中国科学技术大学

学士学位论文

中国科学技术大学

基于搜索引擎的短文本视觉表达的研究

作者姓名：　　黄红艳

学科专业：计算机科学与技术

导师姓名：　孙广中 副教授

完成时间：　二〇一七年五月

University of Science and Technology of China

A dissertation for bachelor's degree

Visual Expression for Short Text Base on Search Engine

Author's Name:	Hongyan Huang
Speciality:	Computer Science and Technology
Supervisor:	A.P. Guangzhong Sun
Finished Time:	May, 2017

致　谢

在研究学习期间，我有幸得到了三位老师的教导，他们是：我的导师，中国科大孙广中研究员，微软亚洲研究院谢幸老师和宋睿华老师。三位深厚的学术功底，严谨的工作态度和敏锐的科学洞察力使我受益良多。衷心感谢他们多年来给予我的悉心教导和热情帮助。

感谢宋睿华老师在实验方面的指导和帮助。科大的李顶龙同学和陈仲夏同学参与了部分试验工作，在此深表谢意。

目 录

摘　要

短文本囿于字数极少，这使得人在理解时会联系到各种感官经验，这其中就包括视觉。比如“猴子爬树”，我们在理解时会联想到树木的图像。本课题旨在利用搜索引擎，对一段短文本能否使用图像表达进行探究。本课题试图解决以下几个问题：第一，定义短文本的视觉表达；第二，通过调查问卷，人工标记具体短文本是否能视觉表达；第三，通过监督学习的方法，自动识别能视觉表达的短文本。

关键词：近似搜索　图片搜索　可视化　询问分类

Abstract

The short text is limited by the few words, which makes it possible for people to understand the sensory experience, which includes vision. For example, ”monkeys climb trees”, we will understand the tree when the image. This topic aims to use the search engine, a short text can use the image expression to explore. This question attempts to solve the following questions: First, the definition of short text of the visual expression; Second, through the questionnaire, manually marked the specific short text can be visual expression; Third, through the supervision of learning methods, automatic identification of visual expression The short text.

Key Words: Similarity Search, Image Retrieval, Visualization, query classification

图目录

表目录

算法索引

第一章　简介

短文本与图像关联是当下热门且富有意义的研究方向。短文本分词研究已较为深入，卷积神经网络成熟运用使得图像识别有了成熟的商业产品，而短文本与图像的等价表达仍有待研究。随着大数据时代的来临，互联网多媒体资源丰富，文字与图像的关系更为紧密。这使得短文本的视觉表达成为可能。

本研究的动力来自于参与 NTCIR-13[1]短文本对话 STC-2[2]比赛。比赛中文部分由日本国立情报学研究所 NII 和华为诺亚方舟实验室联合组织，比赛内容是给定一个微博 (在本文中用 Post 表示) 内容，构建一个检索或者生成系统来得到连贯的、有意义的回复 (本文中用 Cmt 表示)。陈仲夏师兄在去年 STC-1 比赛的主要想法是通过 TF-IDF 和 Word2Vec 算法在语料库中寻找相似的微博，用相似的微博语料库中对应的评论作为此微博 Post 的评论 Cmt。本研究主要贡献在寻找相似的微博内容的时候，在传统的文字相似性上添加视觉相似性来寻找和重新评价其关联性。

随着多媒体时代的到来，我们无论是在自己发微博写朋友圈的时候，还是浏览新闻看公众号文章都会配上相关的图片，甚至近些年流行的表情包都表明图片在信息交流传播中占据了重要的一部分。这一方面极大的充实了图片搜索引擎的内容，另一方面也说明了文字在再创造和演变的过程中将视觉、听觉等感官高度编码，变成一个同意的利于传播的格式。但是人在理解短文本的内容时往往会带有自身的经验和联想。例如，有一个微博说到“今天又老一岁了”，根据我们的常识我们能推理出发这个微博的博主今天过生日，因此“生日快乐!”是合适的回答。但是如今的计算机还并不具备常识，目前的主流的方法中它只能通过微博内容文字与文字之间的相似性来表示微博语义上的相似性。但是在短文本中的语义往往需要较多的背景知识，单从文字角度挖掘微博之间的相关性会存在局限性。我们想法是，通过过搜索引擎的图片搜索功能，使将微博的文字“解码”, 扩充其视觉感官的部分，扩充计算机在理解微博语义部分的背景知识。

[1]日本国立情报学研究所 NII 组织，是信息检索国际性评估会议

[2]该比赛分为中文和日文两个部分，中文部分语料库来自微博，日文部分来自 Tiwwer

第二章　短文本对话

表 2.1　新浪微博一个典型帖子 Post 和它的评论 Comment

Post	创新工场三年庆，在我们的「智慧树」会议室。 Today is the 3-year anniversary of Innovation Works. We are in the meeting room named Tree of Wisdom.
Comment 1	时间过得真快，创新工场都 3 年了！周年庆快乐！ How time flies; Innovative Works is three years old! Happy Anniversary!
Comment 2	小小智慧树，快乐做游戏，耶！ Little Wisdom Tree, happy games, yeah!
Comment 3	会议室挺气派，顶一个！ The meeting room is quite impressive, the best one!

人与电脑之间的自然语言交流是最具挑战性的 AI 问题之一，涉及语义理解，推理和常识知识的运用。尽管过去几十年来对人机交互研究工作进行了大量的努力，但令人遗憾的是，这个问题的进展非常有限。其中一个主要原因是缺乏大量的真实对话数据。

在短文本对话任务中，我们只考虑一个简化版本的问题：由两个短文组成的一轮对话，前者是用户的初始帖子，后者是计算机给出的评论。我们把它称为短文本对话（STC）。由于在 Twitter 和微博等社交媒体上提供了大量短文本对话数据，我们预计在使用大数据的问题研究中可以取得重大进展，就像在机器翻译，社区问答等领域一样。随着社交媒体的出现和移动设备的广泛传播，通过短消息的对话已成为重要的沟通方式。许多现实中应用可以从短文本对话的研究中受益，例如手机上的自动消息回复，Siri 等语音助手以及用于智能家居设备上各种聊天机器人。

在 NTCIR-12 上短文本对话的作为试点任务提出，让有兴趣的自然语言对话的研究人员聚在一起。在 NTCIR-12，短文本对话（STC）被认为是一个信息检索（IR）问题，通过在日语子任务中保持一个大量的 Twitter 中的子博客 Twitter 和 Twitter Twitter 的留言对，然后找到一个聪明的方法来重用这些现有的评论来回应新的帖子。在中文子任务中，语料库来自于微博。

在今年我参加的 NTCIR-13 中，除了基于检索的方法之外，主办方还考虑了基于生成的方法来生成“新”评论。基于生成的方法已经成为一个热点研究课题，近年来受到最多的关注，而基于检索的方法是否完全被替代或者与基于生成的短文本对话（STC）任务的方法组合在一起仍然是一个开放的问题。NTCIR-13 的短文本对话任务提供了一个透明的平台，通过进行综合评估来比较基于检索和基于新生成方法。此外，主办方鼓励参与者探索一些有效的方式来结合两种方法来获得更智能的聊天机器人。

一、任务描述

在本文的研究中，我们将短文本对话定义为一个信息检索问题，即基于检索的短文本对话。存储库我们也称为语料库，是有来自于微博的帖子-评论（Post-Cmt）对。在表2.1中显示了一个微博帖子和三个对应它的评论。每个参赛队伍都会收到主办方事先准备好的语料库。值得注意的是，不仅一个帖子会对应多个评论，在对评论进行去重复处理编号后，会有不少评论对应了多个帖子。

在**训练**期间，主办方会发布训练数据，他们是被人工标记评级过的帖子-评论对。关于人工标记评级，将在下一节阐述。我们要使用之间收到的语料库和这些被标记过的帖子-评论对作为训练数据，构造一个会话系统。

在**测试**期间，每个参赛队伍对收到一个测试集，由一下微博帖子组成，每个帖子都被保证不在语料库中。我们需要为每个测试帖子提供十个结果（评论）的排名列表，而且这些评论必须来自于语料库。

在**评估**期间，所有参赛团队提交的结果都会被匿名集合到一起人工标记。评估会使用信息检索分级相关的测量方法。

在表2.2显示了中文子任务中数据集的统计结果。在语料库中有差不多 20 万条微博帖子，每个帖子平均有 30 条评论，但是大概有 100 万帖子-评论对中的评论重复。主办方标记了 225 个帖子的作为训练数据，每个帖子平均大约有 30 个候选评论。有 100 个帖子作为测试数据，这些用于测试的帖子保证不再语料库中，参赛者需要搜索自己的会话系统，为每个测试帖子寻找 10 个评论和其排名作为结果。微博上原始网页的文本是中文的，为了帮助非华人参赛者，主办方提供了英文版本，所有的英文翻译都来自于机器翻译。华人参赛者也能收到英文版本作为参考。

表 2.2　STC 中文子任务数据集统计

语料库 Retrival Repository	#Posts	196,495
	#Comments	4,637,926
	#original pairs	5,648,128
标记数据 Labeled Data	#Posts	225
	#comments	6,017
	#labeled pairs	6,017
测试数据 Test Data	#query posts	100

二、研究想法

短文本对话（STC）的一个简单方法，也许是大部分人第一种想要尝试的方法，就是将其作为信息检索（IR）问题：维护一个大型的短文对话数据库，并开发主要基于信息检索（IR）技术的会话系统。给定一个初始帖子（Post）A，系统搜索语料库并返回最合适的评论（Cmt）。存储库中的评论（Cmt）最初是针对 Post A 以外的一些帖子发布的，但我们假设语料库足够大，包含了所有可能存在的帖子（Post），因此我们假设可以将其重新用于对 A 的合理评论（Cmt）。也就是说，我们处理更简单的基于检索的短文本对话（STC），而不是追求基于生成的短文本对话（STC），即从用户的初始帖子（Post）生成适当的评论（Cmt）。利用先进的信息检索（IR）技术和大数据，即使是基于检索的短文本对话（STC）系统也可能在每一轮会话中最终都会像人类一样表现出来。

因此，我们想研究的关键问题是：给定一个新的帖子（Post），如何搜索语料库来返回适当的（即类似人的）评论?

公认的一个原则是，合适的评论应该跟原帖子谈论相同的话题。为了追求这样的原则性，在之前，Ji et al.[1]的研究者，他们重点关注一个给定的帖子和每个可能的评论之间的相关性。也就是说，一个给定的帖子和可能的评论有相同的一些词或者短语。我们可以大胆地推断出它们和给定的帖子在谈论同一个话题，如表2.1中所示的评论，评论可以从周年庆、智慧树、会议室几个方面展开，因此他们与帖子“创新工场三年庆，在我们的「智慧树」会议室。”有相同的一些词和短语。但是，与原帖子有同样或者相似的字词并不是一个合适的评论的必要

[1]Z. Ji, Z. Lu, and H. Li. An information retrieval approach to short text conversation. CoRR, abs/1408.6988, 2014.

条件，比如一条微博谈论去香港的旅游计划，一个合适的评论可能是“羡慕你”。在这个例子中，原微博和评论在字词上是完全不同的而且是相关的。

在去年 NTCIR-12 上，陈仲夏师兄提出了一种新的思路，如图2.1所示。假设我们的语料库足够大，覆盖到了每个话题的每种表达，那么我们可以将短文本回答看成是一个纯粹的信息检索的问题。在给定一个新帖子时候，我们可以在存储库中找寻找语料库中的帖子，那么他的评论一定是这个新帖子合适的评论。那么一个新帖子在我们语料库中能找到相似的帖子，他对应的评论很可能是新帖子合适的回复。值得注意的是，语料库的原始帖子-评论对有 5,648,128 条，而去重后的评论数量只有 4,637,926 条，这意味着大约有一百万条的评论被重复使用。我们对语料库原始帖子-评论对（#original pairs）中的评论出现的频率进行了统计。与我们常识相符，越是能够通用的评论的频数越高，例如“羡慕你”这个评论在帖子-评论对中出现了 412 次。

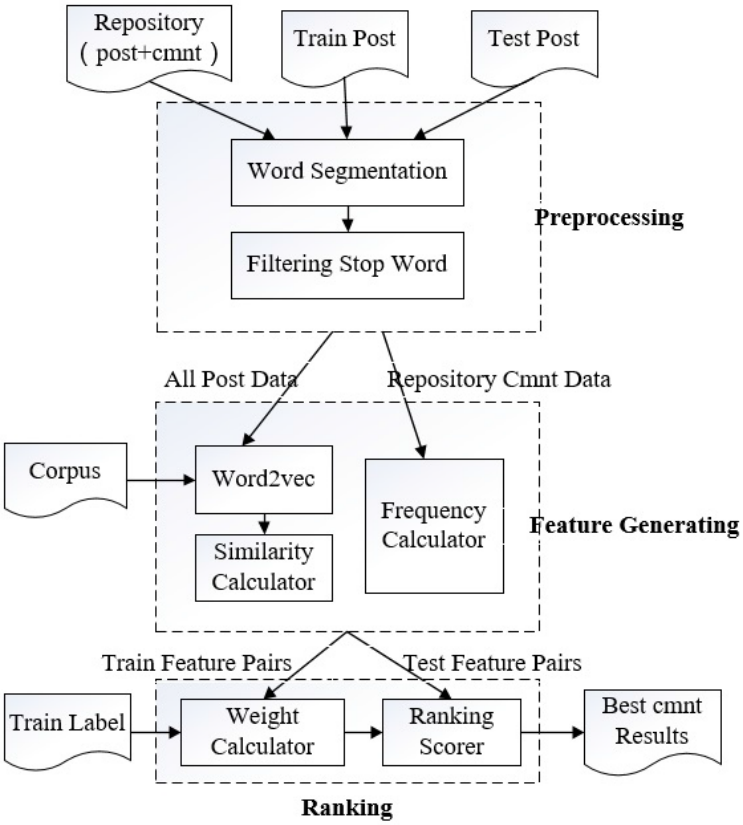


图 2.1　短对话系统架构

如图2.1所示陈仲夏师兄的系统架构，整个系统分为三个部分，一是将帖子、评论做分词，去停顿词预处理；二是通过的文字上的相似为新帖子在语料库中找出相似的帖子集；三是收集相似帖子集对应的评论集，综合考虑帖子的相似度、

评论的频率为候选评论排序选出 10 个评论并排序。

我们今年的短对话系统继承于陈仲夏师兄的设计，主要区别是在第二部分为新帖子找的相似的帖子上，这也是本文研究的主要目标。除了陈仲夏师兄利用文字上的相似性来关联相似的帖子，我们还想利用图片搜索引擎将短文本视觉化，利用视觉上的相似性来寻找和重新排序相似的帖子。考虑的今年的短文本对话比赛过滤掉了大部分高频的部分，李顶龙同学会对第三部分评论排序部分做相关研究和修改。

第三章　视觉表达

短文本对话希望根据一段短文字，计算机模仿人类给出连贯的有意义的回复。人类在对一段文字进行回复的时候会经历：理解问题——大脑思考——给出回答，这样的一个过程。但是怎么判断别人（或者计算机）是否真正理解了你的意思？这个问题似乎很难回答。在目前自然语言理解领域上，对其有两个定义，一个是基于行为的，一个是基于表示的。对于前者，如果你说“给我拿一杯茶来”，别人（机器人）真的按你说的做了（行为），就认为他了解了你的意思; 而对于后者，如果你说“猴子爬树”，别人把它联系到了大脑中猴子爬树的概念（表示），就认为他理解了你的意思。

十几年前，脑科学研究中有一个有趣的发现。当把电极插到猴子的大脑前运动皮质 (pre-motor cortex) 时，有一个脑细胞会在猴子自己吃香蕉和看别人吃香蕉时，同样处于兴奋状态，也就是说对猴子来说这个脑细胞对应着“吃香蕉”的概念。（猴子和人的运动都是由小脑控制，但大脑的前运动皮质也与运动有关。）后来对人脑做类似的实验，但使用功能磁共振。让人实际做和想象做各种动作，比如跑步和想象跑步，爬树和想象爬树。结果发现，对同一动作，实际做和想象做大脑的前运动皮质中发生反应的部位完全一致。现在一个得到广泛支持的理论认为，对于同一个概念，大脑用固定的脑细胞去记忆，人理解语言的过程，就是激活相关概念的脑细胞，并关联这些概念的过程表示同一个概念的脑细胞，可以通过不同的方式被激活。例如，有一个细胞表示人在喝水，当你看到人在喝水的时候，或者当你从书中读到人在喝水的时候，这个脑细胞同样会被激活。这也能解释为什么我们在读小说的时候常常有身临其境的感觉。

每个人把自己经历的事件进行编码，存储记忆在脑细胞中，在与外界的交互中这些脑细胞被激活，相关的记忆被唤醒。所以，不同人对同样的语言会有不同的理解，因为他们的经历不同。但也有许多共性，因为大家在交流过程中，相互激活对方脑中的表示相同内容的细胞。发明比喻的时候，大脑中表示两个不同概念的部位都开始兴奋，相关的脑细胞之间产生新的连接，概念之间产生关联，这个过程被称为神经结合 (neural binding)，是现在脑科学研究的重要课题[1]。

语言的理解实际上是一个非常复杂的过程，动用了各种各样的感官，动用了大脑中所有的相关知识。我们的研究想让计算机读到一个短文本的时候能模拟

[1]Lakoff G. What Studying the Brain Tells Us About Arts Education, 2013.

人读到它的时候大脑产生的概念。这个过程就像是一个刚刚渡过婴儿期开始定义这个世界的孩子，她一开始一定没有“树”概念，她是在见过许多现实中的树，照片中的树，图画中的树之后才有了“树”的概念。也就是说，这样的概念其实是来自于这个孩子的视觉感受，当她在产生“树”的概念时是这些视觉图像的浓缩。正如“树”在脑海中会浮现出一个“树”的图片，我们在看到一段短文本对其产生概念，会在脑海中产生视觉图像，当然这样的视觉图像并不是在所有的短文本上就会产生，比如“昧着良心”。因此我们认为，**当我们读到一段短文字，对其产生的概念是否能用图像表示定义为短文本的视觉表达**。对于对话系统来说，

图 3.1　单词视觉表达问卷调查

它在看到一段短文本时，产生概念是什么？对于计算机来说。我们可以通过图片搜索引擎轻易地将短文本转化成图像集合。但是怎么判断这些图像集合能否等价于其视觉表达呢？我们将每个进行图片搜索的短文本称为一个查询用 Query 表示，将图片搜索得到的结果进行了人工标注如图3.1。标注员能够在一个页面里看到 Query（查询）的短文本，和 Query（查询）通过图片搜索得到的图片集合。标注员需要回答 Query（查询）是否能用上述图片视觉表达以及勾选出符合标注员对短文本产生概念的视觉表达的图像。经过反复的人工标记，我们得到了以下准则：

- Query（查询）经过图片搜索得到的图片能够与其查询的短文本在人脑海中联系的的视觉表达相关，视为图片能作为 Query(查询) 的视觉表达。
- 当 Query（查询）经过图片搜索得到的图片能作为 Query(查询) 的视觉表达的数量足够多时，视为与 Query 的意思能用图片视觉表达，数量不够的时候视为 Query 的意思能用上述图片部分视觉表达
- 当搜索得到的图片集有多个相同的语义时，当短文本的语义存在多个时，以及当符合 Query（查询）视觉表达的图片数量不够多时，都视为 Query 能用搜索得到的图片集部分视觉表达
- 当搜索得到的图片集数量不足，或者图片杂乱无章时，视为 Query（查询）不能用图片视觉表达。

在反复标记的过程中我们发现很多有趣的东西。除了我们很容易联想到的一些生活中的物品比如笔，树有视觉表达，在图3.1中，我们定义的一些在现实中没有的虚无的东西，也找了一致表达。在这个例子中图片搜索引擎返回给我们了很多自由女神像的图片，这跟我们人类的常识相符，我们使用自由女神像来代表自由，这是我们想要的在文字之外的关联。在另一个层面说，我们想要通过搜索引擎来寻找短文本一致的视觉表达。我们认为如果一个 Query（查询）返回的图片越多一致的图片，这个图片是其人在脑海中的视觉表达概率越高。

另一个有意思的例子是 Query：“玉米”的消费力，这样的一个 Query（查询）在图片搜索后得到的结果是李宇春演唱会的图片。“玉米”除了是我们知道的食物之外，还是李宇春粉丝团的名字。在经过图片搜索引擎视觉表达后，像是经过了一轮人的思考和联系之后得到的视觉表达。同样的例子还出现在，在对帖子进行视觉表达后聚类中，关于食物话题的帖子的视觉表达会聚类在一起。我们认为经过图像表达可以比文字更能表示贴近我们脑海中的形象。

为了区别于 Post（帖子），我们特意使用了 Query（查询）来表示用于图片搜索的短文本。在本文的研究中，我们为了使 Post（帖子）能够更好的进行图像表达，后续的相关章节会讨论关于帖子怎么变成 Query（查询）。

在我们想设计的对话系统想要模拟人在理解短文本过程中是否利用到了我们的视觉感官系统。我们通过将短文本（帖子）变成一些相关的 Query(查询)，将 Query（查询）得到图片集定义为计算机能够理解的短文本的视觉表达。

一、图片搜索引擎

由于搜索引擎已经花费了大量精力来完善他们官方使用的图像检索服务，我们可以很轻易收集标签图像。我们可以通过搜索图像标题、文件名和周围文本等方式来检索图像。现在各大图片搜索引擎都提供了搜索图片的 API，要实现自动检索图像，我们可以将一个单词或短语（本文我们统一称为 Query 查询）作为 HTTP 查询提供给搜索引擎，并直接下载所返回的均匀缩略图，而不是直接下载原图像。一个是因为源网站图像要返回源网站下载，所以不如缩略图稳定；二是最后可能涉及的查询 Query 数量很多，考虑到存储空间；三是为图像大小相对均匀，这样在提取特征和问卷调查时候需要考虑图片差别。

各大图片搜索引擎提供的自动检索的接口都设置了流量限制，而我们的研究和对话系统数量较大，我们的研究经费难以承担。在宋睿华老师的帮助下，我们的难题得到了解决。因此本文的研究都使用 Bing 的图像搜索作为搜索引擎（https://www.bing.com/images）。我们的自动检索接口检索所得图片和人工检索结果一致。

二、图片集数量

我们希望通过图片搜索后的图片的集合来代表 Query 查询的图像表达。但是，显然我们不能将所有的图片缩略图都下载下来。一个是程序运行时间和存储限制，第二个是人工标记也不能顾上。我们当然希望能用最少的图片集数量来判断一个查询是否能够用视觉代表。

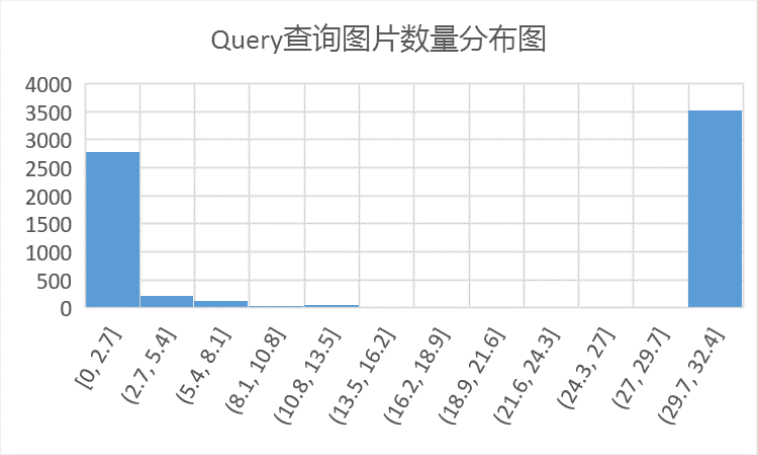


图 3.2　查询图像数量统计

因此我们测试帖子进行分词后将帖子所有可能长度进行了分割。也是说，如

果一个帖子有 N 个单词，我们可以将帖子分成 N*(N-1)/2 个 Query（查询）。因此，现在的查询集合包含了所有可能长度。我们将图片搜索引擎的返回图片数量设置到 30，记录每条 Query 查询的返回图片数量。100 个测试帖子共分成了 6,872 个 Query（查询），查询的最大词数是 58，最小词数为 1。返回图片数量如上图3.2所示，图片数量集中在最小值 0 和最大值 30，在图片数量在 10 附件可以涵盖大部分 Query 查询。为了尽可能少的下载图片数量，本文中搜索引擎中一个 Query 查询取回的图片最大数量为 10。

第四章　词的视觉表达

文字在演变过程中是将我们看到的感受到的记录下来，在最早的文字中我们用圆圈来代表太阳，然后经过演变发展成我们今天的“日”。说到视觉表达，我们第一直觉反应就是我们看见的物体形状。所以我们先研究短文本短的极限，单词，我们直觉上他们是短文本中最能视觉表达的部分。在我们的常识中，一个句子越长它的含义越丰富，因而越难寻找到一个图像准确的表达它的语义。为了最后能训练更好的短文本视觉表达自动分类器，我们希望找到在更短的文本中找到视觉表达的词。这章我们来讨论短文本的特例，词的视觉表达。

一个句子的语义成分或者说是关键成分在人脑海中有一些固定概念，相反一些辅助成分我们对其没有概念，它们只是帮助我们理解语义。因此，我们大胆推测在句子中起辅助作用的成分在计算机中也没有视觉表达。换句话说，这些成分的视觉表达很大概率在对话系统中不起作用。例如，我们对“?”进行图片检索，搜索的图片大部分都是问号形状的图片，但是这个图像表达对我们帖子间的关联没有意义。因此，词性可以帮助我们筛去一些大概率没有视觉表达的部分。我们只选择名词和形容词作为我们的实验对象。

一、中文常用词

我们使用了商务印书馆2008年出版的图书《现代汉语常用词表》中的56,008个现代词语以及频率。对其进行词性标注以后筛选了其中的名词后还有23,742个词。我们选出了其中300个进行了标记，记为常用词组。

二、诗集中的词

我们收集了1920年以前的519位诗人的诗词，并将其利用结巴分词进行分词。将诗句分词后进行词频统计和词性标注。词性标注后，名词有11,771个，形容词有1,390个，我们抽选了最高频的150个名词和50个形容词作为高频组，随机选取的150个名词和50个形容词作为随机组，两个组合并称为诗集组。对400个词进行图片搜索后组成问卷调查，进行如图3.1所示的人工标记。

因为本文的研究利用到了大量的人工标记，为了人工标记的可信度，我们使用Fleiss’Kappa算法来测量这些人工标记的一致度。算法统计了标记人员在每个

表 4.1　诗词视觉表达调查问卷的人工标记一致度 (Fleiss'Kappa)

计算方法	诗集组	高频组	随机组
A,B,C	0.345198	0.346088	0.337107
A,(BC)	0.421180	0.534748	0.335548
(AB),C	0.400946	0.381584	0.409502
(AC),B	0.240052	0.222053	0.252251

小题上的选择数量，然后综合每个小题得到所以标注人员之间的一致度。卡帕值在 [-1,1]，卡帕值越高说明人们的选择越一致，卡帕值越低说明他们对问题的理解存在分歧，当所有人在所有的问题都选了同一个选项时，卡帕值为 1，相反如果卡帕值小于 0 说明标注员之间的对这个问题的分歧不如随机选择。

如表4.1所示，表中的数值对应标注员的卡帕值。计算方法中 A、B、C 对应问卷调查中的问题“Query（查询）是否能够用图片视觉表达”中的选项。A：不能视觉表达；B：部分能视觉表达；C：能视觉表达。在没有合并的算法（A,B,C）中，我们欣喜的看到他们都取得了比较高的一致度。说明我们的标注结果是可靠的。我们试图确定人们对能视觉表达一致度更高，还是对不能视觉表达一致度更高。我们将三个选项两两合并，将 AC 合并的选项是没有道理的因此它的卡帕值下降是合理的，另外两个选项交不合并的算法都有明显提高，说明这种合并在选择一致度上是合理的。虽然高频组和低频组对合并算法各有偏好，在所有组中，在 A,(BC) 算法中取得了更高的一致度，它表示人们对不能图像表达有更高的一致性，倾向于混淆部分能表达（B）和完全能表达（C）。我们在训练视觉表达分类器时，也会用到这种算法。

三、小结

表 4.2　词视觉表达调查问卷结果

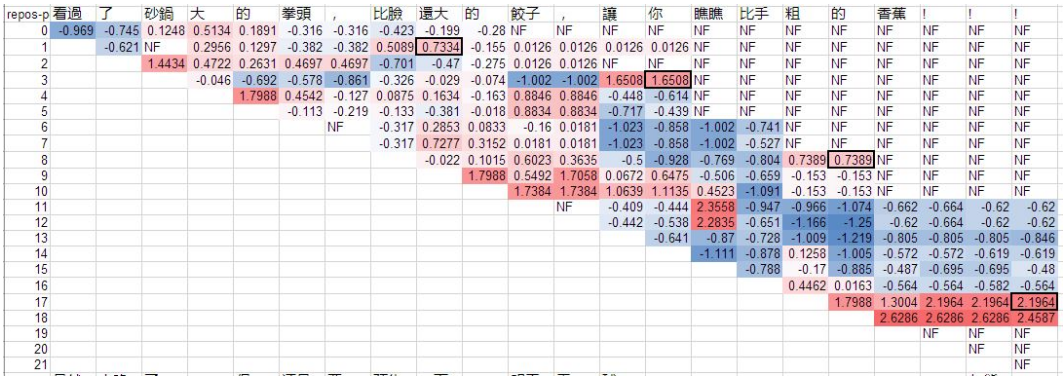
查询集	常用词组	诗集组	高频组	随机组
Wquery 得分	0.683333	0.77625	0.815	0.7376

在表5.1中记录了标注结果。我们在调查问卷中的问题“Query（查询）是否能够用图片视觉表达”中的选项。A：不能视觉表达标记为 0 分；B：部分能视

觉表达标记为 1 分；C：能视觉表达标记为 2 分。归一化后算得如上表中的得分。在高频的词汇中我们能看到在高频的词汇中能够有大部分的单词可以被判定为视觉表达。现代汉语常用词组和诗集组的比较中，我们能够看到，诗集中的常用词更能够视觉表达。这可能与诗在用词的时候更考虑意象，喜欢用“太阳”“世界”“朋友”这样的词汇，而常用词更多的是“问题”“关系”“社会”这样的单词。在高频词和随机组的比较中，我们发现高频词和低频词在视觉表达上存在差距，我们认为这部分原因来自于搜索引擎，如果一个短语出现的频率很少，那么网上对它的资料更倾向于杂乱无章。还有一部分也符合我们的认知，如果一个东西的频率越低，我们对其越没有清晰的印象。

第五章　短文本的视觉表达

在我们的对话系统中,[?] 我们希望通过相似的视觉表达来关联相似的 Post（帖子）。在早几年的微博中，对帖子（微博) 的长度做了限制，最长不超过 140 个词。在样例帖子集（Sampled Post，我们在语料库中随机选取了 34 个帖子作为样例帖子集）中单词数量最高的帖子达到了 58 个词，这并不适合现在标签化的图片搜索引擎。一般来说，一段短文本单词数目越多，他表达的含义越丰富，也就越难用一个图像完整细致的表达。在分割的尝试过程中，我们也体会到了这种规律。



repos-p	看過	了	砂鍋	大	的	拳頭	,	比臉	還大	的	餃子	,	讓	你	瞧瞧	比手	粗	的	香蕉	!	!	!
0	-0.969	-0.745	0.1248	0.5134	0.1891	-0.316	-0.316	-0.423	-0.199	-0.28	NF	NF	NF	NF	NF	NF	NF	NF	NF	NF	NF	NF
1		-0.621	NF	0.2956	0.1297	-0.382	-0.382	0.5089	0.7334	-0.155	0.0126	0.0126	0.0126	0.0126	NF	NF	NF	NF	NF	NF	NF	NF
2			1.4434	0.4722	0.2631	0.4697	0.4697	-0.701	-0.47	-0.275	0.0126	0.0126	NF	NF	NF	NF	NF	NF	NF	NF	NF	NF
3				-0.046	-0.692	-0.578	-0.861	-0.326	-0.029	-0.074	-1.002	-1.002	1.6508	1.6508	NF	NF	NF	NF	NF	NF	NF	NF
4					1.7988	0.4542	-0.127	0.0875	0.1634	-0.163	0.8846	0.8846	-0.448	-0.614	NF	NF	NF	NF	NF	NF	NF	NF
5						-0.113	-0.219	-0.133	-0.381	-0.018	0.8834	0.8834	-0.717	-0.439	NF	NF	NF	NF	NF	NF	NF	NF
6							NF	-0.317	0.2853	0.0833	-0.16	0.0181	-1.023	-0.858	-1.002	-0.741	NF	NF	NF	NF	NF	NF
7								-0.317	0.7277	0.3152	0.0181	0.0181	-1.023	-0.858	-1.002	-0.527	NF	NF	NF	NF	NF	NF
8									-0.022	0.1015	0.6023	0.3635	-0.5	-0.928	-0.769	-0.804	0.7389	0.7389	NF	NF	NF	NF
9										1.7988	0.5492	1.7058	0.0672	0.6475	-0.506	-0.659	-0.153	-0.153	NF	NF	NF	NF
10											1.7384	1.7384	1.0639	1.1135	0.4523	-1.091	-0.153	-0.153	NF	NF	NF	NF
11												NF	-0.409	-0.444	2.3558	-0.947	-0.966	-1.074	-0.662	-0.664	-0.62	-0.62
12													-0.442	-0.538	2.2835	-0.651	-1.166	-1.25	-0.62	-0.664	-0.62	-0.62
13														-0.641	-0.87	-0.728	-1.009	-1.219	-0.805	-0.805	-0.805	-0.846
14															-1.111	-0.878	0.1258	-1.005	-0.572	-0.572	-0.619	-0.619
15																-0.788	-0.17	-0.885	-0.487	-0.695	-0.695	-0.48
16																	0.4462	0.0163	-0.564	-0.564	-0.582	-0.564
17																		1.7988	1.3004	2.1964	2.1964	2.1964
18																			2.6286	2.6286	2.6286	2.4587
19																				NF	NF	NF
20																					NF	NF
21																						NF

图 5.1　帖子的全切割视觉表达预测热力图

如图5.1是一个样例帖子分词后全分割尝试后的对其查询进行是否能视觉表达预测后的热力图。表格中第 i 行第 j 列，$P_{i,j}$ 代表着 $Query_{i,j}$ 的在视觉表达分类器中预测的结果，$Query_{i,j}$ 是帖子分词后第 i 个单词到第 j 个单词的子帖子，作为一个查询。比如第 1 行第 4 个数字 $P_{0,3}$ 为 0.513356，代表着查询“看過了砂鍋大”能否视觉表达的可能性。图中 $P_{i,j}$ 数值越大，颜色越红表示 $Query_{i,j}$ 能视觉表达的可能性越大，反之颜色越蓝表示不能视觉表达的概率越大。在视觉分类器中将 0 作为分割线，$P_{i,j}$ 为正数表示其能视觉表达，$P_{i,j}$ 为负数表示它不能视觉表达。$Query_{i,j}$ 为 NF（not found）表示查询没有搜索到图片，在样例中我们能观察到“NF”集中在右上角。这说明查询的长度越大，图片搜索引擎找到图片的可能性越小。因此我们很有必要对帖子做一些分割，使其能够更好的视觉表达。

我们想通过将帖子进行合理分割，使其变成的能视觉表达的帖子集。我们在图5.1观察到，当查询长度稍短一些时，更容易视觉表达，但也会因此与句子意思偏离。我们还是想让查询在保存帖子原意的情况下，让查询能够视觉表达。我们尝试了许多种方法发现: **长度**、**标点符号**、**分词**、**去停顿词**都是合理的分割手

段。考虑到顺序切割的末尾部分，我们都对尾部做了向前填充处理。

第一节　独立分割

一、基于长度的分割

在常识和实验中都表明了短文本的长度对其视觉表达有直接的影响。我们将帖子只基于长度的分割且不重复的分割成查询集，N 代表了 Query 的长度。下面是详细的算法5.1

```
Data: Post,N
Result: Query[]
initialization;
while i < Post.length do
    if i+N < Post.length then
        Query[k]=Post.substr(i,i+N);
    else
        Query[k]=Post.substr(Post.length-N,Post.length);
    end
    k++;
    i+=N;
end
```

算法 5.1: SplitbyN

二、基于分词的长度分割

我们在研究样例帖子集的时候，发现只基于长度分割的会对产生很多错误分割，比如“阿森纳”是一个著名的足球俱乐部名称，但是如果长度恰好分割到之间，比如分为“阿森”和“纳”则与句子原语义产生较大的偏离。于是我们规定句子分离成分最小单位是单词而不是字，并按其最大字数 N 将其分离。具体算法详见**??**, 算法主要是将词序列拼凑成尽可能接近于字数 N 的查询。并考虑了尾部补全和单词长度大于字数要求的情况，比如”中华人民共和国”。

```
Data: PostWords[](Split by Jieba),N
Result: Query[]
initialization;
while i < PostWords.Count() do
    j=0;
    Query[k]=NULL;
    while  i<PostWords.Count() and Query[k].length+PostWords[i].length<=N
     do
        Query[k]+=PostWords[i];
        i++;
    end
    if i == PostWords.Count() then
        Query[k]=NULL;
        i–;
        while i>=0 and Query[k].length+PostWords[i].length<=N do
            Query[k]+=PostWords[i];
            i–;
        end
        i=PostWords.Count();
    end
    if  i<PostWords.Count() and j==0 and Spost[i].length>N then
        Query[k]=PostWords[i];
        i++;
    end
    k++;
end
```

算法 **5.2:** JiebaMN

算法名称来自于我们所使用的分词算法，结巴（Jieba）分词。结巴分词号称做最好的 Python 中文分词组件。结巴中文分词采用的算法大致可以分为三个部分：首先基于 Trie 树结构高效地将词图扫描，生成句子中汉字所有可能成词情况，并由此构成一个有向无环图；然后利用动态规划查找最大概率路径, 找出基于词频的最大切分组合；最后对于未出现过的词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。目前结巴分词支持三种分词模型：精确模式、全模式、和搜索引擎模式。我们使用了精确模式将我们的帖子分成了单词序列。

三、基于标点的长度分割

我们发现标点符号可以对其进行天然的语义分割，但是在我们的对话系统中，不是所有的标点都适合作为语义分割。比如书名号可以强调一本书名，但是如果将其强制作为语义分割符号，剩下的语义部分将会不完整。换句话说，有些标点符号会将帖子变成更小的语义单元，这几乎等同于中文分词。但是在较长的短文本中，我们还是希望将语义分割成相互独立的段落。一些断句所用的标点符号才是我们希望的使用的分割标识，比如句号、分号、问号、感叹号等。具体的分割算法件算法5.3

四、实验数据

我们对上述分割算法和长度都进行了尝试，并将其查询集经过自动检索的方法搜到的图片进行了人工标记，如图5.2。与图3.1单词的视觉表达想对应的是，我们添加了一个问题“Query（查询）表达的意思是否符合 Post（帖子）”来判断分割成子帖子之后的语义变化。

如表5.1所示，我们分析一些标记结果。我们在调查问卷中的问题“Query（查询）是否能够用图片视觉表达”中的选项，A：不能视觉表达标记为 0 分；B：部分能视觉表达标记为 1 分；C：能视觉表达标记为 2 分。归一化后算得如下表中的 Wquery 得分。同样的我们将问题“Query（查询）表达的意思是否符合 Post（帖子）”选项，A：不符合 Post 标记为 0 分，B：部分符合 Post 标记为 1 分，C：符合 Post 标记为 2 分。归一化后算得如下表 Wpost 中得分。将两个分数相乘为 Wq*Wq 得分。在（a）表中 Wpost 参数中我们能清晰的发现查询字数和帖子原意的表达直接相关。Splitby10 和 Splitby12 有大于 10% 查询不返回图片，明显的高于其他方法。因此在人工标记后我们认为 JiebaM8 是独立分割方法中最能视

```
Data: PostWords[](Split by Jieba),N，SymbolSet
Result: Query[]
initialization;
while i < PostWords.Count() do
    if (i == PostWords.Count() or PostWords[i] in SymbolSet) and
     (Query[k]!=NULL) then
        k++;
        Query[k]=NULL;
    else
        if Query[k].length+PostWords[i]>N and Query[k]!=NULL then
            k++;
            Query[k]=NULL ;
        end
        Query[k]+=PostWords[i];
    end
    i++;
end
```

算法 5.3: SymbolMN

图 5.2　帖子视觉表达问卷调查

觉表达的分割算法。从人工标记的结果来看，短文本的视觉表达还是远不如单词的视觉表达。然而从表（a）中的结果来看，我们不能因为追求视觉表达而与原始帖子产生偏离。

第二节　关联分割

人在理解一个帖子，并对其产生回复的时候很多时候并不需要对其产生很全面的回复，可能就是几个关键词。在我们最开始的例子“创新工场三年庆，在我们的「智慧树」会议室。”评论都是针一个或几个方面展开。其中的关键词并不是相互独立的，我们可以对“创新工场、年庆”产生回复，也能对”年庆会议室”产生回复。

我们将帖子结巴分词后，使用 simHash 算法过滤掉停顿词，剩下的单词作为标签，进行 WxCy 组合。WxCy 组合是指，每 x 个关键词（标签），重复 y 个。我们使用视觉表达分类器后预测统计如下表5.2

在表中，pre_+ 表示视觉分类预测的值为正数即查询能视觉表达，pre_- 表示视觉分类器预测的值为负数，pre_{NF} 表示查询没有返回图片。在上表中 W3C2 和

表 5.1　词视觉表达调查问卷结果

查询集	Wquery 得分	Wpost 得分	Wq*Wp 得分
Splitby4	0.5	0.372146	0.221461
Splitby8	0.5	0.565789	0.291667
Splitby12	0.45	0.725	0.298701

注：词视觉表达调查问卷结果 (a)

查询集	Wquery 得分	Wpost 得分	Wq*Wp 得分
Splitby6	0.24	0.35	0.156667
Splitby8	0.276316	0.346491	0.157895
Splitby10	0.3535	0.378788	0. 209596

注：词视觉表达调查问卷结果 (b)

查询集	Wquery 得分	Wpost 得分	Wq*Wp 得分
JiebaM8	0.401709	0.487179	0.373932
Splitby8	0.355263	0.416667	0.322368
SymbolM8	0.3375	0.391667	0.35625

注：词视觉表达调查问卷结果 (C)

表 5.2　WxCy 视觉表达预测统计

funName	pre_{+}	pre_{-}	pre_{NF}	all	$pred_{+}$%	pre_{-}%	pre_{NF}%
Sampled_W3C1	44	139	3	186	0.236559	0.747312	0.016129
Sampled_W3C2	128	228	5	361	0.354571	0.621579	0.013850
Sampled_W4C1	29	83	8	120	0.241667	0.691667	0.066667
Sampled_W4C2	43	124	8	175	0.245714	0.708571	0.045714
Sampled_W5C1	17	64	4	85	0.200000	0.752941	0.047059
Sampled_W5C2	18	88	7	113	0.159292	0.778761	0.061947
Sampled_W6C1	19	43	3	65	0.292308	0.661538	0.046154
Sampled_W6C2	14	60	3	77	0.181818	0.779221	0.038961
Sampled_W7C1	9	37	5	51	0.176471	0.725490	0.098039
Sampled_W7C2	14	39	7	60	0.233333	0.650000	0.116667
Sampled_W8C1	17	19	8	44	0.386364	0.431818	0.181818
Sampled_W8C2	19	22	9	50	0.380000	0.440000	0.180000
Sampled_W9C1	11	12	9	32	0.343750	0.375000	0.281250
Sampled_W9C2	14	11	9	34	0.411765	0.323529	0.264706

W6C1 方法的查询集能视觉表达的概率更高且只有很小的概率搜索不到图片。

第三节　对话系统

我们将 Splitby12、JiebaM8、W6C1、W3C2 四种方法在对话系统中做了尝试。我们每个查询后取得的图片进行特征提取后变成一个 256 维的向量。我们将查询图片集的向量进行 K-means 聚类后取出最大类的中心作为查询的视觉表达。每个帖子被分割成一个查询集，通过查询集的视觉表达的来寻找相似的帖子。对话系统寻找评论的部分由李顶龙同学完成，这里只贴出跟我的方法选择相关的实验结果。

表 5.3　对话系统的人工标记结果

方法	image256	image256-char256	char256
Splitby12	1.0067	1.058	1.0318
JiebaM8	1.0513	1.0407	1.0226
W3C2	1.0067	1.125	1.0226
W6C1	1.0067	1.0208	1.0226

在本论文的书写的时候 NTCIR － 13 STC － 2 项目刚刚公布了测试数据，所以我们的研究大部分基于 NTCIR － 12 STC － 1 中的数据。我们采取了人工标记，并使用 NTCIR 官方的排序算法得到了结果，如表5.3所示。image256 表示单纯用图片关联相关帖子的结果，Char256 表示用文字关联的结果，image256-char256 表示两种方法的结合。表中的值是对话系统在测试帖子上提交的结果。虽然最高值在 W3C2 方法中的图像和文字方法中取得。总的来说，图片对文字关联方法只有微弱的提升。

第六章　视觉表达分类器

由于查询集数量十分庞大，我们想做一个视觉表达分类器自动地判定一个查询是否能视觉表达。我们对查询经过图片搜索，所得到的图片集，用图片集来判断一个查询是否能视觉表达。利用现有的 TLC 图像识别技术，将图片放入卷积神经网络中学习出它的特征特征向量。图像识别技术依据这个图片特征向量进行图像识别，调用特征提取接口将这个向量经过降维、归一化后将图片转化为一个 256 维的图片向量。

我们将查询集的图片向量来判断一个查询是否能视觉表达。我们以图片集的数量，图片向量间的欧拉平均距离以及它的标准差，余弦平均距离以及标准差，聚类后熵、最大类的数量、最大类的半径为作为分类特征，使用线性分类算法训练分类模型。我们开始直接使用短文本的人工标记作为训练数据，但是因为短文本的负例太多，我们后来部分添加了单词的人工标记，修正了正负例的比例。最后使用了 882 个查询作为训练数据，分类器参数如下表6.1所示

表 6.1　视觉表达分类器

正确率	精确率	回归率	F1
73.94	75.56	51.52	61.26

第七章　结论与思考

虽然我们的想法来源与短文本理解时的概念飙升，但是在我们的对话系统上加上图片视觉表达关联的相似的帖子之后我们只取得了微弱的提升。我们在短文本的一个特例——词的视觉表达上特别是在诗集上和高频率的词上的视觉表达取得了很好的效果，也因此训练出了一个效果良好的视觉分类器。在几次的结果中来看，我们的加入图像寻找相似的帖子能取得微弱的提升。如果仅仅从寻找相似的帖子这个字过程看，我们的确能寻找到很多正面的典型的例子。但是为什么文字的关联加上图片关联比单纯的文字关联只取得了微弱进步？一方面在微博数据中能视觉表达的帖子的比例太少，我们可能不倾向于写一个描述性的或者说容易视觉表达的帖子，我们更喜欢议论一些东西，吐槽一些东西，比如“不悔梦归处，只因太匆匆”。一方面可能上我们还没有找到将短文本在计算机中视觉表达的合适方法，比如“自由”作为单词的时候我们可以用“自由女神”表示，但是在一个谈论自由的帖子中就不那么合适。我们可能不如 Word2Vec 类似文字在图像中关联算法，是我们需要图片搜索引擎搜索图片来完成视觉表达这样一个过程，而不是直接地相互关联。

我们在这篇文章中提出了一些关于短文本相似度计算新的想法，少数关于这方面的资料用的图像识别的方法还停留在 SIFTs 时代。这几年随着深度学习的发展，图像识别技术在卷积神经网络中的巨大进步，图像识别的精度大大提升，另一方面图像和文本的相关联的资料也还在增多。期待在这个方向上看到更好的发展。

参考文献