基于神经主题模型的字幕对话文本聚类研究

（申请清华大学工学硕士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 张 镭 镧

指 导 教 师 ： 周　强　副研究员

二〇二〇年五月

Research on Clustering for Subtitle Dialogue Text Based on Neural Topic Model

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Computer Science and Technology

by

Leilan Zhang

Thesis Supervisor: Professor, Qiang Zhou

May, 2020

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

如今，数据驱动的对话生成技术极大地依赖于训练数据的规模和质量，对大规模的高质量对话语料产生了巨大需求。然而，目前公开的大规模中文真实对话语料数量却十分匮乏。值得关注的是，影视剧字幕通常具有特定场景，风格贴近人类日常对话，非常适合构建多轮对话语料。但字幕中说话人和场景边界信息的缺失，使得字幕难以被直接利用。一种自然的想法是，将字幕数据以话题线索为单位，按照主题进行组织，形成多轮对话数据。字幕数据以流的形式呈现，其中所包含的不同话题线索之间并没有明显边界，若直接用于主题建模则会因主题混杂而不易取得理想效果。如果能将字幕对话流依据主题进行分割，进而就能在分割后的对话片段上进行主题建模，并据此对对话片段进行组织。因此，本文将字幕对话文本的聚类问题分解为两个主要步骤：1）首先利用主题分割技术，将字幕对话流分割为具有不同主题的对话片段；2）然后在切分出的对话片段上进行主题建模，并依据得到的主题表示对对话片段进行聚类。

因此，本文主要研究面向影视剧字幕对话文本的主题分割和主题建模问题。对话文本长度极短，其稀疏性问题非常严重，尽管已有一系列成熟的主题分割和主题建模方法，但这些方法通常适用于新闻等长文本，而在对话文本上则往往难以到达理想的效果。对于字幕对话流的主题分割问题，本文提出序列标注架构下的主题分割模型：利用 BERT 抽取语义表示并采用时序卷积网络检测对话流中的主题变换点；对于字幕对话片段的主题建模问题，本文以神经主题模型为框架，设计了以高斯混合分布为先验假设的主题模型，以实现更好的对话文本聚类效果。

本文的主要研究成果包括以下几个方面：

- 以中英文影视剧字幕和英文剧本为基础，提出了一种自动标注场景和说话人的方法，并据此自动构建了包含 26 万条话语消息的中英文对话基础标注库。
- 设计了基于预训练语言模型 BERT 和时序卷积网络的主题分割模型，可有效地将字幕流数据分割为以话题线索为单位的对话片段。
- 通过对比分析已有的神经主题模型，设计了基于高斯混合先验的神经主题模型，并在字幕流数据集上验证了该模型的有效性。
- 基于所提出的主题模型，对大规模字幕流数据进行了话题线索自动聚类分析，并对聚类的结果进行了初步评估。

关键词：主题模型；高斯混合模型；变分自编码器；主题分割；对话语料

Abstract

Nowadays, the data-driven dialogue generation technology heavily depends on the scale and quality of the training data, creating a huge demand for large-scale high-quality dialogue corpus. Nevertheless, the large-scale Chinese dialogue corpora currently available are very scarce. It is noteworthy that the TV and movie subtitles usually have specific scenes, and have a style closing to the human daily conversations, which is suitable for a multi-turn dialogue corpus construction. However, the lack of speaker and scene boundary information makes it difficult to utilize subtitles directly. An intuitive idea is to organize the subtitle data according to their topics in the unit of topic thread, to form a multi-turn dialogue corpus. But the subtitle data is presented in the form of a stream, in which there is no obvious boundary between two adjacent topic threads. If it is directly used for topic modeling, it would be hard to achieve an expected effect due to the mixture of topics. If the conversation stream in subtitles can be segmented according to the topics, then the topic modeling and clustering can be applied to these dialogue segments. Therefore, using subtitles to construct a multi-turn dialogue corpus can be decomposed into two main steps: 1) First, divide the subtitle stream into segments with topic segmentation algorithm; 2) Second, apply topic modeling on the divided dialogue segments to extract the topics of these dialogue segments, and cluster them based on the obtained topic representation.

This thesis mainly studies the topic segmentation and topic modeling tasks for the dialogue text of subtitles. The dialogue text is usually short in length, which leads to a severe sparsity problem.Although there exists a series of mature topic segmentation and topic modeling methods, they usually suitable for long texts like news, but can not achieve as good performance on dialogue text as expected. For the topic segmentation of dialogue texts, this thesis proposes a sequence-labeling based topic segmentation model: utilizing BERT to extract the semantic representation and adopting a Temporal Convolutional Network to detect the topic change point; for the topic modeling task on the subtitle dialogue segments, this thesis designs a neural topic model with Gaussian mixture distribution as a priori hypothesis to achieve good cluster performance on dialogue texts.

The main research results of this thesis include the followings:

- Propose a method to annotate scene and speaker tags automatically based on Chi-

nese and English subtitles and English scripts, and constructed a Chinese and English dialogue corpus containing 260,000 utterances accordingly.

- Design a topic segmentation model based on the pre-trained language model BERT and the Temporal Convolutional Network, which can segment the subtitle stream data into dialogue segments in the unit of topic threads effectively.
- Propose a neural topic model with a Gaussian mixture prior distribution by comparing the existing neural topic models, and verify the validity of the model on the subtitle stream data.
- Apply the topic clustering algorithms on the large-scale subtitle data based on the proposed topic model, and preliminarily evaluate the clustering results.

Key Words: Topic Model; Gaussian Mixture Model; Variational Auto-encoder; Topic Segmentation; Dialogue Corpus

目 录

主要符号对照表

$\boldsymbol{D}$	文档集
$\boldsymbol{W}$	词汇表
θ	文档-主题分布矩阵
$\boldsymbol{\Phi}$	主题-词分布矩阵
d	文档
BOW	文档词袋表示
TCN	时序卷积网络
CNN	卷积神经网络
BiLSTM	双向长短时记忆网络
ReLU	ReLU 激活函数
Softmax	Softmax 归一化函数，其作用于向量 v 后的第 i 个分量为：$\mathrm{Softmax}(v)\vert_i = \frac{e^{v_i}}{\Sigma_j e^{v_j}}$
$N(\mu, \sigma^2)$	以 μ 为均值、σ^2 为方差的高斯分布
$\mathbb{E}f(z)$	$f(z)$ 的数学期望
$\inf f(x)$	$f(x)$ 的下确界
$\|x - y\|$	x 与 y 的欧式距离

第 1 章　绪论

1.1　研究背景与意义

深度学习领域的突飞猛进，促进了人工智能的迅速发展，也掀起了国内外对智能对话技术的研究热潮。智能对话技术是人机交互研究的热点，其发展影响及推动着语音识别与合成、自然语言理解、对话管理以及自然语言生成等领域的进展；许多大型互联网公司都推出了智能对话技术相关的产品，如微软的小冰语音助手、苹果公司的 Siri、阿里巴巴的小蜜智能客服等，这些行业巨头也将智能对话技术作为其重点研发方向。对话生成技术作为智能对话领域的重要研究子方向之一，在人工智能发展的各个阶段都受到了人们大量的关注。

如今，大规模的真实对话语料库已经成为提升数据驱动的对话生成系统性能的必不可少的组成因素。Al-Rfou 等[1] 开始尝试使用 Reddit 论坛数据进行多人、多轮次的对话系统研究。利用从中下载的 1 亿多条论坛对话数据可以训练出具有更大语境和参与者历史信息的识别模型，在他们定义的答复选择排序任务中取得了更好的实验效果。这项研究初步证明了更大规模、更符合自然对话的数据对提升对话生成模型性能的重要性。

然而公开的大规模的真实对话语料数量较少，其中中文领域尤其匮乏。传统的如 CASIA-CASSIL 对话数据[2] 规模较小且面向特定任务，而从社交媒体中获取的对话数据，如新浪微博[3]、百度贴吧[4] 等，则是以消息-回复对的形式组成，与人们日常对话相差较大。表1.1列出了目前主要的对话语料库及其特点。

表 1.1　主要中英对话语料库统计

对话语料库	数据规模	特点
Ubuntu Dialogue[5]	930K dialogues	英语，单轮，任务相关
Switch Board[6]	650 dialogues	英语，多轮，开放域，稀疏
Resturant comments[7]	6618 dialogues	英语，多轮，开放域，稀疏
DailyDialog[8]	13118 dialogues	英语，多轮，开放域
Weibo[3]	4.4M pairs	中文，消息-回复，开放域
Tieba[4]	82K pairs	中文，消息-回复，开放域
Douban[9]	2M pairs	中文，问答，开放域

为了获取更多符合特定需求的真实对话数据，近年来研究人员开始尝试使用众包策略人工构建多话轮对话数据集。Zhang 等人[10] 构建了基于特定人物角色的

PERSONA-CHAT 对话数据集：首先众包构建 1155 个包含至少 5 个句子的人物角色描述，然后招募两组众包者分别扮演不同人物角色完成了包含 164,356 条话语消息的 10,981 个角色相关对话片段。

采用类似的策略，Zhou 等人[11] 开发了基于特定文档的对话数据集 CMUDoG，Moghe 等人[12] 则构建了基于背景知识的电影相关的对话数据集。这些数据集为探索更深入的多轮对话生成模型打下了很好的基础。

值得关注的是，影视剧字幕提供了对话数据的丰富资源，相比于微博、论坛等社交媒体用语，其风格更接近正常的人类日常对话，且公开可获取的字幕具有十分巨大的数据量，Lison 等人[13] 分析了免费网站 OpenSubtitles.org 中 2018 全年共 60 种语言的字幕文件，词量达到了 22.2G。其中英文规模最大，达到 447K 字幕文件、441M 字幕句子和 3.2G 总词次；简体汉语的相应规模也分别达到了 29.1K、31.2M 和 191M。比 2016 年的相关数据增长了 30% 以上。巨大的研究价值使得字幕数据受到了越来越多研究人员的关注。

尽管字幕数据蕴藏了巨大的潜在价值，但由于字幕中说话人和场景信息的缺失，使得字幕中话轮关系不明确，且前后相邻的话语可能在语义上发生较大的跳跃，导致字幕难以直接作为对话语料用于对话生成研究。一种可行的方法是，将字幕对话流以其中的话题线索为单位、按照主题相似度进行组织。字幕中的对话以流的形式呈现，其中往往包含多个话题线索，且相邻的话题线索之间没有明显分界。若对字幕流数据直接进行主题建模，则由于其中杂糅了多个主题而使主题建模的效果不够理想，因此，有必要将字幕根据主题的延续性进行分割，再将分割后的片段通过主题建模获得其主题表示，并按照主题的相似程度进行聚类，则由此产生的对话片段库，无论是直接用于对话生成训练，还是后续进一步提炼，都比原始对话流有着巨大的优势。其中关键的环节是 1）对字幕流的主题分割，将连续的字幕流分割为具有不同主题的对话片段；2）对分割后的对话片段进行主题建模，挖掘出对话片段潜在的主题信息。

由于字幕中对话文本具有长度极短、稀疏性严重的特点，而已有主题分割方法和主题模型大多针对长文本而设计，在对话文本上难以达到理想效果，而这两个步骤对字幕对话文本的聚类效果有直接的影响，因此本文主要研究在字幕流对话文本场景下的主题分割和主题建模方法。

1.2　国内外研究现状

本节分为两小节，分别简要介绍了主题分割和主题建模的典型研究工作。

1.2.1　主题分割相关研究

主题分割旨在将一篇文档分割为若干个主题片段，使得每一个主题片段内的句子拥有相近的主题。主题分割为诸如文本摘要，信息检索，对话分析等任务提供了基础。例如，在长文本中查找特定部分的情况下（例如会议记录或字幕），除非搜索整个文档，否则很难找到感兴趣的片段的起始点。但是，如果将文档组织为主题片段，则检索起来会容易很多。

已有的主题分割方法包括无监督的方法和有监督的方法。

无监督的方法利用了主题和词汇用法之间的强相关性，可以粗略地分为两类：基于相似度的方法和概率生成方法[14]。基于相似度的方法假定相同主题片段中的句子比不同主题片段中的句子有更高的相似度。因此，可以通过相邻句子之间的相似度变化来检测主题转换。代表的模型包括 Texttiling[15]，C99[16]，LCSeg[17]。概率生成方法假定文档由一系列隐藏的主题变量组成，并且每个主题都有自己的分布。因此，可以通过单词分布的变化推测主题转换，这类方法以 HMM 和 LDA 为代表。Purver 等人[18] 提出的 PLDA 通过计算了两个相邻段落之间的共有的主题分布量来推断主题的变换。Misra[19] 和 Riedl 等人[20] 则提出了基于 LDA 来计算句子的相似度，进而推断主题变换。

有监督的方法可以利用更多特征（例如提示短语，长度和相似度得分），大致可分为基于决策树的分类器[21] 和概率模型[22][23]。Hakkani 等人[21] 结合了词汇特征（如词汇相似度）和会话特征（如长时间停顿，语速变化，沉默等），以进行主题和子主题的分割。Hernault 等人[24] 在基于 CRF 的分割模型中集成了词汇和句法功能。有监督的方法通常比无监督的方法具有更好的性能，但是它们依赖于大量的标记数据和手工设计的特征。

近年来，一些研究人员探索了神经网络方法在主题分割任务中的应用。Wang 等人[25] 首次提出了一种基于 BiLSTM 和 CNN 的序列标记架构用于主题分割。Wang 等人[26] 提出了一个基于 CNN 的模型，通过学习段落之间的偏序关系来对语义一致性进行排序。Badjatiya 等人[27] 使用 CNN 和 BiLSTM 分别对句子和上下文进行编码，并引入注意力机制来解决 BiLSTM 的远程依赖问题，该模型通过对当前句子是否是主题转换点进行分类来完成主题分割任务。Sehikh 等人[28] 提出了一种基于 RNN 的模型，用于对语音识别生成的文本进行主题分割。Koshorek 等人[29] 基于双层 BiLSTM 的构造了一个分割模型，低层对句子的语义进行编码，而高层对上下文信息进行编码。通过对句子序列进行标注来实现主题分割。该模型与本文所提出的模型有相似的思路，但是 BiLSTM 不擅长处理远程依赖关系，这使得它的句子语义表示通常不如 Transformer。

1.2.2　主题建模相关研究

神经主题模型以神经网络为基础，通过刻画文本生成过程来表征文本的潜在主题信息。在此类模型中，输入一般采用文档的词袋模型，并添加相应的词向量层和其他网络层以生成文档。模型通常使用反向传播算法来逐层更新参数。

早期神经主题模型以前馈神经网络为基础构建，利用其权重矩阵分别来表示文档-主题分布和主题-词分布。在此之后，具有高斯先验分布隐空间结构的变分自编码器（VAE）被用于构建主题模型。

Miao 等人[30] 首先提出基于 VAE 的神经变分文档模型 (NVDM)，该模型通过无监督的训练模式，试图从文档的词嵌入空间中提取潜在主题特征，其架构符合标准 VAE 的网络结构，并假设隐空间中主题表示的先验分布为高斯分布，通过多层感知机将每个文档编码为隐空间中主题的后验分布 $q(z|d) = ReLU(\mu_d + \epsilon\sigma_d)$；解码器则利用重参数采样来生成文档。在 NVDM 中，仅仅采用了权重矩阵来表示主题-词分布，因此，NVDM 的主题一致性往往不如 LDA。

Ding 等人[31] 则针对这一问题，提出使用预训练词向量来度量主题词间的语义相似性，将其纳入优化目标，迫使模型关注主题一致性。与 NVDM 相比，该方法的确有效提高了主题一致性。

Miao 等人[32] 以 NVDM 为基础，在 VAE 的结构下提出了一系列形式相近的神经主题模型，并以主题权重为研究点，着重考查了隐空间主题的表示方式。对于给定的输入文档 d，这类模型通过多层感知机将其映射为隐空间的正态分布 $N(\mu_d, \sigma_d)$，并从该分布中采样得到隐变量 z，不同于 NVDM，z 并不直接用作解码器的输入，而是通过以下三种不同的变换方式得到文档-主题分布 $\theta = q(z|d)$：（1）Gaussian Softmax Construction (GSM)：即通过 Softmax 函数进行归一化，得到主题分布 $\theta = Softmax(W^T z)$；（2）Gaussian Stick Breaking Construction (GSB)：通过折棍模型和 Sigmoid 函数来得到主题分布 $\theta = f_{SB}(Sigmoid(W^T z))$，理论上可证明，折棍模型产生的分布服从狄利克雷分布，因此该模型实际上以近似的狄利克雷分布作先验；（3）Reccurent Stick Breaking Construction (RSB)：采用折棍模型和循环神经网络 RNN 构建主题分布 $\theta = f_{SB}(f_{RNN}(W^T z))$。

Srivastava 等人[33] 基于 VAE 提出了另一种主题模型 AVITM。该模型摒弃了多元高斯和 Dirichlet 分布的先验假设，而是提出了 logistic-正态分布作为文档-主题分布的先验假设，这使得模型既可以避开变分自动推断引起的主题同质化问题，又可以借助 VAE 的重参数技巧加快训练速度。相比于 NVDM 和变分推断下的 LDA，AVITM 具有更高的主题一致性。

Nan 等人[34] 认为 VAE 所高度依赖的重参数技巧仅适用于“均值-方差”分布

族，这极大地限制了先验分布的选择范围，另一方面，VAE 所采用的 KL 散度迫使所有后验分布都逼近标准高斯分布，本质上使得后验分布独立于文档输入，造成后验坍塌问题。Nan 等人进而提出基于 WAE 来构建主题模型，WAE 采用 Wasserstein 距离来刻画分布之间的差异，该距离比 KL 散度具有更好的稳定性。他们选择了同 LDA 相同的狄利克雷先验假设，将模型称作 WLDA，实验表明 WLDA 在长文本上取得了比 LDA 和此前的神经主题模型更高的主题一致性。

Dieng 等人[35] 提出嵌入主题模型 ETM，该模型同样基于 VAE 的框架，在训练主题表示的同时，联合训练词向量，且以主题-词分布的向量作为主题的嵌入表示，将主题向量与词向量的内积加入优化目标中，得到的主题向量具有较直观的可解释性，他们的实验表明，其主题一致性超过了 LDA 和 NVDM。

1.3　本文研究内容与章节组织

本文以字幕流数据为基础，以对话文本主题聚类为目标，主要研究了字幕流对话文本场景下的主题分割与主题建模方法。

与新闻、独白等长文本不同，对话文本有以下特点：

- 每个文档仅包含极少乃至一个句子，且所含单词数量极少；
- 每个文档所覆盖的单词类别仅占整个词汇表的很小一部分。

这两个特点使得基于词袋模型构建的“文档-单词”矩阵仅含有极少的非零项，即所谓的稀疏性问题。

由于本文基于有监督学习模式设计主题分割模型，需要足够的带标注的训练数据，因此本文首先以公开可得的英文剧本和中英文双语字幕为基础，提出了一套自动标注算法，将剧本中的场景和说话人信息标注到字幕中，由此获得带标注的基础字幕对话语料库 Sub，该数据集为下一步主题分割提供了训练和测试数据的基础。

主题模型是通过词与词在文档中的共现关系来推断主题-词分布，因此所包含主题越单一的文档越有利于主题模型习得真实的主题-词分布，由此可知，主题分割越准确，即分割后的主题片段内部具有更单一的主题分布而尽可能避免杂糅多个主题，则提供给下一步主题建模的对话片段的主题内聚性越高，进而越有利于得到准确反映真实主题分布的的主题模型。本文基于 BERT 和时序卷积网络设计了序列标注架构的主题分割模型，并基于上述自动标注所得的 Sub 数据集构造了扩充的 Sub 训练集和测试集，同时为了更全面地评估该主题分割模型在不同类型的文本下的性能，本文还测评了该模型在新闻类数据集 Weibo 和两人日常会话数据集 DAct[36] 上的性能。

利用训练好的主题分割模型，将更多没有对应剧本的字幕流切分为对话片段，并对这些对话片段进行主题建模。此前的主题模型大都针对长文本进行建模，然而研究表明，文档的长度对于主题模型有着较大的影响，长度过短的文档将极大地降低 LDA 等概率主题模型的性能。因此，传统的长文本主题模型在对话文本中不再适用，有必要设计更为适合的主题模型。本文对比分析了若干具有代表性的主题模型，针对对话文本的特点提出了基于高斯混合先验的神经主题模型 GMNTM，设计实验测评了其在基础字幕数据集 Sub 和扩展字幕数据集 SubX 上的性能。此外，本文还在较大的两人日常对话数据集 zhdd[8] 和新闻短标题数据集 cnews 上测试了该模型，以更全面地评估该模型在短文上的主题建模效果。

利用训练好的主题模型，抽取出分割得到的对话片段的主题表示，本文进而对比了不同聚类算法在抽取出的主题表示上的聚类性能，并展示了自动聚类形成的类簇中的若干话题线索。

图1.1展示了本文的总体流程框架：1）最初通过自动对齐剧本和字幕，为下一步主题分割提供了标注数据；2）由于主题建模是建立在文档基础上，因而主题分割切分出的话题片段，为下一步的主题建模提供了良好的文档划分基础；3）有效的主题建模为下一步的主题聚类提供了较好的主题表示。

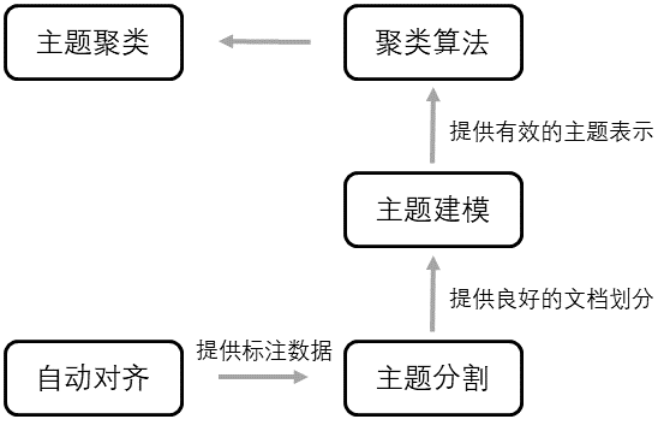


图 1.1　对话文本主题聚类总体流程

本文的主要贡献包括：

- 以影视剧字幕为基础，提出了一种自动标注场景和说话人的方法，并构建了包含 26 万个话语消息（utterance）的对话语料库。
- 基于 BERT 和时序卷积网络提出了一种针对对话流的主题分割模型。
- 通过对比分析已有的神经主题模型，提出基于高斯混合先验的主题模型，并在字幕对话数据集上验证了该模型的有效性。
- 基于所提出的主题模型，对大规模的未标注对话文本进行了聚类，并对聚类

的结果进行了评估。

本文具体研究内容与各章节组织安排如下：

第 1 章：绪论。本章主要阐述本论文的研究背景及意义；概述了国内外相关课题的研究进展；概括了本文的主要贡献和创新点；简要介绍了本论文的章节组织结构。

第 2 章：字幕对话场景标注。本章主要说明字幕数据的预处理步骤，通过对齐字幕和其对应的剧本，来为字幕对话标注场景和说话人信息，介绍了字幕与剧本的粗对齐模型，以及对齐错误的检测和纠正模型。

第 3 章：字幕对话主题分割。本章介绍了传统的主题分割方法，并介绍了本文提出的基于 BERT 和时序卷积网络的主题分割模型的设计，并通过模拟数据集实验来证明了该模型的有效性。

第 4 章：对话文本主题建模。本章主要阐述了主题模型相关的基础知识和评价指标，介绍了经典概率主题模型 LDA，介绍了目前最新的 3 种神经主题模型，包括基于标准 VAE 的主题模型 NTM-GSM，基于 WAE 的的主题模型 WLDA，基于 Embeded Topic space 的主题模型 ETM，介绍了本文提出的具有高斯混合先验的神经主题模型 GMNTM，通过实验对比不同神经主题模型的性能差异，并验证了所提出的 GMNTM 的有效性。

第 5 章：对话文本主题聚类。本章主要介绍了聚类结果的评价方法，介绍了所采用的聚类方法，以及实际的聚类结果及示例。

第 2 章　字幕对话场景标注

2.1　本章概述

如绪论所述，字幕数据中说话人和场景标记等结构信息的缺失，导致很难判断两个连续的字幕行是否属于同一说话人，因此难以直接从字幕中提取出可用的对话数据。

另一方面，影视剧剧本具有丰富的结构信息，包括说话人和场景边界。值得指出的是，公开可获取的中文影视剧剧本数量极少，不能直接以此作为构建对话语料的数据来源，但是，互联网上却存在相当数量的英语剧本可以公开获得。由于中英双语字幕中的中文和英文句子已天然对齐，并且字幕通常与它们对应的剧本共享大部分对话内容，因此字幕与对应的剧本存在对齐的可能性，在对齐完成后，可利用剧本中的结构标签为字幕数据标注所需的结构信息。因此本文在句子级别对齐这两种资源（即英文剧本和双语字幕），并通过为字幕标注说话者和场景边界标签构成一个基础的中文对话语料库。

上述标注过程可分为两个主要步骤：1）粗对齐：对于每个字幕句子，利用信息检索技术找到相应剧本中最佳匹配的话语 (utterance)，并将该话语的说话人和场景标签映射到字幕句子；2）错误检测和纠正：基于时序卷积网络识别第一个步骤中产生的对齐错误，并采用启发式策略来纠正这些对齐错误，以提高标注质量。

2.2　字幕与剧本粗对齐

本文从互联网上收集了 4 部美剧的剧本（*Castle*，*Friends*，*House* 和 *The Big Bang Theory (TBBT)*）及其对应的字幕。由于剧本是半结构化文本，必须先将它们解析后然后才能提取出其中的元素。每个美剧的剧本都有自己的格式规范，因此需要针对不同的美剧设计对应的解析器。通常，美剧剧本中包含三种元素：1）场景标题，通常包含诸如“Scene”，“INT”，“EXT”等特殊字符串，用以标记场景的开始并可以用来指示场景的边界；2）说话人，通常出现在一行的开头，后跟冒号；3）对话内容，通常紧跟在说话人后面，并一直持续到该行的结尾。

在本文中，将场景定义为两个相邻场景标记之间的所有内容，话语则定义为说话人一次对话所包含的内容，图2.1显示了 TBBT 第 2 季第 8 集（S02E08）的剧本片段。

```
Scene: The apartment.
Sheldon: Oh look, Saturn 3 is on.
Raj: I don' t want to watch Saturn 3. Deep Space Nine is better.
Sheldon: How is Deep Space Nine better than Saturn 3?
Raj: Simple subtraction will tell you it' s six better.
Leonard: Compromise. Watch Babylon 5.
Sheldon: In what sense is that a compromise?
Leonard: Well, five is partway between three... Never mind.
Raj: I' ll tell you what, how about we go rock-paper-scissors?
```

图 2.1　原始剧本片段

根据这些特点，本文设计了解析器来提取上述元素。解析算法的主要流程为：首先，过滤掉所有动作指令和场景描述（通常用括号括起来）。然后，逐行扫描剧本并使用不同的模式来检测特定元素。一旦检测到场景标题，则结束上一个场景并开始一个新场景。最后，将所提取出的话语和其他元素将转换为 XML（可扩展标记语言）的格式输出。

```
<scene id="1">
  <utterance uid="1-1">
    <speaker>Sheldon</speaker>
    <content>Oh look, Saturn 3 is on.</content>
  </utterance>
  <utterance uid="1-2">
    <speaker>Raj</speaker>
    <content>
      I don' t want to watch Saturn 3. Deep Space Nine is better.
    </content>
  </utterance>
  <utterance uid="1-3">
    <speaker>Sheldon</speaker>
    <content>
       How is Deep Space Nine better than Saturn 3?
    </content>
  </utterance>
  <utterance uid="1-4">
    <speaker>Raj</speaker>
    <content>
       Simple subtraction will tell you it' s six better.
    </content>
  </utterance>
  <utterance uid="1-5">
    <speaker>Leonard</speaker>
    <content>Compromise. Watch Babylon 5.</content>
  </utterance>
```

图 2.2　解析后的剧本片段

图2.2显示了处理后的剧本格式，其中 <scene> 标签代表场景边界，id 属性表示其索引号。<utterance> 标签标记话语消息，其 uid 属性指示其在场景中的顺序号，例如 uid =’4-3’，表示该话语是第四个场景的第三个话语。因此，若两个话语具有相同的 uid 前缀（例如 5-6 和 5-2），表明它们位于同一个场景中。剧本中，每个话语拥有唯一的 uid，每个 uid 属性均可用于确定话语。因此，一旦知道字幕行对应话语的 uid，则可以查找到说话人的姓名并将其映射到字幕中。

字幕文件由连续的字幕块组成，每个字幕块由文本内容行和时间戳行组成。在本文中，字幕行是指字幕块中的台词内容。一般而言，剧本中的话语不一定总能

与字幕行完全匹配。影视作品拍摄过程中的修改、删除或是演员的即兴表演都可能使得剧本话语和字幕行之间产生差异。此外，剧本中较长的话语可能分为几个较短的字幕行，或者剧本中两个较短的话语可能合并在单个字幕行中，因此，从剧本到字幕的对应关系可能是一对一，一对多或多对一。

本文采用信息检索技术来处理字幕和剧本的对齐问题。将剧本中的话语视作文档，而每个字幕行则被视作查询，则对齐任务转换成为字幕行（查询）选择最匹配的话语（文档），并将话语对应的 uid 标注到字幕行中。本文采用 BM25 指标作为衡量文档和查询之间的相关性的评分函数。

在实践中，本文选择 ElasticSearch（开放源代码搜索引擎）完成索引和搜索的任务，并将剧本中匹配度最高的话语的 uid 标注到字幕行中。图2.3展示了两个已标注有 uid 标签的字幕片段（分别取自 Friends S02E08 和 TBBT S02E11），其中每个字幕行的行首由两个尖括号包裹的即为 uid 标签。

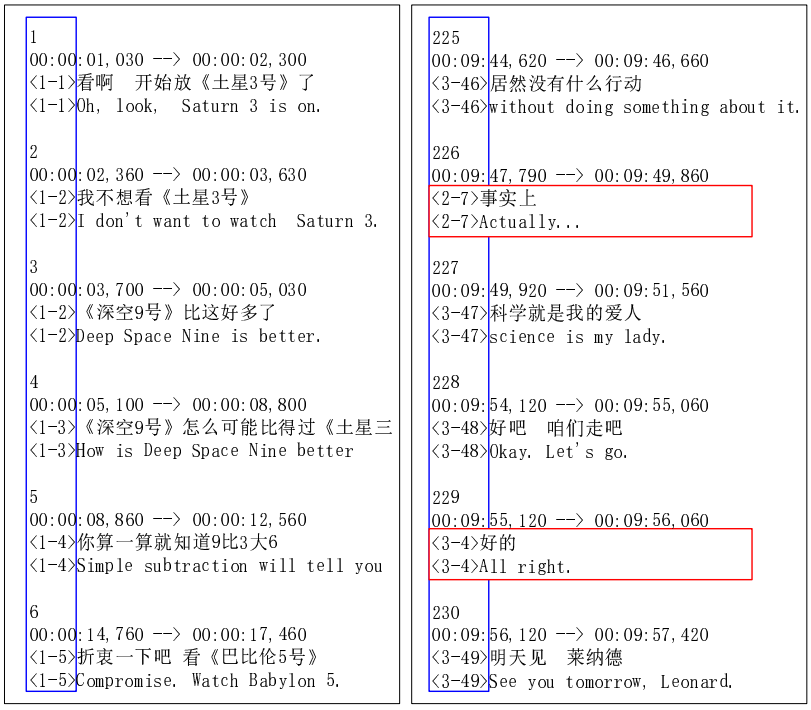


图 2.3　粗标注后的剧本片段

2.3　错误识别和纠正

2.3.1　错误识别模型

如上节所述，字幕及其对应剧本之间在内容和格式上的差异都将增加对齐的难度，诸如表达方式的变化，对话内容的重复，文本过短等现象都可能导致映射过程中产生对齐错误。

一般情况下，引发对齐错误的原因是在剧本中存在这样的话语，由于其长度和所包含词汇的缘故，会与本不应匹配的字幕行拥有最高的相似度。例如，当剧本中的较长的话语在字幕中分成若干行，且剧本中存在另一个相似但较短的话语时，较短的话语可能会与所查询的字幕行具有更高的相似度，因为较短的话语会增加词的频率。图2.4显示了此类错误情况：左蓝色框中的第一行字幕（“Peter?”）本应该与右下蓝色框中包含 6 个词汇的的话语“Peter! Peter! That guy’s pretty huge”（uid＝“14-7”）相匹配，但实际的试验中却被对齐到右上角红色框中的话语（uid＝“7-7”），因为后者仅含有一个词汇“Peter”。（取自 *Friends* S03E24）

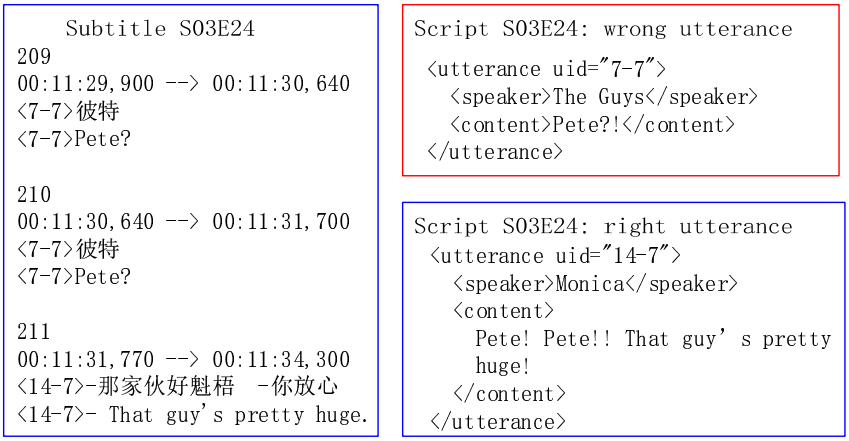


图 2.4　对齐错误情形一

这些映射错误将导致将错误的 uid 被穿插放置在正确标注的 uid 之间。如图2.3中左侧的片段所示，蓝框中的 uid 标签表明字幕的整个标签序列实际上构成大部分有序的数对序列。而映射错误将会破坏这种一致性：图2.3右侧片段中的蓝框为含有对齐错误的标记序列，由红框标出了标注错误的 uid，整个序列为 [⟨3-46⟩,⟨2-7⟩,⟨3-47⟩,⟨3-48⟩,⟨3-4⟩,⟨3-49⟩]，根据上下文，出现在第一个红框中的 uid 应该是 ⟨3-46⟩ 或 ⟨3-47⟩，而实际上却是 ⟨2-7⟩，并且 uid 出现在第二个红框中的值应该是 ⟨3-48⟩ 或 ⟨3-49⟩，而实际上却显示为 ⟨3-4⟩。因此，从粗标注的字幕中提取的 uid 标签可被视为递增的点对序列，而通常破坏一致性的映射错误则可被视为该序列中的异常点。纠正这些错误的标注可以分为两个主要步骤来完成：第一步，构建一个异常检测模型，将标签序列输入模型中，检测到异常后将其替换为特殊标签 ⟨0,0⟩。第二步，根据这些特殊标签的上下文采用一些启发式策略进行还原。

由于 uid 标签序列是几乎有序的序列，因此一种自然的方法是计算其差分序列，并检测差值远大于其他差值的位置。本文探索了加入一个滑动窗口来处理这个任务。但是，窗口的大小成为一个问题：如果窗口设置得太小，则其可能会被一段连续的错误 uid 标签覆盖，无法纠正其中的任何一个；如果将窗口设置得太大，

则其中可能会有几段错误的 uid 标签，无法分辨出哪些是异常值。该问题的难点在于应该根据正确的标签检测出错误的标签，但是，除非知道序列的模式，否则无法识别哪些是正确的标签。因此，本文将问题转换为序列建模任务来处理。

本文采用时序卷积网络（TCN）来对 uid 标签序列进行建模。TCN 是 Bai 等人[37] 提出用于建模时序问题的一种模型，在 10 个标准序列建模任务上都超过了 RNN。TCN 使用卷积运算来处理序列建模问题。与 RNN 相比，卷积的结构允许 TCN 并行计算，从而具有更快的速度。TCN 的设计具有两个特点：1）输出与输入的长度相同；2）仅使用过去的信息，未来的信息不会泄漏到过去。这些特点使其适用于 uid 序列中的异常检测任务，异常检测模型为基于 TCN 的分类器，可对序列中的 uid 标签是正常标签还是异常标签进行分类。

图2.5展示了本章异常检测模型的整体架构，输入为一个 uid 序列，uid 的每个维度都作为输入层的一个通道。该网络有 8 个隐藏层，根据经验，连续异常段的长度通常不超过 7，因此将卷积核的大小设置为 7。输出是 $\{0,1\}$ 二值序列，其长度与 uid 标签序列相同，异常将被标记为 1。例如，在图2.5中，在输入序列中以红色突出显示的 uid ⟨3-2⟩ 是一个异常标签，因此网络输出 1 来表明检测结果。

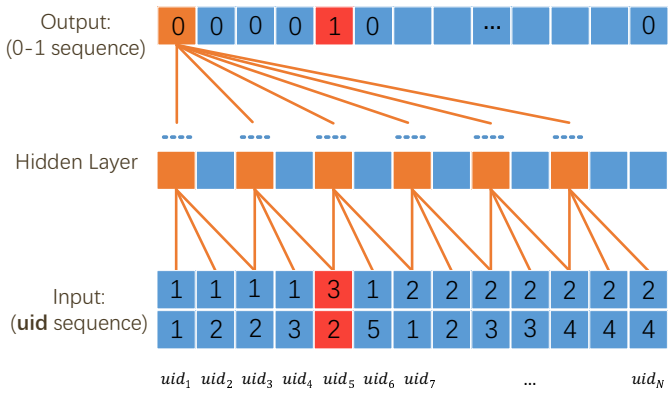


图 2.5　对齐错误识别模型架构

由于 TCN 属于监督学习，需要带标签的训练数据。为了解决训练数据缺失的问题，本文采用人工生成训练数据的方法，要求所生成的模拟数据尽可能地接近真实数据，确保它们在标签序列中场景的数量和长度上具有相同的分布。由于 uid 的第一维和第二维分别代表标记序列中场景的索引和场景中话语的索引，因此该第一维的最大值等于标记序列中场景的数目。场景的长度等于该场景中 uid 的第二维最大值。因此，在给定序列中的场景数量和每个场景的长度范围后，则可以自动生成有序标签序列。本文首先对所处理的 4 部美剧中 uid 的分布情况进行了统计。

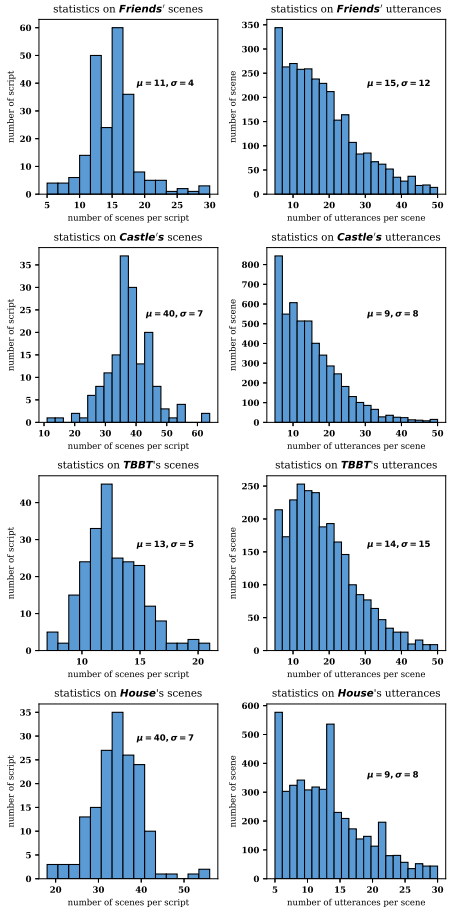


图 2.6　4 部美剧中场景数量和长度的分布统计

图2.6显示了统计结果。第一列的图对应于每个剧本的场景数量，第二列的图对应于场景的长度（每个场景的话语数量）。结果表明，这四部美剧的上述两个指标都近似于正态分布，因此在有序标签自动生成算法中使用正态分布进行采样会较为合理。根据图2.6中所示的场景数量和长度的平均值和标准方差，可以使用真实标签序列生成具有相同统计特征的序列。算法1展示了自动生成有序标签序列的步骤，该算法接收 4 个参数：$s\mu$，$s\sigma$，$u\mu$，$u\sigma$ 分别表示场景数量和场景长度的平均值和标准方差。该算法中的两个 while 循环用于确保其产生的序列包含两个以上的场景。

例如，假定 $(AvgCnt, StdCnt, AvgLen, StdLen) = (s\mu, s\sigma, u\mu, u\sigma)$，采样出一个正态随机变量 $\xi \sim N(s\mu, s\sigma)$ 的值（例如 $sMax$）来表示场景数量，给定一个场景后，采样出另一个随机变量 $\zeta \sim N(u\mu, u\sigma)$ 的值（例如 $uMax$）来控制该场景的长度，假设当前场景的 **id** 属性为 sid，接下来生成一个形如 $[\langle sid, 1\rangle, \cdots, \langle sid, uMax\rangle]$ 的序列来表示该场景的 **uid** 序列，并将该序列添加到整个标签序列中。

与真实标签数据相比，生成的序列保持有序而缺乏异常点。因此，应该将异

Algorithm 1 Generate Ordered Sequence

Input: $s\mu$,$s\sigma$,$u\mu$,$u\sigma$ **Output**: Tag Sequence with Order

令 $lst = [\,]$，$sMax = 0$
采样 $sMax \sim Normal(s\mu, s\sigma)$，$sMax$ 需不小于 2
for $sid = 0$ to $sMax - 1$ **do**
  $sid = sid + 2$，$uMax = 0$
  采样 $uMax \sim Normal(u\mu, u\sigma)$，$sMax$ 需不小于 2
  令 $tmp = [\,]$
  for $uid = 0$ to $uMax - 1$ **do**
    $tmp = tmp + [[sid, uid + 1]]$
  end for
  $uLst = uLst + tmp$
end for
return $uLst$

常点添加到目前的有序序列中。跟据第上一小节的分析，标记错误意味着某些 uid 被映射到了错误的位置，其中有些甚至多次错误映射。

为了模仿该过程，本文采用了诸如切换两个随机选择的 uid，多次重复插入一个 uid 等策略来模仿真实标签数据中的异常。算法同时构造了对应的标签：将异常添加到序列中后，该异常的对应标签将设置为一个 1。对于每个美剧，我们构建了包含 200K 个序列的训练集来训练错误检测模型。

2.3.2　错误纠正模型

为了纠正序列中检测到的异常，使用与上述训练错误检测模型相似的方法建立了另一个错误纠正模型，然而实验结果表明，由于神经网络的连续性，该错误纠正模型可能会将某些正确的 uid 更改为错误的 uid。因此，本文提出一种启发式算法来重建检测出异常的序列。该算法使用多种策略来重建已识别的错误：

- 如果异常标签是一个孤立点，则根据其位于场景的边界还是内部区域，将其替换为其前后点的 uid；
- 如果异常标签形成一个完全位于场景内部的连续片段，则将它们直接替换为使用线性插值算法得到的相同数量的点；
- 如果异常标签形成一个连续的片段，并且该片段跨越两个场景的边界，则边界后面的部分将由递减的点序列取代，然后通过插值算法来计算边界前面的部分。

图2.7列举了几种情况及其使用这些策略的纠正结果。

图 2.7　纠正算法处理的三种情形

将训练完成的 TCN 和纠正算法结合起来，即可检测和纠正映射错误。之后可将说话者标签和 uid 序列都标注到字幕行中。如果在原始剧本中找不到由纠正算法创建的 uid，则该字幕行的说话人将设置为前一个字幕行的说话人。图2.8显示了纠正后的字幕。

```
1
00:00:01,030 --> 00:00:02,300
<1-1,Sheldon> 看啊 开始放《土星3号》了
<1-1,Sheldon> Oh, look, Saturn 3 is on.

2
00:00:02,360 --> 00:00:03,630
<1-2,Raj>我不想看《土星3号》
<1-2,Raj>I don't want to watch Saturn 3

3
00:00:03,700 --> 00:00:05,030
<1-2,Raj>《深空9号》比这好多了
<1-2,Raj>Deep Space Nine is better.

4
00:00:05,100 --> 00:00:08,800
<1-3,Sheldon>《深空9号》怎么可能比得过
<1-3,Sheldon> How is Deep Space Nine
```

```
225
00:09:44,620 --> 00:09:46,660
<3-46,David> 居然没有什么行动
<3-46,David> without doing something

226
00:09:47,790 --> 00:09:49,860
<3-47,Leonard> 事实上
<3-47,Leonard> Actually...

227
00:09:49,920 --> 00:09:51,560
<3-47,Leonard> 科学就是我的爱人
<3-47,Leonard> science is my lady.

228
00:09:54,120 --> 00:09:55,060
<3-48,Penny> 好吧　咱们走吧
<3-48,Penny> Okay. Let's go.
```

图 2.8　字幕片段纠正后的标注结果

2.4　实验结果与分析

本文首先评估了错误检测模型的性能。在相同的生成算法下构造了一个独立的测试集。评估表明，该模型可以在测试集中获得 0.97 的 F1 分数。本文还以恢复

准确度评估了上述启发式算法的能力：$acc = \frac{n}{len}$，其中 n 是已恢复序列与原始序列之间相同 uid 标签的数量，len 是标签序列的长度。实验表明，该算法的平均恢复精度为 0.95。

将该方法应用于这 4 部美剧的所有字幕中，得到了 779 个结构化剧本和 779 个带标注的中英文字幕，其中包含 18129 个场景。表2.1和表2.2列出了主要统计数据。

表 2.1　4 部美剧对话语料基本统计

Item	Size
Total num of structured scripts	779
Total num of scenes	18129
Total num of utterances	260674
Average num of scenes per script	23
Average num of utterances per scene	14

表 2.2　4 部美剧剧本统计

	tbbt	house	friends	castle
num of episodes	225	164	227	163
num of scenes	2839	5652	3499	6139
num of uttrances	50161	69380	59859	81274
num of speakers	484	1039	691	2074
spkrs per scene	3.56	3.00	3.47	3.30
avg uttr length	11.23	11.37	10.13	12.29

验证数据集的结构如下：对于每个美剧，本文从每一季中选择一个字幕，并根据相应的剧本手动为每个字幕文件标注 100 行，并带有说话人和 uid 标签。例如，对于美剧 TBBT，选择了 S01E01，S02E02，...，S10E10 字幕，对于第 i 个字幕 S0iE0i，手动标记了从 1 到 100 的行。然后将自动标注结果与这些手动标注的标签进行比较，以分别评估话语和场景边界的准确性。实验结果详细显示在表2.3和表2.4中。

在表2.3中，值得注意的一点是，仅使用 BM25，正确映射的标签的比率都高于 80%，这为根据正确的标记纠正映射错误提供了可能性。尽管本文的方法可以正确地将大多数字幕行与其对应的话语对齐，但是仍然无法处理将两个简短话语合并为一个字幕行的情况，这种情况在所有字幕行中都占 0.037。该方法不能完全纠正所有映射错误的其他可能原因是：1）人工生成的训练数据的分布与真实标签

序列的分布之间存在差异；2）恢复策略是启发式的，可能与实际情况有所不同。

Wang 等人[38] 使用 TF-IDF 作为加权因子以及移动窗口策略来对齐 Friends 的剧本话语和字幕行，话语和场景注释的准确率分别可以达到 81.79% 和 98.64%。本文在同一个美剧 *Friends* 上进行的实验表明，仅使用 BM25 作为排名函数而没有错误纠正程序或其他策略，其话语准确度可以达到 85.2%，场景边界的准确度可以达到 93.3%，此外，本文还对其他三个美剧进行了此方法的评估，其话语平均准确度为 84.4%。

表 2.3　说话人平均标注准确率

	TBBT	Friends	Castle	House
TFIDF	0.825	0.818	0.793	0.776
BM25	0.887	0.852	0.812	0.809
BM25+TCN	**0.949**	**0.933**	**0.952**	**0.951**

表 2.4　场景边界平均标准准确率

	TBBT	Friends	Castle	House
TFIDF	0.952	0.986	0.936	0.932
BM25	0.943	0.962	0.926	0.934
BM25+TCN	**0.992**	**0.989**	**0.975**	**0.983**

说话人标注准确率的计算方法为 $\frac{m_u}{n_u}$，其中 m_u 表示右边带注释的说话的数量，n_u 表示字幕中说话的总数。场景边界的准确性计算为 $\frac{m_s}{n_s}$，其中 m_s 和 n_s 分别表示字幕中正确注释的场景边界的数目和场景边界的总数。

结果表明，经过纠错处理的标签序列可以达到 94.62% 的话语准确率，比不进行降噪处理的话语准确率高 10.62%，说明了本章的纠错方法的有效性。

2.5　本章小结

本章通过对齐字幕和对应的剧本，为字幕自动标注了场景和说话人。本文假设同一场景下的对话内容为同一主题，而相邻场景下的对话其主题会发生转换。因此，场景边界标记可为主题分割提供划分依据。通过本章的处理，最终获得了 26 万条带标注的话语消息，形成了基础的字幕对话语料库 Sub（subtitle），为下一章的主题分割和后续的主题建模提供了有较高质量保证的“种子”数据。

第 3 章　字幕对话主题分割

3.1　本章概述

在字幕对话流的主题分割问题上，已有主题分割模型大多利用词法特征，对当前话语是否发生主题变化进行分类，然而这类方法存在两个主要缺点：1. 只着眼于当前话语，而没有充分利用上下文信息；2. 需要手工设计特征，并不断调整合适的阈值，在稀疏性严重的对话文本中往往效果不佳。因此有必要针对对话文本设计更有效的主题分割模型。

近年来的研究表明，基于神经网络的主题分割模型能够取得比传统基于概率的主题分割模型更好的效果[27][28][29]。主题分割神经模型有两种主要架构。第一种是在大小为 k 的窗口中引入上下文，并对当前句子是否为主题片段边界进行分类。第二种是句子序列进行标注以指示主题转换点。以分类为基础进行主题分割的架构通常更重视局部信息，而序列标注架构则可以掌握文档的全局结构。因此，本文将主题分割问题表述为序列标注任务：

- 输入：长度为 M 的场景片段，其中包含话语：$\{S_1, S_2, \cdots, S_M\}$，并由若干个主题片段 $T_1, T_2, ..., T_N$ 组成。每个主题片段包含若干个话语，且与某个主题相关。场景片段是通过上一章的预处理从字幕对话流中获得的片段。
- 输出：相同长度 M 的标签序列：$\{y_1, y_2, ..., y_M\}$，其中 $y_i \in \{0, 1\}$，表示 S_i 是否是新主题片段的开始。

3.2　模型设计

本文的主题分割模型包括两个主要步骤：句子表示和主题分割。句子表示模块用于将句子映射为向量；主题分割模块接收这些向量并检测主题转换边界。其中句子表示模块基于 BERT 将句子映射为其向量表示。

3.2.1　句子表示

预训练语言模型 BERT[39] 为多层双向网络结构，其中每一层都是一个 Transformer 网络。给定一个句子 $[w_1, w_2, ..., w_m]$，E_i 为词 w_i 对应的输入表示，它是通过将 w_i 对应的的原始词向量，段向量和位置向量相加而构成。BERT 提供了一个超参数为 $L = 12$、$H = 768$、$A = 12$ 的中文语言模型，其中 L 是层数（即 transformer 的数量），H 表示隐藏层的大小，A 表示多头注意力机制中的 Self-Attention 的数

量。通过执行“掩码语言模型”和“下一个句子预测”任务，BERT 在大型文本语料库上进行了预训练。中文 BERT 语言模型使用基于字的标记。因此，给定一个包含 N 个单词的句子，BERT 将为每个字符输出大小为 H 的特征向量，并且将整个句子表示为 $N * H$ 的矩阵。在 BERT 的内部，每一层都在前一层的输出上添加自注意力机制，并输出形为 $[N, H]$ 的张量。

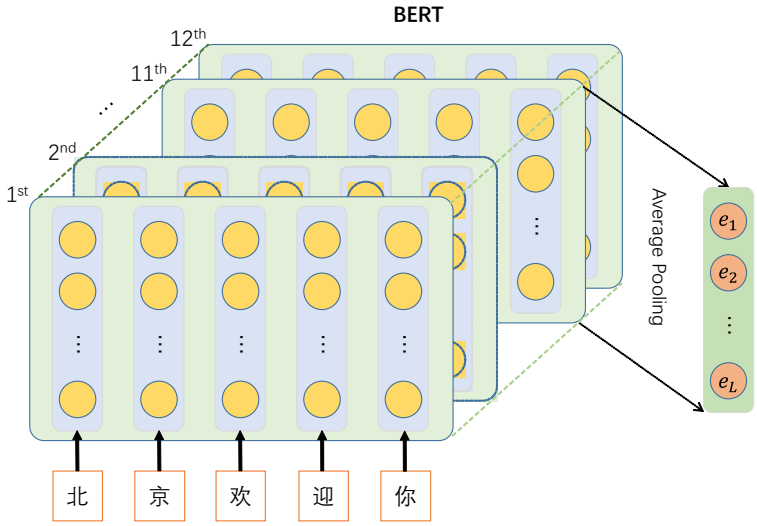


图 3.1　从 BERT 中抽取句子表示向量

像多层 LSTM 一样，网络高层中的权重通常包含与任务相关的信息。Zeiler 等人[40] 指出，网络的顶层包含高层次信息，而底层包含低层次特征，因此，在迁移学习中，高层次特征通常被丢弃，只保留低层次特征并迁移到下游模型。由于 BERT 在预训练中进行了“掩码语言模型”和“下一句预测”两个任务的训练，在 BERT 不同层的网络中，距离最后一层越近，权重就越倾向于预训练任务中的两个目标；距离第一层越近，权重与原始词向量越相似，但包含的高级语义信息却更少。因此，考虑到语义表示能力和计算复杂性，本文从 BERT 的倒数第二层提取输出特征，并对其应用平均池化策略，以生成大小为 [H] 的向量作为输入句子的特征向量。图3.1展示了该过程。

3.2.2　主题分割

序列标注任务通常采用带有 CRF 的双向 LSTM 架构来完成。但在实践中，由于 LSTM 处理远程依赖性能较弱，对于较长的文档，LSTM 并不擅长掌握其全局结构。而且，由于 LSTM 无法并行计算，导致收敛速度较慢。因此，本文选择了 TCN 作为主题分割模块的基础。

TCN 被设计用于序列建模任务。对于大小为 N 的序列数据，TCN 将产生相同

大小的预测序列。TCN 的最显着特征是空洞卷积。它可以确保 TCN 的每个隐藏层都具有与输入序列相同的大小，并且感受野大于具有相同层数的一维 CNN 的感受野。TCN 使用因果卷积来确保时间步长 t 的预测仅依赖于时间步长 $t-1$ 之前的信息，并且不会存在从未来到过去的信息“泄漏”。由于对话流按照时间顺序发展，因此 TCN 的该特性非常适合对话流的主题分割任务。TCN 中还采用了残差卷积，因此可以将底层的要素直接送入顶层，以提高网络的性能。这些属性使得 TCN 可以更好地了解序列的整体结构。另外，与 RNN 相比，TCN 可以并行计算，大大提高了训练和预测速度。

对于句子序列 $S_1, S_2, ..., S_M$，本文使用 BERT 作为编码器来获取它们的语义表示 $E_1, E_2, ..., E_M$。然后将这些句子向量送入 TCN，以输出 0/1 标记序列，其中标记 1 指示主题片段的边界。分割模型的总体架构如图3.2所示。

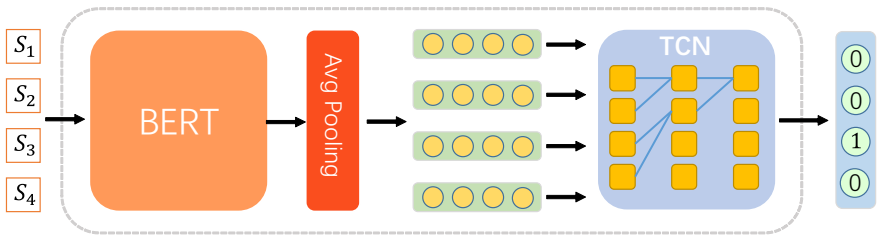


图 3.2　主题分割模型总体架构

由于主题边界的稀疏性，标签序列中 0 和 1 的分布极为不平衡，这导致模型训练过程极易产生偏差。本文使用 Focal Loss[41] 作为损失函数来优化 TCN，以解决类别不平衡的问题。Focal Loss 的计算方法如下：

$$L_{fl} = \begin{cases} -(1-\hat{y})^{\gamma} log\hat{y}, if\ \ y=1 \\ -\hat{y}^{\gamma} log(1-\hat{y}), if\ \ y=0 \end{cases} \tag{3-1}$$

其中 γ 是一个正数，通常经验性地设置为 2。Focal Loss 的出发点是使用不同的罚分来对不同的类别产生不同关注程度。例如，当数据集中的负样本（$y=0$）比正样本（$y=1$）多得多时，模型倾向于将样本分类为负样本（$\hat{y}=0$）。当 $y=0$（负样本）时，$\hat{y}$ 和 $log(1-\hat{y})$ 都较小，因此模型无需对这些样本进行太多调整；当 $y=1$（正样本）时，$(1-\hat{y})^{\gamma}$ 和 $log\hat{y}$ 都很大，模型需要对这些样本进行很多调整。因此，原本数量较少的正样本对模型的影响反而较大，从而有效地解决了类别不平衡问题。

3.3　实验结果与分析

本小节将介绍评估本节主题模型所用的数据集，评估指标，实验结果和分析，以及对现有模型的几种改进措施。

3.3.1　数据准备

本节设计实验评测所提出的主题分割模型在 Sub 数据集上的性能。此外，为了更全面地评估模型在不同类型的文本上的分割性能，并作为字幕对话文本主题分割的参照，本文还选用数据集 Weibo，DAct[36] 进行了实验。其中，Weibo 是从社交媒体收集的新闻短讯，DAct 是两人对话文本，而 Sub 为多人对话文本。通过上一章的预处理，Sub 数据集已经标注了说话人和场景标签，可由人工参考这些标签在对话流中标注主题边界；DAct 数据集则已由人工标注了话题线索和主题边界；对于 Weibo 数据集，本文假设每个微博新闻都是关于某一主题的，因此一条新闻可以被视为一个主题片段。由此 Sub，DAct 和 Weibo 都具有了原始主题片段，本文假设连续的两个主题片段之间存在着主题转换。

由于原始 Sub 数据集的规模较小，因此本文使用简单的扩充策略来自动扩展训练数据集：随机选择主题片段并将其拼接起来以形成新的场景片段，其中拼接点即是主题转换点。使用这种扩充策略，可以自动生成大量的训练数据。表3.1列出了扩充后的数据集的基本统计信息。（简明起见，扩充后的数据集仍采用与原数据集相同的名称，在不致混淆的情况下，此后本章中 Sub 均指扩充后的数据集，其余类似。）

表 3.1　主题分割的三个数据集基本统计

	Weibo	Sub	DAct
Num of scene segments (trainset)	20000	20000	20000
Num of scene segments (testset)	4000	4000	4000
Mean of scene segments’ lengths	12.97	20	26.08
Std of scene segments’ lengths	3.9	5	10
Mean of utterances’ lengths	23.75	9.39	10.24
Std of utterances’ lengths	11.99	4.64	4.94

将每个数据集扩充为 24000 个场景段，并按照 5：1 的比例分为训练集和测试集。从表3.1可以看出，Weibo 中的平均话语长度大于 Sub 和 DAct 中的平均话语长度，这表明 Weibo 中的句子比 Sub 和 DAct 中包含的单词更多。这种差异可能会对分割结果产生影响，因为长句子可能包含比短句子更多的与主题相关的单词。

图3.3显示了从 Sub 数据集中随机选择的一个片段。“id”列是每个话语的索引。如果当前说话人与前一个说话人不同，则“spkr”列设置为 1，否则设置为 0。例如，id 为 2 的话语的“spkr”标签为 1，这表示该话语的说话人与 id 为 1 的话语的说话人不同；id 为 10 的话语的“spkr”标签为 0，表示此话语与 ID 为 9 的说话人相同。“content”列展示了话语内容，“ref”列为参考标签，用于指示对应话语是否是主题转换点。例如，id 为 0-8 的第一段与医学有关，而紧接着 id 为 9-14 的片段与侦探有关，并且这两个段之间存在明显的主题转换，因此 id 为 9 处的“ref”标签设置为 1。“pred”列是模型的预测结果（此列未出现在数据集中）。

idx	spkr	content	label	pred
0	1	他还活着吗?	0	0
1	1	是的	0	0
2	1	给他用肝素，静脉推注免疫球蛋白	0	0
3	1	治疗恶性萎缩性丘疹?	0	0
4	1	心脏停搏并不是那种恶心的停搏	0	0
5	0	是冠脉病变，冠状动脉是大血管	0	0
6	0	也就是说，这不是恶性萎缩性丘疹	0	0
7	1	但是组织活检证实...	0	0
8	1	亚瑟·柯南·道尔也曾是秘密会员	1	1
9	1	而且，只是传说	0	0
10	0	是老私家侦探告诉菜鸟的童话	0	0
11	0	好骗他们干苦活儿	0	0
12	1	你就是吃不着葡萄说葡萄酸	0	0
13	1	好吧，不管真假	0	0
14	1	那个酒吧认识的男的呢	1	1
15	0	你给他电话号码的那一个	0	1
16	1	你是怎么知道的	0	0
17	1	因为他打来找你	0	0
18	0	所以别跟我说，你只是亲了一次男同事	0	0
19	0	因为你又在酒吧，又在阳台上	0	0
20	0	已经一个多月了	0	0
21	0	你都没有礼貌性和我说一声	0	0
22	1	为什么我没有拿到留言	0	0
23	1	但我总觉得有些尴尬	1	0
24	1	别这么想	0	1
25	0	这身是奶奶的茶壶套吗	0	0

图 3.3　Sub 数据集切分结果片段

3.3.2　评测指标

本文采用 F1 值和 WinDiff[42] 两个指标来评估所提出的主题分割模型在话语级别上的分割准确度。此外设计了“span”指标来衡量模型在话语区间级别的分割效果。

1）F_1 和 WinDiff：设 R 为参考分割，F_1 得分仅着重于分割点，其定义为 $\frac{2*p*r}{p+r}$，其中 p 是标签 1 的预测准确率，定义为预测标签 1 中真实标签 1 的比例，r 是标签 1 的召回率，定义为参考标签 1 中预测标签 1 的比例。WinDiff 引入了大小为 k 的滑动窗口，以将预测的分割 H 与 R 进行比较，其中 k 通常设置为 R 中平均分割长度的一半。

WinDiff 的定义为：

$$WinDiff = \frac{1}{N-k}\sum_{i=0}^{N-k}\left(\left|R_{i,i+k} - C_{i,i+k}\right| \neq 0\right)$$

其中 $R_{i,i+k}$ 是窗口中从位置 i 到 $i+k$ 的参考边界的数量，$C_{i,i+k}$ 是同一窗口中的预测边界的数量。作为测量分段错误的概率度量，WinDiff 的值在 0 到 1 之间。WinDiff 的值越小，分割 H 越接近分割 R。当它们彼此相同时，WinDiff 等于 0。

2）span：为了衡量模型在区间级别的分割效果，本文引入了 span 评估指标。

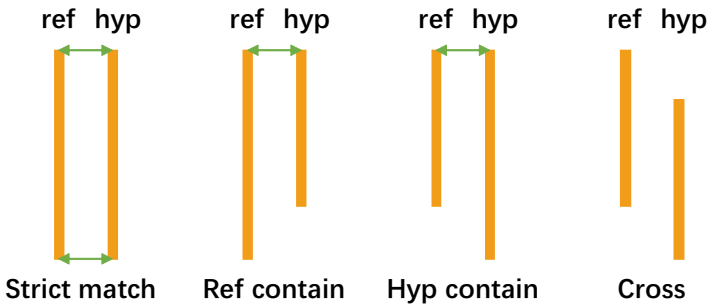


图 3.4　Span 指标的四种情形

如图3.4所示，span 指标在相同位置考察由参考分割（“ref”）和预测的分割（“hyp”）生成的两个分段。共有四种情况：

- 严格匹配：ref 和 hyp 的起点和终点完全相同；
- Ref 包含：ref 和 hyp 只有一个分割点，hyp 较短，因此它包含在 ref 中，这表明模型将某个位置归类为主题转化点，而实际上却没有；
- Hyp 包含：ref 和 hyp 只有一个相同的分割点，并且 hyp 较长，因此包含 ref，这表明模型错过了分割点；
- 交叉：ref 和 hyp 的起点和终点都没有不同，但是有交叉部分，这是一个严重的分割错误。

Span 指标由结果中这四个类别的比例确定。

3.3.3　实验结果

本文在三个数据集上对所提出的模型与 BiLSTM[29] 进行了比较。根据表3.2，本文提出的模型在三个数据集上均取得了最佳的结果。与 BiLSTM 相比，Weibo，DAct 和 Sub 数据集的 F1 分数分别增加了 0.08、0.135 和 0.166。WinDiff 有一定程度的减少。

表 3.2　模型在三个数据集上的分割后的 F1 和 Windiff 评测

	Weibo		DAct		Sub	
	WinDiff	F1	WinDiff	F1	WinDiff	F1
2L BiLSTM	0.2770	0.82	0.2962	0.675	0.3237	0.544
BERT+TCN	**0.1267**	**0.90**	**0.1957**	**0.81**	**0.2408**	**0.71**

另一方面，本文所提出的模型在这三个数据集上的表现有非常大的差异。Weibo 数据上的 F1 得分达到 0.9，而 Sub 数据上的 F1 得分仅为 0.71，这表明主题分割的难度随文档类型不同而变化。由于 Weibo 数据集由新闻文本组成，新闻文本由与主题相关的长句和词汇组成。相比之下，Sub 数据集为对话文本，长度较短且所含无意义词较多，因此，在对话文本中检测主题片段边界更加困难。

表 3.3　模型在三个数据集上分割的 Span 得分

	Stric	Ref Contain	Hyp Contain	Cross
Weibo	0.7999	0.0979	0.1005	0.0015
DAct	0.6438	0.1549	0.1967	0.0046
Sub	0.4729	0.2757	0.2383	0.0130

表3.3显示了模型在三个数据集上主题分割的 span 得分。可以看出，情形 I 的比例与模型的 F1 得分呈正相关。这表明 Span 指标与 F1 分数兼容并且定义明确，情形 I 的比例较高意味着 hyp 和 ref 分段之间的距离很近，而情形 IV（交叉）的比例高则意味着模型在分段级别上的性能很差。

可以看到，情形 I 的比例在各数据集上的趋势与 F1 分数的趋势一致，在 Weibo 上最高，在 Sub 上最低。此外，在每个数据集上 Ref 包含和 Hyp 包含的比率非常接近，这表明模型的分割误差分布是均匀的，过度分割和过度保守的倾向性很小。

3.3.4　模型改进

Sub 和 DAct 数据集包含说话人信息，说话人信息可能对分割结果有一定影响。本节将利用说话人信息改进主题分割的效果。图3.3中，模型将两个连续的话语（id

为 14 和 15）预测为分割点，而实际上应该只有一个（在 id 为 14 的话语之前）。如果可以滤除此类错误，则分割结果应该更加准确。因此，本文引入了说话人标签，并在 TCN 输出之后添加了 CRF 层。如果当前说话的说话人与前一个说话人不同，则将说话人标签设置为 1，否则将其设置为 0。

从表3.4中可以看出，在引入说话人信息之后，Sub 和 DAct 上的 F1 分数分别提高了 10% 和 5%，准确度和召回率都得到了极大的提高。表3.5展示了 Sub 测试集中参考分割和说话人标签组合的统计信息。符号 r0s1 表示参考分割标签和说话人标签分别为 0 和 1 的情形的统计次数。可以看出，r1s0 的比例相对较小（439/80000），这意味着如果说话人不切换，主题转换通常不太可能发生。

表 3.4　改进后的模型在三个数据集上的评测结果

	DAct			Sub		
	F1	Precision	Recall	F1	Precision	Recall
baseline	0.81	0.85	0.77	0.71	0.73	0.69
+ Speaker	**0.86**	0.9	**0.82**	0.81	0.82	**0.81**
+ CRF	0.82	0.87	0.77	0.75	0.81	0.69
+ Speaker + CRF	**0.86**	**0.91**	**0.82**	**0.82**	**0.84**	**0.81**

表 3.5　参考分割和说话人标签组合统计

r0s0	r0s1	r1s0	r1s1
36992	30910	439	11659

表3.6比较了添加说话人信息之前和之后的分割结果。符号 r0p1s0 表示话语的参考标签、预测标签和说话人标签分别为 0,1,0，它表示参考标签显示主题转换没有发生（0），而模型则给出了存在主题转换的预测（1），当前说话人与上一个说话人相同（0）。

表 3.6　说话人信息的引入对分割结果的影响

	reference ≠ prediction				reference=prediction			
	r0p1s0	r0p1s1	r1p0s0	r1p0s1	r0p0s0	r0p0s1	r1p1s0	r1p1s1
baseline	1584	1532	156	3633	35408	29378	283	8026
+ Speaker	4	2135	436	1899	36988	28775	3	9760
diff	-1580	+603	+280	-1734	+1580	-603	-280	+1734

观察表3.6的差值行（diff），其中正的差值反映了 $ref \neq pred$（或 $ref = pred$）

情况下模型的性能降低（或改进），相反，负的差值则反映了“$ref \neq pred$”（或“$ref = pred$”）情况下模型的性能提高（或降低）。还可以观察到，当且仅当 rxpisj 中的 $i = j$ 时，对应的列才具有正的差值；而当且仅当 rxpisj 中的 $i \neq j$ 时，列具有负的差值。这意味着模型的预测与说话人标签具有相同的趋势。例如，在引入说话人信息之后，r0p1s0 从 1584 降低到 4，这意味着如果说话人标签显示没有说话人转换，则主题边界的预测频数会降低 1580；r0p1s1 从 1532 增加到 2135，这意味着如果说话人标签显示有说话人切换（尽管没有进行主题转换），则模型预测发生主题转换的次数增加了 603。这表明说话人切换的信息可以提高分割的可能性，而说话人保持的信息可以降低分割的可能性。因此，说话人信息实际上形成了分割点位置的约束。

id	spkr	content	ref
10	Director	非常不错大伙们 Very nice	0
11	Castle	有没有什么特殊的人 Was there anyone special in his life?	1
12	Castle	他倒想呢但是没有 Oh, he wished, but no.	0
13	Castle	将其放大测试 have it amplified, tested.	1
14	Castle	而且我肯定你们的证据存储科 And I’m sure your evidence storage	0
15	Castle	完好的保留了你的海豚 has kept your dolphin well preserved.	0
16	Chief	泰迪 Teddy	0

图 3.5　错误分割示例

这种约束可能会带来一些意外的结果。例如，r1p1s0 从 283 减少到 3，表明此时并没有发生说话人转换，该模型在没有说话人标签的情况下能够正确预测主题转换，而在引入说话人标签的情况下却未能检测到这些主题转换。这种奇怪的现象是由于数据构造过程中产生的错误所致。图3.5显示了 Sub 的一个片段，从 id 11 到 id 15 的所有话语都由角色“Castle”说出，但是，id 为 11 到 12 的所有话语来自一个主题片段，而 id 为 13 到 15 的所有话语则来自另一个场景。因此，尽管这些话语属于同一说话人，但它们并非来自同一场景，并且确实存在话题转换。因此，由于说话人标签显示并未发生说话人转换，模型会错误地判断主题边界，从而导致性能下降。这种错误在整个数据集中只占很小的比例（接近 280/20000 = 1.4%）。

由于在说话人切换时更有可能发生主题转换，因此引入说话人信息能够提高主题分割的准确度。

3.4　本章小结

本章介绍了所提出的主题分割模型，针对字幕对话文本稀疏性严重的问题，采用了 BERT 进行句子的语义表示，并在序列标注架构下利用 TCN 进行主题分割。实验结果显示了该模型能够有效地对对话文本进行主题分割。采用人工对字幕进行主题片段的标注成本过高，而且相比于能够找到对应剧本并进行标注的 Sub 数据集，更多字幕并没有可以用以参照的剧本，而利用本章所提出的主题分割模型则可以将大量未标注的字幕数据分割为主题片段，并能基于这些片段进行主题建模。本章的研究结果能提供了大量分割合理的主题片段，为进一步开展字幕对话文本的主题建模研究提供了条件。

第 4 章　对话文本主题建模

4.1　本章概述

基于上一章的主题分割模型，大量的未标注字幕可以被分割为主题片段。这些主题片段进而可被用于训练主题模型，进行对话文本的主题建模。主题模型以无监督的方式挖掘出隐藏在文本中的潜在主题信息，适合于从海量非结构化文本中抽取结构化信息。主题模型从提出到今天已发展了近二十年，但相关领域的研究仍然方兴未艾，新的主题模型和算法不断涌现。近年来深度学习的迅猛发展，也为主题模型的研究注入了新的活力，促成了许多基于神经网络的主题模型的诞生。

给定由若干文档组成的文档集 D，设该文档集中的文档数为 $|D|$；记组成文档集的词汇表为 W，设其中含有的总的单词数为 $|W|$。主题模型假设每个文档包含若干主题，每个主题在每篇文档中有各自的概率大小，形成文档-主题分布；词表中的每个单词在各主题中也有各自的概率大小，形成主题-词分布。对于每一个文档 $d(d \in D)$，主题模型的目标是给出 d 所对应的主题的概率分布，并以概率分布最高的一组单词来表示每个主题。在给定了主题数量 K 之后，主题模型利用单词在文档中的共现关系，推断得到文档的主题分布矩阵 θ（大小为 $|D| * K$），以及主题的词分布矩阵 ϕ（大小为 $K * |W|$）。在对话文本的场景下，每个话语往往只包含很少的单词，而词汇表却具有很大的长度，因此话语的词袋表示中只有极少的非零项，稀疏性非常严重，给主题建模带来了巨大的挑战[43]。

绪论中 1.2.2 节介绍了神经主题模型的发展脉络，其中 NTM-GSM、W-LDA 分别是目前在以高斯分布和以 Dirichlet 分布为先验分布的主题模型中，主题一致性最高的模型，ETM 则将主题向量并入词向量空间，在稀疏文本中有较好的表现，因此本章重点介绍上述几种神经主题模型，并针对对话文本的特点，在综合分析了已有的神经主题模型特点的基础上，具体介绍了本章的模型设计——基于高斯混合的神经主题模型（GMNTM）。本章最后通过一组实验对比了不同神经主题模型的性能差异，并展示和分析了实验结果。

4.2　主题模型代表算法

4.2.1　LDA

隐狄利克雷分布 LDA[44] 是基于统计推断的经典主题模型，在许多长文本场景的主题建模任务中均有不错的表现，是后续许多主题模型改进和扩展的基础。LDA

将文档的生成过程建模为两个分布：“文档-主题”分布 $p(z|d)$ 和“主题-词”分布 $p(w|z)$，并假设这两个分布分别服从以 $\vec{\alpha}$ 和 $\vec{\beta}$ 为先验概率超参数的 Dirichlet 分布。

LDA 所建模的文档生成过程可概括为：

Algorithm 2 LDA process

Input: documents D **Parameter**: $\vec{\alpha}, \vec{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}$

for $k \in [1, K]$ **do**
采样对应的词分布 $\vec{\phi}_k \sim Dirichlet(\vec{\beta})$
end for
for $d \in D$ **do**
采样对应的主题分布 $\vec{\theta}_d \sim Dirichlet(\vec{\alpha})$
for $n \in [1, N_d]$，N_d 为 d 中词数 **do**
从对应主题分布中采样词的主题 $z_{d,n} \sim Multi(\vec{\theta}_d)$
从对应词分布中采样当前词 $w_{d,n} \sim Multi(\vec{\phi}_{z_{d,n}})$
end for
end for

LDA 通过上述方式重复采样，直至生成一个完整的文档。

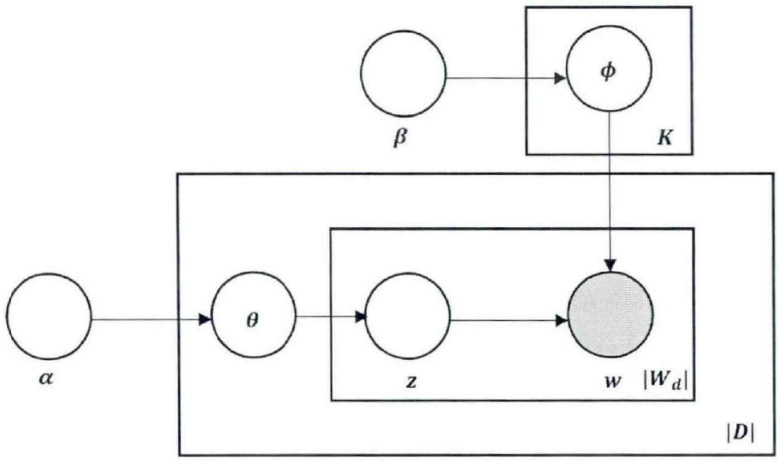


图 4.1　LDA 概率图模型

LDA 的概率图模型如图4.1所示，其中白色圆圈为隐变量，阴影圆圈为可观测变量 w，每个矩形右下角的数字表示矩形内对应的过程需要重复的次数。隐变量中 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$ 为需要学习的参数，分别对应“文档-主题”分布和“主题-词”分布，可通过吉布斯采样（Gibbs Sampling）进行求解。当文档和单词给定后，主题分布的

条件概率计算公式为：

$$p\left(z|z_{-i},w_i,d,\alpha,\beta\right)\propto\frac{n_{-w_i,z|d}+\alpha_z}{\sum_{k=1}^{K}n_{-w_i\cdot z_k|d}+\alpha_{z_k}}\frac{n_{-w_i,w|z}+\beta_i}{\sum_{j=1}^{|W|}n_{-w_j,w|z}+\beta_j} \tag{4-1}$$

其中，$n_{-w_i,w|z}$ 表示主题 z 中，除去当前样本词 w_i 外，单词 w 出现的次数；$n_{-w_i,z|d}$ 表示文档 d 中，除去当前样本词 w_i 外，属于主题 z 的单词数。初始时，算法给各个词随机分配一个主题，之后迭代一定轮次，当吉布斯算法收敛后，即可得到每一文档对应的主题分布 $\vec{\theta}_d$ 以及每一主题对应的词分布 $\vec{\phi}_k$：

$$\begin{aligned}\theta_{z_j}^{d}&=\frac{n_{z_j|d}+\alpha}{\sum_{k=1}^{K}n_{z_k|d}+K\alpha}\\ \phi_i^{z_j}&=\frac{n_{w_i|z_j}+\beta}{\sum_{w}n_{w|z_j}+|W|\beta}\end{aligned} \tag{4-2}$$

在新闻、独白等长文本任务中，LDA 一般能取得较好的效果，但在对话文本任务中，由于文本长度非常短，其词袋表示存在严重的稀疏性问题，LDA 的效果会显著变差，因此 LDA 及以 LDA 为基础进行改进的模型通常不适用于对话文本中的主题建模任务。

4.2.2　NTM-GSM

NTM-GSM 是基于标准变分自编码器的神经主题模型，由 Miao 等人[32] 提出。变分自编码器的目标是对隐变量 z 的真实后验分布 $p(z|x)$ 进行建模，通过贝叶斯法则（Bayes law）可以将后验概率由似然 $p(x|z)$，先验分布 $p(z)$ 和 x 的边际分布 $p(x)$ 表示为：

$$p(z|x)=\frac{p(x|z)p(z)}{p(x)} \tag{4-3}$$

其中分母原则上可通过式（4-4）

$$p(x)=\int p(z)p(x|z)\mathrm{d}z \tag{4-4}$$

来进行计算。然而由于式（4-4）中积分需要遍历 z 的所有取值，在高维空间中往往难以计算。因此，一般不直接求解真实后验分布 $p(z|x)$，而是求解变分后验分布 $q_\phi(z|x)$（ϕ 为变分参数），并不断缩小 $q_\phi(z|x)$ 与 $p(z|x)$ 之间的差异来达到逼近 $p(z|x)$ 的目的。$q_\phi(x)$ 通常选择易于计算的分布族，如高斯分布。本质上，这种方法是将推断问题转换为优化问题，通过最小化变分分布 $q_\phi(z|x)$ 和真实分布 $p(z|x)$ 之间的 KL 散度来求解 $p(z|x)$。分布 $q_\phi(z|x)$ 和 $p(z|x)$ 之间的 KL 散度定义为：

$$D_{\mathrm{KL}}\left[q_\phi(z|x)\|p(z|x)\right]=\sum_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z|x)}=\mathbb{E}_{q_\phi(z|x)}\left[\log\frac{q_\phi(z|x)}{p(z|x)}\right] \tag{4-5}$$

将式（4-5）中后验分布利用贝叶斯法则替换后得到：

$$\begin{aligned}D_{\mathrm{KL}}\left[q_\phi(z|x)\|p(z|x)\right]&=\mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x)-\log\frac{p(x|z)p(z)}{p(x)}\right]\\&=\mathbb{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x)-\log p(x|z)-\log p(z)+\log p(x)\right]\\&=\log p(x)-\underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p(x,z)-\log q_\phi(z|x)\right]}_{\mathrm{ELBO}(\phi)}\end{aligned} \tag{4-6}$$

式（4-6）中右侧括弧中部分为 $logp(x)$ 的变分下界，记为 ELBO（Evidence Lower BOund）。在给定数据 x 之后，x 的分布 $p(x)$ 可视作常数，因此最小化式（4-6）左侧的 KL 散度的优化目标，等价于最大化 ELBO。使用插项的技巧，式（4-6）中 ELBO 可重写为：

$$\begin{aligned}\mathrm{ELBO}(\phi)&=-\mathbb{E}\left[\log p(x,z)-p(z)+p(z)-\log q_\phi(z|x)\right]\\&=\mathbb{E}[\log p(x|z)]-D_{\mathrm{KL}}\left[q_\phi(z|x)\|p(z)\right]\end{aligned} \tag{4-7}$$

因此，最大化 ELBO 等价于最大化式（4-7）右侧，其中第一项为似然函数，将迫使解码器将生成样本 x' 尽可能还原为输入样本 x，通常采用交叉熵进行度量；第二项为关于 z 的分布的正则项，将迫使变分后验分布 $q_\phi(z|x)$ 逼近先验分布 $p(z)$。

在实践中，式（4-7）中的后验分布 $q_\phi(z|x)$ 与先验分布 $p(z)$ 都需要确定为具体的分布才能进行优化，常用的假设是将这两个分布都选定为多元高斯分布，其中后验分布 $q_\phi(z|x)$ 的均值和方差分别假定为 $\mu(x)$ 和 $\Sigma(x)$，并假定其协方差阵为对角阵，先验分布 $p(z)$ 则通常取标准正态分布 $\mathcal{N}(0,1)$。因此，实际的优化目标 $D_{\mathrm{KL}}\left[q_\phi(z|x)\|p(z|x)\right]$ 为：

$$\begin{aligned}D_{\mathrm{KL}}\left[q_\phi(z|x)\|p(z)\right]&=D_{\mathrm{KL}}[\mathcal{N}(\mu(x),\Sigma(x))\|\mathcal{N}(0,1)]\\&=\frac{1}{2}\left(\mathrm{tr}(\Sigma(x))+\mu(x)^T\mu(x)-d-\log\det(\Sigma(x))\right)\end{aligned} \tag{4-8}$$

其中，d 为 z 的维数，$tr(.)$ 为迹运算。图4.2展示了 VAE 的架构，其中，$q_\phi(z|x)$ 作为编码器，将数据 x 映射为隐变量 z 的分布的均值 $\mu(x)$ 和方差 $\sigma(x)$，从该分布中采样得到 $z\sim\mathcal{N}(\mu(x),\Sigma(x))$，$p_\rho(x|z)$ 则用作解码器，通过隐变量生成样本 x'，分布中的参数 ϕ 与 ρ 分别对应编码器和解码器网络中的权重参数。在对隐变量 z 的采样操作中，直接的采样操作并不可导，网络参数难以更新，为此，VAE 中采用

了重参数[45] 的技巧，取 $z=\mu+\epsilon*\sigma$，$\epsilon\sim N(0,I)$，则仍有 $z\sim N(\mu,\sigma)$，在保证分布不变的同时也满足了 z 对 μ 和 σ 的可导性。

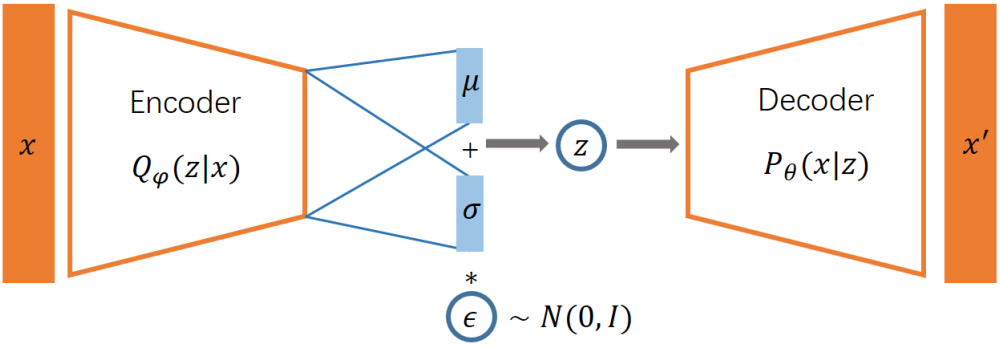


图 4.2　VAE 的网络架构

主题模型 NTM-GSM（Gaussian-Softmax）采用了上述标准 VAE 的架构，以文档的词袋表示 BOW（Bag Of Word）作为输入。考虑到分布需满足规范性，因此从隐空间采样得到高斯变量 z 后，还需要将 z 归一化才能作为主题分布，NTM-GSM 采取的方法是使 z 通过 Softmax 层，即

$$\begin{aligned} z &\sim \mathcal{N}\left(\mu(x),\sigma(x)^2\right) \\ \theta &= \text{Softmax}\left(W_1^T z\right) \end{aligned} \tag{4-9}$$

其中 W_1 为 $L*K$ 的矩阵，L 为 z 的维数，K 为主题数。由此求得的归一化向量 θ（K 维）作为文档的主题分布向量，在导入解码器 $P(x|z)$ 后得到重构文档 x'。解码器中的权重参数 ρ_{K*V} 即为主题-词分布矩阵，令 θ 取第 k 维为 1 的 One-hot 向量并导入解码器，即可得到第 k 个主题的主题-词分布。

4.2.3　W-LDA

为了解决 VAE 的后验坍塌问题，同时能够利用 Dirichlet 分布作为隐空间的先验分布，Tolstikhin 等人[46] 提出基于 WAE（Wasserstein Auto-Encoder）构建主题模型。

WAE 基于 Wasserstein 距离来度量先验分布和后验分布的差异。相比于 KL 散度，Wassertein 距离在两个分布没有重叠时仍然能保持连续并反映两个分布的远近。分布 P_x 与 P_y 的 Wasserstein 距离定义为：

$$W\left(P_x,P_y\right)=\inf_{\gamma\in\Pi\left(P_x,P_y\right)}\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|] \tag{4-10}$$

式（4-10）中 $\Pi\left(P_x, P_y\right)$ 为 (x, y) 的联合分布的集合，满足 $x \sim P_x, y \sim P_y$，对于其中 x 与 y 的某个联合分布 γ，可以求得 γ 下所有 x 与 y 的距离的期望，所有期望的下确界（infimum）即定义为 P_x 和 P_y 的 Wasserstein 距离。

作为 Wasserstein 的定义式，式（4-10）难以直接用于计算，Tolstikhin 等人[46]通过推导，提出可采用式（4-11）作为实践中 WAE 的优化目标，即：

$$D_{WAE}\left(P_x, P_G\right) := \inf_{Q(z|x)\in Q} \mathbb{E}_{P_x}\mathbb{E}_{Q(z|x)}[\|x - G(z)\|] + \lambda \cdot \mathcal{D}_z\left(Q_z, P_z\right) \tag{4-11}$$

其中，G 为解码器，Q_z 为经编码器 Q 映射后的边缘分布，与 VAE 不同的是，WAE 可以使用确定性的编码器 $P(z|x)$，而不需通过采样的方法得到隐变量 z，因为不需为每一个样本 x 在隐空间中求得对应的一个分布，只需要 z 的边缘分布接近先验分布即可。G 将隐变量映射为生成样本 x'，因此式（4-11）第一项为重构误差，度量了生成样本与输入样本之间的差异，第二项 $D_z(Q_z, P_z)$ 则选用 MMD（Maximum Mean Discrepancy）度量先验分布和变分分布的差异。

WAE 的整体算法为：

Algorithm 3 WAE-MMD process

Parameter: 编码器 Q 参数 ϕ, 解码器 G 参数 θ, 正定核函数 k

1: **while** 停止条件不满足 **do**

2: 从训练集中采样 $\{x_1, x_2, ..., x_n\}$

3: 从先验分布 P_z 中采样 $\{z_1, z_2, ..., z_n\}$

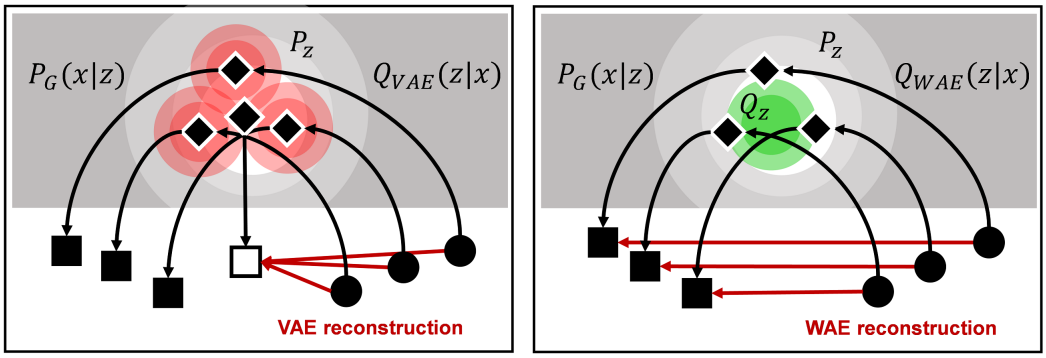
4: 从后验分布 $Q_\phi(z|x_i)$ 中采样 $\tilde{z}_i, (i = 1, 2, ..., n)$

5: 通过最小化式（4-12）来更新 Q_ϕ 和 G_θ：

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\|x_i, G_\theta\left(\tilde{z}_i\right)\| + \frac{\lambda}{n(n-1)}\sum_{\ell\neq j}k\left(z_\ell, z_j\right)\\&+\frac{\lambda}{n(n-1)}\sum_{\ell\neq j}k\left(\tilde{z}_\ell, \tilde{z}_j\right) - \frac{2\lambda}{n^2}\sum_{\ell,j}k\left(z_\ell, \tilde{z}_j\right)\end{aligned} \tag{4-12}$$

6: **end while**

式（4-12）中后三项为 MMD 的离散形式，在分别从先验分布和后验分布中采样后，通过样本的 MMD 值来估计分布 Q_z 与 P_z 之间的实际 MMD 值。



(a) Architecture of VAE　　(b) Architecture of WAE

图 4.3　VAE 与 WAE 的差异对比

图4.3对比了 VAE 与 WAE 的不同点。WAE 与 VAE 的优化目标都由重构误差，以及隐空间中先验分布 P_z 及变分后验分布 $Q(z|x)$ 间的差异构成的正则项组成。在 VAE 中（图4.3(a)），每个输入样本对应一个分布（图4.3(a) 中每一个红色的小圆圈），VAE 的优化目标是使这些分布都趋近标准高斯分布（图4.3(a) 中白色的大圆圈所示），造成的结果是各个分布之间出现层叠，在层叠区域，一个隐变量 z 会对应多个输入样本，使得重构样本实际上是这些输入样本的平均值，正是由于这个原因，VAE 所生成的图像容易出现模糊；在 WAE 中，每个输入样本对应的是一个隐变量 z 而非一个分布，WAE 的优化目标是使后验的边际分布 $Q(z) = \int P(z|x)dP_x$（图4.3(b) 中绿色圆圈）趋近先验分布 P_Z（图4.3(b) 中白色大圆圈所示），各输入样本对应的 z 在隐空间中可以合理分布，避免了分布层叠的问题，可以生成更加清晰的图像。

基于 WAE 构建的主题模型 W-LDA，以文档的词袋表示 BOW 作为输入，编码器由多层感知机构成，并经 Softmax 层生成隐变量 θ 作为主题分布。与 NTM-GSM 中不同，W-LDA 中 θ 由确定性映射得到，无需经过采样操作。

与 LDA 一致，W-LDA 主题模型采用 Dirichlet 分布作为主题的先验分布；由于 θ 需满足归一化约束，即 θ 的可行解空间构成单纯形，因此要求核函数在单纯形上有较好的度量意义，W-LDA 选取了信息扩散核作为 MMD 的核函数：

$$\mathbf{k}\left(\theta,\theta'\right)=\exp\left(-\arccos^2\left(\sum_{i=1}^{d}\sqrt{\theta_i\theta_i'}\right)\right) \tag{4-13}$$

该核函数度量了将不同向量映射到球面后的像之间的测地线长，相较于常见的 L^2 范数，该核函数对位于单纯形边界处的点更为灵敏，因此更适合于数据稀疏的场景。与 NTM-GSM 相同，解码器的权重参数 β 构成了主题-词分布，因此，令 θ 取遍不同的 One-hot 向量并导入解码器，即可得到对应主题的主题-词分布。

4.2.4　ETM

现有主题模型通常通过主题-词分布来描述主题，为了更细致地刻画所挖掘的主题的可解释性，Dieng 等人[35] 提出了 ETM（Embedded Topic Model），在神经主题模型中引入了词向量和同样维数的主题向量，使主题与词可在同一个嵌入空间进行表示。

ETM 中定义了一个词向量矩阵 ρ，大小为 $L*V$，其中 L 是词向量的维数，V 是词汇表的大小；同时定义了一个主题向量矩阵 α，大小为 $L*K$，其中 K 为主题数。与 NTM-GSM 类似，在 ETM 中，文档的词袋表示 BOW 由编码器映射到隐空间的高斯分布，经采样得到隐变量 z，经过 Softmax 归一化后得到主题分布，即 $\theta = Softmax(Wz + b), z \sim N(\mu(x), \sigma(x))$；主题-词分布反映的是每个词在各主题中的重要程度，ETM 中将第 k 个主题的主题向量与各个词向量进行内积后，再经 Softmax 归一化的结果作为其主题-词分布 β_k，即：$\beta_k = Softmax(\rho^T \alpha_k)$。

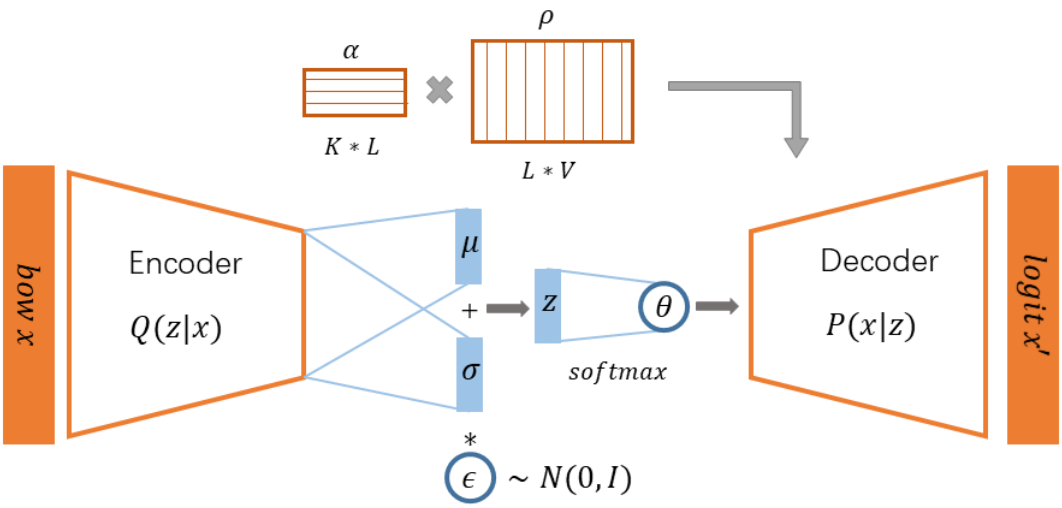


图 4.4　ETM 的网络架构

ETM 的网络架构如图4.4所示。对比图4.2，ETM 与 NTM-GSM 最主要的区别在于，ETM 中主题-词分布矩阵 β_{K*V} 被分解为主题向量矩阵 α_{K*L} 与词向量矩阵 ρ_{L*V}，词向量的训练与主题模型的训练同步进行，使得主题模型能根据词嵌入表示的调整逐步完善。ETM 的目标函数与 NTM-GSM 相同，由重构误差及隐变量的后验分布同先验分布的 KL 散度构成。

ETM 的算法整体流程为：

Algorithm 4 ETM process

Parameter: 主题向量矩阵 α，词向量矩阵 ρ

for $epoch = 1, 2, ..., N$ **do**

对每一主题 k，计算主题-词分布 $\beta_k = Softmax(\rho^T \alpha_k)$

从文档集中选择一个 minibatch **B**

for $d \in$**B do**

计算 d 的归一化词袋表示 x_d

计算均值向量和对数方差向量 $\mu_d = \mu(x_d))$，$log\sigma_d = log\sigma(x_d)$

采样隐变量 $z \sim N(\mu_d, \Sigma_d)$，计算主题分布 $\theta_d = Softmax(Wz + b)$

计算重构样本的生成概率 $p(x^{'}|\theta_d) = \theta_d^T \beta$

end for

计算 ELBO 及其导数，更新 $\alpha_{1:K}$，ρ 及网络参数

end for

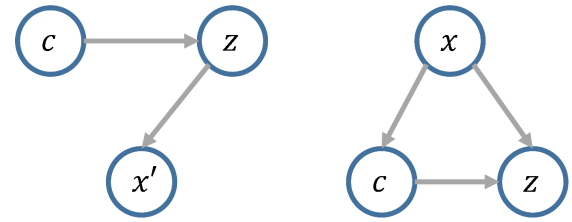
4.3　基于高斯混合先验的神经主题模型 GMNTM

4.3.1　GMNTM

在 VAE 中，尽管先验分布的多元高斯分布可以简化计算，但这种单峰的假设也对隐空间形成限制，同时由于 VAE 将各输入样本对应的隐变量分布都逼近标准高斯分布，这些高度层叠的分布可能会降低模型的性能。从主题模型的角度看，相同主题的文档对应的隐变量应该具有较近的距离，而不同主题的文档则应该被映射到隐空间中相距较远的位置，即相似主题的文档应该在隐空间聚成簇，而不同主题的文档应分布在不同的簇中。而 VAE 将所有分布都逼近单一簇的特性，将使得不同的主题对应的隐变量有混杂的趋势，可能导致不同主题的文档不易区分开，进而造成得到的主题在主题多样性和主题连贯性上都不够理想。

在对话文本中，由于话语长度较短，每个话语所包含的主题较为单一，因此对于属于同一主题的话语，以单峰的高斯分布作为先验较合适；属于不同主题的话语，其对应的主题的先验分布则应有好的区分性，尽可能减少分布层叠的情形。

因此，本文提出用高斯混合分布作为主题隐变量的先验分布，以替代 VAE 中以标准高斯分布为先验分布的假设，并基于高斯混合变分自编码器（GMVAE[47]）来设计主题模型。



(a) *Generative Process*　　(b) *Inference Process*

图 4.5　GMVAE 的概率图模型

GMVAE 引入了离散的类别隐变量 c，用以指示每个输入样本的所属的高斯成分；z 为连续隐变量，产生于 c 所指定的高斯分布。其生成过程如图4.5(a) 所示，设预先给定的高斯成分个数为 K，GMVAE 首先从多项分布 $Multi(\boldsymbol{\pi})$ 中采样得到类别变量 c，由 c 所对应的高斯分布 $\mathcal{N}(\mu_c, \sigma_c^2\boldsymbol{I})$ 中采样得到隐变量 z，z 再经变换 $f(z;\phi)$ 后得到生成样本 x。

根据图4.5(a) 所展示的条件依赖关系，x、c 和 z 的联合概率分布 $p(x,z,c)$ 可分解为：

$$p(x,z,c) = p(c)p(z|c)p(x|z) \tag{4-14}$$

式（4-14）中各概率满足：

$$\begin{aligned} p(c) &= Multi(\boldsymbol{\pi}) \\ p(z|c) &= \mathcal{N}(z|\mu_c, \sigma_c^2\boldsymbol{I}) \\ p(x|z) &= \mathcal{B}er(x|\mu_x) \end{aligned} \tag{4-15}$$

其中 $\mu_x = f(z;\theta)$，$\mathcal{B}er(x|\mu_x)$ 为参数为 μ_x 的 Bernouli 分布。根据上述生成过程，由 log 函数的上凸性及 Jensen 不等式，可求得 x 的对数似然的变分下界：

$$\begin{aligned} log p(x) &= log \int_z \Sigma_c p(x,z,c) dz = log \int_z \Sigma_c q(z,c|x) \frac{p(x,z,c)}{q(z,c|x)} dz \\ &\geqslant \int_z \Sigma_c q(z,c|x) log \frac{p(x,z,c)}{q(z,c|x)} dz = \mathbb{E}_q(z,c|x)[log \frac{p(x,z,c)}{q(z,c|x)}] = \mathcal{L}_{ELBO} \end{aligned} \tag{4-16}$$

其中 $q(z,c|x)$ 为变分后验分布，由图4.5(b) 所展示的推断过程，$q(z,c|x)$ 可分解为

$q(z,c|x)=q(z|x)q(c|x)$，代入式（4-16）则有：

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x})=&E_{q(\mathbf{z},c|\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{z},c)}{q(\mathbf{z},c|\mathbf{x})}\right]\\=&E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x},\mathbf{z},c)-\log q(\mathbf{z},c|\mathbf{x})]\\=&E_{q(\mathbf{z},c|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})+\log p(\mathbf{z}|c)\\&+\log p(c)-\log q(\mathbf{z}|\mathbf{x})-\log q(c|\mathbf{x})]\end{aligned}\tag{4-17}$$

与 VAE 中类似，式（4-17）中的变分后验分布 $q(z|x)$ 设为高斯分布，通过神经网络 $g(x)$ 进行建模，即：

$$\begin{aligned}\left[\tilde{\mu};\log\tilde{\sigma}^2\right]&=g(\mathrm{x};\phi)\\q(\mathrm{z}|\mathrm{x})&=\mathcal{N}\left(\mathrm{z};\tilde{\mu},\tilde{\sigma}^2\mathrm{I}\right)\end{aligned}\tag{4-18}$$

将式（4-15）和式（4-18）代入式（4-17），可得到：

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}=&\frac{1}{M}\sum_{l=1}^{M}\sum_{i=1}^{D}x_i\log\boldsymbol{\mu}_x^{(l)}\Big|_i+\left(1-x_i\right)\log\left(1-\boldsymbol{\mu}_x^{(l)}\Big|_i\right)\\&-\frac{1}{2}\sum_{c=1}^{K}\eta_c\sum_{j=1}^{H}\left(\log\sigma_c^2\big|_j+\frac{\tilde{\boldsymbol{\sigma}}^2\big|_j}{\sigma_c^2\big|_j}+\frac{\left(\tilde{\boldsymbol{\mu}}|_j-\boldsymbol{\mu}_c\big|_j\right)^2}{\sigma_c^2\big|_j}\right)\\&+\sum_{c=1}^{K}\eta_c\log\frac{\pi_c}{\eta_c}+\frac{1}{2}\sum_{j=1}^{H}\left(1+\log\tilde{\sigma}^2\big|_j\right)\end{aligned}\tag{4-19}$$

其中 M 为所采样的 z 的个数，D 为 x 的维数，J 为隐变量 z 的维数，π_c 为第 c 个高斯成分的先验概率，$\eta_c=q(c|x)$，可通过式（4-20）进行估计（详细推导参见附录）：

$$q(c|x)=p(c|z)=\frac{p(c)p(z|c)}{\Sigma_{c'=1}^{K}p(c')p(z|c')}\tag{4-20}$$

GMNTM 的整体网络结构如图4.6所示：

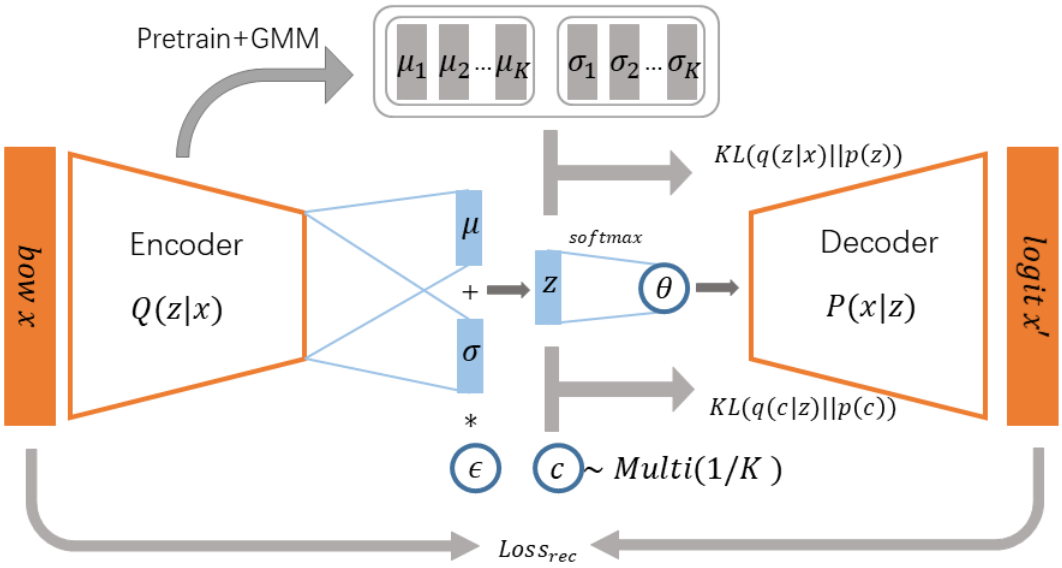


图 4.6　GMVAE 的网络结构

该网络首先需要经过自编码器的重构过程进行预训练，训练结束后使用高斯混合模型 GMM 进行聚类，将所得的各高斯成分的均值 μ_K 和方差 σ_K 作为隐空间中高斯混合分布的初始值，类别隐变量 c 的先验分布设定为均匀的离散分布，即 $Multi(\frac{1}{K})$。对于输入样本 x，由编码器 Q 映射得到隐变量 z 的分布的均值 $\tilde{\mu}$ 和对数方差 $log\tilde{\sigma}^2$，采样得到 z，由各高斯成分的均值和方差及对应的权重，可求得 z 的先验分布 $p(z) = \Sigma_{i=1}^{K}\pi_i\mathcal{N}(z;\mu_i,\sigma_i^2)$，进而得到 z 的变分后验分布 $q(z|x)$ 与先验 $p(z)$ 的 KL 散度；由式（4-20），可求得 c 的变分后验分布 $q(c|x)$ 与先验分布 $p(c) = \frac{1}{K}$ 的 KL 散度；重构误差加上上述 KL 正则项即构成了该网络的优化目标。与 NTM-GSM 类似，基于 GMVAE 构建主题模型时，将隐变量 z 经过 Softmax 层后得到的归一化向量 θ 作为主题分布向量，即 $\theta = Softmax(Wz+b)$；将第 k 个高斯成分的均值经变换后得到的 θ_k 导入解码器，所得的归一化输出即为第 k 个主题-词分布。

4.3.2　实验结果及分析

4.3.2.1　数据准备

在 Sub 数据集的基础上，本文采用第 2 章所描述的方法，通过添加 CSI，Merlin，Seinfeld 等共 18 部美剧的中文字幕将其扩展为 SubX 数据集[48]。利用第三章中提出的主题分割模型对上述两个数据集中的字幕流进行了分割，并以分割后形成的主题片段作为文档。此外，为了评估对话文本的长度对模型的影响，本文将 DailyDialog 英文对话数据集[8] 翻译后构建了中文对话数据集 zhdd 和 zhddline，zhdd 以一个对话片段作为文档，zhddline 则以一条话语作为文档；此外，为了评估模型在新闻类

短文本上的性能，本文从搜狐新闻中随机抽取了 180,000 条短新闻构成了 cnews 数据集。上述数据集的基本统计如下：

表 4.1　主题建模数据集基本统计

	文档数	词表大小	Avg 文档长度	Std 文档长度
Sub	54106	67946	61.2	23.8
SubX	132833	81253	49.4	24.2
zhdd	12321	21917	22.9	18.2
zhddline	84327	21917	13.4	3.0
cnews	180000	86989	27.12	1.81

4.3.2.2　评测指标

本文主要采用主题连贯度及主题多样性两个指标对主题模型的建模结果进行评价。

主题连贯性（Topic Coherence，简记为 TC）是同一主题内任意两个词的互信息的均值，该指标假设主题越连贯，则越倾向于采样经常在同一文档中出现的词，其互信息也会越高。该指标度量了提取出的主题的可解释程度，连贯性指标较高的主题模型其解释性通常较强。主题连贯性的定义为：

$$\mathrm{TC} = \frac{1}{K}\sum_{k=1}^{K}\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} f\left(w_i^{(k)}, w_j^{(k)}\right) \tag{4-21}$$

其中 $w_1^k, w_2^k, ..., w_n^k$ 是第 k 个主题中权重最大的前 n 个词（n 一般取为 10），$f(w_i, w_j)$ 定义为：

$$f\left(w_i, w_j\right) = \frac{\log\frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P\left(w_i, w_j\right)} \tag{4-22}$$

其中 $P(w_i, w_j)$ 是两个词 w_i 和 w_j 在同一篇文档中共现的概率，$P(w_i)$ 是词 w_i 的出现概率，可通过频率来估计。

主题多样性（Topic Diversity，简记为 TD）用于衡量提取出的主题的多样程度，其定义为所有主题中的权重最大的前 25 个词所组成的集合中，去重后的词表占原集合的比例。主题多样性趋于 0 时，表明所提取的主题有较多冗余；主题多样性趋于 1 时，表明所提取的主题有较高的多样性。

4.3.2.3　实验结果

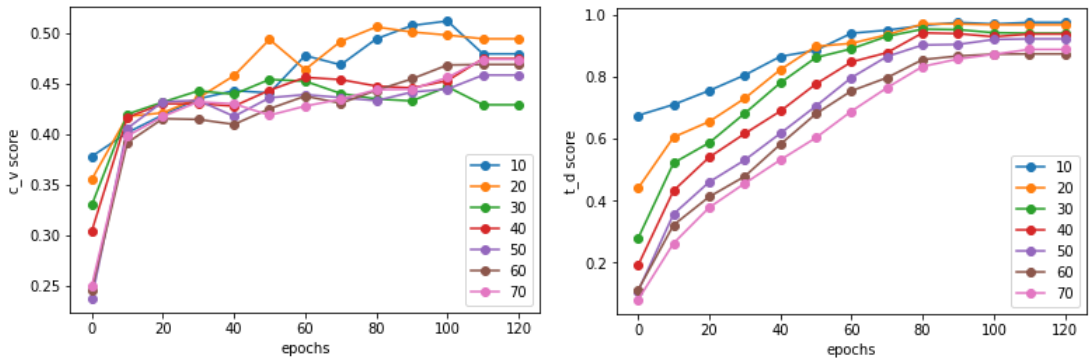


图 4.7　TC 和 TD 指标随训练轮数的变化 (GMNTM on Sub)

图4.7展示了 GMNTM 在 Sub 数据集上主题连贯度 TC 和主题多样性指标 TD 随训练轮数的变化，可以看到随着训练轮数的增加，两个评价指标都在逐步增长并趋于稳定，表明 120 轮次对 GMNTM 在 Sub 数据集上的训练已足够充分，此外 TD 指标比 TC 指标对不同主题数的更加敏感，当主题数越大时，TD 指标越低，这反映出当挖掘的主题数越多时，要保持尽可能少的重复会越困难；比较不同的主题数时 TC 指标的变化，可以看到，当主题数为 20 时，GMNTM 在 Sub 数据集上能取得最高的主题连贯度。

本文将主题数分别设置为 10，20、30、40、50、60、70，对比了不同神经主题模型在各数据集上的训练 120 个 epoch 后的性能。表4.2展示了不同模型在不同主题数下的主题连贯度指标。

从表4.2可以看到，在三个对话数据集 Sub、SubX 和 zhdd 上，对于大多数的主题数设定，GMNTM 都取得了最高的主题连贯度，表明以高斯混合分布建模对话文本中的主题分布的确有效，在新闻数据集 cnews 上，W-LDA 则有更好的表现。一个值得关注的现象是，随着主题数的增多，NTM-GSM 的性能逐步下降，可能的原因是在隐空间中较小的范围内建模了较多的分布，层叠的可能性变大，使得同一个隐变量对应多个主题分布，形成后验坍塌。此外，整体而言，基于神经网络的主题模型在对话文本上的性能都要优于传统概率主题模型。

在字幕对话数据集 Sub 上，不同主题数对主题连贯性的影响如图4.8所示。

表 4.2　不同主题数下各个模型的主题连贯度

主题数		10 TC	20 TC	30 TC	40 TC	50 TC	60 TC	70 TC
cnews	LDA	0.323	0.377	0.380	0.382	0.390	0.414	0.426
	NTM-GSM	0.404	0.442	0.451	0.478	0.423	0.410	0.389
	W-LDA	0.438	**0.491**	0.454	**0.497**	**0.486**	**0.481**	**0.450**
	GMNTM	**0.452**	0.484	**0.486**	0.461	0.471	0.437	0.435
Sub	LDA	0.323	0.329	0.347	0.344	0.318	0.310	0.302
	NTM-GSM	0.403	0.421	0.438	0.416	0.399	0.374	0.368
	W-LDA	**0.462**	0.452	**0.467**	0.442	0.428	0.434	0.451
	GMNTM	**0.462**	**0.498**	0.454	**0.475**	**0.458**	**0.469**	**0.474**
SubX	LDA	0.332	0.346	0.376	0.365	0.361	0.350	0.342
	NTM-GSM	0.402	0.415	0.454	0.401	0.412	0.412	0.405
	W-LDA	0.413	0.410	0.456	0.461	**0.427**	0.401	0.387
	GMNTM	**0.431**	**0.461**	**0.476**	**0.454**	0.421	**0.424**	**0.412**
zhdd	LDA	0.342	0.351	0.384	0.381	0.376	0.362	0.352
	NTM-GSM	0.382	0.395	0.412	0.393	0.392	0.382	0.367
	W-LDA	**0.423**	0.434	0.457	0.465	0.449	**0.431**	0.427
	GMNTM	0.421	**0.441**	**0.475**	**0.470**	**0.461**	0.424	**0.432**

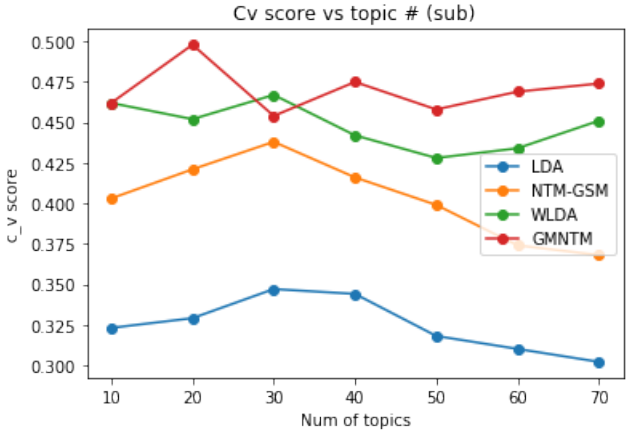


图 4.8　主题连贯性随主题数的变化

可以看到，在 Sub 数据集上，各模型的主题连贯性随着主题数的增加，总体呈现先上升后下降的趋势，各模型在主题数为 20 至 30 的时候能取得最高的主题连贯度。

以下展示了各模型在主题数为 10 时的前 10 个主题词，

从各模型提取到的前 10 个主题词可以看到，主题连贯性最低的 LDA 模型，其

表 4.3　不同主题数下各个模型的主题多样性

主题数		10 TD	20 TD	30 TD	40 TD	50 TD	60 TD	70 TD
cnews	LDA	0.84	0.83	0.95	0.92	0.89	0.88	0.89
	NTM-GSM	0.875	0.835	0.817	0.91	0.85	0.842	0.812
	W-LDA	**0.97**	**0.95**	0.95	**0.98**	**0.97**	**0.96**	**0.96**
	GMNTM	**0.97**	0.93	**0.98**	0.92	0.93	0.95	0.95
Sub	LDA	0.84	0.85	0.85	0.81	0.76	0.77	0.79
	NTM-GSM	**0.94**	0.85	0.92	0.93	0.93	0.91	0.90
	W-LDA	0.86	**0.92**	**0.95**	0.92	0.91	**0.94**	**0.94**
	GMNTM	0.87	**0.92**	0.89	**0.95**	**0.95**	0.93	**0.94**
SubX	LDA	0.92	0.92	0.92	0.87	0.85	0.87	0.85
	NTM-GSM	**0.94**	0.92	0.93	0.92	0.89	0.90	0.89
	W-LDA	0.90	**0.93**	0.95	0.90	0.89	0.92	0.93
	GMNTM	0.92	**0.93**	**0.96**	**0.93**	**0.94**	**0.93**	**0.95**
zhdd	LDA	0.87	0.89	0.86	0.85	0.84	0.82	0.82
	NTM-GSM	**0.92**	0.91	0.91	0.92	0.90	0.89	0.88
	W-LDA	0.91	0.91	**0.95**	0.90	0.91	0.91	0.93
	GMNTM	0.91	**0.92**	0.94	**0.95**	**0.95**	**0.92**	**0.95**

Topic 9 中同时包含了“约会”、“宝宝”、“病毒”、“炸弹”等主题词，不能清晰地描述所要表达的主题；NTM-GSM 提取得到的主题有了一定改进，如其 Topic 4 中包含的单词大部分与医疗主题相关，但 NTM-GSM 存在与 LDA 同样的问题，许多主题词表达并不明晰，在某些相对明显的主题中，仍然混杂着与该主题相关性不大的主题词，如 Topic 8 中超过一般的词与“车”有关，但仍然出现如“区”、“组”、“酒筵”等相关性较小的词语；而 W-LDA 则有了较大的提升，例如 Topic 1、Topic 4 与 Topic 10 可以使人较为容易地分别联想到“刑侦”、“情感”和“交际用语”；GMNTM 提取到的主题词同样具有较好的表达力，而且可以看出 GMVAE 提取到的主题更为细致，如 Topic 5 与“情感”相关性较大，而 Topic 10 则与“婚礼”相关性较大。这表明主题连贯性指标与人的主观感受相符，同时也验证了本文所提出的 GMVAE 主题模型在对话文本主题建模上的有效性。

SubX 数据集取自影视剧中的特定场景下的对话，对话双方由于有共同的上下文环境，加之影视剧所要追求的艺术效果，因此其话语消息中会有较多的省略或双关等修辞手法，也因此引入了较多噪声。与之形成对比的是数据集 zhdd，该数据集中的对话通过众包的方式由人工在线编撰完成，由于不存在共同的时空环境，

表 4.4　LDA 在 SubX 中提取得到的前 10 个主题词

LDA (K=20)										
Topic 1	喝	陛下	杯	莱	加油	啤酒	电影	那种	上帝	喝酒
Topic 2	孩子们	明天	钥匙	号	结婚	父母	快乐	宝贝	还好	瑞秋
Topic 3	星际	之门	伊	眼睛	莲	上帝	灵魂	魔法	身体	石头
Topic 4	警察	儿子	选择	父亲	谋杀	调查	世界	保护	事实上	威胁
Topic 5	惊喜	幸运	送	礼物	痛苦	准备好	车	求	昨晚	衣服
Topic 6	块	证明	送	手机	婚礼	派对	美元	在场	偷	花
Topic 7	杰瑞	画	猫	穿	戒指	求婚	节目	衣服	跳	约会
Topic 8	丈夫	爸	女孩	妈	还好	父母	父亲	昨晚	男孩	女儿
Topic 9	约会	尸体	宝宝	赢	提	病毒	炸弹	报告	结束	消失
Topic 10	莱	佩妮	婚姻	局	么	真相	确实	结婚	联调	妻子

表 4.5　NTM-GSM 在 SubX 中提取得到的前 10 个主题词

NTM-GSM (K=20)										
Topic 1	猫	臭臭	唱到	赚钱	唱	不在	最爱	故意	带你去	乖乖
Topic 2	旗帜	博士	有趣	回忆	背后	学	收看	节目	面	特辑
Topic 3	长官	之门	成功	星际	洞	象形文字	虫	毁灭	建立	译码
Topic 4	戊	肝炎	型	公司	治疗	红斑狼疮	说谎	恶化	疫苗	骗
Topic 5	上帝	还好	娜	入睡	魔法	好久不见	尽头	离去	害怕	嘉
Topic 6	还好	怎么	救	没问题	懂	准备好	活着	特	天哪	听见
Topic 7	剪刀	史	石头	布	蜥蜴	斩首	布包	踩	同义	毒死
Topic 8	区	组	货车	轮胎	鉴证	车辙	车	凶手	酒筵	尸体
Topic 9	研究	奖	冠军	学校	项目	宝宝	大学	美元	比赛	周
Topic 10	瑞秋	首席	选	解雇	伴娘	婚礼	答应	赢	在乎	无辜

因此其中的话语消息相对更完整，所涉及的主题也与日常生活关联性较大，因此本文同样考察了各个模型在数据集 zhdd（以对话片段为单位）和 zhddline（以话语消息为单位）上的性能。

相较于 SubX，W-LDA 和 GMNTM 两个模型从数据集 zhdd 中抽取出的主题都更加明晰，这表明 zhdd 中噪声更少的假设成立。在这两个数据集上，GMNTM 都取得了最好的性能表现。值得注意的是，GMNTM 在 zhddline 上的主题连贯性得分相对较高，且实际抽取的主题词经过人工评判后，也与其主题具有很高的相关性，表明 GMVAE 在非常短的对话文本上（utterance 级）也可以提取到有意义的主题，这进一步验证了本文所提出的主题模型在对话文本上的有效性。

表 4.6　W-LDA 在 SubX 中提取得到的主题词

W-LDA (K=20)										
Topic 1	找到	凶手	发现	指纹	孩子	尸体	也许	鉴证	证据	谋杀
Topic 2	再见	很好	再来	对了	今晚	太好了	开心	性感	咖啡	喜欢
Topic 3	记录	证据	证明	电话	监控	调查	找到	发现	录像	在场
Topic 4	约会	朋友	派对	参加	喜欢	男人	真的	女人	开心	婚礼
Topic 5	电影	没错	星际	名字	迷航	故事	博士	游戏	美国	粉丝
Topic 6	凶手	东西	有人	监控	时间	地方	肯定	录像	系统	线索
Topic 7	瑞秋	天啊	你好	真的	谢谢	亲爱的	抱歉	太好了	对不起	回来
Topic 8	东西	抱歉	干嘛	没错	佩妮	怎么	想要	房间	道歉	听到
Topic 9	快点	晚安	准备好	放下	拜托	干什么	钥匙	该死	把手	天哪
Topic 10	谢谢	不行	不用	东西	真的	时间	客气	感觉	地方	玩笑

表 4.7　GMNTM 在 SubX 中提取得到的主题词

GMNTM (K=20)										
Topic 1	凶手	毛和	伤口	死者	死因	痕迹	刺伤	击碎	发现	擦净
Topic 2	心脏	出血	感染	导致	呼吸	检查	肝脏	血压	解释	症状
Topic 3	求婚	结婚	体育	纪念日	孩子	订婚	住	孕妇装	伴郎	项链
Topic 4	实验	苦求	周相	消音	干涉	右脑	难题	错误	观测	积分
Topic 5	真的	喜欢	约会	我会	男人	没错	抱歉	名字	结婚	第一
Topic 6	工作	生活	改变	这份	想要	接受	痛苦	喜欢	努力	放弃
Topic 7	计划	想要	警察	戒指	打电话	明白	回来	一点	伙计	喜欢
Topic 8	凶手	找到	发现	线索	死者	指纹	证据	鉴证	手机	尸体
Topic 9	电话	手机	昨晚	打给	记得	离开	打来	办公室	名字	号码
Topic 10	参加	婚礼	瑞秋	晚上	开心	结婚	爸爸	邀请	记得	妈妈

表 4.8　W-LDA 在 zhdd 中提取得到的主题词

W-LDA on zhdd (K=20)										
Topic 1	国家	中国	比赛	奥运会	世界	有趣	发生	美国	明白	环境
Topic 2	孩子	父母	家庭	美国	丈夫	青少年	学习	孩子	太多	中国
Topic 3	美元	兑换	豪斯	价格	菲	力	钱	五件	可恶	推过去
Topic 4	床垫	电影	石头	迟到	明天	洗衣机	记得	递送	硬币	几样
Topic 5	台风	包装	砖	摄像机	盒子	砖头	美元	蠢	中国	新
Topic 6	公交	路	辆	下车	站	坐	美元	赶上	分钟	请问
Topic 7	女孩	高	又瘦	后街	共同点	长	男孩	深刻	一头	鼓舞人心
Topic 8	账户	支票	食物	美元	年轻	钱	想要	父母	取消	乐意
Topic 9	听到	消息	孩子	老师	骨肉	很难	听说	回嘴	父母	安
Topic 10	处方	医生	配药	药	小姐	补充	药物	添	现在	感觉

表 4.9　GMNTM 在 zhdd 中提取得到的主题词

GMNTM on zhdd (K=20)										
Topic 1	喝	饮料	选择	啤酒	酒	巧克力	开胃菜	鸡蛋	奶酪	热狗
Topic 2	天气	城市	地方	住在	夏天	比赛	冬天	希望	下雨	公寓
Topic 3	条	路	座位	分钟	感谢	停车	汽车	错过	车	明白
Topic 4	公司	份	员工	面试	职位	希望	经验	机会	这份	老板
Topic 5	想要	咖啡	杯	喝	份	蔬菜	饮料	啤酒	味道	酒
Topic 6	衣服	漂亮	颜色	想要	衬衫	音乐	条	跳舞	头发	适合
Topic 7	电影	公交车	辆	邀请	出租车	今晚	坐	部	聚会	站
Topic 8	账户	钱	银行	价格	支付	提供	服务	合同	订单	明白
Topic 9	美元	钱	元	付	想要	英镑	花	价格	现金	公寓
Topic 10	明天	下午	早上	时间	点钟	到时	星期五	有空	今晚	周末

表 4.10　W-LDA 在 zhddline 中提取得到的主题词

W-LDA on zhddline (K=20)										
Topic 1	公司	价格	学习	希望	一家	英语	份	销售	时间	选择
Topic 2	好运	祝	处于	旱季	考虑	特别	花园	下雨	钱	肯
Topic 3	煮	鸡蛋	煮熟	蔬菜	肉	炒菜	奶酪	洋葱	稍微	起着
Topic 4	想要	妈妈	特别	试试	没问题	太棒	明白	好了	哪种	好主意
Topic 5	明天	回来	电话	希望	迟到	准备	太好了	时间	呆	打算
Topic 6	电影	想要	中国	新	买了	衣服	也许	部	车	很棒
Topic 7	衣服	想要	漂亮	中国	食物	颜色	设计	特别	最喜欢	想买
Topic 8	衣服	电影	漂亮	想要	买了	音乐	食物	新	很棒	时尚
Topic 9	想要	妈妈	钱	明白	爸爸	好了	特别	太棒	没问题	感谢
Topic 10	间	卧室	客厅	浴室	客房	厨房	街区	女儿	层	房间

表 4.11　GMNTM 在 zhddline 中提取得到的主题词

GMNTM on zhddline (K=20)										
Topic 1	专业	适合	更好	打字	原因	关心	技能	方法	简单	领域
Topic 2	公司	员工	工资	投票	加薪	奖金	提供	职位	每月	带薪
Topic 3	账户	美元	钱	银行	服务	卡	支付	信用卡	明白	价格
Topic 4	房间	大床	单床	旅馆	下午	晚安	无烟	修理	房	预订
Topic 5	沙拉	份	甜点	咖啡	点菜	牛排	牛肉	瓶	开胃菜	酒’
Topic 6	头发	选择	油腻	洗	穿衣服	太棒	掉	早上	头天	熨平
Topic 7	中国	企业	经济	国内	政策	发表	影响	国家	秒钟	发生
Topic 8	是因为	浪漫	空气	很棒	短发	清新	搭档	试试	约会	见过
Topic 9	汽车	辆	路	坐	下车	站	巴士	赶上	分钟	车站
Topic 10	感谢	什么	开心	砖头	大忙	到时	客气	想见	女士	邀请

4.4　本章小结

本章对比分析了已有的神经主题模型，并针对对话文本提出了基于高斯混合先验的神经主题模型，实验表明，所提出的模型在字幕对话数据集 Sub 和 SubX 上都取得了较现有主题神经模型更高的主题一致性，从所提取的主题词的对比来看，该评估指标也能与人的主观判断大致吻合。本章所提出的 GMNTM，对从字幕对话流中切分出的主题片段可以给出较理想的主题表示，为下一步基于主题表示进行聚类奠定了基础。

第 5 章　对话文本主题聚类

5.1　本章概述

通过对对话文本进行主题建模，对话文本可经由模型映射到主题空间，通过主题分布向量进行表示。在此基础上，本章将通过对隐空间的主题分布向量进行聚类，实现对对应文本按照主题进行划分和组织的目标。本章首先简要综述了常见的聚类算法；接着整理常见的聚类评价指标；实验部分，为了对模型进行全面、客观的评价，分别针对有类别标签的新闻标题数据集 cnews，和无类别标签的对话数据集 SubX 以及 zhdd，给出了基于外部的评价和基于内部的评价，最后对聚类出的主题簇进行了主题描述，对不同主题展示了具有代表性的例子。

5.2　聚类方法

本节将简要介绍 K-means、Mean Shift、DBSCAN 三种聚类方法。

5.2.1　K-means

K-means[49] 是基于划分的聚类方法，其核心思想是通过迭代计算类簇均值来更新类簇中心，直到满足某些收敛标准为止。K-means 的一个基本假设是：对每个类簇都可以找到一个类簇中心，使得该类簇中的所有点到该类簇中心的距离小于到其他类簇中心的距离。由这个假设可以推得，K-means 的优化目标是最小化

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} \lambda_{nk} \left\| x_n - \mu_k \right\|^2 \tag{5-1}$$

其中 N 为数据点的个数，K 为类簇的个数，μ_k 为类簇 k 的中心，λ_{nk} 在数据点 n 被划分为类簇 k 时为 1，其余情况为 0。K-means 采用迭代的方法来优化 L，第一步先固定 μ_k，寻找最优 λ_{nk}，可知只需将数据点划分到距离最近的 μ_k 即可；第二步则固定 λ_{nk}，寻找最优的 μ_k，求导取零可知，此时 μ_k 的取值为类簇 k 中数据点的均值。由于每轮迭代都在搜索 L 的极小值，总体上 L 会持续减小或不变，因而最终会收敛到一个极小值。K-means 的整体步骤为：

K-means 是一个简单而且高效的聚类算法，在许多场景下都取得了较好的效果，但 K-means 只能用于发现球形类簇，其聚类效果会受类簇中心的初始值影响，初始值选取的不好，可能导致 K-means 收敛到局部最优解。

Algorithm 5 K-means process

Input: 数据点集 D，K **Output**: 类簇 C

初始化 K 个类簇的中心 $\{\mu_k\}$

while 中止条件未满足（最大迭代次数、最小变化误差等） **do**

for 每个数据点 x_i **do**

计算 x_i 到各类簇中心 $\{\mu_k\}$ 的距离，将其划分为距离最小的类簇中心的类中

end for

for $k \in 1, 2, \cdots, K$ **do**

$\mu_k = \frac{1}{N_k} \Sigma_{i \in cluster_k} x_i$

end for

end while

5.2.2　Mean Shift 聚类

Mean Shift 算法[50]（均值漂移）是一种非参数聚类算法。给定 n 个 d 维数据点，对于 x 点，Mean Shift 会以当前点为球心，以 h 为半径作一个球形滑窗。以当前点 x 为起点，滑窗内的所有点形成的向量的加权和作为偏移均值向量，在下一步中，将以该偏移向量的终点为球心作新的球形滑窗，此后迭代上述步骤，直至收敛或满足其他中止条件。

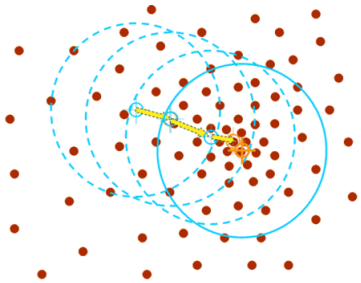


图 5.1　均值漂移过程

Mean Shift 聚类的步骤是，对于每一个数据点 x，以该点为起点执行均值漂移，不断重复直至收敛，把该数据点与收敛位置的元素标记为一类，因而收敛到相同位置的数据点则聚为同一类。

一般而言，滑窗中不同的数据点对被偏移点 x 的作用是不同的，应该将这种差异对均值偏移向量的影响纳入考量。Mean Shift 算法通过引入核函数来度量滑窗中数据点与被偏移点之间的不同距离对均值偏移向量的不同贡献，常用的是高斯

核函数：

$$K(x) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{x^2}{2h^2}} \tag{5-2}$$

其中 h 为高斯核的带宽，因而均值漂移向量可表示为：

$$M_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{\sigma_i}\right)(x_i - x)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{\sigma_i}\right)} \tag{5-3}$$

图5.2展示了 Mean Shift 聚类的过程：

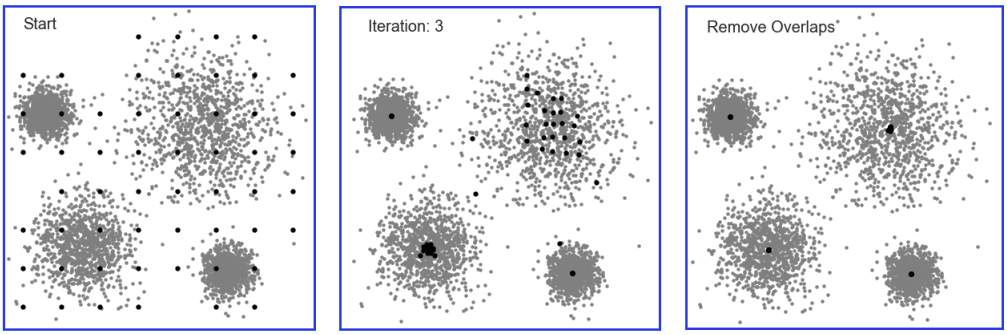


图 5.2　Mean Shift 聚类示例

相较于 K-means，Mean Shift 不需要预先设定类簇数量，也不会受限于类簇的形态，然而，Mean Shift 最终聚类效果的好坏，与滑窗半径的大小有很大关系。

5.2.3　DBSCAN

DBSCAN[51] 基于密度进行聚类。DBSCAN 首先统计每个数据点的 ϵ-邻域内的数据样本个数，如果其数量不小于 $MinPts$（最小包含点数），则将该数据点作为核心对象。DBSCAN 的相关概念如下：

- ϵ-邻域 $N_\epsilon(x)$：数据集 D 中所有与 x 的距离不超过 ϵ 的数据点组成 x 的 ϵ-邻域，即 $N_\epsilon(x) = \{x_i \in D | d(x_i, x) \leqslant \epsilon\}$，其大小记为 $|N_\epsilon(x)|$。
- 核心对象：D 中满足 $|N_\epsilon(x)| \geqslant MinPts$ 的所有数据点。
- 密度直达：若 x 为核心对象，$y \in N_\epsilon(x)$，则 y 可由 x 密度直达。
- 密度可达：若 x 与 y 之间存在数据点序列 $\{p_i\}_{i=1,\cdots,T}$，满足 p_{i+1} 可由 p_i 密度直达，$\{p_i\}_{i=1,\cdots,T}$ 为核心对象，则称 y 由 x 密度可达。
- 密度相连：若 x 与 y 之间存在核心对象 z，满足 x 和 y 均由 z 密度可达，则称 x 与 y 密度相连。

DBSCAN 的基本思想是从核心对象出发，通过密度可达找到最大的密度相连

的数据集合，即构成一个类簇，并继续为尚未聚类的核心对象选择类簇，直至所有的核心对象都有类别。

DBSCAN 的算法流程为：

Algorithm 6 DBSCAN process

Input: ϵ，$MinPts$，数据集 D **Output**: 类簇 C

初始化簇 ID：$k = 1$，构建核心对象集 $\Omega = \{x||N_\epsilon(x)| \geqslant MinPts\}$

随机选取核心对象 $p \in \Omega$，更新簇 ID：$k = k + 1$，构建当前簇核心对象队列 $Q = \{p\}$，构建当前簇数据集 $C_k = \{p\}$

while $Q \neq \varnothing$ **do**

取走一个核心对象 $p' \in Q$，将 $N_\epsilon(p')$ 中所有未标记访问的点放入当前簇 C_k 中，并将这些点标为已访问。

end while

输出簇划分 $\{C_1, \cdots, C_k\}$

相较于 K-means，DBSCAN 无需预先设定类簇的数量 K，且可发掘任意形状的类簇，同时对数据集中的异常点不敏感。但是，DBSCAN 对于类间差距很大的数据，聚类效果往往不理想，且不同的 $(\epsilon, MinPts)$ 参数组合，对于最终的聚类质量有较大影响。

5.3　评价指标

由于聚类算法是无监督学习方法，对于聚类结果的质量优劣，没有一种绝对的评价方法。通常认为，理想的聚类算法，应该做到：（1）类内相似度尽可能高；（2）类间相似度尽可能低。目前针对聚类算法的评价已有大量的研究，总结起来可以分为两类：（1）内部评价指标；（2）外部评价指标。本节将分别介绍这两类指标中的代表性方法。

5.3.1　内部评价指标

内部评价指标基于样本集自身的特征来对聚类结果的优劣进行评价。这类指标主要依据类内平均相似度和类间平均相似度来衡量聚类的有效性，目前常用的指标包括轮廓系数、CH 指标、类簇凝聚度 SSE 等。

轮廓系数 (Silhouette Coefficient,SC)：轮廓系数[52] 基于样本之间的距离来计算相似度，每个样本点都有各自的轮廓系数，由两部分组成：

- W：当前样本点与类内其他样本点的平均距离

• B：当前样本点与距离最近的其他类簇中所有样本点的平均距离

当前样本点的轮廓系数定义为：

$$s = \frac{B - W}{max(W, B)} \tag{5-4}$$

整个样本集的轮廓系数是各样本点的轮廓系数的均值。轮廓系数 SC 的取值范围是 $[-1, 1]$，当类簇分离越大且密度越高，轮廓系数越高。然而，轮廓系数倾向于对凸簇给出更高的得分，因此不适合基于密度的聚类方法（如 DBSCAN）。

CH 指标 (Calinski-Harabasz Index)：CH 指标[53] 以类内样本点与类中心的距离的平方和来度量类内紧密度，以各类簇中心与样本集中心的距离的平方和来度量类间分离度，并将两者比值作为评估指标。CH 指标的定义为：

$$CH(K) = \frac{\text{tr}(B)/(K-1)}{\text{tr}(W)/(N-K)} \tag{5-5}$$

其中 K 为类簇的个数，N 为样本总数，B 为类间协方差阵，W 为类内协方差阵。CH 指标越高，表明聚类的性能越好。然而与轮廓系数类似，CH 指标倾向于给凸簇会比其他类型的簇更高的得分，因此不适合基于密度的聚类算法（如 DBSCAN）。

类簇凝聚度：类簇的凝聚度通常以类内点对的平均距离来衡量，定义为类内方差和：

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \cdots, K \tag{5-6}$$

其中 $\bar{x}_i$ 为第 i 个类簇的类中心，n_i 为第 i 类的样本点数量。SSE 越接近 0，则类簇的凝聚度越高，聚类效果越好。

5.3.2　外部评价指标

外部评价指标通产需要获得样本点的标记信息，然而在实际聚类应用中，往往难以获得真实的标记，因此外部评价指标通常用于为特定数据集选择合适的聚类算法。

兰德系数 (RI). 对于由 N 个样本点组成的样本集，兰德系数[54] 考查每一样本对是否应放入同一类簇中，其定义为：

$$RI = \frac{TP + TN}{C_N^2} \tag{5-7}$$

其中，TP 是将标签相同的样本对划入同一类簇的数量，TN 是将标签不同的样本对划入不同簇的数量，C_N^2 为可能的样本对个数。RI 的取值范围为 $[0, 1]$，值

越大表明聚类结果与外部标签越接近。其缺点是对随机分配的类簇标签，RI 并不能保证接近 0。

Word2vec[55]. 基于词向量的评价方法，使用在大规模语料上预训练的词向量，计算类内词所对应词向量之间的相似程度，直觉上，同一主题下的词之间的词向量相似度应比不同主题之间的词向量相似度更高，因此能以主题进行聚类的算法，其类内词向量应拥有较高的平均相似度，类间则有较低的平均相似度。

5.4　聚类结果

本文分别采用 K-means、Mean Shift 和 DBSCAN 以及取最高概率主题的方法，对 SubX、zhdd 以及 cnews 数据集的主题分布向量（由 GMVAE 提取）进行了聚类实验。其中 cnews 为带标签的短新闻数据，共包含 12 个新闻类别，其聚类结果可以使用外部评价指标进行评估。聚类实验中选择 $K = 20$ 的主题模型对应的隐变量进行聚类，类簇数量与主题数目同样设为 20。（下表中，Argmax 指标表示直接选择其文档-主题分布中权重最大的主题类比作为文档的类别。）

表 5.1　SubX 上的聚类质量评估

	Silhouette Coef	CH Index	SSE
K-means	**0.41**	**52.1**	0.29
Mean Shift	0.35	17.3	0.17
DBSCAN	0.17	13.5	**0.13**
Argmax	0.23	34.7	0.19

表 5.2　zhdd 上的聚类质量评估

	Silhouette Coef	CH Index	SSE
K-means	**0.30**	18.4	0.19
Mean Shift	0.22	23.6	0.13
DBSCAN	0.18	10.7	**0.06**
Argmax	0.25	**27.1**	0.11

从表5.3可以看出，DBSCAN 在 cnews 数据集上取得了最高的外部评价得分，而在 subX 和 zhdd 数据集上，又取得最好的类簇凝聚度得分。因此，本文采用 DBSCAN 算法进行文本聚类。图5.3展示了 SubX 在不同类簇数下的 SSE 得分。可以看到，对数据集 Sub 和 zhdd 来说，聚类的最佳类簇数目均为 20；对 cnews 数据集而言，最佳类簇数为 40。

表 5.3　cnews 上的聚类质量评估

	Silhouette Coef	CH Index	SSE	Rand Index
K-means	**0.49**	**71**	0.28	0.423
Mean Shift	0.26	38	0.22	0.437
DBSCAN	0.13	24	**0.17**	**0.493**
Argmax	0.41	51	0.2	0.282

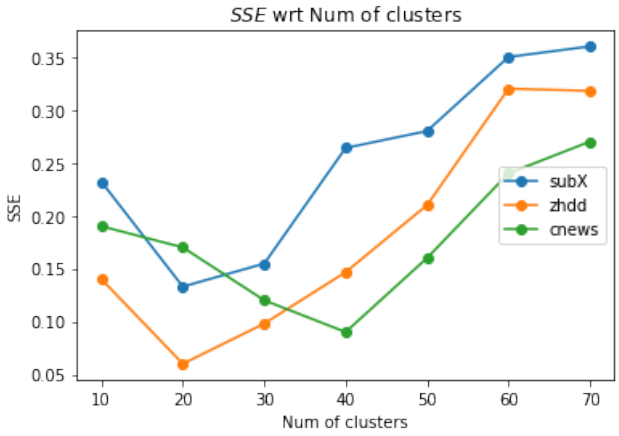


图 5.3　不同类簇数目对应的类簇凝聚度 SSE

本文对 SubX 聚类后（$K = 10$）的每个类簇中，随机抽取了 6 个话题线索作为该类簇的代表，附于附录 B 中。从所展示的话题线索来看，各类簇都有较为鲜明的主题，例如类簇 10 与经济金融类的主题有较明显的联系，表明了所选取的聚类算法的合理性和主题模型给出的主题表示的有效性。

第 6 章　总结与展望

6.1　本文研究总结

为推动解决大规模中文对话语料库缺乏的问题，本文提出利用影视剧字幕作为数据来源，构造按主题组织的对话语料库。尽管字幕数据数量丰富且风格多样，但字幕中说话人和场景信息的缺失却阻碍了其被直接用于对话语料构建。字幕的流式数据的特点使得其中包含的多个话题线索边界不明，这导致难以将主题建模应用于字幕数据上。为此，本文提出了先对字幕流数据进行主题分割，再基于分割后的主题片段进行主题建模的方法。由于主题建模以文档为单位，利用文档中词的共现关系来推断主题分布，因此，主题分割效果的好坏将会影响主题建模的结果，而主题表示的好坏则对主题聚类有直接的影响。近年来的研究与本文初期的试验结果均表明，基于神经网络的主题分割模型比传统统计模型具有更好的性能，因此本文基于神经网络设计针对对话文本的主题分割模型。

由于神经网络为有监督的学习模式，需要足量的带标注的训练数据（即标注了主题变换点的字幕流数据），因此本文首先利用公开的英文剧本和中英字幕，设计了一套自动对齐和标注算法，为 4 部美剧的字幕标注了场景和说话人信息，获得了高质量的基础对话数据集 Sub，为标注主题变换点提供了参考。

根据 Sub 中标注的说话人和场景信息，本文人工标注了原始的主题片段，并使用拼接的扩充策略扩展了原始数据集。本文在构造的模拟数据集上评估了所提出的基于 BERT 和 TCN 的主题分割模型，验证了其对于字幕对话流文本的有效性；该模型进而被应用于更多未标注的字幕流数据，得到了大量的字幕对话主题片段。

本文对比分析了已有的神经主题模型，并基于“单个主题片段具有较集中的主题分布”的假设，改进了基于标准正态分布先验假设的 NTM-GSM，并设计了具有高斯混合先验的神经主题模型，在上述切分出的字幕对话主题片段上进行了主题建模，并通过对比实验验证了该模型对于字幕对话文本的有效性。

基于 GMNTM 所抽取的主题表示，本文对比了经典的若干种聚类算法的主题聚类性能，并通过随机抽取出的各类簇中的一组话题线索展示了实际的主题聚类效果，验证了本文所提出的方法的有效性。

6.2　未来工作展望

由于第 3 章中所提出的主题分割模型训练于扩展的模拟数据集，该数据集未必能较全面地反映字幕对话流数据的分布特点，因此所得到的主题分割模型应用于更多真实字幕数据时，其性能会有所下降，而第 3 章的实验表明，引入说话人信息可以提升主题分割模型的性能，因此未来工作中的一项是为无标注的字幕预测其中的话轮关系，从而可将预测结果作为说话人信息引入以提升分割效果。

此外，基于所构建的多轮对话语料库和本文所提出的 GMNTM，将对话生成技术与主题模型结合，探索利用主题模型导向的对话生成技术也是本文未来工作的一个要点。

参考文献

[1] Al-Rfou R, Pickett M, Snaider J, et al. Conversational contextual cues: The case of personalization and history for response ranking. CoRR, 2016, abs/1606.00372.

[2] Zhou K, Li A, Yin Z, et al. CASIA-CASSIL: a chinese telephone conversation corpus in real scenarios with multi-leveled annotation // Proceedings of the International Conference on Language Resources and Evaluation, LREC. European Language Resources Association, 2010.

[3] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. The Association for Computer Linguistics, 2015: 1577-1586.

[4] Li J, Song Y, Zhang H, et al. A manually annotated chinese corpus for non-task-oriented dialogue systems. CoRR, 2018, abs/1805.05542.

[5] Lowe R, Pow N, Serban I, et al. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems // Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. The Association for Computer Linguistics, 2015: 285-294.

[6] Petukhova V, Gropp M, Klakow D, et al. The DBOX corpus collection of spoken human-human and human-machine dialogues // Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC. European Language Resources Association (ELRA), 2014: 252-258.

[7] Bordes A, Boureau Y, Weston J. Learning end-to-end goal-oriented dialog // 5th International Conference on Learning Representations, ICLR. 2017.

[8] Li Y, Su H, Shen X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset // Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP. Asian Federation of Natural Language Processing, 2017: 986-995.

[9] Wu Y, Wu W, Xing C, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2017: 496-505.

[10] Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too? // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2018: 2204-2213.

[11] Zhou K, Prabhumoye S, Black A W. A dataset for document grounded conversations // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 708-713.

[12] Moghe N, Arora S, Banerjee S, et al. Towards Exploiting Background Knowledge for Building Conversation Systems // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 2322-2332.

[13] Lison P, Tiedemann J. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles // Proceedings of the 10th International Conference on Language Resources and Evaluation LREC. European Language Resources Association (ELRA), 2016.

[14] Li J, Sun A, Joty S R. Segbot: A generic neural text segmentation model with pointer network // Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI. 2018: 4166-4172.

[15] Hearst M A. Texttiling: Segmenting text into multi-paragraph subtopic passages. Comput. Linguistics, 1997, 23(1):33-64.

[16] Choi F Y Y. Advances in domain independent linear text segmentation // 6th Applied Natural Language Processing Conference, ANLP. ACL, 2000: 26-33.

[17] Galley M, McKeown K R. Improving word sense disambiguation in lexical chaining // Proceedings of the 18th International Joint Conference on Artificial Intelligence. 2003: 1486-1488.

[18] Purver M, Griffiths T L, Körding K P, et al. Unsupervised topic modelling for multi-party spoken discourse // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 17-24.

[19] Misra H, Yvon F, Jose J M, et al. Text segmentation via topic modeling: an analytical study // Proceedings of the 18th ACM conference on Information and knowledge management. 2009: 1553-1556.

[20] Riedl M, Biemann C. TopicTiling: A text segmentation algorithm based on LDA // Proceedings of ACL 2012 Student Research Workshop. Association for Computational Linguistics, 2012: 37-42.

[21] Tür G, Hakkani-Tür D, Stolcke A, et al. Integrating prosodic and lexical cues for automatic topic segmentation. Comput. Linguistics, 2001, 27(1):31-57.

[22] Reynar J C. Statistical models for topic segmentation // Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 1999: 357-364.

[23] Utiyama M, Isahara H. A statistical model for domain-independent text segmentation. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, 2002.

[24] Hernault H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation // Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing, 2010. Proceedings: volume 6008. Springer, 2010: 315-326.

[25] Wang L, Li S, Xiao X, et al. Topic segmentation of web documents with automatic cue phrase identification and blstm-cnn // Lin C Y, Xue N, Zhao D, et al. Natural Language Understanding and Intelligent Applications. Cham: Springer International Publishing, 2016: 177-188.

[26] Wang L, Li S, Lv Y, et al. Learning to rank semantic coherence for topic segmentation // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1340-1344.

[27] Badjatiya P, Kurisinkel L J, Gupta M, et al. Attention-based neural text segmentation. 2018, 10772:180-193.

[28] Sehikh I, Fohr D, Illina I. Topic segmentation in ASR transcripts using bidirectional RNNS for change detection // 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU. IEEE, 2017: 512-518.

[29] Koshorek O, Cohen A, Mor N, et al. Text segmentation as a supervised learning task // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. Association for Computational Linguistics, 2018: 469-473.

[30] Miao Y, Yu L, Blunsom P. Neural variational inference for text processing // Proceedings of the 33nd International Conference on Machine Learning, ICML. 2016: 1727-1736.

[31] Ding R, Nallapati R, Xiang B. Coherence-aware neural topic modeling // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 830-836.

[32] Miao Y, Grefenstette E, Blunsom P. Discovering discrete latent topics with neural variational inference // Proceedings of the 34th International Conference on Machine Learning, ICML. PMLR, 2017: 2410-2419.

[33] Srivastava A, Sutton C A. Autoencoding variational inference for topic models // Proceedings of the 5th International Conference on Learning Representations, ICLR 2017. 2017.

[34] Nan F, Ding R, Nallapati R, et al. Topic Modeling with Wasserstein Autoencoders // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019: 6345-6381.

[35] Dieng A B, Ruiz F J R, Blei D M. Topic modeling in embedding spaces. CoRR, 2019, abs/1907.04907.

[36] Zhou Q. Dialog act annotation for chinese daily conversation. Journal of Chinese Information Processing, 2017, 31(6):75.

[37] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. CoRR, 2018, abs/1803.01271.

[38] Wang L, Zhang X, Tu Z, et al. Automatic construction of discourse corpora for dialogue translation // Proceedings of the 10th International Conference on Language Resources and Evaluation LREC. European Language Resources Association (ELRA), 2016.

[39] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. Association for Computational Linguistics, 2019: 4171-4186.

[40] Zeiler M, Taylor G, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning // : volume 2011. 2011: 2018-2025.

[41] Lin T, Goyal P, Girshick R B, et al. Focal loss for dense object detection // IEEE International Conference on Computer Vision, ICCV. IEEE Computer Society, 2017: 2999-3007.

[42] Pevzner L, Hearst M A. A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 2002, 28(1):19-36.

[43] 陆恒杨. 基于语义信息辅助的短文本主题模型研究 [博士学位论文]. 南京大学, 2019.

[44] Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research, 2013, 3:993.

[45] Kingma D P, Welling M. Auto-encoding variational bayes // 2nd International Conference on Learning Representations, ICLR. 2014.

[46] Tolstikhin I O, Bousquet O, Gelly S, et al. Wasserstein auto-encoders // Proceedings of the 6th International Conference on Learning Representations, ICLR. OpenReview.net, 2018.

[47] Dilokthanakul N, Mediano P A M, Garnelo M, et al. Deep unsupervised clustering with gaussian mixture variational autoencoders. CoRR, 2016, abs/1611.02648.

[48] Liang Y, Zhou Q. Automatically build the basic annotated daily conversation corpus based on the subtitles and scripts of tv plays // The Eighteenth China National Conference on Computational Linguistics, CCL 2019. CIPS, 2019: 18.

[49] MacQueen J, et al. Some methods for classification and analysis of multivariate observations // Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: volume 1. 1967: 281-297.

[50] Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell., 1995, 17(8):790-799.

[51] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96. AAAI Press, 1996: 226-231.

[52] Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 1987, 20(1):53−65.

[53] Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 1974, 3(1):1-27.

[54] Rand W M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 1971, 66(336):846-850.

[55] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems. 2013: 3111-3119.

致　谢

书林学海，三载匆匆。转眼之间，硕士生涯就将结束了。

感谢我的导师周强老师，周老师为我营造了一个良好的科研环境，总是愿意腾出自己的宝贵时间来对我的研究进行耐心地指导，使我从三年前对自然语言处理的一无所知逐渐步向正轨，周老师严谨的治学态度、渊博的学识、正直谦和的为人都潜移默化地影响着我，使我终身受益。周老师，感谢您三年来对我学术上的谆谆教导和生活上的关心。

感谢实验室的王东老师，王老师对待科研的热情和学术上的深厚功底都让我敬佩，每次和王老师聊天或参加王老师组织的学术分享会都让我获益匪浅。感谢王东老师三年来的帮助、教诲和指点。

硕士学习期间，实验室的师兄和同学也给了我极大的帮助。感谢李蓝天师兄、汤志远师兄、张勇大哥对我平时的关照，无论是生活中还是学习上遇到问题，都给了我很大的助力。感谢好友蔡云麒博士，平时交流中给了我很多灵感和中肯的建议，替我厘清了许多困惑的概念。感谢刘逸博师妹对我的鼓励和发自远方的问候。

感谢家人一直以来对我的关心，最想要感谢父母的无私付出和支持，是你们的关爱让我在迷茫时不自怨自艾，是你们的鼓励让我能够砥砺前行。

最后，感谢在百忙之中参与审阅、评议本论文的各位老师，谢谢你们。

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：_____

附录 A　第 4.3.1 节 $q(c|x) = \mathbb{E}_{q(z|x)}[p(c|z)]$ 的证明

由于 $\mathcal{L}_{ELBO}$ 可表示为：

$$
\begin{aligned}
\mathcal{L}_{\text{ELBO}}(\mathbf{x}) &= E_{q(\mathbf{z},\mathbf{c}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z},c)}{q(\mathbf{z},c|\mathbf{x})}\right] \\
&= \int_{\mathbf{z}} \sum_{\mathbf{c}} q(\mathbf{z},c|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|c)p(c)}{q(\mathbf{z},c|\mathbf{x})} d\mathbf{z} \\
&= \int_{\mathbf{z}} \sum_{c} q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(c|\mathbf{z})p(\mathbf{z})}{q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x})} d\mathbf{z} \\
&= \int_{\mathbf{z}} \sum_{c} q(c|\mathbf{x})q(\mathbf{z}|\mathbf{x}) \left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \log \frac{p(c|\mathbf{z})}{q(c|\mathbf{x})}\right] d\mathbf{z} \\
&= \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \sum_{\mathbf{c}} q(c|\mathbf{x}) \log \frac{q(c|\mathbf{x})}{p(c|\mathbf{z})} d\mathbf{z} \\
&= \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) D_{KL}(q(c|\mathbf{x}) \| p(c|\mathbf{z})) d\mathbf{z}
\end{aligned}
\tag{A-1}
$$

式 (A-1) 中第一项与 c 无关，而第二项非负，因此关于 $q(c|x)$ 最大化 $\mathcal{L}_{ELBO}$ 意味着 $D_{KL}(q(c|x)||p(c|z)) = 0$，因此有

$$
\frac{q(c|x)}{p(c|z)} = u \tag{A-2}
$$

其中 u 为常数。又因为 $\Sigma_c q(c|x) = 1$ 且 $\Sigma_c p(c|z) = 1$，则有

$$
\frac{q(c|x)}{p(c|z)} = 1 \tag{A-3}
$$

两边同时取期望，则有 $q(c|x) = \mathbb{E}_{q(z|x)}[p(c|z)]$。

附录 B　SubX 聚类所得各类簇话题线索示例

表 B.1　Topic threads of cluster 1

关键词:　信号　手机　探测　机主　连接　关掉　信息　网络　号码　名单	
A: 你说得对， A: 每个雇员的电脑里， A: 都有程序是在偷偷运行的， A: 复制文件, 电子邮件，所有按键... A: 然后把数据传到阔恩的电脑里。 B: 好家伙，网络是公司的， B: 一切资料也是它的！ A: 虽然卑鄙, 但却合法... B: 知识就是力量！	A: 这是改性芯片，这次是硬件入侵吗？ B: 没错，这些芯片覆写了系统里的软件， B: 黑客就可以用器械为所欲为， B: 在本案中就是把恶意软件， B: 装到了整个医院的网络中。 A: 这点我懂了， A: 为什么用智能电视呢？ A: 它本身就有硬盘目标， A: 干嘛不直接把恶意软件装到上面呢？
A: 他们为啥还用 DDR 3 内存啊， A: 他们脑子坏了吗？ B: 就是啊，我也是这么想的。 A: 但他们还加了嵌入式静态存储器... B: 等等，先停一下，” 他们” 是指谁啊？ A: 你在开玩笑？ B: 啊？我真的不是， B: 而这个嵌入式静态存储器， B: 应该能够弥补两家用的两种内存之间， B: 每秒 100 G 比特的带宽差距啊！	A: 就会把所有点击者的 IP 地址发给我。 B: 但这不是一个公共网站吗？ B: 一定会有千万个用户看到这个网页。 A: 从亮闪闪小队的知名度来看， A: 至少会有几十万人看， A: 但是超级缓存数据的魅力在于， A: 在发送每个 IP 地址给我们之后， A: 它会附着到对方的浏览器里， A: 监视他们的网络流通。
A: 本来航班是绝不会靠近那团风暴的。 B: 其他航班有遇到这问题吗？ A: 没有，但我们在复查， A: 如果黑客已经侵入了航空系统， A: 那他很可能有后门入侵网络存储器， A: 所以现在每架在飞的航班都有危险。 B: 拜托！克鲁米！明明有车干嘛走路找虐！ A: 瞧瞧我，你是在约皮尔斯探员吗？	A: 我之前一直都想错了。 B: 这听起来并不好。 A: 不，很好，因为我意识到， A: 这段代码不能帮我们找到罗米思。 B: 我还在等着听好的部分呢，伙计。 A: 这段代码是如何被植入手机的才是关键， A: 这是配对记录， B: 对，是手机最后一次连接的设备的记录。 A: 这些手机都跟同一台设备配对过。

表 B.2　Topic threads of cluster 2

关键词:　恶性　导致　症状　搏　血管　丘疹　冠状　抗　萎缩性　肝素	
A: 她呼吸衰竭，心脏和肝脏已经不行了。 B: 她疼吗? 痛苦吗? A: 我不知道。 B: 血流稳定, 没有血块, 没有扰动。 A: 该死, 关掉吧, 我们在浪费时间。 B: 看右心室！ A: 不是川崎病？ B: 不!	A: 低血糖是真的， A: 她可能患有糖尿病， A: 胰岛素服用过量。 B: 2032 室的病人的随身物品在这里吗？ B: 听着，痉挛的发作能让她在医生检查时有个睡觉的地方， B: 或许还有几顿免费餐。 A: 20 块赌里面有胰岛素。 B: 哇！求你收起来吧！ A: 抽搐怎么说? 她胳膊动了。 B: 为何要装抽搐呢？ B: 痉挛还不够严重吗？
A: 快把探针抽出来。 B: 胎儿的心率刚刚降到 50， B: 子宫开始收缩了。 A: 早产给她吊一针叔丁喘宁。 B: 我们用母育酚安胎阻止了早产。 B: 宫缩现在暂时停止了。 C: 肝脏活检结果为阴性， C: 这肯定是镜式反应综合征， C: 由肚子里的胎儿造成的。	A: 长期超量使用， A: 加上长期过量饮用这个， A: 导致了甲状腺功能减退， A: 直到表现为精神异常才被诊断。 B: 急诊使用的镇静剂导致你发生粘液性水肿。 A: 我还真没想过还有这种可能。 B: 豪斯说生活充满讽刺。 A: 救我的不是豪斯， B: 而是我的储物柜钥匙。
A: 多辛苦啊！ A: 但你挺过来了！ A: 这次也一定能做到！ B: 进入膀胱了，开始注入盐水。 B: 我要到处探测一下， B: 然后把管子拉出来， B: 有点紧。 A: 有压迫感是正常的， A: 再坚持几分钟就行了。 C: 不是，那里是我的胸口， C: 呼吸不了！呼吸不了！ B: 他没事吧? B: 心音模糊，颈静脉扩张。	A: 那样能让我快点出院吗？ B: 不行，不大可能。 B: 不过泰勒会帮你拍个胸部 X 光片， B: 然后我们再做个胸痛检查， B: 我们会尽快的。 C: 谢谢格蕾医生，非常感谢。 D: 看来这还真算件事儿， D: 再跟我说下我为何要拒绝凯丽呀， A: 既然你在这儿， A: 帮我看下这个心电图。

表 B.3　Topic threads of cluster 3

关键字:　真的　喜欢　约会　名字　戒指　男人　伴郎　项链　结婚　孩子	
A: 现在是凌晨 1 点，她已经睡了。 A: 我要向她求婚。 B: 什么？ A: 我觉得我会幸福的。 A: 餐馆赚了些小钱， A: 我刚买了部水上摩托艇， A: 胆固醇也降下来了， A: 可没有她一切都没意义。 B: 莫蒂，我觉得这不太可能， B: 她已经想重新开始了， B: 你先坐吧。	A: 糟糕的妈妈、糟糕的人都一个意思， A: 因为无论她做了什么， A: 只要她是个不称职的母亲， A: 她就是个失败者，对吧？ B: 听我说，当我怀着崔沃斯的时候， B: 我就很清楚自己的情况， B: 所以查尔斯跟我分开时， B: 我放弃了孩子的监护权， B: 因为我希望孩子能在最好的环境下成长， B: 这样看来我不是一个好妈妈。
A: 听着菲比， A: 我... 我非常爱你， A: 但我不想再结婚了， A: 只是我的第一次婚姻太... 太糟糕了， A: 我已经对婚姻失去信心了。 B: 有那么糟吗？ A: 到最后... 有一次她故意便便在我的... B: 哦！那真糟！ B: 但你不认为换个人会不一样吗？	A: 接下来几个月得在这住了。 B: 但我们都有工作， B: 哪有空一天到晚在这照顾她？ A: 我们把年假休掉啊。 B: 我的假期想留给夏威夷，不是地狱！ A: 那我真是没招了。 B: 华仔，我爱你， B: 作为你的妻子，你妈妈的问题 我责无旁贷， B: 所以... 我想离婚。
A: 我很遗憾我们没有... 结婚生子， A: 一起变老， A: 我曾是那么的渴望， A: 我曾那么渴望当你的老婆不顾一切... B: 为什么要告诉我这些? A: 我要试着学会放手， A: 你的灵魂才会安息。 B: 可我已经安息了，Izzie。 A: 我回来全是为了你， A: 永别了，Denny。	A: 说我搬出去和男友住是错误的决定。 A: 他说的对，我现在全都明白了。 A: 为什么那时就不明白呢？ B: 你那时身处爱情迷雾之中， B: 爱情迷雾爱情就像毒品， B: 会让聪明人做出愚蠢的事情。 B: 当迷雾都散开， B: 你瞧瞧周围， B: 你便会想，我这之前怎么想的！

表 B.4　Topic threads of cluster 4

关键词:　凶手　死因　发现　伤口　死者　痕迹　刺伤　击碎　擦净　爪痕	
A: 凶手很专业， A: 双枪夺命， A: 一枪击中脑后， A: 她倒下再开一枪， A: 以防万一， A: 再把她埋在沙子里， A: 只等着水泥车早上来， A: 完成毁尸灭迹。 B: 处决型凶杀， B: 水泥车... B: 有没有人觉得像是黑帮刺杀？ B: 这里可有” 坚实” 的证据啊。	A: 刚刚跟分局警督通了电话， A: 他只知道他们发现货车从昨晚一直停到现在。 A: 贝克特一有消息就会联系你。 B: 货车在绿点。 B: 她在一小时前就应该到了， B: 她还没联系我的唯一原因是... A: 别说了，理查德。
A: 我们在案发现场发现中国香烟。 B: 她给中国人工作吗？ B: 你要么现在告诉我们真相， B: 要么我们把情况交给媒体。 A: 你觉得他们会发表吗？ B: 他们也许不会， B: 但我知道很多博主愿意发表， B: 更别提我那 50 多万微博上的书迷。 A: 我说的绝对不能泄露出去！	A: 你能否把他的头摁稳? A: 我数三下一…二…三! A: 头皮处的伤口毫无规律… A: 颅顶骨合缝上的伤口也是如此， A: 伤口一直延伸到大脑镰处。 B: 大脑没有被刺穿， B: 他之所以能开车都因为这根木棍填塞效应?
A: 我们开工吧。 B: 我该做什么? A: 证物照片记录，记下我们所收集到好吗? C: 几个骑自行车的发现这尸体， C: 验尸官刚到。 B: 派瑞斯，她口袋里的学生证 13 岁。 A: 我讨厌这类的案子， A: 身体的位置和沾到的泥土不一致， A: 在她死后， A: 她的胸前留下血液并沾上泥土。	A: 有目击者吗? B: 没有， B: 这些人都只是听到了声响， B: 一个路人在救护人员赶来前试图帮忙但... A: 车里还有别人吗? B: 没了，就他自己。 A: 你有什么发现? B: 看来他开车的时候在发短信。
	A: 对着第二个歹徒你一共开了几枪？ B: 可能是 5 枪。 B: 在这种情况下很难说清楚。 A: 你确定？ B: 我不确定。 B: 这就是为什么我说可能...

表 B.5　Topic threads of cluster 5

关键词:　工作　生活　改变　痛苦　办公　想要　接受　喜欢　努力　放弃	
A: 升职涨工资。 B: 兄弟，怎么突然说起这个？ A: 珍妮又怀了，我们昨晚得知的。 B: 太棒了！恭喜啊！ B: 是好事吧？ A: 当然太棒了，但我又要多糊一张口了， A: 还得买衣服， A: 更别提大学学费了。 A: 凯文！你能不能就消停一会儿！	A: 你为什么不去？ B: 我去不成，我要准备新工作的面试。 A: 你不是已经有工作了吗？ B: 大家竟然还说你不太关心别人， B: 这份工作比之前好很多， B: 职位是公司副总裁。 B: 这家公司为其他公司做 数据重构和统计保理。 A: 你怎么可能懂这些东西？ B: 那就是我现在的工作。
A: 我没准就在帮《纽约客》写稿了， 拿钱搞笑。 A: 现在的工作也很有趣啦。 A: 明天... 我不必打领带， B: 要是我答应了去美林证券上班呢？ A: 什么？美林证券？ B: 对，我有个客人在那里上班， B: 他说我有炒股票的头脑。 A: 那你怎么不去？ A: 因为当时我不相信。	A: 我们正在研究合同内容， A: 我得挂了，我也爱你。 B: 付你工钱是要你来打扫， B: 不是让你逛商场的，韦斯奇。 A: 你说我是贼吗？ B: 放松点，不就开个玩笑嘛。 A: 该死的，这里冷死了。 B: 再冷也比满屋子的血腥味强， B: 之前总以为我的工作够糟的了。
A: 杰瑞，生意啊，我太激动了！ B: 有一堆的公司都是这样的。 B: 我搞不懂为什么 15 年来 他们就没赢过。 A: 这都多亏了那辆车，你看， A: 史坦布瑞纳基本上是 黎明一破晓就第一个上班的人， A: 他看到了我的车， A: 他以为我是第一个来的， A: 最后走的人是威廉， A: 他也看到了我的车， A: 他以为我一直在熬夜加班， A: 他们以为我一天工作 18 个小时。 A: 把车钥匙锁在车里是 最好的升职方法啊！	A: 她爸妈重修了小木屋， B: 那你就不工作了？ A: 是啊，呃，他们不会知道的， A: 我的车在那儿。 B: 这是个好主意吗， B: 你都还在升职的边缘？ A: 我在那个办公室里才会影响我的运气。 C: 苏・艾伦・米施克要见你。 C: 苏・艾伦・米施克？好吧，让她进来 D: 嗨，伊莲你好。

表 B.6　Topic threads of cluster 6

关键词:　灯节　油　圣诞　讲述　德州　节日　唱　犹太　快乐　故事	
A: 还觉得我是个傻蛋吗？ A: 看！今天是圣诞树的点灯之夜！ B: 那是两周前的报纸。 A: 耶？是谁老把旧报纸丢垃圾桶的？ B: 我好想带凯西去看， B: 没想到居然错过了。 A: 至少你还有伴， A: 我讨厌过节时孤伶伶的情人节。 B: 我的生日转眼就到， B: 然后他们又要点亮那棵圣诞树。 A: 我想找个伴。	A: 但现在是早上 5 点 30， A: 我们已经准你睡懒觉啦。 B: 好嘛，既然我起来了爸， B: 我给你买了生日礼物， B: 这儿，生日快乐！ A: 噢，杰瑞，我该给你买一份礼物。 B: 什么意思？ C: 让你爸清净，是送给他的最好的礼物。 A: 噢！是雷达探测器！ C: 雷达探测器？ C: 我从没有见过你一小时走出过 20 英里， C: 你就像玫瑰碗游行的大元帅。
A: 好，等下，等下！庆祝下！ B: 要不要来点香槟？ A: 香槟？来点吧！ A: 订婚可不是常有的事，来吧！ B: 嗯，好吧，你猜怎么着？ B: 没有香槟...	A: 不想打破你的美梦卡塞尔， A: 不过圣诞老人是不存在的。 B: 不再存在了， B: 你得承认那一脸胡须加大肚腩， B: 这位圣诞老人还挺货真价实的。 A: 如果说” 货真价实” ， A: 是指一位穿着红衣的体重超标人士， A: 希望他带着身份证。 B: 带着呢！还有许多拐杖糖的包装纸。
A: 因为不到一个星期， A: 就是世纪新年狂欢节了。 B: 世纪新年？是千禧年， B: 我去不了，我要和父母去太阳谷渡假。 A: 拜托，说实话... 雪，松树，家庭... A: 那根本就不是节日聚会， A: 船，疯狂的中国烟火超人... A: 我现在都能感受到， A: 好好想想吧！ A: 哦，还有， A: 如果你不来是因为这里的紧张气氛， A: 也许那儿可能有一点过去的热情吧。	A: 看起来烟花禁令不会影响 A: 南方公园的节日庆典， A: 南方公园的湖边人头攒动， A: 居民们和游客们正在准备点燃 A: 世界上最大的火焰蛇， A: 这真是空前绝后的大事。 B: 你们也许都记得火焰蛇， B: 就是这些一点燃就会长出尾巴的小圆片， B: 但是今天这个直径足有一英里， B: 而且有 20 层楼那么高。

表 B.7　Topic threads of cluster 7

关键词:　碧欧泉　馅饼　乳　蜂蜜　蛋糕　酒　没关系　脆饼　松糕　乳酪	
A: 我要给他们提供法式美食， A: 你的肉汁、乳酪、薯条... A: 哈维尔，你要知道， A: 法裔加拿大人对作为巴黎文化的后裔， A: 感到非常自豪， A: 我要给他们带伯纳丁餐厅的油酥面点。 B: 卡塞尔，那是全世界最好的餐厅之一， B: 他们不外卖。 A: 对吕克·德拉库阿也许如此。	A: 我告诉你... 我做得到。 B: 嗯，他做得到。 B: 看起来太美味了， B: 那火鸡，那填料，那酸梅酱呢！ C: 够了！猴子都会做！ B: 听我说，我们真得很抱歉... B: 他又来那什么” 目光接触” 了， B: 别看他。
A: 是的，伙计，很好吃的。 B: 事实上，我不能吃牛肉， B: 我的肠膜有退化性疾病， B: 吃牛肉会让我胀气。 A: 抱歉 Kyle，我该为你做什么? B: 不，我... 我不想麻烦你。 A: 我可以为你做意大利面，或者是冻鱼片。 B: 如果不麻烦的话，我想吃冻鱼片。 A: 我马上就把它放微波炉， A: 怎么了 Kyle 2 号？	A: 没什么， A: 只是感觉应该给我的好朋友， A: 带点中国菜。 B: 你记得我叫你点的鸡里面， B: 莴苣要切块不要切丝？ A: 记得。 B: 即使菜单明确说要切丝。 A: 记得。 B: 红烧饭不要白烧? A: 记得。 B: 在韩国杂货店买质量好的热芥末？ A: 记得。 B: 在超市买低钠酱油? A: 记得。
A: 你还写说我们店里接受信用卡， A: 我们其实没有。 B: 好啊，那段话我可以收回， B: 但我坚持我的观点。 B: 我知道怎么做菜， B: 但决不是你们那样。 B: 你们的大蒜番茄酱， B: 简直跟番茄汁没两样， B: 应该要配伏特加和芹菜茎。 A: 我以店里那酱汁为荣，美味极了。 B: 你是开意大利餐厅的， B: 竟然觉得那样叫做美味？	A: 但你不是总念叨着， A: 要赢个诺贝尔奖吗？ B: 我也念叨过不要在我朋友面前吐槽我， B: 这你丫怎么就没记得呢？ C: 好了，你们想去的话就告诉我。 B: 我愿意， B: 我就很享受陈酿黑皮诺红酒的口感。 C: 我相信那红酒一定很搭你这份炸鸡块。 B: 听起来会很棒，我们应该去。 A: 又是酒吗？

表 B.8　Topic threads of cluster 8

关键字:　动力　发现　系统　实验　外壳　观测　米　下降　加热　干涉	
A: 让我参加日内瓦研究之旅， A: 观测欧洲原子核研究委员会的 超级对撞机！ B: 真不公平！你甚至不是物理学家！ A: 好了，有两个角度看到这个问题。 B: 出去！ A: 拜！你有提高。 B: 谢谢！真人实战确实有助于提高。	A: 请稍等，你这就算问完了吗？ B: 那你还有什么相关的信息要提供吗？ A: 天呐多得去了！ A: 比如说，我笔记本电脑里有五种思维实验， A: 其中四种都是关于量子测量课题的 强有力的重述。 B: 这对他们有何帮助？ A: 这样他们就可以监控科学出版社。
A: 我在用自由电子激光束 A: 做 X 射线散射实验。 B: 是不是激光束碰巧把你视网膜烤焦了？ A: 没。 B: 那你就能开车了？快走。 A: 我不是告诉你了么， A: 我现在是晚上工作， A: 你要另外想办法了。 B: 是啊，所以我没办法。	A: 主要是我所认识的物理学家 A: 都又白又宅， B: 我不是宅男， B: 我的防晒霜很好用， B: 所以黑色素沉淀得很少。 A: 那你是在和莱纳德一起做实验么？ C: 事实上，我们在... B: 我们在检测光电倍增管的辐射强度， B: 用在一种新的暗物质探测器上面的。
A: 不会留下任何活口， A: 这个地方并不安全。 B: 我猜这是个实验室。 C: 你怎么知道? B: 这些是他们从星际之门 带回来研究的东西， B: 都有标签。 B: 这是特坎尼的面具， B: 这是拉葛许的黏土圆锥， B: 上面有楔形文字。 A: 卡特, 丹尼尔! A: 找到纪念品店了，我们该走了！	A: 说真的，恭喜你谢尔顿。 A: 我在网上看到了你的论文， A: 创造一种新的重元素的方法， A: 相当启发人。 B: 谢谢！ A: 信不信由你，我刚得知， A: 中国湖北核子物理研究所的研究小组， A: 在回旋加速器上进行了测试， A: 结果相当喜人！ A: 谢尔顿，太不可思议了！ B: 我知道。

表 B.9　Topic threads of cluster 9

关键词:　召唤　背叛　女巫　族群　魔法　主人　复仇　选择　怜悯　百年	
A: 召唤时间神不是问题， A: 但召唤的时候， A: 迪恩得在他旁边， A: 严格地说，要把手放那家伙身上， A: 就能跟着他一起回来了。 B: 不是自动就能回来吗？ A: 不是，我们需要找准他们那边的时间， A: 差一秒都不行， A: 不然回来的就是愤怒的大神 而不是你哥了。 B: 而我哥就永远困那儿了？	A: 哪怕只有万分之一的可能， A: 她也必须死， A: 否则整个王国将会灭亡。 B: 我无法理解。 A: 有朝一日，你会成为一名国王， A: 那时你就会明白了， A: 你不得不作出这样的决定， A: 黑暗的力量正威胁着这个王国。 B: 父王，我明白巫术是邪恶的， B: 但失去公正也是错误的。 A: 没错，我也不知道自己 会成为怎样的国王。
A: 能够证明我们对付的是个恶灵的证据， A: 在死人附近生长的草， A: 特别是没有安息的死者， A: 别处可没看到这种草， A: 除了被谋杀的神父的墓碑上。 A: 就是他，Sam！ B: 也许... A: 好吧，想要更多证据? A: 那我就给你更多证据。 B: 怎么? A: 召唤 Gregory 的鬼魂。 B: 什么?	A: 你说整个灵媒镇有多少颗水晶球？ B: 大概在五十到全世界所有之间吧。 A: 水晶可以充当召唤幽灵的天线，对吗？ B: 这也是灵媒会开始使用水晶球的原因。 A: 也就是说，这镇上的每一个灵媒店， A: 都装了一个鬼魂卫星天线？ B: 没错，这地方到处是招鬼的人。
A: 那个囚犯应该对他们的死负全责， A: 我要立刻处决他，在您传令之前。 B: 还是稍等片刻吧。 A: 恐怕您所看到的伊森爵士， 以及奥斯瓦德爵士， A: 都不是他们的真面目巫术。 B: 好吧，我又欠了高汶一条命。 B: 你参与混战的事情， B: 国王打算当作不知道。 A: 太棒了！谢谢你，亚瑟！	A: 你觉得她能挺过来吗？ B: 不行，绝不容易。 B: 哦，你看上去刚从地狱回来。 C: 你去试试整晚驱魔会有什么感觉。 C: 那个美女和胖家伙他们活着， C: 虽然后半生都要用来还治疗费了， 但毕竟… A: 什么样的刀竟能杀死恶魔？ B: 我还是得再问一遍… B: 那个神秘女孩是谁？ B: 只要能帮你摆脱这个噩梦， B: 需要什么尽管说! A: 还放出了什么更可怕的恶魔？

表 B.10　Topic threads of cluster 10

关键词:　出价　买了　钱　账户　支票	美元　信用卡　元　账单　慷慨
A: 我联系了银行， A: 提醒了他们这事， A: 他们正在做内部调查， A: 但假如我没找到什么， A: 他们也找不到。 B: 我都不记得， B: 上次用这张借记卡是什么时候了。 A: 你上次用是什么时候？ B: 最近没用，大概是几个月前吧。	A: 相信我，我也想啊。 B: 经济低迷时期， B: 人们的超前消费行为你觉得该由谁负责？ A: 谢谢，我接受大多数经济学家的说法， A: 超前消费是可以接受的， A: 只要消费的对象是公共投资， A: 比如基础设施、国防建设、教育研究等。 B: 哇塞！这么专业！哪学来的啊？ A: 纽约大学一学期的经济学理论课， A: 这问题又是从何而来？
A: 那些月光族， A: 通常都是最后才填最大一笔支票。 B: 是啊，比如房租， B: 我把钱攒起来， B: 然后付房租， B: 指望我的薪水， B: 能在房东把我的房租支票存起来之前付清。 A: 是的，但是你不知道的是， A: 银行在处理大额租金支票前， A: 根本不会去支付小额生活缴费， A: 这就是陷阱。	A: 好吧，要不... B: 要是拥有意大利的餐馆怎么样？ A: 什么？ B: 它就既不是钱，也不是礼物， B: 而是投资你将有 15 % 的股份。 A: 是个够诱人的提议。 B: 要是经营得好你还能分到部分利润， B: 要是做垮了都算我们的， B: 你投资的两万还是你的。
A: 罗宾汉已经占领你们的网络了， A: 他想什么时候解冻账户都可以。 B: 看！伊利亚，看到了吗？ B: 你的名字也在名单上！ A: 不是我。 B: 你收到罗宾汉给的银行退款了吗？ A: 没有，我账户里有远远不止 200 块， A: 好尴尬， A: 最近两周花了四千元买花。	A: 今年卖了三万个， A: 一个二十刀， A: 在拉斯维加斯的黑帽大会上都卖脱销了。 B: 你有买了这些面罩的人的信用卡记录吗？ A: 信用卡？你说笑呢！ A: 把他们的宝贵的电子信息泄露出来， A: 不可能绝不。 A:” 禁止人肉” 只用现金交易， A: 我的客户都以匿名为荣。

个人简历、在学期间发表的学术论文与研究成果

个人简历

1993 年 12 月 08 日出生于四川省乐山市。

2012 年 9 月考入四川大学数学系基地班专业，2016 年 7 月本科毕业并获得理学学士学位。

2017 年 9 月进入清华大学计算机系攻读工学硕士学位至今。

发表的学术论文

[1] Leilan Zhang, Qiang Zhou. Automatically Annotate TV Series Subtitles for Dialogue Corpus Construction. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1029-1035. (EI 收录, EI 检索号:20201308362207)

[2] Leilan Zhang, Qiang Zhou. Topic Segmentation for Dialogue Stream.2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1036-1043. (EI 收录, EI 检索号:20201308362204)