自动构建基于电视剧字幕和剧本的日常会话基础标注库

梁宇海[1]，周强[2]

（1.北京邮电大学，北京 100876；

2.清华大学信息技术研究院语音和语言技术中心，北京信息科学与技术国家研究中心，北京 100084）

摘要：真实人类对话数据量不足已经成为限制数据驱动的对话生成系统性能提升的主要因素，尤其是汉语语料。为了获得丰富的日常会话语料，可以利用字幕时间戳信息把英语电视剧的英语字幕及其对应的汉语字幕进行同步，从而生成了大量的汉英双语同步字幕。然后通过信息检索的方法把双语同步字幕的英文句子跟英语剧本的演员台词进行自动对齐，从而将剧本中的场景和说话者信息映射到双语字幕中，最后得到含有场景及说话者标注的汉英双语日常会话库。利用这种方法，我们自动构建了包含 978,109 对双语话语消息的接近人类日常会话的多轮会话数据库 CEDAC。经过抽样分析，场景边界的标注准确率达到 97.0%，而说话者的标注准确率也达到 91.57%。该标注库为后续进行影视剧字幕说话者自动标注和多轮会话自动生成研究打下了很好的基础。

关键词：日常会话语料；电视剧剧本解析；双语字幕同步；剧本和字幕的自动对齐

中图分类号：TP391　　**文献标识码：**A

Automatically Build the Basic Annotated Daily Conversation Corpus Based on the Subtitles and Scripts of TV Plays

Yuhai Liang[1], Qiang Zhou[2]

（1.Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. CSLT, RIIT, BNList, Tsinghua University, Beijing 100084, China）

Abstract: The insufficient human dialogue corpus has been a key factor restricting the performance of dialogue generation system, especially the Chinese dialogue corpus. In order to obtain rich corpus, time-stamps can be used to synchronize English subtitles and corresponding Chinese subtitles, so that abundant Chinese-English bilingual subtitles can be generated. Then, align the bilingual subtitles and the utterances in the corresponding English scripts, so that the tags of speaker and scene in the scripts can be mapped to each pair of sentences in the bilingual subtitles. Through this method, CEDAC, which is a multi-turn dialogue corpus and is approximate human daily conversation, was generated with 978,109 pairs of Chinese-English bilingual utterances. The experimental result shows it achieves the accuracy of 97.0% on scene boundary annotations and 91.57% on speaker annotations. The corpus lays a good foundation for the following research on automatically annotating speakers of subtitles and multi-turn dialogue automatic generation system.

Key words: Daily dialogue corpus; Parsing of TV play scripts; Synchronization of subtitles; Automatic alignment between scripts and subtitles

[1] 该论文是第一作者在清华大学实习时完成的工作

[2] 论文通讯作者是周强

相关的NSFC资助信息：国家自然科学基金资助项目 (61433018,61373075)

0 引言

随着对话生成技术的发展，人们迫切希望得到大规模的多轮会话语料，但是目前公开的人类日常会话语料并不多，汉语语料尤为稀缺。

目前学术界最常用的是从社交媒体上获取的语料，例如微博[1]和贴吧[2],它们是以消息-答复对的形式组织的，跟人类日常会话有很大差别。除此之外，其它开放的会话语料要么是任务导向型的，如 Medical DS[3]是关于医疗诊断的会话、CASIA-CASSIL[4]是电话会话、Ubuntu[5]是技术咨询会话等等，要么数据量比较小，如 DailyDialog[6]。

而电视剧字幕比较符合我们“真实会话”和“数据量大”的需求。因为它们记录了演员在特定场景下的对话信息，其描述风格比微博、论坛等社交媒体用语更接近正常的人类日常会话。此外，它数量巨大，Lison P 等人[7]研发的 OpenSubtitles2018 数据集收集了多达 208,000 个电影和电视剧剧集的字幕，其中的汉语话语消息多达 3120 万条。

但在字幕文件中，每一条字幕数据仅由字幕 id、字幕起止时间戳和话语消息组成。虽然其数量庞大，但说话者信息的缺失使研究人员无法判断两句连续的字幕句子是否属于同一个人说的，而场景信息的缺失则使研究人员无法确定两句连续的字幕句子是否属于同一个场景下的。所以从中提取到的对话片段经常带有噪声，无法单独由字幕文件得到理想的多轮对话数据。

对此，Lison P 等人[8]提出了配套剧本信息对字幕自动添加说话者标注的处理方法，使用句子对齐工具将剧本的说话者标签添加到字幕上。但是，这套方法却不适用于汉语字幕。因为高质量的汉语剧本资源非常缺乏，即使有大量的汉语字幕文件，仍然无法找到相应的可利用的汉语剧本。

为此，我们提出了基于英语剧本和汉英双语同步字幕的汉英双语日常会话标注库的构建方法，可以总结为三步：1）解析英语剧本，提取话语消息和场景、说话者标注信息。2）同步汉语字幕和英语字幕，生成汉英双语同步字幕。3）对齐汉英双语同步字幕和剧本话语消息，把剧本中的说话者和场景标注信息映射到汉英双语同步字幕上。

在第二步中，如果剧本已经存在对应的汉英双语字幕，就不需要进行字幕同步了。但是大部分的英语影视剧不存在汉英双语字幕，只存在单英语字幕及其翻译字幕（单汉语字幕），所以要先进行第二步操作。

这样，我们就可以利用网络上丰富的英语剧本获得大规模的汉英双语会话标注库。

在下面的几节中，第 1 节对相关研究进展进行简要评述，第 2 节介绍我们的基本思路，给出相关术语定义和会话标注库构建的总体流程，第 3 节详细说明三个主要步骤的处理细节，第 4 节是实验结果及统计分析，最后的第 5 节给出论文工作的总结和展望。

1 相关工作

虽然目前研究人员收集了很多会话语料，但是数据量庞大的真实日常会话语料并不多。Wang H 等人[1]从新浪微博上收集了大量的消息-答复对语料。其中绝大部分是单轮对话，与人类日常的多轮会话有很大差别。而且，不同的消息-答复对通常形成不同的话题结构，这也有别于多人真实会话中的话题延续和转换情况。这些原因都使它不适合用于多轮对话生成系统训练。其他社交媒体的语料，例如英语的 Reddit 数据集[9]，汉语的豆瓣数据集[10]、百度贴吧数据集[2]，都跟它大同小异。

此外，还有很多数据量庞大的任务导向型数据集，例如英语的 Ubuntu 数据集[5],汉语的 Medical DS[3]等。这些数据集话题比较单一，且多以问答的形式推进，只适用于目标导向的对话生成系统，并不适合用于通用对话生成系统。

DailyDialog 数据集[6]是基于日常真实会话构建的英语数据集，有丰富的话题和多种交流形式，噪音较少。但它是基于手工标注开发的，目前数据规模约 10 万条话语消息，难以进一步扩增。值得注意的是，Li Y 等人[6]指出真实的日常会话通常有 3 个或多个话题，而 DailyDialog 每个会话平均有 8 个话轮，适合用于训练紧凑的多

轮对话生成模型。因此我们可以把会话平均话轮数作为判断多轮会话库数据分布是否合理的一个参考指标。

为了得到数据量庞大的接近人类日常会话的语料，研究人员注意到了影视剧字幕和剧本的潜在应用价值。OpenSubtitles2016[11]是具有多达 60 种语言的字幕数据集，其中汉语话语消息有多达 2480 万条。在 OpenSubtitles2018[7]中，汉语话语消息已经增加到 3120 万条。可见，字幕数据量不仅庞大，而且还在逐年快速增长。

但是，OpenSubtitles 字幕数据集没有场景划分标注，通常在一个字幕文件中有多达 150 个的话轮，相比于 DailyDialog 数据集的平均 8 个话轮，不太适合用来直接训练对话生成模型。

在剧本方面，Banchs R E[12]从 753 部英文电影的剧本中提取了对话，包含了 764,146 条话语消息。但是剧本的对话表达相对生硬和书面化，有时候不易理解，所以仍旧是经过演员演绎的字幕更符合日常会话。而且，除英语以外的其他语言剧本都比较稀缺，包括汉语剧本。

因此，研究人员开始将剧本和字幕的信息进行融合。Lison P 等人[8]使用句子对齐工具将剧本的说话者标签添加到字幕上，从而获得多种语言的平行标注语料，Wang L 等人[13]使用了信息检索的方法完成了这个工作，然而目前他们都没有公开数据集，而且他们只利用了网络上的双语字幕，更多的单语字幕却没有被利用，所以他们形成的数据集规模都不大。值得借鉴的是，在 Wang L 等人的研究中，使用了 TF-IDF 构造句子特征，然后在滑动窗口下计算字幕句子和剧本句子之间的余弦相似度，由此实现剧本与字幕的对齐。最终在电视剧“Friends”上得到 81.8%的说话者标注准确率，98.6%的场景边界标注准确率。可见，说话者标注准确率还有较大的改进余地。

而我们在计算字幕句子和剧本句子之间的相似度上采用 BM25 算法[14]，它在信息检索中有着比 TF-IDF 更好的性能。同样，为了提高对齐的准确率，我们也使用了滑动窗口。最终在电视剧“Friends”上得到了更好的说话者标注准确率，为后续对说话者自动标注的研究打下很好的基础。

此外，带有标注的汉英双语会话数据集还有 NUS-SMS-Corpus[15]，但它来源于短信息，会话简短，基本是单轮会话，不适合用于会话生成系统。

2 总体方案

在网络上公开的汉语剧本大部分是不完整的，而且格式上也不统一，难以收集和提取完整的会话数据。所以即使我们有大量的汉语字幕，仍然无法像构建英语会话库那样可以直接融合剧本信息。但我们另辟蹊径：既然英语影视剧剧本比较丰富，那么只要先同步影视剧相应的英语字幕及其翻译字幕（汉语字幕），再对齐对应的英语剧本话语消息，就能把英语剧本中相关的说话者和场景信息映射到汉英字幕上，从而得到具有丰富标注信息的汉英双语会话标注库。

因为英语字幕和汉语字幕都需要与相应的视频播放画面同步，所以它们的播放时间大致上应该相同。利用字幕文件中每条字幕的播放起止时间戳信息，就可以很好地同步每一条英语字幕和其对应的汉语字幕，形成汉英双语同步字幕。

在对齐剧本与字幕方面，我们使用了信息检索技术，它主要衡量一个查询语句跟多个文档之间的相关性。我们则把每一对汉英双语同步字幕中的英语句子当作查询语句，英语剧本中的话语消息当作相关文档，计算两者相关性得分。从中选择相关度最大的剧本话语消息作为该对汉英双语同步字幕的匹配预测，从而实现两者的自动对齐。最后把该剧本上的场景和说话者标签映射到这对汉英双语同步字幕上就可以实现对汉英字幕消息的自动标注。

下面对上述任务描述中涉及的几个主要术语概念给出如下简要定义描述：

- 剧本/脚本(Script)：是剧作者记录不同场景内容的书面文本形式。其描述格式相对规范，以场景内容组织，每条对话内容行一般附有说话者 id。
- 字幕(Subtitle)：实时记录演员的对话表演内容的书面文本形式。其基本单位由记录视频起止时间的时间行和记录对话信息的内容行组成。
- 播放时间(Playtime)：某字幕内容行在视频中出现的起止时间间隔。

- 场景(Scene)：是在一个特定时空中发生的一个或多个动作或事件。
- 话语消息(Utterance)：会话库中以文本形式记录的会话参与者在会话过程中发出的一条言语信息。大致对应剧本中的一条对话内容行和字幕中的一条字幕内容行。
- 说话者(Speaker)：发出话语消息的人，标注信息之一。
- 话轮(Turn)：是指会话过程中某一时刻获得说话机会的某个说话者所说的话。大致可以表示为：说话者+话语消息。

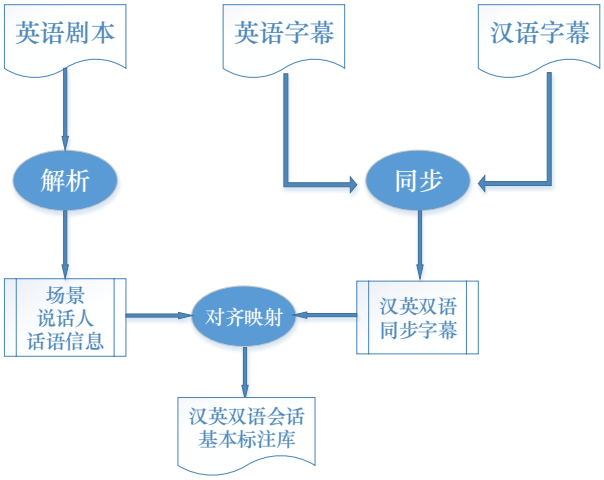


图 1 汉英双语会话基础标注库构建流程图

构建会话标注库的流程如图 1 所示，我们的目标是获得带场景信息和说话者标注的汉英双语会话标注库。原始材料是英语剧本、英语字幕和汉语字幕，经过剧本的解析得到标注信息和用于对齐的话语消息，经过字幕的同步得到了汉英双语同步字幕，最后通过两者共同的英语句子作为联系桥梁，把标注信息映射到双语字幕，实现最终的会话标注库构建目标。

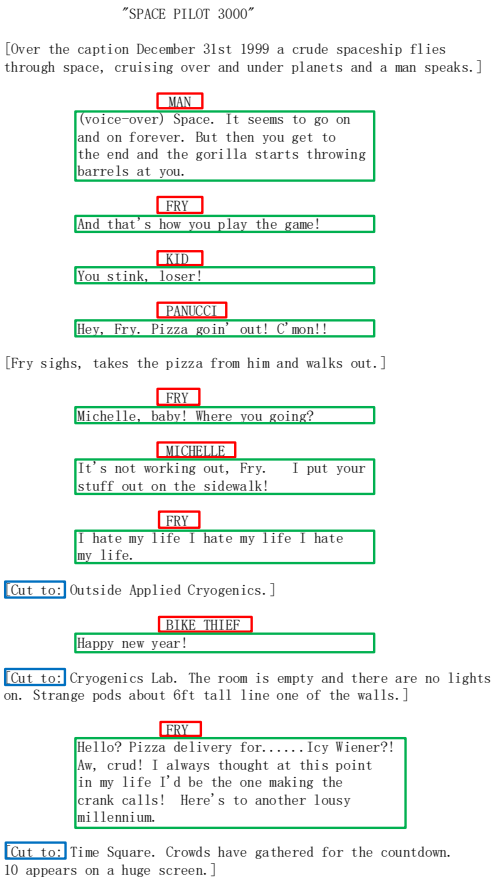
最终得到的汉英双语会话标注库设计成文件、场景和话语消息三个层次的标注结构。首先，我们将电视剧的一集组织为一个文件，便于与原始数据内容对接。在文件层面包含一个或多个场景，用场景在剧集内部的序号作为场景 id(SceneID)，并保留场景的时空内容描述信息(SceneDescription)；在场景层面包含一个或多个话语消息(Utterance)，用话语消息在该场景内部的序号作为话语消息 id(UtteranceID)，同时标注该话语消息的说话者(Speaker）。

为便于记忆，我们将把构建完成的汉英日常对话库简记为 CEDAC(Chinese-English DAily Conversation Corpus)。

3 构建方法

这一节我们详细介绍汉英日常会话标注库构建各步骤的主要算法。首先介绍剧本的收集和解析的方法，然后说明同步字幕的方法，最后介绍剧本与字幕对齐的算法和策略。

3.1 剧本收集与解析


"SPACE PILOT 3000"

[Over the caption December 31st 1999 a crude spaceship flies through space, cruising over and under planets and a man speaks.]

MAN
(voice-over) Space. It seems to go on and on forever. But then you get to the end and the gorilla starts throwing barrels at you.

FRY
And that's how you play the game!

KID
You stink, loser!

PANUCCI
Hey, Fry. Pizza goin' out! C'mon!!

[Fry sighs, takes the pizza from him and walks out.]

FRY
Michelle, baby! Where you going?

MICHELLE
It's not working out, Fry. I put your stuff out on the sidewalk!

FRY
I hate my life I hate my life I hate my life.

Cut to: Outside Applied Cryogenics.]

BIKE THIEF
Happy new year!

Cut to: Cryogenics Lab. The room is empty and there are no lights on. Strange pods about 6ft tall line one of the walls.]

FRY
Hello? Pizza delivery for.....Icy Wiener?! Aw, crud! I always thought at this point in my life I'd be the one making the crank calls! Here's to another lousy millennium.

Cut to: Time Square. Crowds have gathered for the countdown. 10 appears on a huge screen.]

图 2 剧本示例

我们选择电视剧为主要数据源，原因主要有两点：1)电视剧剧集的剧本格式比较统一，相比之下，不同电影之间剧本格式往往不太一样，解析电影剧本工作量会很大，难以快速地扩增我们的会话标注库；2)可以获得英语剧本的电影通常是早期的电影，没有对应的汉语字幕，而近期的电影，又通常缺乏剧本。与此相反，能找到剧本的电视剧，基本上都能找到汉语字幕，这得益于网络上的各个电视剧翻译组和喜爱者的贡献。

我们的电视剧英语剧本爬取于 IMSDB[3]、livejourna[4]等网站。一个电视剧有多个剧集，每个剧集对应一个剧本文件，每个剧本包含一个或多个场景，每个场景包括一个或多个话轮，如图 2 所示。

图 2 是电视剧“Futurama”的“Space Pilot 3000”剧集的剧本，蓝色框的内容为场景切换标签，后面一般紧接场景内容描述。红色框的内容为说话者，绿色框的内容是话语消息。

虽然并不是所有剧本都具有一样的格式，但总得来说，都是使用正则表达式匹配场景切换标签、说话者和话语消息。它们的特征如下：

1) 场景切换标签：[EXT. SceneDescription]，[INT. SceneDescription]，[Cut to: SceneDescription]，(FLASHBACK)等，两个切换标签间的内容作为一个场景内容。
2) 说话者：单独作为一行并居中；跟话语消息同一行并后接冒号，如“Tony：I want to go”；全部字母大写，如“TONY”。
3) 话语消息：紧接在说话者下一行，所以两个说话者中间的内容即为前面说话者的话语消息；与说话者同一行，位于冒号后面的内容。需要注意的是，话语消息里掺杂的说话者动作神情的描述需要过滤掉，这些描述通常用小括号或者中括号对括起来。

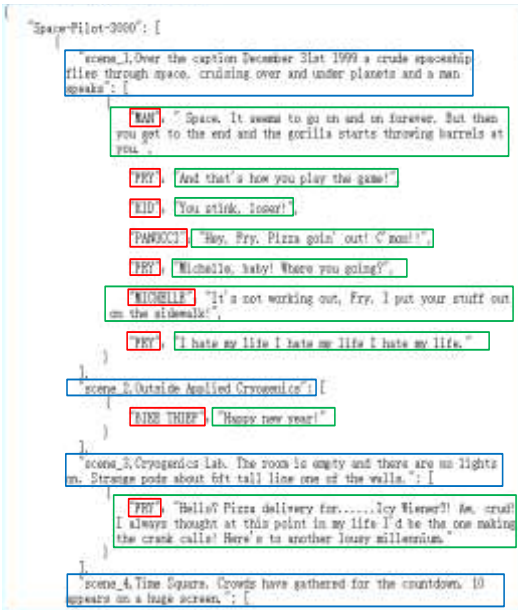


图 3 解析后的剧本信息示例

最终，根据以上特征提取相应的场景内容描述(SceneDescription)、说话者(Speaker)及其话语消息(Utterance)，并以场景 id(SceneID)为顺序组织起来得到解析后的剧本，如图 3 所示。蓝色框的内容即为场景信息，包括场景 id 和场景内容描述，红色框的内容为说话者，绿色框的内容是话语消息。例如，场景 id 为“scene_2”，描述为“Outside Applied Cryogenics”的场景下有名为“BIKE THIEF”的说话者，他的话语消息为“Happy new year!”。

3.2 字幕收集与同步

汉语字幕主要来自于射手网（伪）[5]。如果字幕原本就是双语字幕，我们可以省去同步字幕的步骤，但大部分电视剧没有双语字幕，因此我们不仅需要收集该电视剧的汉语字幕，而且需要收集相应的英语字幕以进行字幕同步。英语字幕的收集途径除了射手网（伪），还有 opensubtitles.org 网站[6]。

原始的字幕文件中，每条话语消息都有其起止时间戳，如图 4 所示。其中子图（a）是英语字幕，子图（b）是其对应的汉语字幕。我们可以看到，对应的两条字幕的播放时间是完全一致的，比如，子图（a）中的“Happy New Year!”的播放时间和子图（b）中的“新年快乐”的播放时间是完全一致的，这是因为它们都需要跟播放画面同步。基于此信息，我们可以对它们实现对齐。

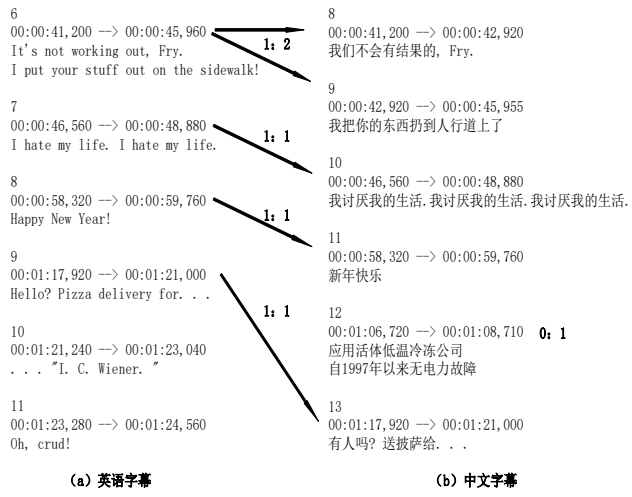


图 4 单语字幕示例

[3] https://www.imsdb.com/
[4] https://scriptline.livejournal.com
[5] http://assrt.net/
[6] https://www.opensubtitles.org/en/search/subs

但是，在翻译字幕当中，由于语言表达差异或者字幕格式的原因，通常还存在其他不太理想的情况，比如子图（a）中的第一条字幕对应子图（b）中的第一和第二条字幕，这是 1:2 的对齐类型；此外，子图（b）中的 id 为“12”的字幕是电视剧中的旁白，而在子图（a）中并没有收录，所以找不到对应的句子，这是 0：1 的对齐类型。此外还有 1：0 型，2：1 型等。理想的对齐类型是 1：1，即一条英语字幕对应一条汉语字幕，这也是正常情况下大部分句子的对齐类型。

除了有多种对齐类型外，很多对应的汉英字幕的播放时间也并不完全一致，可能会错开一个时间差，如图 5 的子图（a）所示，或者两者的播放时长并不相等。如图 5 的子图（b）所示。因此，我们并不能简单地用时间戳是否相等来判断句子之间是否匹配。

但是我们可以使用两条句子播放时间的重叠程度来判断这两条句子是否对应。只要这两个句子时间重叠的部分足够大，那么这两个汉英字幕大概率是对应的。我们设置两个阈值参数α和β来判断重叠部分是否足够大，如果重叠部分在该英语句子播放时间中占比满足阈值α，在该汉语句子播放时间中占比满足阈值β，则把这两个句子进行同步。

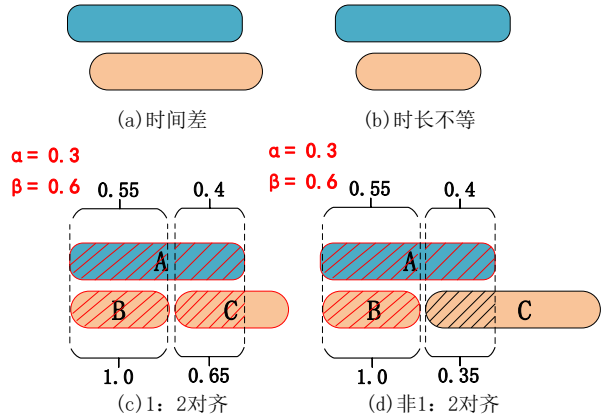


图 5 字幕播放时间示意图（蓝条代表英语字幕播放时间，红条代表汉语字幕播放时间）

经过实验验证，对阈值进行如下设置会取得较好的同步效果。设置两组α和β阈值，一组阈值是$\alpha = 0.3$，$\beta = 0.6$，另一组阈值是$\alpha = 0.6$，$\beta = 0.3$，两组中的任何一组被满足，就同步该对字幕。原因如下：

（1）一般来说，1：1 类型的两个句子时间重叠比率都接近 1，可以轻易地满足两组阈值，从而实现同步。

（2）1：0，0：1 类型的句子通常是噪声，需要过滤掉。通常它跟对应字幕里的句子的重叠比率都为 0 或者接近 0，远远不会满足两组阈值，从而会被过滤掉。

（3）1：2 对齐类型也可以很好地实现。以图 5 中的子图（c）为例子，句子 A 与句子 B 的重叠部分占句子 A 时长的 0.55，占句子 B 时长的 1.0，满足第一组参数，同时，句子 A 与句子 C 的重叠部分占句子 A 时长的 0.4，占句子 B 时长的 0.65，也满足第一组参数，所以句子 A 可以与句子 B 和 C 实现 1：2 对齐。

需要说明的是，较大的β能有效地防止错误对齐，因为 1：2 对齐意味着一个英语句子对应两个汉语句子，那么这两个汉语句子都应该相对较短，β取值较大的目的就是为了选出短句子，过滤掉长句子。如子图（d）所示，长句子 C 并不满足 0.6 的β阈值，所以会被过滤掉。

（4）同理，2：1 对齐类型的句子都会满足第二组参数，且保证一个长汉语句子对应两个短英语句子。

总得来说，其中一个阈值设置得较小，是为了捕捉 1：2 类型和 2：1 类型的“2”，把另一个阈值设置较大是为了保证“2”是两条短句子。从而保证 1：2 类型和 2：1 类型的准确率。


```
"00:00:06,600 --> 00:00:10,280": [
  "Space, it seems to go on and on forever.",
  "宇宙，看似永无边际"
],
"00:00:10,640 --> 00:00:13,800": [
  "But at the end, a gorilla throws barrels at you.",
  "但你会到达尽头，然后大猩猩开始向你扔水桶"
],
"00:00:19,840 --> 00:00:24,560": [
  "-That's how you play the game. -You stink, loser.",
  "这个游戏就是这么玩的 -你水平真臭，失败者"
],
"00:00:24,800 --> 00:00:28,200": [
  "Fry! Pizza going out. Come on!",
  "Fry，披萨好了，快点！"
],
"00:00:38,560 --> 00:00:40,800": [
  "Michelle! Baby! Where you going?",
  "Michelle，宝贝！你要去哪儿？"
],
"00:00:41,200 --> 00:00:45,960": [
  "It's not working out, Fry. I put your stuff out on the sidewalk!",
  "我们不会有结果的，Fry，我把你的东西扔到人行道上了"
],
"00:00:46,560 --> 00:00:48,880": [
  "I hate my life. I hate my life.",
  "我讨厌我的生活，我讨厌我的生活，我讨厌我的生活。"
],
"00:00:58,320 --> 00:00:59,760": [
  "Happy New Year!",
  "新年快乐"
],
"00:01:17,920 --> 00:01:21,000": [
  "Hello? Pizza delivery for. . .",
  "有人吗？送披萨给。。。"
```


图 6 汉英双语同步字幕示例

一般情况下，该算法可以得到较好的结果，但有时候，英语字幕和汉语字幕从一开始就存在一个较大的时间差，也就是说，后面所有对应的句子都完全不重叠，这时候它们匹配到的句子基本上是错误的句子。对此的解决办法是，找到两个参考点，计算出这个时间差，然后给英语或者汉语字幕中的句子补上这个时间差即可。虽然目前还需要手动地选取参考点，但是通过少量的人工干预，我们就可以得到高质量的对齐字幕，如图 6 所示。这是构建会话标注库中很关键的一步。

这里需要说明的是，是否需要手工矫正主要依赖于该汉英字幕的质量。可以分为如下情况：

(1)如果汉英字幕时间戳完全一致，显然不需要手工矫正；

(2)如果汉英字幕时间戳存在较小的时间差，也不需要手工矫正。因为阈值条件给预了时间戳一定的“容错”，只要时间差不足以让对应的汉英字幕失匹，可以不进行手工矫正时间戳。

(3)如果时间戳存在时间差过大，致使对应的汉英字幕的播放时间完全不对应，那么我们就需要手工矫正，弥补时间差。

因此，我们应该尽量找到时间戳相互对应的高质量汉英字幕，这样可以减少人工干预的工作量。目前在整个制作基础标注库的过程中，涉及手工矫正的字幕有 521 个，占比 29.49%。

3.3 剧本和字幕的自动对齐

从 3.1 和 3.2 中，我们分别得到了解析后的剧本和汉英双语同步字幕，接下来，我们用信息检索的方法实现两者话语消息的自动对齐，然后将剧本的标签信息映射到汉英双语同步字幕即可。

通常，剧本的话语消息跟字幕文件里演员的表述不尽相同。剧本中的长句子会因为屏幕大小的限制，被切分成多条较短的字幕句子来表达。有些剧本中的短句子也可能被演员临时发挥，口语化地表达为多条字幕。因此，字幕和剧本的对齐可能包括一对多、多对一或者多对多的情况。

信息检索的技术适合用来解决这个问题。信息检索技术可以衡量一个查询语句跟多个文档之间的相关性。于是，每对汉英双语同步字幕的英语字幕可以作为查询语句输入，剧本话语消息可以作为多个文档，返回剧本中与这条英语字幕相似度最高的一条话语消息作为输出即可实现对齐。

具体地，我们用 BM25 算法[14]计算相似度大小，选取相似度最大的剧本话语消息作为结果，将其对应的场景 id 和说话者标签映射到与该对汉英双语同步字幕上。

BM25 相似度计算是跟词频有关的，比如一条字幕“Okay.”跟剧本句子“Okay, Okay.”的相似度得分会比跟剧本句子“Okay.”大，因为查询的单词在前者的出现频率比较大。此时如果“Okay.”才是本应对齐的句子，那么由于映射的是“Okay，Okay.”的标注，就会造成错误。

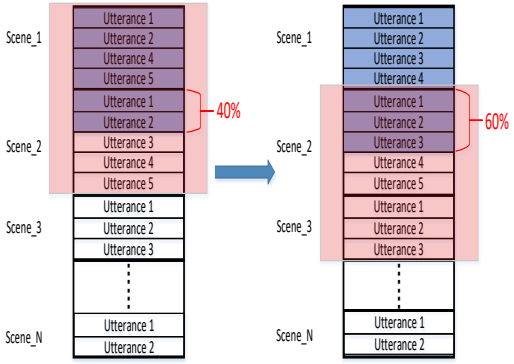


图 7 窗口滑动示意图

为了提高准确率，我们使用了一个策略，在计算一条字幕跟剧本句子的相似度时，在剧本上添加了一个滑动窗口，只计算窗口中的剧本话语消息而不是剧本中所有的话语消息，从而减小了对齐错误出现的可能性。

窗口的大小是两场景内的所有句子，滑动的条件是当窗口内的第二个场景的 50%以上的句子被匹配了，那么就向后滑动一个场景，如图 7 所示。通常来说，影视剧相邻两个场景内的话题是有一定差异的，所以一般不会在相邻的两个场景内出现相似的句子，但是可能存在诸如“Yes.”、“Okay.”这种高频句子同时出现在两个相邻场景内的情况，从而可能会造成错误的匹配。所以触发窗口滑动的条件需要更加严格，即要第二个场景内的 50%的句子被匹配了才滑动窗口。

最终，把标注信息映射到同步字幕上，得到汉英双语会话基础标注语料，如图 8 所示。中括号对[]内是场景内容描述，尖括号对<>内的是话语消息的标注信息：第一个数为场景 id；第二个数为该场景内的话语消息 id，最后还有该话语消息的说话者 id。

```
1
[Scene Description: Over the caption December 31st 1999 a crude spaceship
flies through space, cruising over and under planets and a man speaks]
00:00:06,600 --> 00:00:10,280
<1-1,MAN>Space, it seems to go on and on forever.
<1-1,MAN>宇宙,看似永无边际

2
00:00:10,640 --> 00:00:13,800
<1-2,MAN>But at the end, a gorilla throws barrels at you.
<1-2,MAN>但你会到达尽头,然后大猩猩开始向你扔水桶

3
00:00:19,840 --> 00:00:24,560
<1-3,FRY>That's how you play the game.
<1-3,FRY>这个游戏就是这么玩的

4
00:00:19,840 --> 00:00:24,560
<1-4,KID>You stink, loser.
<1-4,KID>你水平真臭,失败者

5
00:00:24,800 --> 00:00:28,200
<1-5,PANUCCI>Fry! Pizza going out. Come on!
<1-5,PANUCCI>Fry.披萨好了.快点!

6
00:00:38,560 --> 00:00:40,800
<1-6,FRY>Michelle! Baby! Where you going?
<1-6,FRY>Michelle, 宝贝! 你要去哪儿?

7
00:00:41,200 --> 00:00:45,960
<1-7,MICHELLE>It's not working out, Fry. I put your stuff out on the
sidewalk!
<1-7,MICHELLE>我们不会有结果的, Fry.我把你的东西扔到人行道上了

8
00:00:46,560 --> 00:00:48,880
<1-8,FRY>I hate my life. I hate my life.
<1-8,FRY>我讨厌我的生活.我讨厌我的生活.我讨厌我的生活.

9
[Scene Description: Outside Applied Cryogenics]
00:00:58,320 --> 00:00:59,760
<2-1,BIKE THIEF>Happy New Year!
<2-1,BIKE THIEF>新年快乐

10
[Scene Description: Cryogenics Lab. The room is empty and there are no
lights on. Strange pods about 6ft tall line one of the walls.]
00:01:17,920 --> 00:01:21,000
<3-1,FRY>Hello? Pizza delivery for. . .
<3-1,FRY>有人吗? 送披萨给. . .
```

图 8 汉英双语会话基础标注语料示意

4 实验结果

我们从两个方面对构建的会话标注库进行了初步分析。4.1 节进行统计分析， 4.2 节进行性能分析。

4.1 标注库的统计分析

我们目前收集了 17 部电视剧（其中 *New Supplement* 是新增的 6 部电视剧，包括“Crime Scene Investigation”、“24 Hours”、“Grey's Anatomy”、“Growing Pain”、“Supernatural”和“Veronica Mars”），每部电视剧的统计信息如表 1 所示。

收集的电视剧主要是生活喜剧，内容上较贴近日常会话，也收集了科幻、悬疑、冒险等类型的电视剧，话题比较丰富。有些电视剧是不完全收集的，比如 “Futurama”，只有 7 个剧集，有待进一步的扩充。

最终总共有 40,541 个场景会话，978,109 对汉英双语话语消息，涉及 10937 个说话者，场景平均说话者数为 2—5 个。

表 2 为会话标注库的综合统计信息。从中可以看出，场景平均话语消息数是 12.23，而基于日常真实会话的 DailyDialog 数据集的会话平均话语消息为 8，两者在平均话轮数上比较接近。会话标注库的剧集内场景数频率分布、场景内话语消息数频率分布、场景内说话者数频率分布如图 9 所示。

子图（a）是剧集内场景数的频率分布直方图，可以看到有两个比较集中的区间，这是因为“The Big Bang Theory”、“Friends”和“South Park”这三个电视剧剧集内场景数集中在 10-15，而且收集的“The Big Bang Theory”和“Friends”的剧集数最多，分别为 225 和 227，所以造成 10-15 区间的高峰。

而其他电视剧场景切换较多，每个剧集内场景数一般为 30-40，而且剧集数都在 170 以下，

表 1 各电视剧的统计信息

电视剧	类型	剧集数	场景数	话语消息数	说话者数	场景平均说话者数
House M.D.	悬疑	164	5652	125421	1039	3.00
The Big Bang Theory	生活喜剧	225	2839	92990	484	3.56
Friends	生活喜剧	227	3499	88364	691	3.47
Desperate Housewife	生活情景剧	110	3271	78670	791	2.89
Seinfeld	生活喜剧	167	6173	70269	684	2.45
South Park	动画喜剧	125	2228	48883	2508	5.97
Futurama	动画喜剧	7	110	2060	125	4.27
Castle	悬疑	163	6139	158265	2074	3.30
Stargate SG-1	科幻	167	4783	90977	1625	2.46
Merlin	冒险	51	1758	25473	192	2.33
Lost	科幻	43	548	25137	200	3.33
New Supplement	/	298	3541	171600	524	3.41
合计（合计平均）	/	1747	40541	978109	10937	3.37

表 2 基础标注库的综合信息

项目	数据
每个剧集的平均场景数	26.11
每个场景的平均话语消息数	12.23
每个场景的平均说话者数	2.40

所以在 30-40 的区间形成第二个较矮的峰值。

子图（b）是场景内话语消息数频率分布直方图，可以看出，场景内话语消息数为 6 到 14 条的频率最高，两边频率逐渐递减。Li Y 等人[6]的研究指出，真实日常会话一般包含 3 个或 3 个以上的话题，因此，场景内话语消息数分布集中在 6-14 是比较合理的。

除此之外，其他区间也有一定的频率分布，甚至较高的区间内仍然存在少量的频率分布。这是因为电视剧经常出现多人会话的场景，多人会话场景的话语消息数较多，随着会话标注库规模的扩大，不同人数的多人会话越来越丰富，不同话语消息数的场景也随着越来越丰富，但是，话语消息数较大的场景会较少，所以最终促成长尾效应的出现。

子图（c）是场景内说话者数频率分布直方图，可以看出，最多的是二人对话，其次是三人对话，值得注意的是一人会话也有不少，这是因为电视剧经常出现独白，如自言自语、说话者独自抒发感情或表达个人愿望等；也有可能在某些剧集中出现旁白，如说话者以回顾往事的方式展开情节。与其频率相当的还有五人对话，也存在少量五人以上的对话，甚至出现十人以上的对话场景，这也正是图（b）出现的长尾效应的原因。

4.2 标注库的性能评测

我们需要进行两个检测。一个是检测汉英双语同步字幕的准确率，另一个是检测剧本与汉英双语同步字幕的对齐准确率。

首先检测汉英双语同步字幕的准确率。我们随机从每个电视剧中抽取了 5 个样本，为了检测不同位置的同步情况，对每个汉英双语同步字幕样本的前 10 对话，中间 10 对话以及最后 10 对话进行人工检测，检测结果如表 3 所示[7]，“Correct”代表完全正确，即同步的两个句子能够完全互译；“Partially”代表部分正确，即两个句子只能部分互译，不能完全互译，通常是由于某些本应该为 1：2 或 2：1 对齐的句子没能满足阈值被丢失，或者某些本应该为 1：1 对齐的句子，错误地与前后句子实现了 1：2 对齐，从而造成部分错误；“Wrong”代表同步的两条句子完全不能互译。

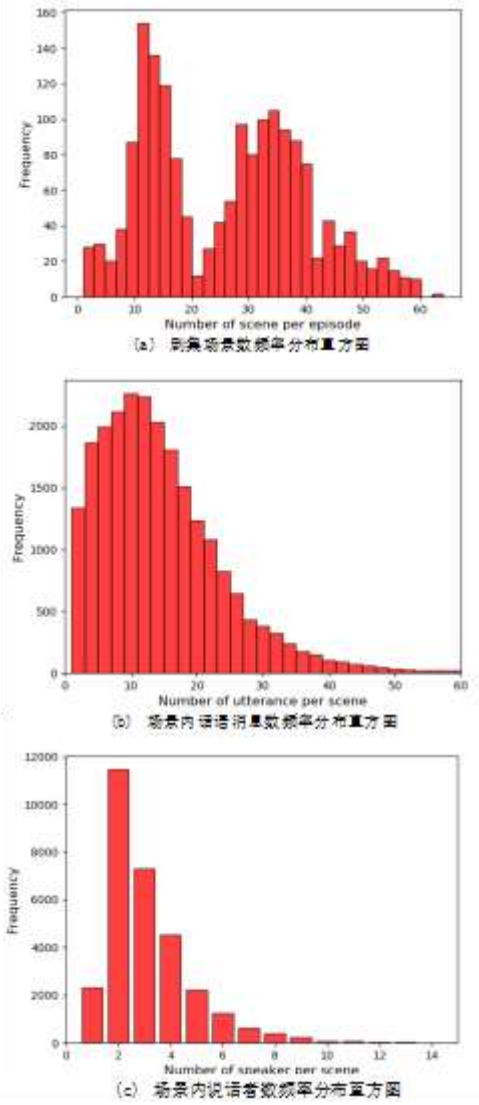


图 9 频率分布图（子图（a）为剧集内场景数的频率分布直方图，子图（b）为场景内话语消息数频率分布直方图，子图（c）为场景内说话者数频率分布直方图）

最终平均同步正确率为 94.0%。少量错误的同步字幕对后续的工作影响不大。

值得注意的是有的电视剧正确率较低，如“South Park（SP）”，为 86.8%，有的正确率很高，如“Merlin”，为 100%，这是因为一个剧集的英语字幕可以找到多个版本的汉语字幕（网络上有多个翻译组，用途也不一），而不同版本

[7] “House M.D.”、“The Big Bang Theory”、“Friends”、“Castle”这四部电视剧的字幕本身为双语字幕，不需要检测。

表 3 同步字幕检测结果

	DH	Seinfeld	SP	Futurama	SG-1	Merlin	Lost	*NS*
Correct	93.5%	92.5%	86.8%	98.5%	91.4%	100%	97.0%	92.3%
Partially	0.0%	7.5%	10.3%	0.0%	3.5%	0.0%	0.0%	4.6%
Wrong	6.5%	0.0%	2.9%	1.5%	5.1%	0.0%	3.0%	3.1%

表 4 各电视剧说话者标注平均准确率

	House	TBBT	Friends	DH	Snfld	SP	Ftrm	Cstle	SG-1	Mrln	Lost	*NS*
TFIDF+w	0.776	0.825	0.818	0.812	0.840	0.811	0.790	0.793	0.781	0.760	0.770	0.793
BM25	0.809	0.887	0.852	0.870	0.888	0.867	0.870	0.812	0.864	0.810	0.870	0.884
BM25+w	0.931	0.924	**0.913**	0.902	0.911	0.917	0.910	0.931	0.917	0.904	0.920	0.909

表 5 各电视剧场景边界标注平均准确率

	House	TBBT	Friends	DH	Snfld	SP	Ftrm	Cstle	SG-1	Mrln	Lost	*NS*
TFIDF+w	0.932	0.952	0.986	0.946	0.933	0.940	0.909	0.936	0.926	0.907	0.875	0.917
BM25	0.934	0.943	0.962	0.959	0.963	0.957	0.909	0.926	0.916	0.933	0.916	0.935
BM25+w	0.976	0.961	**0.989**	0.972	0.985	0.966	1.000	0.964	0.954	0.960	0.958	0.956

汉语字幕时间信息有很大的偏差，所以我们要尽量选择跟英语字幕时间信息一致的汉语字幕，越一致，正确率越高。

第二个检测的是剧本与同步双语字幕对齐的准确率。首先需要制作验证集，我们在每一部电视剧的每一季中随机选取一个剧集的汉英双语同步字幕作为样本，然后依据对应的剧本信息对样本的第 1 到 100 行进行人工标注。将自动对齐得到的标注跟验证集比对，计算每一部电视剧的平均标注准确率。

我们与 Wang L 等人[13]的方法进行了对比，结果如表 4 和表 5 所示。Wang L 等人的方法在电视剧“Friends”上得到的说话者标注正确率为 81.79%，场景边界正确率为 98.64%，生成了大约 100,000 条话语消息的汉语标注库。而我们的工作在电视剧“Friends”上得到的说话者标注正确率为 91.30%，提高了 9.5%；场景边界标注准确率为 98.90%，提高了 0.26%；在其他电视剧上的准确率也相对较高。其中“Futurama（Ftrm）”检测样本只有一个，所以差异较大。最后，会话标注库说话者标注平均准确率为 91.57%，会话库场景边界标注平均准确率为 97.00%。错误率可能来源于在滑动窗口内出现相似的剧本话语消息，如高频句子；也可能来源于某些字幕句子本身与剧本句子差异较大。在数据量方面，共有 978,109 对汉英双语同步话语消息，相比之下，大约扩增了 10 倍。

5 结论

本文提出了一个基于电视剧剧本和字幕数据构建日常会话标注库的可行方法。我们利用字幕的时间戳信息同步汉英单语字幕，使用信息检索方法对齐英文剧本和字幕，从剧本中抽取相关标注信息，最终获得汉英日常会话标注库。该会话标注库是基于 17 部电视剧字幕的多轮会话库，话题丰富，内容接近人类真实日常会话，数据分布合理，包含 978,109 对汉英双语话语消息，噪音较少，适合用于训练对话生成模型。我们已在 https://github.com/LiangYuHai/CEDAC 开源了部分数据，并将持续扩充我们的会话标注库。

由于同步字幕需要人工干涉，给更大规模地扩增会话标注库造成困难。在未来的工中，我们将对此进行改进，提高会话标注库构建效率。同时将基于此会话基础标注库，开展影视剧字幕说话者自动标注和多轮会话自动生成模型的研究。

参考文献

[1] Wang H, Lu Z, Li H, et al. A dataset for research on short-text conversations[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 935-945.

[2] Xing C, Wu W, Wu Y, et al. Topic aware neural response generation[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[3] Wei Z, Liu Q, Peng B, et al. Task-oriented

dialogue system for automatic diagnosis[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 201-207.
[4] Keyan Zhou, Aijun Li, Zhigang Yin, et al. Casia-cassil:a chinese telephone conversation corpus in real scenarios with multi-leveled annotation[C]//Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010.
[5] Lowe R, Pow N, Serban I, et al. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems[J]. arXiv preprint arXiv:1506.08909, 2015.
[6] Li Y, Su H, Shen X, et al. Dailydialog: A manually labelled multi-turn dialogue dataset[J]. arXiv preprint arXiv:1710.03957, 2017.
[7] Lison P, Tiedemann J, Kouylekov M. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018.
[8] Lison P, Meena R. Automatic Turn Segmentation for Movie & TV Subtitles[C]// Proceedings of the 2016 IEEE Workshop on Spoken Language Technology IEEE conference. 2016.
[9] Al-Rfou R, Pickett M, Snaider J, et al. Conversational contextual cues: The case of personalization and history for response ranking[J]. arXiv preprint arXiv:1606.00372, 2016.
[10] Wu Y, Wu W, Xing C, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots[J]. arXiv preprint arXiv:1612.01627, 2016.
[11] Lison P, Tiedemann J. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles[J]. 2016.
[12] Banchs R E. Movie-DiC: a movie dialogue corpus for research and development[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012, 2: 203-207.
[13] Wang L, Zhang X, Tu Z, et al. Automatic construction of discourse corpora for dialogue translation[J]. arXiv preprint arXiv:1605.06770, 2016.
[14] Robertson S, Zaragoza H, Taylor M. Simple BM25 extension to multiple weighted fields[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004: 42-49.
[15] Tao C, Min-Yen K. Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus[C]//Language Resources and Evaluation, 2013: 299-355.

梁宇海（1995——），硕士研究生，主要研究领域为自然语言处理。
E-mail：lyh20180204@163.com

周强（1967--）,博士，研究员，主要研究领域为计算语言学，自然语言理解。
E-mail：zq-lxd@tsinghua.edu.cn