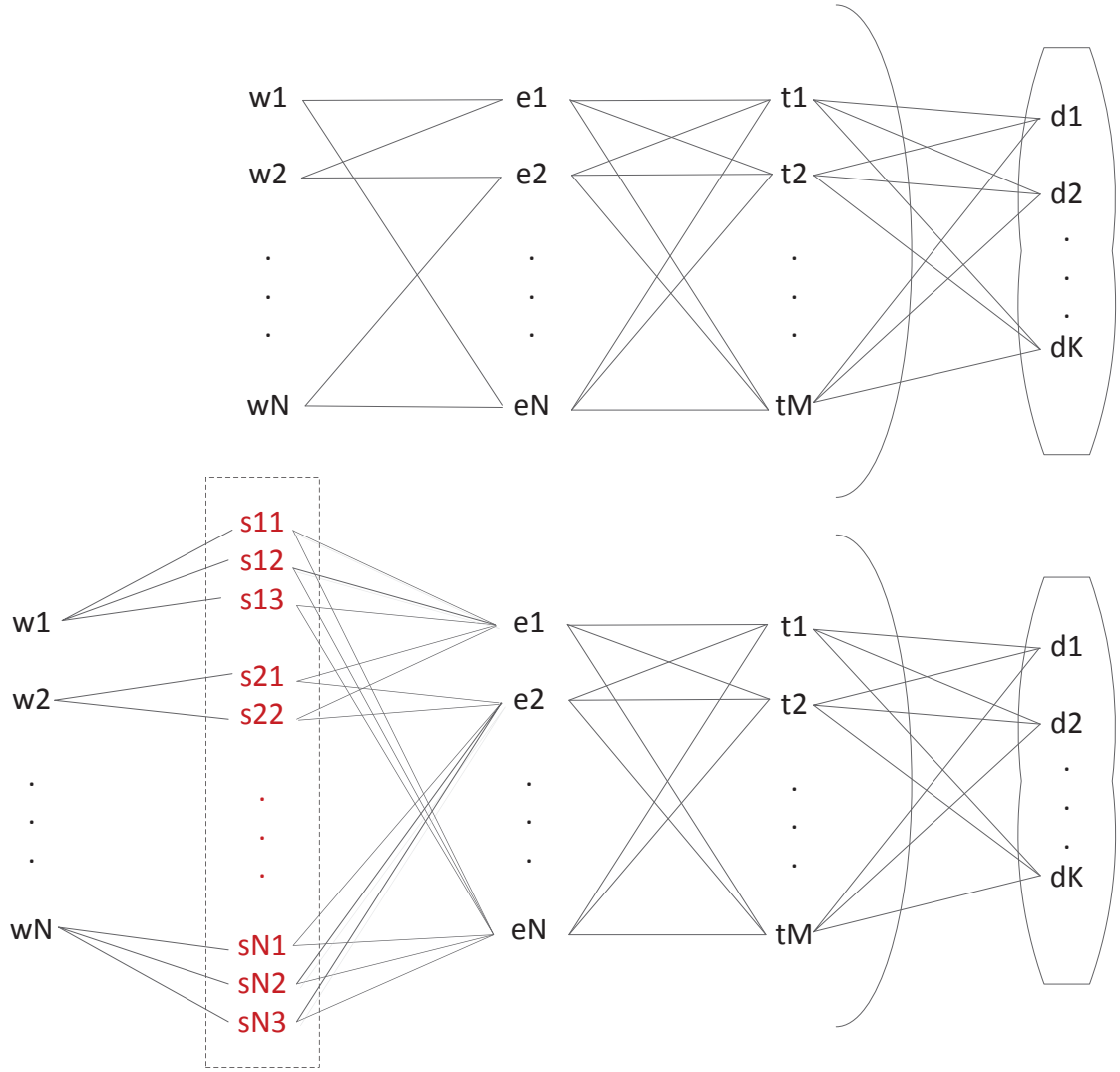


图 5.1　JointLDA 模型（上图）和基于词义的跨语言主题模型（下图）对比示意图。虚线框中的变量为潜变量（即词义）

5.2.1　JointLDA 模型

图5.2表示 JointLDA 的概率模型[85]。其中可见变量词 w_{ij} 和文档 d_i 的语言 l_{d_i} 用阴影表示，隐藏变量 z_{ij} 和 e_{ij} 分别表示词 w_{ij} 的主题和对齐词对。给定文档集 D 和文档中的词集 W，对齐词对集为 E，主题集为 Z，JointLDA 的概率生成模型如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 选择文档的语言 $l_{d_i} \sim Binomial(1/2)$.
   (c) 对于文档 d_i 中的词 w_{ij}:

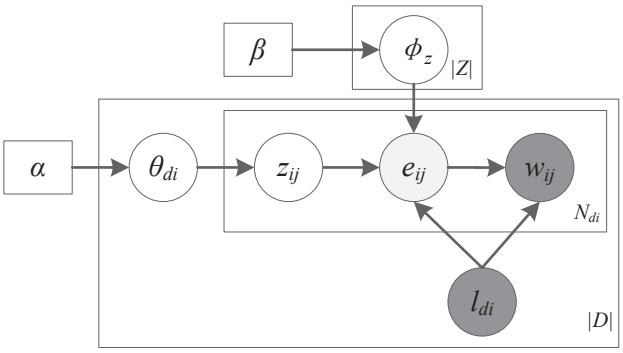


图 5.2　JointLDA 模型

i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.

ii. 选择对齐词对 $e_{ij} \sim Mult(\phi_{z_{ij}}) \cdot \varphi(e_{ij}, l_{d_i})$.

iii. 选择词 $w_{ij} \sim p(w_{ij}|e_{ij}, l_{d_i})$.

其中 d_i 表示文档集 D 中的 i 个文档，w_{ij} 表示文档 d_i 的第 j 个词，z_{ij} 表示词 w_{ij} 的主题，e_{ij} 表示词 w_{ij} 的对齐词对。α, β 是模型的超参数。$\phi_{z_{ij}}$ 和 θ_{d_i} 分别表示主题-词对分布和文档 -主题分布。这两个分布都是从狄利克雷分布抽样出来的。值得注意的是给定对齐词对和语言，只能有一种可能生成词，因此 $p(w_{ij}|e_{ij}, l_{d_i}) = 1$。

JointLDA 对未登录词处理方法是对所有的未登录词都人为的创建一个对齐词对 w : _NA_ 或者 _NA_ : w。人工创建的对齐词对和真正对齐词对的区别是前者只能在一种语言中有词，后者则在两种语言中都有词。因此在 JointLDA 模型生成过程中，如果词对 e_{ij} 在语言 l_{d_i} 中有词，$\varphi(e_{ij}, l_{d_i})$ 的值为 1，否则为 0。

JointLDA 采用吉布斯抽样（Collapsed Gibbs Sampling）学习模型[85]。在迭代过程中，词 $w_{ij} = w$ 的话题和词对条件概率公式按照公式 (5-1) 计算。

$$P(z_{ij} = z, e_{ij} = e|\boldsymbol{z_{-ij}}, \boldsymbol{e_{-ij}}, \boldsymbol{w}, \boldsymbol{l}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{e}_{-ij,z} + \beta}{n_{-ij,z} + W\beta} \times p(w_{ij}|e_{ij}, l_{d_i}) \tag{5-1}$$

在公式 (5-1) 中，$n^{d_i}_{-ij,z}$ 表示文档 d_i 中话题为 z 的词数；$n^{e}_{-ij,z}$ 表示话题为 z 同时 $e_{ij} = e$ 的词数；$n^{d_i}_{-ij}$ 表示文档 d_i 的词数；$n_{-ij,z}$ 表示话题为 z 的词数。上面变量中的 $-ij$ 表示在计数过程中去掉词 w_{ij}。

5.2.2　词对齐词义 LDA

JointLDA 的思想是将词对信息插入到话题和词之间，通过对齐的词来构造跨语言主题模型。词对齐词义 LDA 模型的跨语言处理与 JointLDA 相似，在 SCo-SLDA 模型的基础上，将词对信息插入到话题和词义之间，通过对齐的词来

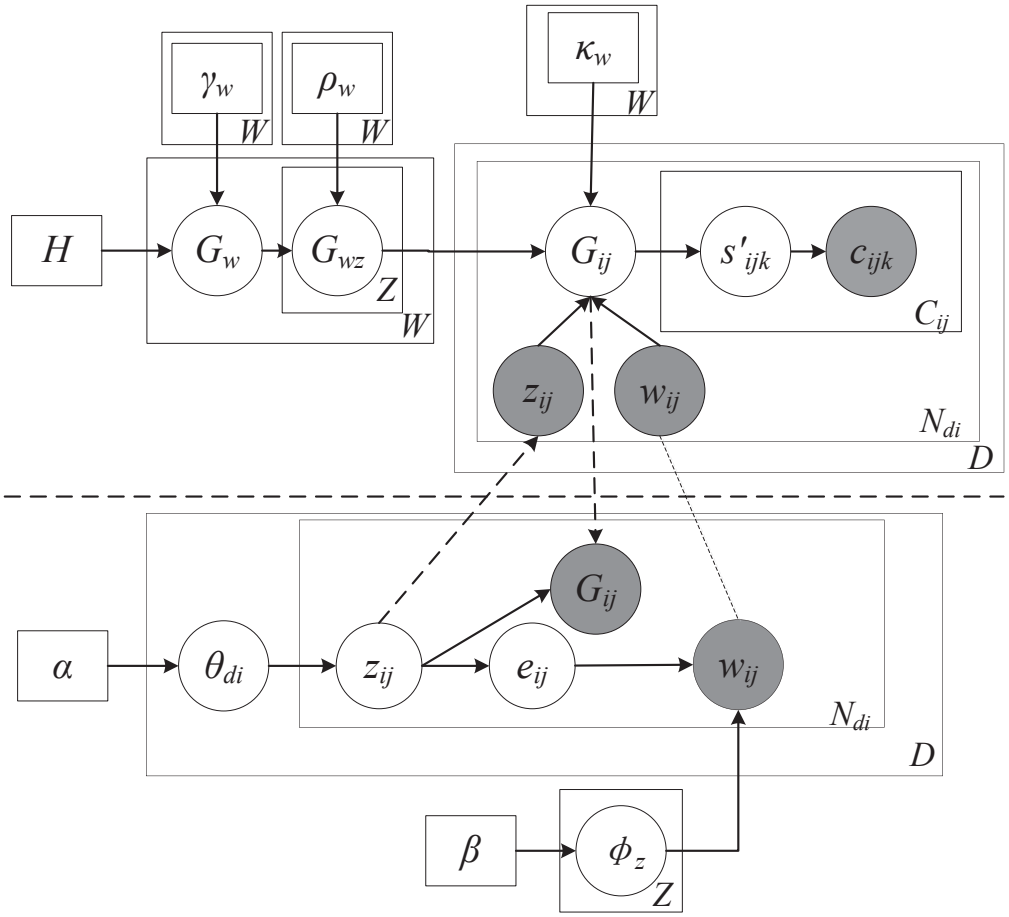


图 5.3　WA-SLDA 模型

解决跨语言障碍。对于未登录词的处理也与 JointLDA 相同，对于所有的未登录词都创建一个对齐词对 w : _NA_ 或者 _NA_ : w。

图5.3表示 WA-SLDA 模型。在图5.3中，虚线之上的部分表示 WSI 步骤，虚线之下的部分表示 DR 步骤。两个步骤的交替过程与 SCo-SLDA 模型相同。

WA-SLDA 的词义归纳的过程与 SCo-SLDA 相同，如下：

1. 对于每个词 w:
   (a) 选择词义分布 $G_w \sim DP(\gamma_w, H_w)$.
   (b) 对于每个主题 z:
      i. 选择词义分布 $G_{wz} \sim DP(\rho_w, G_w)$.
2. 对于每个文档 d_i:
   (a) 对于词 w_{ij} 的上下文窗口 v_{ij} :
      i. 选择词义分布 $G_{ij} \sim DP(\kappa_{w_{ij}z}, G_{w_{ij}z})$.
      ii. 对于目标词 w_{ij} 的每个上下文词 c_{ijk}:
         A. 选择词义 $s_{ijk} \sim G_{ij}$.
         B. 选择上下文词 $c_{ijk} \sim Mult(\eta_{s_{ijk}})$.

WA-SLDA 的文档表示与 SCo-SLDA 的区别在于加入了词对信息，具体过程如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 的每个词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择对齐词对 $e_{ij} \sim Mult(\phi_{z_{ij}}) \cdot \varphi(e_{ij}, l_{d_i})$.
      iii. 选择词 $w_{ij} \sim p(w_{ij}|e_{ij}, l_{d_i})$.
      iv. 选择词义分布 $G_{ij} \sim DP(\kappa_{w_{ij}z}, G_{w_{ij}z})$.

在 WA-SLDA 模型中，词的主题和词对决定于三个因素：

1. 文档的主题分布，
2. 给定主题，生成词对的概率，
3. 给定词的词对和主题，生成词义的概率。

前面两个因素与 JointLDA 相同，而最后一个因素反映了词义的影响。

WA-SLDA 采用吉布斯抽样（Collapsed Gibbs Sampling）进行学习[5]。WA-SLDA 的学习过程与 SCo-SLDA 类似，有两组隐藏变量需要交替估计：

第一组是给定每个词的词义的主题 z_{ij} 和词对 e_{ij}，估计词义分布 G_{ij}。由于词义分布 G_{ij} 与 e_{ij} 无关，G_{ij} 的估计过程与文献 [128] 相同。

第二组是给定词义分布，估计 z_{ij} 和 e_{ij}。z_{ij} 和 e_{ij} 的条件概率公式可以按照公式5-2计算。

$$
\begin{aligned}
&P(z_{ij} = z, e_{ij} = e|\boldsymbol{z_{-ij}}, \boldsymbol{e_{-ij}}, \boldsymbol{w}, \boldsymbol{l}) \\
&\propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{e}_{-ij,z} + \beta}{n_{-ij,z} + W\beta} \times p(w_{ij}|c_{ij}, l_{d_i}) \times \frac{\prod_{s\in\{s_w\}} \prod_{g=0}^{n^{s}_{ij}-1} (\kappa_{wz}\pi^{s}_{wz} + g)}{\prod_{g=0}^{C_{ij}-1} (\kappa_{wz} + g)}
\end{aligned}
\tag{5-2}
$$

5.2.3　词义对齐词义 LDA

5.2.3.1　从词对齐到词义对齐

WA-SLDA 将词对信息加入到词义主题模型中，构造了主题、词对、词义的协同学习过程。其中词的词义与上下文词以及词义在主题上的分布相关，与词对无关。进一步考虑词对和词义的关系，有两点值得注意：

首先，词对中两个词存在对齐词义。比如在第3章中提到，词 arm 有两个词义：

例 5.1：

{

arm# 1={ limb: 0.159, forelimb: 0.069, sleeve: 0.019 }

arm# 2={weapon: 0.116, war: 0.039, battle: 0.026 }

}

词“手臂”的词义为：

例 5.2：

{

手臂 # 1={ 胳膊: 0.159, 上肢: 0.089, 衣袖: 0.014 }

}

因此词对“arm: 手臂”有一个对齐词义：

例 5.3：

{

arm# 1={ limb: 0.159, forelimb: 0.069, sleeve: 0.019 }

手臂 # 1={ 胳膊: 0.159, 上肢: 0.089, 衣袖: 0.014 }

}

其次，对齐词义中的两个词义存在相互影响。如果一个主题出现 arm# 1 的概率较高，那么它出现手臂 # 1 的概率也较高。

为了表示词对和词义的关系，我们首先定义了词对的词义集合：

定义 5.1：　词对 $e = w_{l_1} : w_{l_2}$ 的词义集合 s_e 由词对中词的词义集合组成，

$$s_e = \{s_e^1; s_e^2\} = \{\boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_1}}, \boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_2}}\} \tag{5-3}$$

其中 $\boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_i}}; i = 1, 2$ 表示词 w_{l_i} 的词义集合。

由此，词对“arm: 手臂”的词义集合可以表示为：

例 5.4：

{

arm# 1={ limb: 0.159, forelimb: 0.069, sleeve: 0.019 }；

arm# 2={weapon: 0.116, war: 0.039, battle: 0.026 }；

手臂 # 1={ 胳膊: 0.159, 上肢: 0.089, 衣袖: 0.014 }；

}

为了表示词对中的对齐词对，我们定义了词对的对齐词义集合：

定义 5.2：　词对 $e = w_{l_1} : w_{l_2}$ 的对齐词义集合 s'_e 由词对中的词的词义集合映射而成，

$$s'_e = \{s'^1_e : s'^2_e\} = \{\boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_1}} \cdot M^{l_1}_e : \boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_2}} \cdot M^{l_2}_e\} \tag{5-4}$$

其中 $\boldsymbol{s}_{\boldsymbol{w}_{\boldsymbol{l}_i}}; i = 1,2$ 表示词 w_{l_i} 的词义集合，$M^{l_i}_e; i = 1,2$ 是词 -词对的词义映射矩阵，$M^{l_i}_e$ 中每个元素定义如公式 (5-5)。

$$m_{gh} = \begin{cases} 1 & \text{词 } w_{l_i} \text{ 的第 } g \text{ 个词义对映词对 } e \text{ 的第 } g \text{ 个词义;} \\ 0 & \text{其他;} \end{cases} \tag{5-5}$$

其中 $k = 1, \cdots S_{w_{l_i}}$，$S_{w_{l_i}}$ 表示词 w_{l_i} 的词义个数；$h = 1, \cdots S_e$，S_e 表示词对 e 的词义个数。

由此，词对“arm: 手臂”的对齐词义集合可以表示为：

例 5.5：

{

arm#1: 手臂 #1

arm#2: _NA_

}

其中 _NA_ 表示没有词义对应。

5.2.3.2　词义对齐词义 LDA

在 WSA-SLDA 模型中，不仅考虑了主题和词对，主题和词义的相互作用；还需要考虑词对和词义的相互影响。

图5.4表示 WSA-SLDA 模型。在图5.4中，虚线之上的部分表示 WSI 步骤，虚线之下的部分表示 DR 步骤。两个步骤的交替过程与 WA-SLDA 模型类似，区别在于 WA-SLDA 中词义和词对互不相关，而 WSA-SLDA 模型考虑了词义和词对的相互作用。我们认为词的词义分布 G_{ij} 与这个词的词对在主题中的分布 G_{ez} 相关。在 WSA-SLDA 模型中，我们假设词义分布 G_{ij} 是从 G_{ez} 为基分布的狄利克雷过程中抽样出来的，即公式 (5-6)，其中 l_w 为词 w 的语言[①]。

$$G_{ij} \sim {M^{l_w}_e}^{-1} \cdot DP(\kappa_{ez}, G_{ez}) \tag{5-6}$$

① $M^{l_w}_e$ 其实是一个置换矩阵，所以它一定可逆。

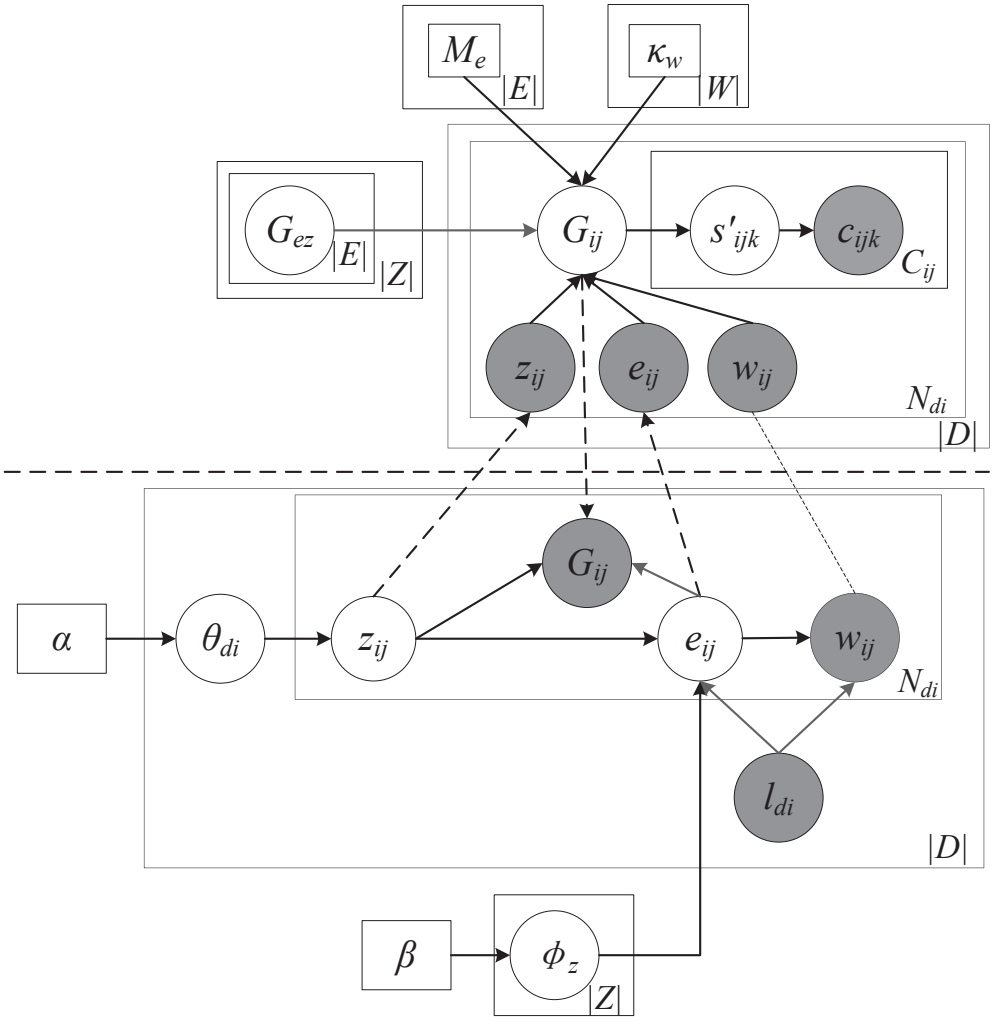


图 5.4　WSA-SLDA 模型

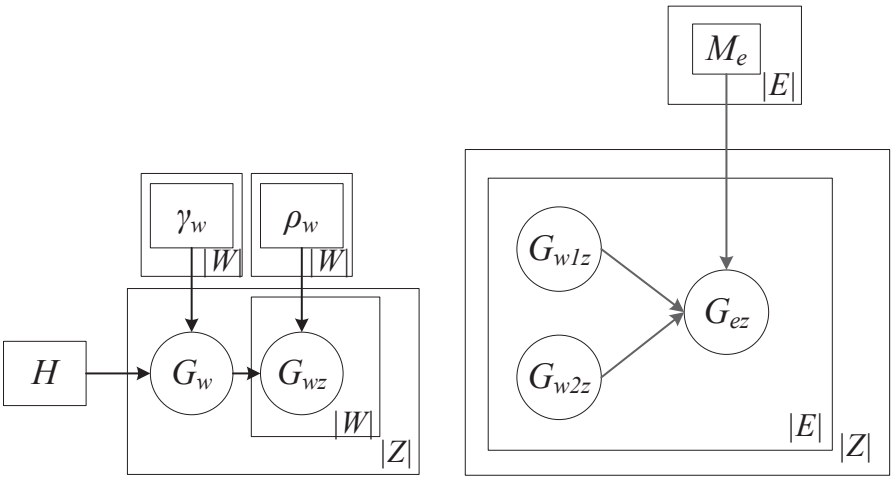
图5.5表示图5.4中变量 G_{wz} 和 G_{ez} 的生成过程。与 WA-SLDA 相同，词 w 的 G_{wz} 定义了一个 HDP 先验。而词对 e 的 G_{ez} 则决定于词对中的词 w_{l_1} 和 w_{l_2}。我们这里将 G_{ez} 简单的定义为 $G_{w_{l_1}z}$ 和 $G_{w_{l_2}z}$ 的平均分布，即公式 (5-7)；同样 G_{ez} 的参数 κ_{ez} 定义如公式 (5-8)。

$$G_{ez} = \frac{\varphi(e,l_1)G_{w_{l_1}z} \cdot M_e^{l_1} + \varphi(e,l_2)G_{w_{l_2}z} \cdot M_e^{l_2}}{\varphi(e,l_1) + \varphi(e,l_2)} \tag{5-7}$$

$$\kappa_e z = \frac{\varphi(e,l_1)\kappa_{w_{l_1}z} + \varphi(e,l_2)\kappa_{w_{l_2}z}}{\varphi(e,l_1) + \varphi(e,l_2)} \tag{5-8}$$

对于未登录词，创建词对 w : _NA_ 或者 _NA_ : w，它的词义分布和参数与词 w 相同。

WSA-SLDA 的词义归纳过程如下：



(a) G_{wz}　　(b) G_{ez}

图 5.5　G_{wz} 和 G_{ez} 的生成过程

1. 对于每个词 w:
   (a) 选择词义分布 $G_w \sim DP(\gamma_w, H_w)$.
   (b) 对于每个主题 z:
      i. 选择词义分布 $G_{wz} \sim DP(\rho_w, G_w)$.
2. 对于每个词对 e:
   (a) 根据公式 (5-7) 指定词义分布.
3. 对于每个文档 d_i:
   (a) 对于词 $w_{ij} = w$ 的上下文窗口 v_{ij} :
      i. 选择词义分布 $G_{ij} \sim M_{e_{ij}}^{l_{w_{ij}}-1} \cdot DP(\kappa_{e_{ij}z_{ij}}, G_{e_{ij}z_{ij}})$.
      ii. 对于目标词 w_{ij} 的每个上下文词 c_{ijk}:
         A. 选择词义 $s_{ijk} \sim G_{ij}$.
         B. 选择上下文词 $c_{ijk} \sim Mult(\eta_{s_{ijk}})$.

WA-SLDA 的文档建模过程如下：

1. 对于每个主题 z:
   (a) 选择 $\phi_z \sim Dir(\beta)$.
2. 对于每个文档 d_i:
   (a) 选择 $\theta_{d_i} \sim Dir(\alpha)$.
   (b) 对于文档 d_i 的每个词 w_{ij}:
      i. 选择主题 $z_{ij} \sim Mult(\theta_{d_i})$.
      ii. 选择对齐词对 $e_{ij} \sim Mult(\phi_{z_{ij}}) \cdot \varphi(e_{ij}, l_{d_i})$.
      iii. 选择词 $w_{ij} \sim p(w_{ij}|e_{ij}, l_{d_i})$.
      iv. 选择词义分布 $G_{ij} \sim M_{e_{ij}}^{l_{w_{ij}}-1} \cdot DP(\kappa_{e_{ij}z_{ij}}, G_{e_{ij}z_{ij}})$.

在 WSA-SLDA 模型中，词的主题和词对决定于三个因素：

1. 文档的主题分布，
2. 给定主题生成词对的概率，
3. 给定词对和主题, 生成词义的概率。

前面两个因素与 JointLDA 相同，最后一个因素反映了词义的影响。最后一个因素与 WA-SLDA 的区别在于 WA-SLDA 中词义与词对无关，而 WSA-SLDA 中词对影响词义的概率分布。分析其本质，词对对词义的影响其实是词义的对齐词义的影响。例如词对 $e =$ "arm: 手臂"，对齐词义集合如例5.5所示，假设 $w_{ij} =$ "arm"，那么根据公式5-6，词义分布 G_{ij} 与 G_{ez} 相关，也就是说词义条件概率 $p(\text{"arm\#1"}|\text{"arm"}, z_{ij}, e_{ij})$ 与不仅与概率 $p(\text{"arm\#1"}|\text{"arm"}, z_{ij})$ 正相关，还与概率 $p(\text{" 手臂 \#1"}|\text{" 手臂"}, z_{ij})$ 正相关。同理在词义分布已知的情况下，词对条件概率 $p(\text{"arm: 手臂"}|\text{"arm"}, z_{ij}, G_{ij})$ 不仅与概率 $p(\text{"arm\#1"}|\text{"arm"}, z_{ij})$ 和概率 $p(\text{"arm\#2"}|\text{"arm"}, z_{ij})$ 相关，还与概率 $p(\text{" 手臂 \#1"}|\text{" 手臂"}, z_{ij})$ 相关。

5.2.3.3 参数估计

WSA-SLDA 模型需要估计如下三组参数：

1. 每个词的词义分布;
2. 每个词的主题和词对;
3. 每个词对的词 -词对的词义映射矩阵;

我们采用随机 EM（stochasitc EM）算法[145] 进行后验参数估计。参数估计有三个过程：抽样词义分布（第一组参数），抽样主题和词对（第二组参数），找到最大后验映射矩阵（第三组参数）。

词义分布、主题和词对的抽样采用吉布斯抽样（Collapsed Gibbs Sampling）[5]。

首先介绍给定每个词的主题和词对，估计词义分布 G_{ij}。词义分布 G_{ij} 的估计过程与文献 [128] 类似，其中 $\{G_w\}$，$\{G_{wz}\}$ 的抽样与文献 [128] 相同。G_{ez} 可以表示为公式 (5-9)。

$$G_{ez} = \Sigma_{s_e} \pi_{ez}^{s_e} \delta_{\eta_{s_e}} + \pi_{ez}^{u} G_{ez}^{u}, \tag{5-9}$$

根据公式 (5-7)，$\{\pi_{ez}^{s_e}\}$ 是词义混合，可以按照公式 (5-10) 计算，其中 $s'_{w_{l_1}}$ 和 s'_{w_2} 分别表示词对 $e = \{w_{l_1} : w_{l_2}\}$ 中词 w_{l_1} 和词 w_{l_2} 中通过矩阵 $M_e^{l_1}$ 和 $M_e^{l_2}$ 映射后与 s_e 对应的词义。G_{ez}^{u} 的分布为混合狄利克雷过程，可以表示为公式5-11。

$$pi_{ez}^{s_e} = \frac{pi_{w_{l_1} z}^{s'_{w_{l_1}}} * \varphi(e, l_1) + pi_{w_{l_2} z}^{s'_{w_2}} * \varphi(e, l_2)}{\varphi(e, l_1) + \varphi(e, l_2)} \tag{5-10}$$

$$G^{u}_{ez} = \frac{\varphi(e,l_1)DP(\gamma_{w_{l_1}}, H_{w_{l_1}}) + \varphi(e,l_2)DP(\gamma_{w_{l_2}}, H_{w_{l_2}})}{\varphi(e,l_1)+\varphi(e,l_2)} \tag{5-11}$$

得到 G_{ez} 和 G_{wz} 后，根据公式 (5-6)，词义 s_{ijk} 的条件概率可以按照公式 (5-12) 计算，其中 $C_{w_{ij}}$ 为词 w_{ij} 的上下文词的个数；$n^{v_{ij}}_{-ijk,s}$ 表示上下文窗口 v_{ij} 中词义为 s 的个数；$n^{c_{ijk}}_{-ijk,s}$ 表示词义为 s 的上下文词 c_{ijk} 的个数；$n_{-ijk,s}$ 表示词义为 s 的上下文词的个数，其中 $-ijk$ 表示在技术过程中排除上下文词 c_{ijk}。

$$\begin{aligned} &P(s_{ijk} = s|\boldsymbol{s_{-ijk}}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{e}) \\ &\propto (n^{v_{ij}}_{-ijk,s} + \kappa_{e_{ij}z_{ij}} pi^{s'}_{e_{ij}z_{ij}}) \frac{n^{c_{ijk}}_{-ijk,s} + \epsilon_{w_i j}}{n_{-ijk,s} + C_{w_{ij}}\epsilon_{w_{ij}}} \end{aligned} \tag{5-12}$$

其次介绍给定词义分布，估计话题和词对。词 w_{ij} 的 z_{ij} 和 e_{ij} 的条件概率可以按照公式 (5-13) 计算。

$$\begin{aligned} &P(z_{ij} = z, e_{ij} = e|\boldsymbol{z_{-ij}}, \boldsymbol{e_{-ij}}, \boldsymbol{w}, \boldsymbol{l}) \\ &\propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{e}_{-ij,z} + \beta}{n_{-ij,z} + W\beta} \times p(w_{ij}|c_{ij}, l_{d_i}) \times \\ &\frac{\prod_{s\in\{s_w\}} \prod_{g=0}^{n^{s-1}_{ij}} (\kappa_{ez} pi^{s'}_{ez} + g)}{\prod_{g=0}^{C_{ij}-1} (\kappa_{ez} \sum_{s\in\{s_w\}} pi^{s'}_{ez} + g)} \end{aligned} \tag{5-13}$$

最后介绍如何找到最大后验映射矩阵。映射矩阵可以利用匈牙利算法[146] 最大化似然概率来获得。词 w 中的第 g 个词义与词对 e 中的第 h 个词义映射的对数似然概率为：

$$\begin{aligned} &\mu_{gh} \\ &= logp(\boldsymbol{z}, \boldsymbol{e}, \boldsymbol{s}|m_{gh} = 1, \boldsymbol{M_{-g,-h}}, \boldsymbol{w}, \boldsymbol{G_{wz}}, \boldsymbol{G_{ez}}, \boldsymbol{\theta}, \boldsymbol{\beta}, H) \\ &\propto \sum_{i=1}^{|D|} \sum_{j=1}^{N_{d_i}} \sum_{k=1}^{|C_{ij}|} logp(z_{ij}, e_{ij} = e, s_{ijk} = s^{g}_{w}|m_{gh} = 1, \boldsymbol{M_{-g,-h}}, \boldsymbol{w}, \boldsymbol{G_{wz}}, \boldsymbol{G_{ez}}, \boldsymbol{\theta}, \boldsymbol{\beta}, H) \\ &\propto \sum_{i=1}^{|D|} \sum_{j=1}^{N_{d_i}} \sum_{k=1}^{|C_{ij}|} logp(s_{ijk} = s^{g}_{w}|m_{gh} = 1, w, G_{wz}, G_{ez}, H) \\ &= \frac{n^{v_{ij}}_{s} + \kappa_{e_{ij}z_{ij}} * pi^{s'}_{e_{ij}z_{ij}}}{n^{v_{ij}} + \kappa_{e_{ij}z_{ij}} * \sum_{s\in\{s_{w_{ij}}\}} pi^{s'}_{e_{ij}z_{ij}}} \frac{n^{c_{ijk}}_{s} + \epsilon_{w_{ij}}}{n_s + C_{ij}\epsilon_{w_{ij}}} \end{aligned} \tag{5-14}$$

5.3 实验评测

基于词义的主题模型的评测一共有三个部分：第一，评测跨语言主题；第二，分析所提出模型的词义 -主题分布和文档主题分布；最后，比较了本章提出的文档

建模方法 WA-SLDA，WSA-SLDA 与现有方法以及前几章提出的文档建模方法。

5.3.1　跨语言话题分析

本实验在跨语言话题分析任务上评测模型生成的主题，并且比较所提出模型 WA-SLDA，WSA-SLDA 与 JointLDA 等基线系统的性能。

5.3.1.1　实验设置

实验数据

本章的实验数据（开发集和测试集）与2.5节相同。

词对齐

本章评测了词典和翻译概率两种词对齐信息。我们从 HowNet 中抽取翻译词对作为词典，在开发集中使用 Giza++[112] 获取翻译概率。

评测指标

本章采用与2.5节相同的评测指标，即 F-measure。

5.3.1.2　实验 #1.1：不同的词对齐来源

实验 #1.1 旨在比较不同的词对齐来源对词义主题模型的影响。我们实现了四种不同的跨话题分析系统。

- **JointLDA&DIC**：这个系统利用词典中的对齐信息，使用 JointLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **JointLDA&TRAN**：这个系统利用翻译概率中的对齐信息，使用 JointLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **WA-SLDA&DIC**：这个系统利用词典中的对齐信息，使用 WA-SLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **WA-SLDA&TRAN**：这个系统利用翻译概率中的对齐信息，使用 WA-SLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。

表 5.1 不同词对齐（词典和翻译概率）下的系统性能

系统 \ 数据集	TDT41	TDT42
JointLDA& DIC	0.759	0.678
JointLDA&TRAN	0.766	0.729
WA-SLDA& DIC	0.784	0.737
WA-SLDA&TRAN	**0.847**	**0.747**

系统参数

WA-SLDA 的参数设定如下：

1. 在 WSI 步骤中，每个词的超参数 γ_w, ρ_w , ϵ_w 设定为：$\gamma_w \sim Gamma(8, 0.1)$, $\rho_w \sim Gamma(5, 1), \kappa_w \sim Gamma(0.1, 0.028), \epsilon_w = 0.1$ 。
2. 在 DR 步骤中，$\alpha = 1.5$ ，$\beta = 0.1$ 。

本实验中所有的超参数均是在 TDT42 数据集上进行优化的。文档的主题个数设为该数据集的类簇数。在所有实验中，为了避免吉布斯抽样的随机性，先进行 2000 轮抽样，然后在进行 200 轮，抽取其中的每 20 轮结果做平均。

实验结果如表5.1所示。

实验结果分析

从表5.1中可以看出：

首先，利用翻译概率作为对齐信息的系统 JointLDA&TRAN 和 WA-SLDA&TRAN 的性能好于利用词典作为对齐信息的相应系统 JointLDA&DIC 和 WA-SLDA&DIC 。这是由于从 HowNet 抽取的词典比开发集中的翻译概率的未登录词要多。后面的实验我们均采用翻译概率作为对齐信息。

其次，利用相同对齐信息的系统中，WA-SLDA&TRAN 的性能好于 JointLDA&TRAN 而 WA-SLDA&DIC 的性能好于 JointLDA&DIC 。这是因为 WA-SLDA 是基于 SCo-SLDA 的拓展，而 JointLDA 是基于传统 LDA 的拓展由于 SCo-SLDA 考虑到词义及其与话题的相互影响，性能要好于 LDA，这已经在第4章中得到证明，那么以更好模型作为基础的 WA-SLDA 的性能也要好于 JointLDA。

5.3.1.3 实验 #1.2：不同的对齐方式

实验 #1.2 旨在比较不同的对齐方式（词对齐和词义对齐）对主题模型的影响，实现了两种不同对齐方式的系统。

- **WA-SLDA**：这个系统利用翻译概率中的对齐信息，使用 WA-SLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大

的主题中。这个系统与实验 #1.1 中的 WA-SLDA&TRAN 系统相同。

- **WSA-SLDA**：这个系统利用翻译概率中的对齐信息，使用 WSA-SLDA 对文档中的主题进行建模，然后将文档的主题看成话题并将每个文档分到概率最大的主题中。

系统参数

WSA-SLDA 采用的参数与 WA-SLDA 相同，均采用实验 #1.1 的参数。

实验结果如表5.2所示。

表 5.2 采用不同文档建模模型的话题分析系统的 F-measure

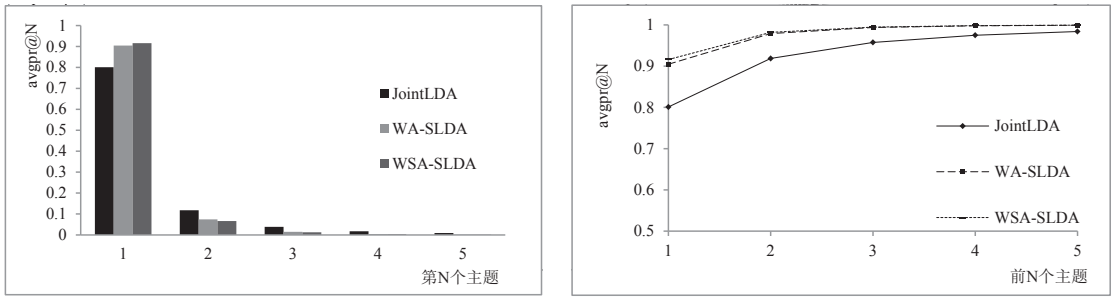
系统 \ 数据集	TDT41	TDT42
WA-SLDA	0.847	0.747
WSA-SLDA	**0.863**	**0.817**

表 5.3 词对“drug: 药物”的词义

drug	药物
_NA_	药物 #1={ 地震: 0.069, 短缺: 0.049, 供应: 0.009 }
drug#1= { patient: 0.075, medically-assisted: 0.057, doctor: 0.009 }	药物 #2={ 病人: 0.069, 医生: 0.049, 生命: 0.009 }
drug#2= { make: 0.055, scientist: 0.043, mechanism: 0.012 }	药物 #3={ 研制: 0.069, 有效: 0.049, 治疗: 0.009 }
drug#3= { marijuana: 0.125,sentence: 0.033, crime: 0.009 }	_NA_

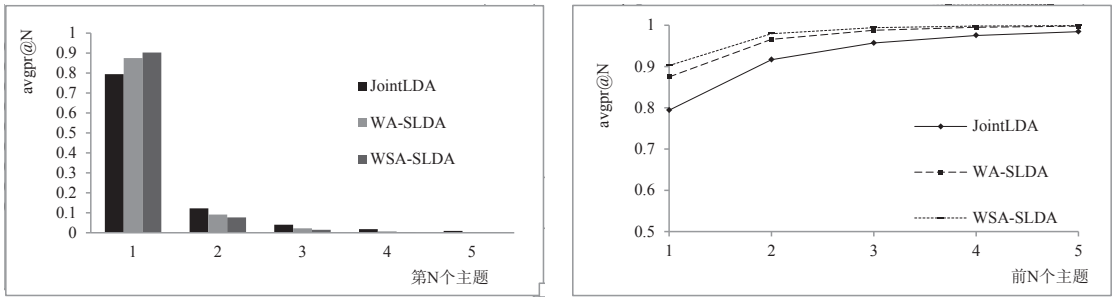
实验结果分析

从表5.2我们可以看出：在提出的两个模型中，WSA-SLDA 的性能在两个数据集上均好于 WA-SLDA。这是由于 WSA-SLDA 模型不仅考虑了词对齐，还考虑了词义层面的对齐。WSA-SLDA 在词义归纳步骤中考虑了词对的词义分布，根据词对词义分布的定义公式 (5-7)，词对中词义的分布还与它的对齐词义有关，这有利于词义获得更准确的词义分布。例如，在 TDT41 数据集归纳出词对“drug: 药物”的词义如表5.3所示。drug#1 和药物 #2 是对齐词义，在有关安乐死的话题“Dutch Parliament Votes to Legalize Euthanasia”中，drug#1 具有较大的概率那么药物 #2 也



(a.1)　　(a.2)

(a) TDT41



(b.1)　　(b.2)

(b) TDT42

图 5.6　前五个主题的词（词义）-主题平均分布 (1) 第 k 个主题的直方图 (2) 前 k 个主题的曲线图

有可能具有较大的概率，反之亦然。WSA-SLDA 根据词义分布估计词对和话题，获得更准确的词义分布可以获得更准确地词对和话题。

5.3.2　概率分布分析

本实验旨在分析模型的词义 -主题以及文档 -主题的平均分布。除了所提出的模型（WA-SLDA 和 WSA-SLDA），本实验还比较了 JointLDA 的词 -主题以及文档 -主题的平均分布情况。

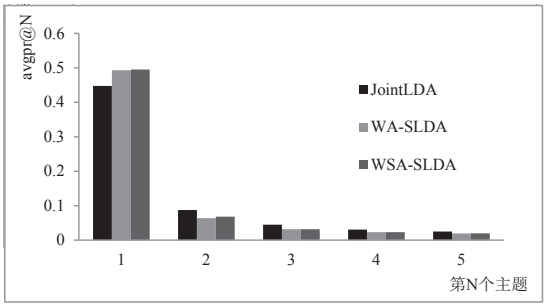
5.3.2.1　实验 #2.1：整体概率分布分析

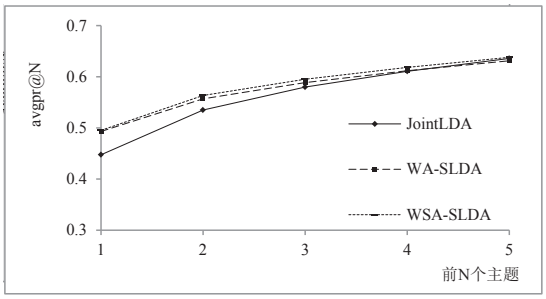
图5.6和图5.7的计算方法与4.3.2.1相同。

实验结果分析

1) 从图5.6中可以看出：

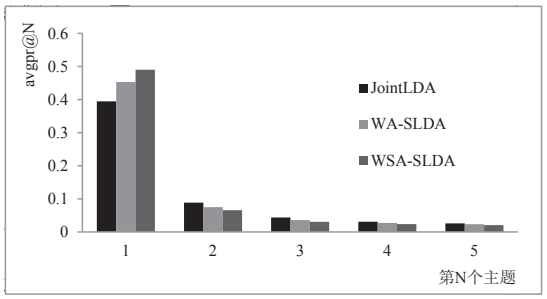
首先，在直方图中，SLDA 系列模型的词义 -主题平均分布与 JointLDA 的词 -主题平均分布相比要更加尖锐；在曲线图中，SLDA 系列模型的曲线在 JointLDA 模型之上。这意味着主题模型中，与词相比，词义可以提供更好的区分性。

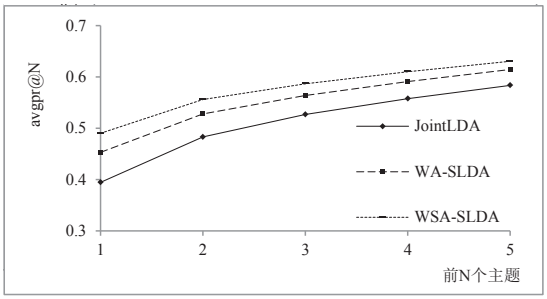




(a.1)　　(a.2)

(a) TDT41





(b.1)　　(b.2)

(b) TDT42

图 5.7　前五个主题的文档 -主题平均分布 (1) 第 k 个主题的直方图 (2) 前 k 个主题的曲线图

其次，直方图中，WSA-SLDA 的词义 -主题平均分布与 WA-SLDA 相比要更加尖锐；而曲线图中，WSA-SLDA 的曲线在 WA-SLDA 模型之上。这得益于 WSA-SLDA 考虑了词义对齐，使得词义的分布更加准确。

2) 从图5.7中可以看出：

首先，SLDA 系列模型与 LDA 相比文档 -主题分布更加尖锐；在曲线图中，SLDA 系列模型的曲线均高于 JointLDA 模型。这意味着 SLDA 模型中，文档集中于更少的主题，这使得 SLDA 模型的主题更具有区分性。

其次，WSA-SLDA 的文档 -主题分布最为尖锐。曲线图中，WSA-SLDA 的曲线在 WA-SLDA 曲线之上。这意味着，考虑词义对齐后获得的词义分布可以为文档中的主题提供更强的后验概率估计。

5.3.2.2　实验 #2.2：基于文档标签的概率分布分析

为了进一步分析所提出的模型，与4.3.2.2节类似，实验 #2.2 以 JointLDA 模型作为基线，比较了 SLDA 系列模型中每个文档概率最大的主题。实验设置与4.3.2.2节相同。实验结果如表5.4，表5.5和表5.6。

表 5.4　百分比 (%)/# 文档，其中文档满足两个条件：1) JointLDA 和 SLDA 模型中话题均识别正确；2）$p_{\text{SLDA}}(\hat{z}|d) > p_{\text{JointLDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42
WA-SLDA	72.9/1447	63.0/659
WSA-SLDA	72.7/1509	74.1/824

表 5.5　百分比 (%)/# 文档，其中文档满足两个条件：1) JointLDA 和 SLDA 模型中话题均识别错误；2）$p_{\text{SLDA}}(\hat{z}|d) > p_{\text{JointLDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42
WA-SLDA	70.5/129	54.3/127
WSA-SLDA	73.1 /104	71.9/135

实验结果分析

从表5.4中可以看出，在 JointLDA 和 SLDA 系列模型都识别正确的文档集中，SLDA 可以在大多数文档上提供一个更大的 $p(\hat{z}|d)$（结果百分比均大于 50%）。这说明了当文档识别正确时，SLDA 系列模型可以增强正确话题的权重。表5.5显示，即使识别错误，与 LDA 模型相比，SLDA 模型在大多数文档中仍然会给正确话题一个较高的权重。换言之，与 LDA 模型相比，SLDA 模型给错误的话题一个较低的权重。表5.6显示，在所有文档中，与 LDA 模型相比，SLDA 模型在大多数文档中正确话题的概率值更高。上面的分析证明了 SLDA 模型可以增加正确话题的概率减少错误话题的概率。因此 SLDA 模型更加适合于一些需要用话题分布的权重作为输入的复杂系统。

5.3.3　文档建模模型性能比较

本实验对比了基于统计词义的跨语言主题模型和现有方法以及前几章提出的跨语言文档建模方法的性能。本实验评测的文档建模方法均采用平行语料获取跨

表 5.6　百分比 (%)/# 文档，其中文档满足 $p_{\text{SLDA}}(\hat{z}|d) > p_{\text{JointLDA}}(\hat{z}|d)$

系统 \ 数据集	TDT41	TDT42
WA-SLDA	72.1/1927	55.1/1177
WSA-SLDA	74.6/1927	75.5/1177

语言信息，具体信息如下：

- **VSM**：采用 VSM 表示文档，余弦相似度计算文档相似度，并从平行语料翻译概率获取翻译信息。
- **JointLDA**：采用 JointLDA[85] 表示文档，利用平行语料的翻译概率获取对齐词对，主题个数与文档集的类簇个数相同，将文档的主题看成话题并将每个文档分到概率最大的主题中。
- **PTM**：使用文献 [78] 的方法。用 PTM 在平行语料中获取跨语言主题，然后再测试集中推测主题。主题个数设为 1000。
- **CLGVSM**：本文第2章提出的跨语言广义向量空间向量模型。在测试集中采用 SOCPMI 计算单语言相似度，利用翻译概率计算跨语言相似度。
- **BK-SVSM**：本文第3章提出的基于全局词义的跨语言文档建模方法，采用 Bisecting K-Means 生成全局词义，SVSM 表示文档。
- **WSA-SLDA**：本章提出的基于统计词义的跨语言主题模型。利用翻译概率中的对齐信息，使用 WSA-SLDA 对文档中的主题进行建模。

实验结果如表5.7所示。

表 5.7　采用不同文档建模模型的话题分析系统的 F-measure

系统 \ 数据集	TDT41	TDT42
VSM	0.751	0.732
JointLDA	0.766	0.729
PTM	0.493	0.482
CLGVSM	0.762	0.748
BK-SVSM	0.809	0.791
WSA-SLDA	**0.863**	**0.817**

实验结果分析

从表5.7我们可以看出：

首先，基于统计词义的跨语言主题模型（WSA-SLDA）的性能要好于基于全局词义的文档建模方法（BK-SVSM）。BK-SVSM 是在平行语料中获得词义，通过聚类算法获得词义对齐，然后用对齐的词义表示文档。而本章提出的 WSA-SLDA 模型则是利用平行语料的词对齐信息，在测试集中估计词义，同时估计词义对齐。这样做的好处在于可以避免平行语料和测试集的词义偏差，另外 WSA-SLDA 模型通过词义和话题、词对相互影响，可以学习出话题相关的词义，更有利于确定词的话题。

其次，基于统计词义的跨语言主题模型（WSA-SLDA）的性能要好于跨语言广义空间向量模型（CLGVSM）。CLGVSM 考虑了词相似度，他只能反映词的同义关系和跨语言歧义问题，无法反映多义问题和单语言歧义问题。而基于统计词义的跨语言主题模型则可以同时解决词的同义和多义问题，因此可以同时解决它们导致的跨语言歧义和单语言歧义问题。

第三，基于统计词义的跨语言主题模型（WSA-SLDA）的性能要高于基线系统 VSM、JointLDA 以及 PTM。VSM 没有考虑到翻译歧义问题。JointLDA 是在以词为特征的主题模型。PTM 所构建的语义空间是在固定的平行语料中构建的，因此它没有考虑到目标聚类集的特征的重要性。而基于统计词义的跨语言主题模型则同时考虑到跨语言歧义和单语言歧义，以词义作为特征，对文档中的主题更具有区分性。同时基于统计词义的跨语言主题模型是在目标聚类集中直接构造语义空间，不存在特征偏差问题。

5.3.4　复杂度问题

基于统计词义的跨语言主题模型需要在文档中估计词义、主题以及词义对齐，复杂度较高。该方法需要估计的参数有每个词的词义归纳模型即 HDP 模型的参数，映射矩阵 M 以及文档表示模型 LDA 的参数。由于每个词都有一个词义归纳模型，这导致了基于统计词义的跨语言主题模型的高复杂性。但本文工作专注于模型精度的提高，尚未在降低复杂度方面进行大量探索。本文设想了如下降低复杂度的办法：

- 预先训练获得大多数词汇的词义分布，而在聚类过程中只根据测试文档内容调整词义分布。由于大部分复杂度来自于词义分布的估计，因而可大大降低模型和聚类算法的复杂度。
- 分布式部署。通过部署多台服务器或多个进程，达到运行时间的缩减。这在本文的下一步研究展望中也会提到。

5.4　本章小结

针对跨语言问题，本章在词义主题模型中引入词对齐信息，设计了两个基于词对齐的跨语言词义主题模型：词对齐词义 LDA 和词义对齐词义 LDA。在词对齐词义 LDA 模型中，将词对看作变量，加入到 SCO-SLDA 模型中，解决跨语言的对应问题。在词义对齐词义 LDA 中，除了考虑词对齐信息，还估计了词义对齐，构造了主题、词义、词义对齐的迭代估计过程。本章在跨语言话题分析任务中评测了两个模型。

根据词义 -主题和文档 -主题的平均概率分析，跨语言词义主题模型可以提供更强的后验分布估计。实验结果表明，跨语言词义主题模型可以改进跨语言话题分析的性能。

具体来说基于词义对齐的模型（WSA-SLDA）的性能比仅考虑词对齐的模型（WA-SLDA）性能更好，这是由于对齐的词义分布相互影响，可以获得更准确的词义分布，从而进一步改进跨语言话题分析的性能。

第 6 章　总结和展望

6.1　论文工作总结

跨语言话题分析的任务是从不同语言的文档中识别出文档的话题，并将相同话题的文档进行汇总，其中同一个话题可能包含不同语言的文档。跨语言话题分析的关键问题在于如何获得不同语言的文档建模。本文认为语义是跨语言文档建模的关键，研究了基于语义的文档建模，提出了基于词相似度和基于词义的一系列文档建模模型，并在跨语言话题分析应用上验证了模型的有效性，为跨语言话题分析和跨语言文档建模在未来的研究和应用奠定了基础。

概括来说，本文的工作重点和研究共现主要体现在以下几方面：

1) 考虑了词相似度这一语义信息，将单语广义向量空间模型拓展为**跨语言广义向量空间模型**。词相似度反映了词和词的相似程度。跨语言相似度反映了不同语言之间词和词的相似度。本文提出了一系列计算跨语言相似度的方法，同时利用跨语言相似度将单语广义向量空间模型拓展到跨语言文档建模，提出了适用于广义空间向量模型的特征选择算法。实验结果证明了这个方法在跨语言话题分析这一应用上具有优越性。

2) 进一步考虑词的多义性和跨语言的翻译歧义问题，提出了**一种基于全局词义的跨语言文档建模方法**。与之前的大部分研究不同，本文并未采用语义资源来获得词义定义，而是通过词义归纳算法来获得统计词义。为了更好的表示词义，本文定义了局部词义和全局词义：局部词义反映了同一个词的不同含义，全局词义则反映了不同词之间词义的关系。局部词义可以解决由词的多义性导致的翻译歧义问题，而全局词义则反映了不同语言的对应问题和由翻译选择不同导致的翻译歧义问题。为了获得词义，本文提出了跨语言局部词义归纳算法和全局词义聚类算法。最后对文档进行词义消歧，并利用全局词义表示文档。在跨语言话题分析任务上的实验结果表明了这个方法的性能优于跨语言广义空间向量模型。

3) 为了更好的表示文档，本文提出了**三种基于统计词义的主题模型**。主题模型已经被广泛的用于文档建模，但是传统的主题模型依赖于词的共现来挖掘语义信息。但是词通常在不同的主题上表达不同的含义，在主题模型中使用恰当的词义作为附加特征可以增强区分性。本文不仅仅直接用词义替代词，还提出了进一步的改进模型。本文设计了三种基于词义的 LDA 模型：独立 SLDA（SA-SLDA）, 点估计协同 SLDA（PCo-SLDA），词义混合协同 SLDA（SCo-SLDA）。SA-SLDA

首先用词义归纳算法获得词义，然后在 LDA 中用词义作为元素生成主题。PCo-SLDA 和 SCo-SLDA 则将词义看做主题模型的隐藏变量，协同迭代估计词义和主题。PCo-SLDA 是通过点估计的方法确定词义而 SCo-SLDA 则是考虑词义的分布。实验结果表明了基于词义的主题模型可以改进单语言话题分析的性能；协同学习模型（PCo-SLDA 和 SCo-SLDA）的性能更好；考虑到词义分布的 SCo-SLDA 性能最好。概率分布分析也验证了这一结果：性能好的模型具有更加尖锐的主题分布，可以提供更强的后验信息。本文还评测了 SCo-SLDA 在词义归纳任务上的性能，实验结果验证了使用主题作为词义的伪回馈可以归纳和识别出更准确的词义。

4) 针对跨语言问题，本文在词义混合的协作 SLDA 模型中加入了词对齐信息，提出了**两种基于统计词义的跨语言主题模型**：词对齐词义 LDA（WA-SLDA）和词义对齐词义 LDA（WSA-SLDA）。在 WA-SLDA 模型中，将词对看作变量，加入到 SCo-SLDA 模型中，解决跨语言的对应问题。在 WSA-SLDA 模型中，除了考虑词对齐信息，还估计了词义对齐，构造了主题、词义、词义对齐的迭代估计过程。与仅考虑词对齐的方法相比，更有利于获得准确地词义分布，进一步改进主题的性能。实验结果在跨语言话题分析上表明基于词义的跨语言主题模型的性能好于基于词义的跨语言文档建模方法，其中基于词义对齐的模型（WSA-SLDA）的性能比仅考虑词对齐的模型（WA-SLDA）性能更好。概率分布分析也验证了这一结果。

6.2　下一步研究展望

本文对跨语言话题分析进行了初步的研究，提出了一系列基于语义的跨语言文档建模模型，在跨语言话题分析任务上取得了一定的效果，但距离现实应用还有一定的距离，并存在一些不足之处。针对这些不足之处，对于跨语言话题分析和基于语义的跨语言文档建模方法的进一步研究展望如下：

1) 本文仅使用了新闻类型的语料作为评测数据集。但是目前互联网上还有很多其他类型的语料需要进行话题分析，如 BBS 论坛的帖子、博客等。这些语料在用语上不如新闻语料规范，未登录词较多。还有一些语料如微博除了在规范上存在问题，还存在特征稀疏问题。因此，如何在不规范语料和短文本语料中进行跨语言话题分析值得我们下一步研究。

2) 本研究仅使用统计方法归纳词义，当数据不足时，容易产生过拟合现象。下一步研究可以考虑将语义资源和统计方法结合起来。比如用语义资源的词义作为先验分布，在数据集中估计后验分布。语义资源可以弥补数据不足的时候统计方法产生的过拟合现象，统计方法可以解决语义资源的覆盖率问题。

3) 本文提出四个基于语义的文档建模方法，目前仅在话题分析任务上进行评测。这几种文档建模方法可以应用到其他领域中：如在文档分类中，可以用本文提出的文档建模方法获得文档的特征，进行分类；在情感计算中，用本文提出的文档建模方法可以消除情感词的歧义，提高性能。

4) 本文提出的几种模型，特别是效果最好的基于词义的跨语言主题模型，复杂度较高，在大数据上运算有一定的困难。如何降低模型的复杂度，在大数据上评测模型是下一步研究需要解决的问题。在基于统计词义的跨语言主题模型中，每个词都有一个词义归纳模型，这导致了基于统计词义的跨语言主题模型的高复杂性。下一步研究可以考虑分布算法，将每个词的词义归纳过程进行分布式部署，提高效率；还可以考虑先判断词的类型，仅在多义词上进行词义归纳，减少词义归纳模型的个数，也可以提高效率。

参考文献

[1] 洪宇, 张宇, 刘挺, et al. 话题检测与跟踪的评测及研究综述. 中文信息学报, 2007, 21(6):71–87.

[2] Mori M, Miura T, Shioya I. Topic Detection and Tracking for News Web Pages. Proceedings of Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA: IEEE Computer Society, 2006. 338–342.

[3] Sekiguchi Y, Kawashima H, Okuda H, et al. Topic Detection from Blog Documents Using Users' Interests. Proceedings of Proceedings of the 7th International Conference on Mobile Data Management, Washington, DC, USA: IEEE Computer Society, 2006. 108–108.

[4] Kasiviswanathan S P, Melville P, Banerjee A, et al. Emerging Topic Detection Using Dictionary Learning. Proceedings of Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA: ACM, 2011. 745–754.

[5] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America, 2004, 101(Suppl 1):5228–5235.

[6] 骆卫华, 刘群, 程学旗. 话题检测与跟踪技术的发展与研究. Proceedings of 全国计算语言学联合学术会议 (JSCL-2003) 论文集. 北京: 清华大学出版社, 2003. 560–566.

[7] Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. Commun. ACM, 1975, 18(11):613–620.

[8] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval. Proceedings of Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 1998. 275–281.

[9] Chen Y J, Chen H H. NLP and IR Approaches to Monolingual and Multilingual Link Detection. Proceedings of Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. 1–7.

[10] Clifton C, Cooley R, Rennie J. TopCat: Data Mining for Topic Identification in a Text Corpus. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(8):949–964.

[11] Yang Y, Zhang J, Carbonell J, et al. Topic-conditioned Novelty Detection. Proceedings of Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2002. 688–693.

[12] Ignat C, Steinberger R, Pouliquen B. Navigating multilingual news collections using automatically extracted information. CIT. Journal of computing and information technology, 2005, 13(4):257–264.

[13] Steinberger R, Pouliquen B, Ignat C. NewsExplorer: multilingual news analysis with cross-lingual linking. Information Technology Interfaces, 2005..

[14] Kumaran G, Allan J. Using Names and Topics for New Event Detection. Proceedings of Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. 121–128.

[15] 宋丹, 王卫东, 陈英. 基于改进向量空间模型的话题识别与跟踪. 计算机技术与发展, 2006, 16(9):62–64.

[16] Lo Y Y, Gauvain J L. The LIMSI topic tracking system for TDT2001. Proc. TDT, 2001. 1.

[17] Li Z, Wang B, Li M, et al. A Probabilistic Model for Retrospective News Event Detection. Proceedings of Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2005. 106–113.

[18] Lavrenko V, Allan J, DeGuzman E, et al. Relevance Models for Topic Detection and Tracking. Proceedings of Proceedings of the Second International Conference on Human Language Technology Research, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. 115–121.

[19] Nallapati R. Semantic Language Models for Topic Detection and Tracking. Proceedings of Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop - Volume 3, Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. 1–6.

[20] Papka R, Allan J. On-Line New Event Detection Using Single Pass Clustering TITLE2:. Technical report, Amherst, MA, USA, 1998.

[21] Pouliquen B, Steinberger R, Ignat C, et al. Multilingual and Cross-lingual News Topic Tracking. Proceedings of Proceedings of the 20th International Conference on Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[22] Trieschnigg D, Kraaij W. Scalable Hierarchical Topic Detection: Exploring a Sample Based Approach. Proceedings of Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2005. 655–656.

[23] Tang G, Xia Y. Adaptive Topic Modeling with Probabilistic Pseudo Feedback in Online Topic Detection. Proceedings of Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems, NLDB'10. Berlin, Heidelberg: Springer-Verlag, 2010: 100–108.

[24] 贾自艳, 何清, 张海俊, et al. 一种基于动态进化模型的事件探测和追踪算法. 计算机研究与发展, 2004, 41(7):1273–1280.

[25] 雷震, 吴玲达, 雷蕾, et al. 初始化类中心的增量 K 均值法及其在新闻事件探测中的应用. 情报学报, 2006, 25(3):289–295.

[26] 骆卫华, 于满泉, 许洪波, et al. 基于多策略优化的分治多层聚类算法的话题发现研究. 中文信息学报, 2006, 20(1):29–36.

[27] 赵华, 赵铁军, 张姝, et al. 基于内容分析的话题检测研究. 哈尔滨工业大学学报, 2007, 38(10):1740–1743.

[28] Hatch P, Stokes N, Carthy J. Topic Detection, a new application for lexical chaining. Proceedings of the proceedings of BCS-IRSG, 2000. 94–103.

[29] Stokes N, Hatch P, Carthy J. Lexical Semantic Relatedness and Online New Event Detection (Poster Session). Proceedings of Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2000. 324–325.

[30] Makkonen J, Ahonen-Myka H, Salmenkivi M. Simple Semantics in Topic Detection and Tracking. Inf. Retr., 2004, 7(3-4):347–368.

[31] Lam W, Meng H, Wong K, et al. Using contextual analysis for news event detection. International Journal of Intelligent Systems, 2001, 16(4):525–546.

[32] Landauer T K, Dumais S T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 1997, 104(2):211.

[33] Hofmann T. Probabilistic Latent Semantic Indexing. Proceedings of Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 1999. 50–57.

[34] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. the Journal of machine Learning research, 2003, 3:993–1022.

[35] Blei D M, Griffiths T L, Jordan M I, et al. Hierarchical topic models and the nested Chinese restaurant process. Proceedings of Advances in Neural Information Processing Systems. MIT Press, 2004. 2003.

[36] Blei D M, Griffiths T L, Jordan M I. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. J. ACM, 2010, 57(2):7:1–7:30.

[37] Blei D M, Lafferty J D. Correlated topic models. Proceedings of In Proceedings of the 23rd International Conference on Machine Learning. MIT Press, 2006. 113–120.

[38] Putthividhya D P, Attias H T, Nagarajan S. Independent Factor Topic Models. Proceedings of Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, USA: ACM, 2009. 833–840.

[39] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-topic Model for Authors and Documents. Proceedings of Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, United States: AUAI Press, 2004. 487–494.

[40] Blei D M, Lafferty J D. Dynamic Topic Models. Proceedings of Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA: ACM, 2006. 113–120.

[41] Wang X, McCallum A. Topics over Time: A non-Markov Continuous-time Model of Topical Trends. Proceedings of Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2006. 424–433.

[42] Boyd-Graber J L, Blei D M. Syntactic topic models. Proceedings of Advances in neural information processing systems, 2009. 185–192.

[43] Kitajima R, Kobayashi I. A Latent Topic Extracting Method Based on Events in a Document and Its Application. Proceedings of Proceedings of the ACL 2011 Student Session, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 30–35.

[44] Wallach H M. Topic Modeling: Beyond Bag-of-words. Proceedings of Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA: ACM, 2006. 977–984.

[45] Griffiths T L, Steyvers M, Tenenbaum J B. Topics in semantic representation. Psychological review, 2007, 114(2):211.

[46] Wang X, McCallum A, Wei X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. Proceedings of Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, Washington, DC, USA: IEEE Computer Society, 2007. 697–702.

[47] Kawamae N. Supervised N-gram Topic Model. Proceedings of Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, NY, USA: ACM, 2014. 473–482.

[48] Boyd-Graber J L, Blei D M, Zhu X. A Topic Model for Word Sense Disambiguation. Proceedings of EMNLP-CoNLL, 2007. 1024–1033.

[49] Boyd-Graber J, Blei D. PUTOP: Turning Predominant Senses into a Topic Model for Word Sense Disambiguation. Proceedings of Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. 277–281.

[50] Guo W, Diab M. Semantic Topic Models: Combining Word Distributional Statistics and Dictionary Definitions. Proceedings of Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 552–561.

[51] 张晓艳, 王挺, 梁晓波. LDA 模型在话题追踪中的应用. 计算机科学, 2011, 38(B10):136–139.

[52] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. Proceedings of Proceedings of the 20th International Joint Conference on Artifical Intelligence, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. 1606–1611.

[53] Wang P, Hu J, Zeng H J, et al. Improving Text Classification by Using Encyclopedia Knowledge. Proceedings of Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, Washington, DC, USA: IEEE Computer Society, 2007. 332–341.

[54] Schönhofen P. Identifying Document Topics Using the Wikipedia Category Network. Web Intelli. and Agent Sys., 2009, 7(2):195–207.

[55] Hu J, Fang L, Cao Y, et al. Enhancing Text Clustering by Leveraging Wikipedia Semantics. Proceedings of Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2008. 179–186.

[56] Hotho A, Staab S, Stumme G. Ontologies Improve Text Document Clustering. Proceedings of Proceedings of the Third IEEE International Conference on Data Mining, Washington, DC, USA: IEEE Computer Society, 2003. 541–.

[57] Huang H H, Kuo Y H. Cross-lingual Document Representation and Semantic Similarity Measure: A Fuzzy Set and Rough Set Based Approach. Trans. Fuz Sys., 2010, 18(6):1098–1111.

[58] Wong S K M, Ziarko W, Wong P C N. Generalized Vector Spaces Model in Information Retrieval. Proceedings of Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 1985. 18–25.

[59] Farahat A K, Kamel M S. Statistical Semantics for Enhancing Document Clustering. Knowledge and information systems, 2011, 28(2):365–393.

[60] Dhillon I S. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. Proceedings of Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2001. 269–274.

[61] Pessiot J F, Kim Y M, Amini M R, et al. Improving Document Clustering in a Learned Concept Space. Inf. Process. Manage., 2010, 46(2):180–192.

[62] Purandare A, Pedersen T. SenseClusters: Finding Clusters That Represent Word Senses. Proceedings of Demonstration Papers at HLT-NAACL 2004, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. 26–29.

[63] Pedersen T. Unsupervised corpus-based methods for WSD. Word Sense Disambiguation: Algorithms and Applications, 2006. 133–166.

[64] Pedersen T. Computational approaches to measuring the similarity of short contexts: A review of applications and methods. Technical report, 2008.

[65] Schutze H, Pedersen J O. Information Retrieval Based on Word Senses. 1995. 161–175.

[66] Navigli R, Crisafulli G. Inducing Word Senses to Improve Web Search Result Clustering. Proceedings of Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 116–126.

[67] Leek T, Jin H, Sista S, et al. The BBN crosslingual topic detection and tracking system. Proceedings of Working Notes of the Third Topic Detection and Tracking Workshop. Citeseer, 2000.

[68] Larkey L S, Feng F, Connell M, et al. Language-specific Models in Multilingual Topic Tracking. Proceedings of Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2004. 402–409.

[69] Chen H H, Lin C J. A Multilingual News Summarizer. Proceedings of Proceedings of the 18th Conference on Computational Linguistics - Volume 1, Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 159–165.

[70] 苏伟峰, 李绍滋, 李堂秋, et al. 可分义原向量空间中的跨语种文本过滤模型. Proceedings of 自然语言理解与机器翻译——全国第六届计算语言学联合学术会议论文集, 2001.

[71] Wei C P, Yang C C, Lin C M. A Latent Semantic Indexing-based Approach to Multilingual Document Clustering. Decis. Support Syst., 2008, 45(3):606–620.

[72] Dumais S T, Letsche T A, Littman M L, et al. Automatic cross-language retrieval using latent semantic indexing. Proceedings of AAAI spring symposium on cross-language text and speech retrieval, volume 15, 1997. 21.

[73] Berry M, Young P. Using latent semantic indexing for multilanguage information retrieval. Computers and the Humanities, 1995, 29(6):413–429.

[74] Dumais S T, Letsche T A, Littman M L, et al. Automatic cross-language retrieval using latent semantic indexing. Proceedings of AAAI spring symposium on cross-language text and speech retrieval, volume 15, 1997. 21.

[75] Dumais S, Landauer T K, Littman M L. Automatic cross-linguistic information retrieval using latent semantic indexing. 1997..

[76] Muramatsu T, Mori T. Integration of pLSA into probabilistic CLIR model. Proceedings of Proceedings of NTCIR, volume 4, 2004.

[77] Ni X, Sun J T, Hu J, et al. Mining Multilingual Topics from Wikipedia. Proceedings of Proceedings of the 18th International Conference on World Wide Web, New York, NY, USA: ACM, 2009. 1155–1156.

[78] Mimno D, Wallach H M, Naradowsky J, et al. Polylingual Topic Models. Proceedings of Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 880–889.

[79] De Smet W, Moens M F. Cross-language Linking of News Stories on the Web Using Interlingual Topic Modelling. Proceedings of Proceedings of the 2Nd ACM Workshop on Social Web Search and Mining, New York, NY, USA: ACM, 2009. 57–64.

[80] Tholpadi G, Das M K, Bhattacharyya C, et al. Cluster Labeling for Multilingual Scatter/Gather Using Comparable Corpora. Proceedings of Proceedings of the 34th European Conference on Advances in Information Retrieval, ECIR'12. Berlin, Heidelberg: Springer-Verlag, 2012: 388–400.

[81] Platt J C, Toutanova K, Yih W t. Translingual Document Representations from Discriminative Projections. Proceedings of Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 251–261.

[82] Zhang D, Mei Q, Zhai C. Cross-lingual Latent Topic Extraction. Proceedings of Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 1128–1137.

[83] Boyd-Graber J, Resnik P. Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. Proceedings of Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 45–55.

[84] Boyd-Graber J, Blei D M. Multilingual Topic Models for Unaligned Text. Proceedings of Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, United States: AUAI Press, 2009. 75–82.

[85] Jagarlamudi J, Daumé H. Extracting Multilingual Topics from Unaligned Comparable Corpora. Proceedings of Proceedings of the 32Nd European Conference on Advances in Information Retrieval, Berlin, Heidelberg: Springer-Verlag, 2010. 444–456.

[86] Kumar N K, Santosh K G S, Varma V. Multilingual Document Clustering Using Wikipedia As External Knowledge. Proceedings of Proceedings of the Second International Conference on Multidisciplinary Information Retrieval Facility, Berlin, Heidelberg: Springer-Verlag, 2011. 108–117.

[87] Cimiano P, Schultz A, Sizov S, et al. Explicit Versus Latent Concept Models for Cross-language Information Retrieval. Proceedings of Proceedings of the 21st International Jont Conference on Artifical Intelligence, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009. 1513–1518.

[88] 宁健, 林鸿飞. 基于改进潜在语义分析的跨语言检索. 中文信息学报, 2010, 24(3):105–111.

[89] Pan J, Xue G R, Yu Y, et al. Cross-lingual Sentiment Classification via Bi-view Non-negative Matrix Tri-factorization. Proceedings of Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I, Berlin, Heidelberg: Springer-Verlag, 2011. 289–300.

[90] Wang H, Huang H, Nie F, et al. Cross-language Web Page Classification via Dual Knowledge Transfer Using Nonnegative Matrix Tri-factorization. Proceedings of Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2011. 933–942.

[91] Fortuna B, Rupnik J, Pajntar B, et al. Cross-lingual Search over 22 European Languages. Proceedings of Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM, 2008. 883–883.

[92] Udupa R, Khapra M M. Transliteration Equivalence Using Canonical Correlation Analysis. Proceedings of Proceedings of the 32Nd European Conference on Advances in Information Retrieval, ECIR'2010. Berlin, Heidelberg: Springer-Verlag, 2010: 75–86.

[93] Vinokourov A, Shawe-Taylor J, Cristianini N. Inferring a semantic representation of text via cross-language correlation analysis. Advances in neural information processing systems, 2003. 1497–1504.

[94] Li Y, Shawe-Taylor J. Using KCCA for Japanese—English Cross-language Information Retrieval and Document Classification. J. Intell. Inf. Syst., 2006, 27(2):117–133.

[95] Li Y, Shawe-Taylor J. Advanced Learning Algorithms for Cross-language Patent Retrieval and Classification. Inf. Process. Manage., 2007, 43(5):1183–1199.

[96] Yogatama D, Tanaka-Ishii K. Multilingual Spectral Clustering Using Document Similarity Propagation. Proceedings of Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 871–879.

[97] Wang X, Qian B, Davidson I. Improving Document Clustering Using Automated Machine Translation. Proceedings of Proceedings of the 21st ACM International Conference on Information and Knowledge Management, New York, NY, USA: ACM, 2012. 645–653.

[98] Montalvo S, Martínez R, Casillas A, et al. Multilingual Document Clustering: An Heuristic Approach Based on Cognate Named Entities. Proceedings of Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. 1145–1152.

[99] Kumar N K, Santosh G S K, Varma V. A Language-independent Approach to Identify the Named Entities in Under-resourced Languages and Clustering Multilingual Documents. Proceedings of Proceedings of the Second International Conference on Multilingual and Multimodal Information Access Evaluation, Berlin, Heidelberg: Springer-Verlag, 2011. 74–82.

[100] Mathieu B, Besanon R, Fluhr C. Multilingual document clusters discovery. Proceedings of RIAO. Citeseer, 2004. 116–125.

[101] Lin D. Automatic Retrieval and Clustering of Similar Words. Proceedings of Proceedings of the 17th International Conference on Computational Linguistics - Volume 2, Stroudsburg, PA, USA: Association for Computational Linguistics, 1998. 768–774.

[102] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 1999, 11:95–130.

[103] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学, 2002, 7(2):59–76.

[104] Xia Y, Zhao T, Yao J, et al. Measuring Chinese-English Cross-lingual Word Similarity with HowNet and Parallel Corpus. Proceedings of Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, Berlin, Heidelberg: Springer-Verlag, 2011. 221–233.

[105] Church K W, Hanks P. Word Association Norms, Mutual Information, and Lexicography. Computational linguistics, 1990, 16(1):22–29.

[106] Turney P D. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. Proceedings of Proceedings of the 12th European Conference on Machine Learning, London, UK, UK: Springer-Verlag, 2001. 491–502.

[107] Islam A, Inkpen D. Second order co-occurrence PMI for determining the semantic similarity of words. Proceedings of Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006), 2006. 1033–1038.

[108] 杜飞龙. 知网辟蹊径共享新天地—董振东先生谈知网与知识共享. 微电脑世界, 2000..

[109] Dong Z, Dong Q. Hownet and the Computation of Meaning. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2006.

[110] Dai L, Liu B, Xia Y, et al. Measuring Semantic Similarity Between Words Using HowNet. Proceedings of Proceedings of the 2008 International Conference on Computer Science and Information Technology, Washington, DC, USA: IEEE Computer Society, 2008. 601–605.

[111] Steinbach M, Karypis G, Kumar V, et al. A comparison of document clustering techniques. Proceedings of KDD workshop on text mining, volume 400. Boston, 2000. 525–526.

[112] Och F J, Ney H. Improved Statistical Alignment Models. Proceedings of Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 440–447.

[113] Kong J, Graff D. TDT4 multilingual broadcast news speech corpus. Linguistic Data Consortium, http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp, 2005..

[114] Wang C, Zhang M, Ma S, et al. Automatic Online News Issue Construction in Web Environment. Proceedings of Proceedings of the 17th International Conference on World Wide Web, New York, NY, USA: ACM, 2008. 457–466.

[115] Tufiş D, Koeva S. Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation. Proceedings of Proceedings of the 7th International Workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory, Berlin, Heidelberg: Springer-Verlag, 2007. 447–455.

[116] Stokoe C, Oakes M P, Tait J. Word Sense Disambiguation in Information Retrieval Revisited. Proceedings of Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, New York, NY, USA: ACM, 2003. 159–166.

[117] Harris Z S. Distributional structure. Word, 1954..

[118] Firth J. A synopsis of linguistic theory 1930–1955. 1957. 1–32.

[119] Pantel P, Lin D. Discovering Word Senses from Text. Proceedings of Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2002. 613–619.

[120] Bordag S. Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation. Proceedings of EACL, 2006.

[121] Pinto D, Rosso P, Jiménez-Salazar H. UPV-SI: Word Sense Induction Using Self Term Expansion. Proceedings of Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. 430–433.

[122] Rapp R. A Practical Solution to the Problem of Automatic Word Sense Induction. Proceedings of Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[123] Klapaftis I P, Manandhar S. UOY: A Hypergraph Model for Word Sense Induction & Disambiguation. Proceedings of Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. 414–417.

[124] Klapaftis I P, Manandhar S. Word Sense Induction Using Graphs of Collocations. Proceedings of Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence, Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008. 298–302.

[125] Brody S, Lapata M. Bayesian Word Sense Induction. Proceedings of Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 103–111.

[126] Agirre E, Soroa A. Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. Proceedings of Proceedings of the 4th International Workshop on Semantic Evaluations, Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. 7–12.

[127] Yao X, Van Durme B. Nonparametric Bayesian Word Sense Induction. Proceedings of Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 10–14.

[128] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2004, 101.

[129] Apidianaki M. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. Proceedings of Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 77–85.

[130] 蔡科, 史晓东, 陈毅东, et al. 基于层次聚类的中文词义归纳 (英文). 心智与计算, 2010, 3:003.

[131] Schütze H. Automatic Word Sense Discrimination. Computational linguistics, 1998, 24(1):97–123.

[132] Sanderson M. Word Sense Disambiguation and Information Retrieval. Proceedings of Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: Springer-Verlag New York, Inc., 1994. 142–151.

[133] Mihalcea R, Moldovan D. Semantic Indexing Using WordNet Senses. Proceedings of Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11, Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. 35–45.

[134] Vossen P, Rigau G, Alegria I, et al. Meaningful results for Information Retrieval in the MEANING project. Proceedings of Proc. of the 3rd Global Wordnet Conference, 2006. 22–26.

[135] Dagan I, Itai A, Schwall U. Two Languages Are More Informative Than One. Proceedings of Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 1991. 130–137.

[136] Dagan I, Itai A. Word Sense Disambiguation Using a Second Language Monolingual Corpus. Computational Linguistics, 1994, 20(4):563–596.

[137] Ng H T, Wang B, Chan Y S. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. Proceedings of Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. 455–462.

[138] Li H, Li C. Word Translation Disambiguation Using Bilingual Bootstrapping. Comput. Linguist., 2004, 30(1):1–22.

[139] Diab M T. Word Sense Disambiguation Within a Multilingual Framework[D]. College Park, MD, USA, 2003. AAI3115805.

[140] Bhattacharya I, Getoor L, Bengio Y. Unsupervised Sense Disambiguation Using Bilingual Probabilistic Models. Proceedings of Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[141] Lu Y, Mei Q, Zhai C. Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA. Inf. Retr., 2011, 14(2):178–203.

[142] Lewis D D. Reuters-21578 text categorization test collection, distribution 1.0. http://www. research. att. com/˜ lewis/reuters21578. html, 1997..

[143] Schmid H. Probabilistic part-of-speech tagging using decision trees. Proceedings of Proceedings of international conference on new methods in language processing, volume 12. Manchester, UK, 1994. 44–49.

[144] Hovy E, Marcus M, Palmer M, et al. OntoNotes: The 90% Solution. Proceedings of Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. 57–60.

[145] Gilks W R. Markov Chain Monte Carlo. Proceedings of Encyclopedia of Biostatistics, chapter Stochastic EM: method and application. John Wiley & Sons, Ltd, 2005:.

[146] Lawler E L. Combinatorial optimization: networks and matroids. Courier Dover Publications, 1976.

致　谢

衷心感谢导师郑方研究员近六年的悉心指导和关怀，在清华大学语音和语言中心的这段时间里，郑老师的严谨求实、平易近人，无论在科研上还是生活上都给予了我莫大的关心和帮助，使我受益终生。衷心感谢我的指导老师夏云庆副研究员。夏老师引领我进入了自然语言处理研究领域，悉心传授我做研究写论文的方法。在此，谨向两位老师致以最诚挚的谢意。

感谢新加坡信息通信研究院的张民老师和孙军老师对我在跨语言话题分析研究中给予的指导和帮助。

感谢语音和语言中心的王东老师、周强老师，邬晓钧老师以及其他所有帮助过我的老师，老师们的学识和精神永远值得我学习。

感谢语音和语言中心的王刚师兄、王琳琳师姐以及张陈昊、陈丽欧、张超、别凡虎、王军、李蓝天、王俊俊、邱晗、刘超、龚成、谢仲达等同学和赵欢工程师在学习上和工作上的帮助和支持。

最后，感谢我的家人，他们无私的爱和无条件的支持，让我走过这段值得铭记一生的时光。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1988 年 3 月 24 日出生于江苏省连云港市。

2005 年 9 月考入清华大学计算机科学与技术系，2009 年 7 月本科毕业并获得工学学士学位。

2005 年 9 月免试进入清华大学计算机科学与技术系攻读计算机应用博士至今。

发表的学术论文

[1] Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, Thomas Fang Zheng. Statistical word sense aware topic models. Soft Computing. (已被录用，SCI 源刊)

[2] Guoyu Tang, Yunqing Xia, Erik Cambria, Peng Jin, Thomas Fang Zheng. Document representation with statistical word senses in cross-lingual document clustering. International Journal of Pattern Recognition and Artificial Intelligence. (已被录用，SCI 源刊)

[3] Guoyu Tang, Yunqing Xia, Jun Sun, Min Zhang, Thomas Fang Zheng. Topic Models Incorporating Statistical Word Senses. 15th International Conference, CICLing 2014, p 151-162, Kathmandu, Nepal, April 6-12, 2014. (EI 会议，检索号:20142017719370)

[4] Guoyu Tang, Yunqing Xia, Weizhi Wang, Erik Cambria and Thomas Fang Zheng. Clustering tweets using Wikipedia concepts. The 9th edition of the Language Resources and Evaluation Conference, lrec 2014 , p 2262-2267, Reykjavik, Iceland,26-31 May, 2014

[5] Guoyu Tang, Yunqing Xia, Erik Cambria and Peng Jin. Inducing Word Senses for Cross-lingual Document Clustering. Computational Intelligence and Security (CIS), 2013 9th International Conference on. IEEE, pp 409-414, Leshan, China, 14-15 Dec. 2013

[6] Guoyu Tang, Yunqing Xia, Min Zhang, Haizhou Li, Thomas Fang Zheng. CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. The 5th International Joint Conference on Natural Language Process-

ing, IJCNLP 2011, p 580-588, Chiang Mai, Thailand, 8-13 Nov. 2011.

[7] Guoyu Tang, Yunqing Xia. Adaptive Topic Modeling with Probabilistic Pseudo Feedback in Online Topic Detection. 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010,p 100-108, Cardiff, UK, June 23-25, 2010 (EI 会议，检索号: 20103313149645)

[8] 唐国瑜, 夏云庆, 张民, 郑方. 基于跨语言广义向量空间模型的跨语言文档聚类方法. 中文信息学报, 2012, (02): 116-121.

[9] 唐国瑜, 夏云庆, 张民, 郑方. 基于词义类簇的文本聚类. 中文信息学报. 2013, (05): 113-119.