清华大学

综合论文训练

题目：面向医疗的跨语言信息检索系统

系别：计算机科学与技术系

专业：计算机科学与技术

姓名：陈刚

指导教师：刘洋 副研究员

2016年05月16日

中文摘要

随着社会经济的发展，人们对生活健康问题的关注度越来越高，医疗成为人们日常生活中最重要的问题之一。近年来，随着互联网的高速发展，在线医疗产业得到了迅速的发展，涌现出了一大批的在线医疗平台。由于现有的在线医疗平台都是针对单语种的信息检索，无法满足一些用户的跨语言信息检索需求，所以，开发一个面向医疗的跨语言信息检索系统是非常具有研究和使用价值的。

本文针对跨语言信息检索的问题，实现了面向医疗的中英跨语言信息检索系统框架。这个框架解决了对于输入中文问题返回相关英文医疗问答信息的处理，并且提供详细的问答信息查看，为满足用户面向医疗的跨语言信息检索需求提供了技术支持和基础。

在算法方面，本文主要介绍的是基于词典翻译的特征匹配算法和最小错误率训练算法，在现有的基于词典翻译的跨语言信息检索框架上加入特征匹配的思想，使得检索更为准确和高效。

最后，通过在一定量数据上进行了定量实验，测试了系统在面向医疗的中英跨语言问答信息检索上的性能，并且给出了一些理论分析和解释。

关键词：医疗；跨语言；信息检索；基于词典；特征值；最小错误率

ABSTRACT

Along with the development of social economy, people pay more and more attention to the health of life and health care has become one of the most important problems in people’s daily life. In recent years, with the rapid development of the Internet, online medical industry has been developing rapidly and a large number of online medical platform hava sprung up. Since the existing online medical platforms only use monolingual information retrieval which can’t meet the needs of cross-laguage information retrieval(CLIR) for some users, therefore, to develop a cross-language information retrieval system for medical treatment have a signifcant research value and practical value.

In this paper, for the problem of CLIR, there achievied a framework of Chinese-English CLIR system for medical treatment. The framework has resolved the issue which returns related question and answer information for Chinese input, and provides a detail view of this information. The framwork provides the technical support and foundation to meet the users’ need of medical CLIR.

In the algorithm, this paper uses dictionary based translation feature matching algorithm and the minimum error rate training(MERT) algorithm. Based on the existing framework of dictionary based trainslation CLIR, there added feature matching ideological getting the more accurate and more efficient retrieval.

Finally, quantitative experments on a certain amount of data tests the performance of Chinese-English CLIR system for medical treament and gives the theoretical analysis and explanation.

Keywords: Medical Treament; Cross-Language; Information Retrieval; Dictionary Based; Feature Value; Minimum Error Rate

目录

1. 引言
   1. 研究背景
   2. 相关工作及现状
   3. 主要工作内容
   4. 论文组织结构
2. 面向医疗的跨语言信息检索框架

2.1 面向医疗的跨语言信息检索定义

2.2 英汉双语词典构建

2.2.1 平行语料的抓取

2.2.2 自动构建双语词典

2.3 数据预处理

2.4 检索系统框架

2.4.1 检索模型

2.4.2 检索策略

2.4.3 系统框架

1. 算法

3.1 词语对齐模型

3.1.1 词语对齐模型定义

3.1.2 IBM模型

3.2 特征匹配算法

3.2.1 直接最大熵翻译模型

3.2.2 检索中的特征匹配

3.3 最小错误率训练算法

3.3.1 数据标注

3.3.2 最小错误率训练

1. 实验分析

4.1 实验环境配置

4.1.1 数据准备

4.1.2 环境配置

4.2 评价指标

4.3 实验结果

4.4 实验分析

1. 总结

5.1 工作总结

5.2 工作展望

参考文献

致谢

声明

附录A 书面翻译

1. 引言

1.1 研究背景

随着社会经济的发展，追求高质量的生活成为主流，人们对于生活健康问题的关注度也越来越高；而其中最为突出的就是医疗问题。而且，今年来，随着互联网产业的迅猛发展，一项不同于常规意义上的医院诊断治疗的医疗产业——在线医疗悄然新兴。而这种在线医疗产业的主要形式就是通过成熟的在线医疗平台为互联网用户提供医疗方面的需求，这些有特定医疗需求的用户可以通过这些平台将自己所遇到的问题传达给相关专业人士，然后通过在线方式这些专业人士将问题的回答或者建议反馈给用户，从而解决其问题。这就是当前主流在线医疗产业的运转方式，在线医疗平台在其中起到的就是一个传递信息的作用。

如图1.1所示，直至2015年止，中国在线医疗行业的市场规模已经达到惊人的170亿元。而且从图中我们还可以看出在线医疗产业的发展速度之快，最高增长速率达到200.2%，近乎翻倍的增长也让人看到在线医疗产业的发展潜能。同时Makovsky Health and Kelton最近的调查报告也显示，有超过90%接受采访的美国群众表示在日常生活中使用过互联网查找相关医疗保健等信息；这也反映了不光是国内，乃至国际上，在线医疗产业的发展也呈现一种高速发展的趋势，由此也可以说明我们的研究是符合时代需求的。

图1.1 近年来中国在线医疗产业规模示意图

但是，在在线医疗产业如此迅猛发展的形式下，却出现了一种新的问题。这个问题多体现在国内，近年来，随着中国市场经济的高速发展，人均收入大幅度提升，国内有很多人开始在世界范围内寻求优质的医疗资源，这就使得国人去海外就医成为一种新的趋势。客观的说，国外的某些医疗水平确实比国内要好，在在线医疗产业上体现的尤为显著。所以，追求国际化的医疗方式是当前的一大趋势，但是，就目前的在线医疗产业来说，几乎所有的在线医疗平台针对的都是单语的医疗服务，比如中国的“春雨医生”，国外的“HealthTap”，“Yahoo Health”等等，这些都是比较著名的在线医疗平台，但是它们都只针对单一的语种，对于解决这种跨语言的医疗需求并不能给予很好的帮助。

鉴于这一问题，一个想法就是构建一个跨语言的医疗服务在线平台。当然，一个成熟的在线医疗平台需要较长时间的开发和改进。我们开发的面向医疗的跨语言信息检索系统只是一项较小的课题研究，旨在为更加成熟的跨语言医疗服务平台提供基础。

1.1 相关工作及现状

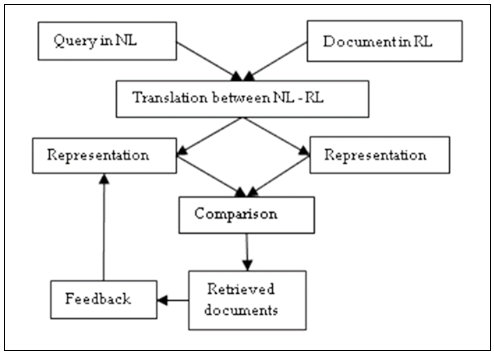
对于跨语言信息检索（cross-language information retrieval）问题，现阶段已经有了一些比较成熟的研究成果。作为信息检索（information retrieval）的一个子领域，跨语言信息检索所面临的问题在于如何构建查询与文档之间的相关性。这一问题的难点又在于不同语言之间的差异性导致处理过程的复杂多变性。Jian-Yun Nie等1提出的跨语言信息检索的框架结构如图1.2所示，这个框架的主要思想就是通过翻译将查询和文档表示成同一个向量空间的两个向量，然后计算两者的相关性，再加入必要的反馈来提升性能。

图1.2 跨语言信息检索框架示例

David A. Hull等2提出了一种基于词典（dictionary-based）的方法用于跨语言信息检索，其主要核心思想在于利用双语词典将查询翻译成目标语言的形式，在结合单语的信息检索手段得到结果。

上述所说的两种方式都是当下针对跨语言信息检索比较流行的框架，但是，事实上这些框架的弊端都存在于翻译的问题，无论是利用第三空间的向量表示来计算查询和文档的相关性还是利用双语词典进行翻译的方法，都要面临的问题就是语言差异性带来的一系列问题，就现在应用最为广泛的基于词典的跨语言信息检索方法来说，存在的问题就有对于复合词的识别和翻译，翻译歧义和超范词汇等问题，这些问题都使得跨语言信息检索的性能无法与单语言的信息检索相比。

Bo Li等32012年发表的论文中提出了一种基于信息的跨语言信息检索模型，该模型致力于将查询分割成若干个单元，并计算每个单元与文档的相关性，最后整合成为查询与文档的相关性，这种基于信息独立的模型也使得跨语言信息检索的性能有所提高。

当然，本文涉及的面向医疗的跨语言信息检索所面临的另一个问题就是目前比较规范的应用于医疗专业的双语词典比较匮乏，需要解决构建医疗领域的特殊双语词典构建问题。

1.2 主要工作内容

这篇文章的主要内容是构建一个面向医疗的跨语言信息检索系统，并通过其测试基于词典（dictionary-based）的特征匹配（feature-matching）算法，基于特征值的对数线性检索模型，以及最小错误率训练算法在信息检索的运用上的性能，然后进行分析评价。

我们所解决的问题主要如下：

**1、医疗双语词典的构建**

我们的双语词典主要针对的领域的是医疗专业，所以采用的方法是从Wikipedia上抓取医疗专业名词的中英平行预料，然后利用GIZA++工具4自动生成中英双语词典，再在此基础之上加入手工调整，并加入普适性词典，最终构成高质量的双语词典。

**2、数据来源及处理**

本文中所涉及的医疗问答数据来源于HealthTap，由于其数据格式由三部分组成，分别是问题（Question），答案（Answer）和标签（Tags），我们对所有数据进行整理，形成数据列表的格式，方便后面的使用。

**3、对数线性检索模型构建**

本文中采用的检索模型是基于特征值的对数线性模型，我们经过一系列的处理之后，对于每个查询-文档对得到固定数目的特征值，通过对每个特征值赋予权重求其线性叠加值作为衡量相关性的标准，最后通过求对数概率的方式得到检索概率，从而得到检索结果。

**4、模型参数训练**

在对数线性检索模型中，最重要的是每个特征值的权重，这是整个模型的参数，也是模型的核心。训练参数的方式是先人工标注一部分数据，利用最小错误率训练算法训练模型的参数，最后通过微调得到最优参数。

**5、系统的搭建**

我们的系统最后以网页的形式展现，类似于搜索引擎，对于给定的中文输入，返回相关的英文医疗问答文档。我们采用python-django框架作为后台，前段通过css+jQuery技术搭建整个系统。

1.3 论文组织结构

本文后续部分的组织结构如下：

第2章详细说明面向医疗的跨语言信息检索框架实现细节，其中内容主要包括双语词典的构建，前期数据的预处理过程，检索模型的构建和系统的运转流程。

第3章主要介绍本文中所利用到的算法及其核心思想，包括构建词典的基于平行预料自动构建算法，模型中用到的特征匹配算法，训练参数所用到的最小错误率训练算法。

第4章主要介绍本文中所涉及的实验内容，包括实验原理，实验的设计和详细步骤以及实验结果分析，主要参数训练实验和系统测试实验。

第5章主要是对前面工作进行总结和评价，包括对后续工作的介绍和研究的未来展望。

1. 面向医疗的跨语言信息检索框架

2.1 面向医疗的跨语言信息检索定义

为了更加直观，准确的描述面向跨语言信息检索的问题，本文在这里对该问题进行一些必要的形式化定义并给出一定的公式推导过程。

首先，通过问题的命名我们可以直观的发现两个限制性条件；第一个是面向医疗的，第二个是跨语言的。我们研究的问题属于信息检索的一个子领域；根据信息检索的定义，我们将查询集合表示为，将文档集合表示为。然后，我们就有了查询-文档组合，信息检索解决的问题就是计算和之间的相关性。针对跨语言的限制条件，本文所研究的目标是中英跨语言信息检索，所以，我们定义，这样我们将查询和文档分别表示为和，这样，我们的目标可以表示为下面的等式：

(2-1)

公式(2-1)表示的是和之间的相关性，其中的表示的是计算查询和文档相关性的函数，其具体细节我们将在第3章详细说明。

当然我们还要考虑到另一个限制条件，那就是面向医疗的，本文研究的内容针对的是医疗问答，我们所获取的数据都是问答类的数据集，所以在前面的基础之上，我们可以将文档的表示由改为，所以我们将公式(2-1)改写为：

(2-2)

至此，我们已经公式化的表示了面向医疗的跨语言信息检索中计算查询和文档相关性的过程。接下来，我们可以形式化的定义检索问题:

(2-3)

公式(2-3)表示的意义就是给定查询和文档集合（QA集合），进过检索返回相关稳定的一个列表，列表的顺序是按照相关性从高到低排序的。根据前面的形式化描述，我们可以将整个问题的过程用图2.1表示。

从图中，我们可以看到，给定查询和文档集合（QA集合），通过计算，得到文档集合中每个文档与查询的相关性，最后通过计算到的相关性得到检索结果，结果表示为m个相关文档的列表形式，按照相关性排序。

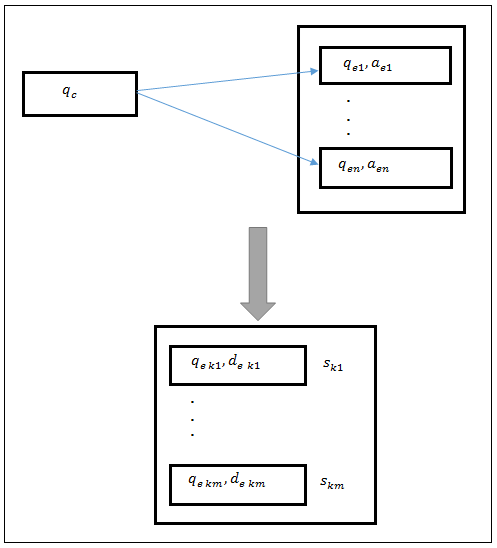
至此，我们完整的对面向医疗的跨语言信息检索问题进行了形式化的分析和描述，并给出必要步骤的公式推导。

图2.1 面向医疗的跨语言信息检索过程示意图

2.2 英汉双语词典构建

2.2.1 平行语料的抓取

为了利用平行语料构建高质量的英汉双语词典，所以需要获取一些高质量的平行语料。我们采用的方法是从Wikipedia上抓取一些弱标志平行数据，这里所谓的弱标志数据是指Wikipedia网页上链接、标题等数据。为了更加符合面向医疗的条件，我们仅仅抓取了Wikipedia上有关医疗的英汉网页内容，然后对其进行解析，从而得到平行预料。

图2.2所示就是Wikipedia弱标志数据，我们从抓取到的网页中抽取这样的平行语句对，整合到一起形成质量较高的平行语料库，而且，我们抓取的内容是针对医疗方面的，所以这些平行语料的领域就比较专一，这也符合我们构建医疗专业领域英汉双语词典的条件。

通过从Wikipedia上抽取出这些高质量的平行语料之后，我们就可以利用这些平行语料进行自动生成双语词典，具体的操作将在下一小节介绍。

图2.2 Wikipedia弱标志数据示例

2.2.2 自动构建双语词典

通过2.2.1我们已经得到了比较高质量的平行预料，接下来要做的就是利用这些平行预料生成双语词典，采用的方法是使用GIZA++工具自动生成。

关于GIZA++工具的使用方法这里将不进行介绍，主要是参照指导说明说进行操作。其大概步骤分为三步，首先下载源码，编译GIZA++和mkcls；第二步，利用从Wikipedia上获取的平行语料，通过mkcls构建word class；最后一步，再次利用平行语料，通过GIZA++进行IBM Model训练，最后得到词语对齐的结果文件。最终得到的文件中内容格式如表2.3和表2.4所示。

表2.3 giza++训练结果文件(e2c)

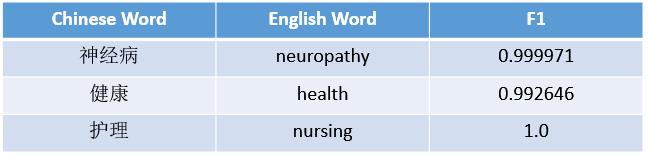
表2.3中所示的是GIZA++训练后English到Chinese的结果文件内容格式，总共三项，第一项为English word，第二项为Chinese word，最后一项是可信度F1。其中F1的值表示的是英文单词和中文词语互为翻译的可信程度，F1值越高，则表示两个词语互为翻译对等体的概率越高，反之则越低。

表2.4 giza++训练结果文件(c2e)

表2.4中所示的是GIZA++训练后Chinese到English的结果文件内容格式，跟表格2.3唯一不同的就是第一项跟第二项位置互换，第一项表示Chinese word，第二项表示English word，F1的意义同表2.3。

得打这样的两个文件之后，就可以生成词典了，我们以表2.3中所示的“health-健康”为例，说明词典生成过程。首先建立双向词典{health:健康}和{健康:health}，并且分别对它们做标记，标记的内容就是表2.3和表2.4中的F1的值，例如{health:健康}(0.680164)和{健康:health}(0.992646)。接下来，我们合并双向词典，得到一个互为翻译的词组<health, 健康>，这就是我们的词典的一个条目。需要说明的是，合并的过程需要考虑到两个词典的标记，即F1的值，为了使得最后的词典质量较高，我们设定的条件是：

(2-4)

(2-5)

公式(2-4)约束了双向翻译可信度的总和需要高于设定阈值，公式(2-5)约束了每个单向翻译的可信度需要高于设定阈值，公式(2-4)的作用是选取高质量的翻译，公式(2-5)的作用是方式仅由公式(2-4)约束而导致的极端误差情况。

至此，我们的英汉双语词典自动构造已经基本完成，但是为了提高词典质量，我们采取了一定的人工矫正手段，例如删除一些质量不好的翻译，或者修正一些不正确的翻译等等；为了解决词典覆盖率问题，我们采取添加普适性词典的方法，向处理好的双语词典中加入规范的普适性词典，这样就形成了最终的高质量、特定领域、高覆盖率的英汉双语词典。

2.3 数据预处理

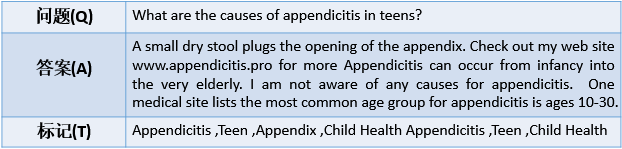
本节讨论的数据预处理针对的是系统所用到的文档数据，即来源于HealthTap上的英文医疗问答数据。首先，我们介绍初始的数据格式：

表2.5 文档数据初始格式

表2.5所示的是初始的数据格式，从表中可以直观的看出，数据主要分为三个组成部分，问题(Q)，答案(A)和标记(T)。其中，问题(Q)是HealthTap平台上用户向医生提出的问题，答案(A)是医生针对病人的问题给出的回答，而标记(T)是根据问答的内容作出的标记，主要用来表示问答内容的主题，核心，范围等等信息。

我们定义由表2.5所示的问题(Q)，答案(A)和标记(T)组成的三元组为一组数据。经过统计，我们的初始数据为426508组。

由于初始数据比较粗糙，会出现重复问题。为了解决这个问题，我们需要对数据进行去重处理。初始数据的重复问题主要出现在问题(Q)和答案(A)上，观察发现，在初始数据中，有大量的相同的问题(Q)和答案(A)组成的重复数据，例如和在初始数据中视为两组不同的数据，但是实际上这两组数据的问题(Q)和答案(A)是相同的，只是由于做的标记(T)不同而导致它们被视为两组不同的数据，明显的，这样的两组数据在检索中应该视为同一个文档，所以我们要做的就是消除这些重复的数据，但是不能丢失标记(T)信息。

我们采取去重合并的方法，首先我们对于初始数据的每一组数据，建立一个映射关系。将问题(Q)和答案(A)组成的二元组作为key值，将标记(T)作为value值，这样我们形成的映射关系形如。经过这样处理之后，我们进行下一步的去重。遍历所有的初始数据，根据映射中key值的唯一性性质，我们迭代更新旧的映射关系从而形成新的映射关系。在更新的过程中，我们要注意的一点是，不能用直接用新建的映射的value值去替换原始的value值，即当我们有一个旧的映射，在用新的映射去更新它时，我们应该得到的是映射而不是映射。因为只有这样，我们才能保证标记(T)信息被完整的保留。经过去重合并处理之后的数据规模经统计为225478组，相比较于初始数据，我们可以看出处理之后的数据精简了很多。

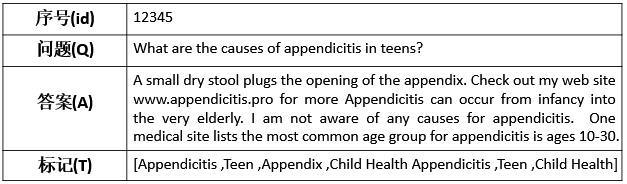
处理完数据重复问题之后，接下来我们要处理的一个问题就是数据的形式化。我们的系统最后的检索结果是这些数据的一个列表形式，那么这个列表的具体形式以及涵盖的信息将会对系统性能有比较大的影响。

表2.6 数据列表形式

表2.6所示就是我们的列表形式，首先，我们对每一个三元组添加一个唯一序号(id)，这个序号的作用用来唯一区分每一组数据。然后，我们将标记(T)表示成为一个列表的形式，目的在于后面做粗检索时可以快速查询。

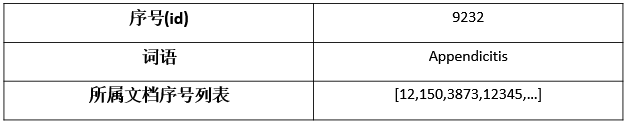
接下来，为了加速检索系统的检索速度，我们还需要对数据进行进一步处理，首先，我们统计所有的标记(T)中出先的所有单词，即利用标记(T)做一个词表。然后，我们对词表中的每一个词进行处理，处理的过程是统计该词语在哪些三元组中出现过，并记下这些三元组的序号(id)。

表2.7 标记词汇处理

表2.7所示的是我们对标记词语所做的处理方式，序号(id)作为唯一区分词语；词语指的是由标记(T)生成词表里的每一个词语；所属文档序号列表记录的是所有标记(T)包含该词语的三元组的序号的集合，它是以列表形式存储，方便后面的检索。

至此，我们已经完成了所有的数据预处理问题，主要解决了数据重复问题和数据形式化问题，这些处理对系统性能的提升都是至关重要的。

2.4 检索系统框架

2.4.1 检索模型

针对我们面向医疗的跨语言信息检索问题，我们采用了一种对数线性模型，这种模型的基础是基于词典翻译的跨语言信息检索技术，采用特征匹配的方式进行检索概率计算。

(2-7)

公式(2-7)完整的表示了整个检索模型。其中的分别表示的是英文问题，中文问题，英文答案和检索问答文档；是模型参数，是计算特征值的函数。这个模型的核心在于对计算好的查询与文档之间的特征值进行线性组合，求出其检索概率。

在本小节，我们只介绍模型的整体构造，对于模型的核心算法以及模型构造的依据，还有特征函数的构造，模型参数的训练，调整等等详细信息我们将在第3章进行阐述。

2.4.2 检索策略

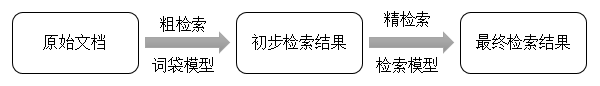
2.4.1节中提到的检索模型虽然是我们采用的核心检索模型，但是如果仅仅只是利用这个检索模型进行最终系统的检索必然会出现检索速率过慢，检索结果稀疏等问题。为了解决这个问题，我们采取了一种由粗到精的检索策略。

图2.8 检索策略示意图

图2.8所示的就是检索策略的流程，第一步，我们根据查询对原始文档进行一次粗检索，粗检索利用的就是传统的词袋模型，找出所有相关的文档，这样做的目的在于在不遗漏相关文档的情况下，缩小检索范围，为后面的精检索做准备；第二步，对第一步检索的结果进行精检索，精检索采用的模型就是2.4.1节介绍的对数线性检索模型。这样做的好处就在于可以大大加速检索的效率，因为从2.4.1节的检索模型可以看出，计算检索概率的操作相对来说比较复杂，如果利用这个模型对所有文档进行计算，会多做许多不必要的计算，影响系统效率。粗检索的词袋模型速度非常快，可以在很少的时间代价内大大缩小检索范围，从而达到加速检索的目的。

2.4.3 系统框架

为了方便系统的使用，我们需要给系统设定一个整体的框架，我们设想的框架类似于一般的搜索引擎，给定输入框输入查询问题，然后根据输入检索得到结果。

图2.9所示的就是系统的一个框架整体，从这个框架中我们直观的看到，当在输入框中输入需要查询的问题时，点击Search，系统将检索结果返回到页面，我们可以看到所需要的英文医疗问答信息，当然，通过点击链接的方式，我们还能查看详细的问答结果。

图2.9 系统框架示意图

1. 算法

3.1 词语对齐模型

本节主要介绍利用平行语料构建双语词典过程中所用到的词语对齐模型。在第2章中，我们已经说明构建双语词典所利用到的工具是GIZA++，所以这里，我们对其中用到的几种重要模型进行说明。

3.1.1 词语对齐的定义

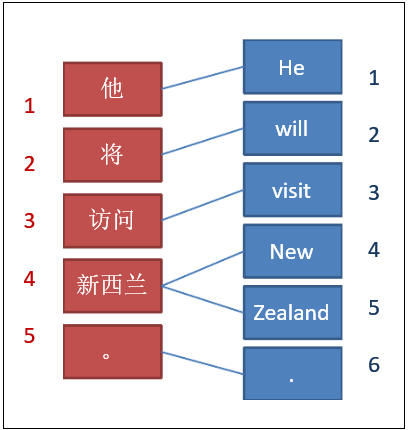
根据Brown5等发表的文章定义，将词语对齐定义为在平行的文本中指明互相对应的词语的过程。图3.1表示的就是一个词语对齐的例子。

图3.1 词语对齐示意图

词语对齐是一个相当复杂的过程。通常来讲，一个词语对齐的效果取决于几个因素的影响，包括重排序，省略，插入以及词对短语对齐。因此，我们需要一个非常概括性的词语对齐表示。形式化地，我们对词语对齐使用如下定义。给定一个源语言（中文）串和一个目标语言（英语）串，对这两个平行串做词语对齐。我们定义这两个词语串的对齐为单词位置的笛卡尔乘积的一个子集，即，一个对齐定义为：

(3-1)

将词语对齐问题建模为源语言和目标语言位置的任意关系是相当有概括性的，随着对齐模型的发展，这种表示方法成为一种一般的形式化表示方法。通常来说，一种模型的提出会在此基础之上强加一些约束条件，例如，一个源语言词语只能对应一个目标语言词语等等。

3.1.2 IBM模型

GIZA++工具主要使用的是IBM Model(1-4)，它们都是统计概率模型，并且是无监督地从语料库中学习。本节主要介绍IBM Model从1到4的一个发展过程，以及模型的架构和原理。

首先是Model 1，这是一个简化的模型，使用了一个均匀分布：

(3-2)

从公式(3-2)我们可以看出Model 1是一个相当简化的模型，它将看成一个均匀分布，而对于一个长度为的源语言（中文）句子和一个长度为的目标语言（英文）句子，其对齐方式总共有种，所以按照均匀分布，得到每种对齐方式出现概率为，最后乘以当前的对齐概率，就得到了IBM Model 1。我们还可以看到，模型1只考虑了两个平行句子的长度因素，而句子中词语的顺序对于最后的对齐概率没有任何影响，由此也能看出模型1是一个相当简化的模型。

IBM Model 2在Model 1的基础之上加入了绝对的调序信息，我们得到：

(3-3)

从公式(3-3)我们可以发现相对于公式(3-2)，这里多了一个参数，我们考虑这个参数信息，其中的和分别是目标语言（英语）句子长度和源语言（中文）句子长度，这个参数计算的是源语言句子中第个单词在目标语言的对齐位置是的概率，由此也可以看出这里考虑到了顺序对应的问题。从形式上我们可以直观的看出Model 2的复杂度比Model 1要大，其考虑的信息更加完整，效果也比Model 1要好。

IBM Model 3在Model 2的基础之上加入了一个新的概念—繁殖子模型，这个模型的定义如下：

(3-4)

公式(3-4)给出的是繁殖子模型的定义，繁殖子模型提出的意义在于，当一个源语言的词语对齐的并不仅仅是一个目标语言词语，而是多个时，Model 1和Model 2是无法解决的，所以就提出了这样一种繁殖子模型，的结果就是源语言句子中第个词语对应于目标语言句子中词语的个数。在此基础之上，定义一种新的对齐的形式如下：

(3-5)

从公式(3-5)可以看出这种新的对齐形式旨在定义一种新的对齐方式，对于源语言句子中的第个单词，与其对应的目标语言句子中的单词位置是所有单词位置的一个子集。这样就解决了Model 1和2不能解决的一对多问题；在此基础之上，IBM Model 3的定义如下：

由公式(3-6)可以看出在Model 3中加入了繁殖子模型的约束，并且在计算公式中项时，忽略了当前词语的对齐对于它前一个词语的对齐的依赖关系，也就是说，对于项的计算如下：

(3-7)

IBM Model 3加入了繁殖子模型之后，增加了模型的复杂度，帮助解决了一对多的对齐情况，使得模型的处理能力得到较大的提升，从Model 2到Model 3是一个比较大的提升。

IBM Model 4跟Model 3的定义形式是一样的，唯一不同的就是在Model 3的基础之加上了相对调序的信息，这里不再过多的叙述。

至此，已经将IBM Model(1-4)介绍完毕，包括形式化的定义和提出的原理。我们在利用GIZA++构建英汉双语词典的过程中，主要使用了这四个模型，利用从Wikipedia上获取的平行语料，进行训练，最后得到对齐关系和对齐概率，根据对齐关系和概率生成所需要的词典。

3.2 特征匹配算法

在第2章的检索模型中，我们使用的核心模型中，利用特征函数计算查询与检索问答文档之间的特征值，最后通过这些计算的特征值的线性组合来直接求查询与文档之间的相关性。这里我们称这个过程为特征匹配。下面详细介绍特征匹配算法的原理和过程。

首先，特征匹配的思想来源于Franz Josef Och6等提出的统计机器翻译的最大熵模型。

3.2.1 直接最大熵翻译模型

Franz Josef Och等人提出的最大熵翻译模型不同于之前的源信道模型，他提出通过构建源语言句子和目标语言句子之间的特征函数来体现两者的翻译可能性。不同于传统语言模型和翻译模型，其形式化定义如下：

(3-8)

从公式(3-8)可以看出最大熵翻译模型的核心在于的构建，只有合理的构造特征函数才能保证计算得到的特征函数值能够表现出源语言句子与目标语言句子之间的翻译概率。

Franz Josef Och在这篇文章中提出了几种特征函数的构造思路，例如特征函数的值等于句子的长度，特征函数是语言模型等等。这个模型的最大好处在于可扩展性强，通过添加特征函数的个数，可以将很多信息考虑进去，只要在训练时参数调整合适，能够得到一个相对较优的结果。

3.2.2 检索中的特征匹配

受3.2.1中最大熵模型的启发，机器翻译的目的在于计算源语言与目标语言之间的翻译概率；而检索的目的在于计算查询与检索文档之间的相关性。我们将检索问题类比到翻译上，将查询看成是源语言端，将检索文档看成是目标语言端，那么我们就可以构造出公式(2-7)的结果。

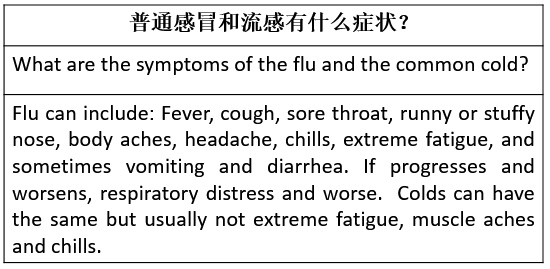
那么如何构造公式(2-7)中的成为检索模型的关键。为了解决这个问题，我们首先看一下我们的检索任务。

图3.1 检索任务示例

图3.1显示的是检索中查询和问答文档的基本格式，从图中我们可以看出，查询问题与文档中的问题几乎互为翻译体，而在文档的答案中也有许多与查询相关的信息。下面，我们以图3.1所示为例，介绍特征函数的构造过程及原理。

首先，我们考虑中文问题和英文问题，忽略英文答案，这两者在长度上很接近，相差不多。所以，自然而然的想到以这两者的长度构造一个特征函数：

(3-9)

其中，是求长度的函数。

从公式(3-9)可以看出，当英文问题与中文问题越接近时，该特征函数的返回值越大，当然，为了避免句子本身长短给结果造成的误差，所以加入了两者长度之和进行矫正，为了避免两者长度相等导致分母为0，所以给分母添加一个偏置量1。

接下来，我们考虑到采用了基于双语词典的方法，那么将查询中的词语经过双语词典进行翻译后得到一些关键词的集合。有了这些关键词，我们可以构造特征函数如下：

(3-10)

其中，是统计所有关键词在文档中出现次数之和的函数，和是两个参数，且都是正实数。

从公式(3-10)可以看出，当关键词在英文问题和答案里面出现的次数越多，则特征函数返回值则越高，也说明英文问题与答案和查询之间的相关性相对而言要高一点。针对英文问题和答案在相关性影响上的差异性，所以给它们分别设定两个参数，即权重，用来区分两者。实验中，。

最后，我们可以考虑信息检索的常用特征Tf-idf，由此构造特征函数如下：

(3-11)

其中和分别是传统的Tf和idf函数。

公式(3-11)加入了传统检索参数Tf-idf，可以使得我们的系统更加的完善，综合考虑多个关键词的信息，可以使得系统在处理一般情况时性能表现得更好。

除了上述三种主要的特征函数之外，我们还加入了中文问题的长度，英文问题的长度，英文答案的长度，文档的词表长度等特征，因为这些因素对查询和文档的相关性都有一定的影响，综合考虑这些信息，并将其整合到系统模型中可以保证模型的完备性，通过后续的参数训练和微调，使得最终的模型处理更加准确可靠。

至此，我们已经将检索核心模型中所用到的特征匹配算法的核心思想以及具体过程原理介绍完毕，此模型的优点在于可根据情况进行修改，有很好的扩展性。

3.3 最小错误率训练算法

在3.2节中，我们介绍了系统核心检索模型使用的特征匹配算法。在检索模型中，我们设定了一些模型参数，这些参数代表的意义是对应特征函数的权重，即对应特征函数对查询和文档相关性的相对影响程度的大小。然而这些参数在建模时并不是最优的，我们需要通过训练数据的训练而得到，才能使得模型的性能得到提升。我们使用的是最小错误率训练算法。

3.3.1 数据标注

模型的训练需要两个元素，一个是训练算法，另一个就是训练数据，在开始训练之前，我们必须有一定量的训练数据，并且训练数据的质量和数量都对模型的训练效果有着至关重要的作用。

我们的训练数据是通过人工标注的方式获得的，首先我们利用系统的粗检索功能找出与查询相关的文档，然后人工对这些文档的相关性进行标注，最后整理成为标准的训练数据，形式化的过程如下：

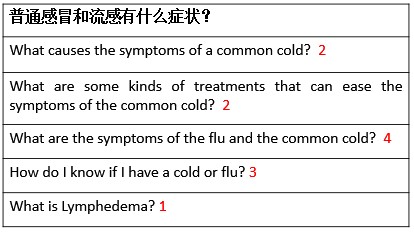
1. 给定中文查询，利用粗检索，返回文档列表。
2. 人工对返回列表进行相关性标记，得到。

图3.2 标注数据示意图

图3.2展示的就是标注数据的过程，其中，我们将文档和查询的相关性划分为4个等级，分别用1、2、3、4表示，数字越大，表示相关性越强，反之，则相关性越弱。

通过标注，我们获得了100个不同中文查询的标注数据结果，在训练时，我们将其中的60组问题的标注数据作为训练集，剩下的40组数据作为测试集。

3.3.2 最小错误率训练

(F. Och. 2003)7提出了一种用于统计机器翻译的对数线性模型参数训练的最小错误率训练算法(MERT)。其实在此之前，也有许多参数训练的算法，它们大多数是以翻译结果的最大翻译概率为优化目标，从而达到求解最优参数的目的。但是问题出现在大多数机器翻译的评价指标是BLEU8而不是翻译概率，所以往往会出现参数训练达到最优，但是翻译效果并不是最优的情况。而MERT合理的解决了这个问题，它的最大成功在于，直接基于评价指标的优化手段，使得参数的优化直接作用于评价指标，这样最优的参数优化就导致最优的评价指标。

虽然MERT算法的提出，是为了解决统计机器翻译中对数线性模型的参数优化问题，但是，在3.2节，我们就讨论过我们的检索模型的来源是最大熵翻译模型，而且我们的检索问题类比到机器翻译上原理是一样的，所以我们可以利用MERT算法来实现检索模型的参数优化。下面就介绍MERT算法在我们的检索模型中的参数优化基本原理。

首先，假设我们的模型参数为，总共有维，现在对其中的第维参数进行优化，根据MERT算法，我们需要固定其他维参数，只有可以自由变化。

现在，假设对于一个中文查询，我们经过粗检索，得到相关的文档集合（为了方便，这里不区分问题和答案，统一称为文档）。根据模型，对于每个相关文档，我们计算特征值的线性组合，作为每个相关文档的得分，对于任意一个相关文档，我们有：

(3-12)

其中是第个特征函数。

由上面的假设，除了第维参数，其他都被固定，所以有：

(3-13)

其中表示的是一个由其他维参数与特征值线性组合的一个固定值，在优化时，我们可以将其看成是固定的值。

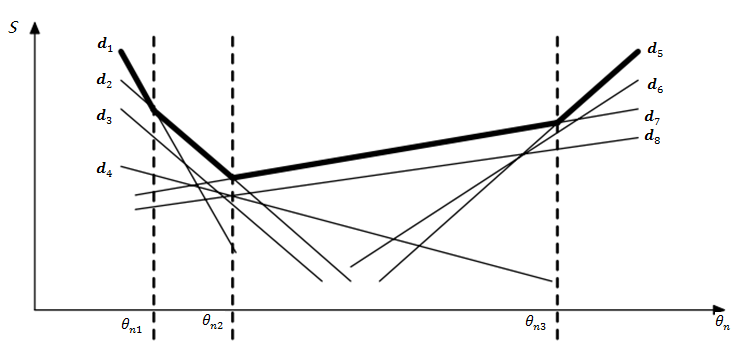
由公式(3-13)我们可以看出，对于每个相关文档的得分，其关于是一个一维线性的关系，可以将其抽象成直线。

图3.3 MERT优化曲线示意图

图3.3给出了相关文档得分随参数的变化而变化的曲线，其中曲线加粗部分是每个区域内的所有相关文档的得分上限。观察曲线，我们发现，随着参数的变化，每个相关文档的得分都在变化，但是如果我们按照每个相关文档的得分给它们进行排序处理，那么显然只有在经过这些曲线的交点处时，排序结果才会发生变化，而MERT的优化目标就是使得每个相关文档的得分排序与实际评价指标的排序结果尽量一致。所以我们只用考虑这些交点的情况即可。

那么我们的优化步骤分为两步：

1. 根据曲线交点，将横轴的取值范围分为若干个区域；
2. 遍历1中所划分的所有区域，计算每个区域中的评价指标，找到使得评价指标最优的区域，将该区域内的值作为的值。

完成了第维参数的优化，其他维的参数优化过程原理是一样的，但是要注意的一点就是，优化过程是迭代进行的，因为，在每一轮优化过程中，其他维的参数跟上一轮相比是有变化的。

在我们的参数优化过程中使用的是Z-mert9工具，由于Z-mert工具里面提供的评价指标只适合机器翻译，所以在使用过程中，需要先改变其评价指标，然后进行参数优化训练，具体过程在下一章进行说明。

第四章 实验

4.1 实验环境设置

4.1.1 数据准备

根据第三章介绍的内容，我们的实验数据来源于人工标注。对于一个中文的问题，我们对系统粗检索返回结果进行标记，得到标记数据用来训练参数的实验。为了方便实验并且结合实际，我们将对于每个中文问题的粗检索结果限定为100个，我们对100个问题进行了标注，所以最终得到的数据规模为100\*100=10000条。表4.1为数据详情。

表4.1 实验数据说明

如表4.1中所示，我们将100个问题分为两组两组，一组数目为60个，另一组为40个；其中60个问题作为训练集，用来训练参数，40个问题作为测试集，用来测试训练效果。

4.1.2 环境配置

本实验利用的是最小错误率训练算法，使用的是Z-mert工具，但是Z-mert工具的原始版本只针对机器翻译的参数训练，对于本实验，需要对Z-mert进行一些必要的修改，以便其适用于本次实验。

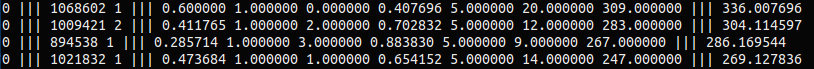
F:\decoder1.png首先第一个是对于解码器的修改，原始的Z-mert工具的解码器针对的是机器翻译，所以在解码时会利用到参考译文选项，对于检索问题，我们的实验没有参考译文，在解码文件中，我们需要的仅仅是候选文档排序列表以及与其对应的特征值。

图4.3 修改后Z-mert解码文件示例

图4.2 原Z-mert解码文件示例

图4.2和4.3所示的分别是修改前后Z-mert解码器生成的解码文件格式，文件中每一行代表一个候选项对应的各项参数，参数用符号“|||”隔开。从前至后每项分别代表的意思是候选项对应的问题序号，候选项，候选项对应的特征值以及当前评分。我们主要修改的部分是将第二项候选项由原来的候选语句替换成粗检索文档对应的编号以及它们各自对应的人工标注的分数，这样做的目的有两个，第一个将候选句子替换成粗检索文档对应的编号是为了方便记录，因为文档的长度可能会很长，直接记录文档太过麻烦，而我们的评价标准是按照人工标注的分数，与文档本身并无关系；第二个就是为了训练时评价方便，所以需要加上人工标注信息。

第二个要修改的部分就是评分程序，在原始的Z-mert工具包里面，包含了一些对于机器翻译的评价指标，例如BLUE等等。但是这些评价指标都不适合信息检索，为了使其能够用于本次试验，所以需要自己编写相应的评分程序，我们采用的评价指标为Precision@N，所以我们的评分程序需要与之一致，我们采用的方法是直接读取人工标注的分数，将其转换为该选项的评分，这样做就能够使得人工评分高的选项排名靠前，而人工评分低的选项排名靠后，即满足Precision@N增加的条件。

本实验使用的Z-mert版本为v1.50，实验环境为Linux操作系统，所有实验均采用单机模式。

4.2 评价指标

对于一个检索系统，如何评判其性能的优劣是一个关键性的问题。传统的信息检索评价指标有很多种，其中代表性的有Precision, Recall, F-score, MAP(mean average precision), P@N(Precision@N)等等。下面简单介绍一下这几个指标。

第一个是Precision, Recall, F-score。这三个指标是互相关联的，通常一起用来表示检索性能的好坏。

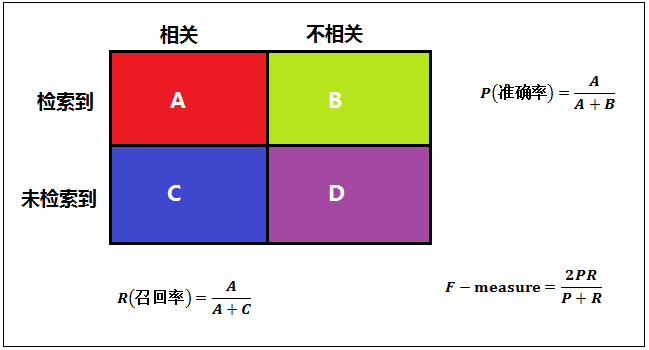


图4.4 Precision, Recall, F-score计算示意图

如图4.4所示，准确率P表示的是系统检索到的相关文档与系统所检索到的所有文档的比值。直观理解就是系统检索的文档中相关性文档所占的比例，召回率R表示的是系统检索到的相关文档与所有相关文档的比值，直观的可以理解为系统检索相关文档在数据库中所有相关文档中所占的比例。通过计算公式我们不难发现，准确率P体现的是系统检索结果的准确率，召回率R体现的是系统检索相关文档的能力。一个理想的检索系统当然是希望这两个指标越高越好；但是实际上一般出现的结果都是一高一低，所以可以利用F-score来限制要求两者都较高。

第二个是MAP(mean average precision)，或者称之为平均准确率。这个评价指标解决了P, R, F-measure三个指标在一些情况下的局限性，而且这个指标充分考虑到了检索结果的排序对性能的影响。实际中，一个检索系统如果能够严格按照相关性的高低来给检索结果排序，那么它的性能就是非常好的。

公式(4-1)、(4-2)、(4-3)所示的就是MAP的计算过程，从中我们可以看出MAP计算的重点在于对待相关性排序上，这种计算方式的理念很符合用户的观念，因为往往用户的希望是在最靠前的位置找到更加相关性的文档，如果我们将一个检索结果按照相关性从小到大排序和从大到小排序，在第一种评价指标(P, R, F-measure)中是无法区别的，但是在MAP评价中却能很清楚的分辨出那种检索结果更加好。

最后一种是P@N，其全称是Precision@N。我们可以将其理解为前N个检索结果的准确率。这个评价指标充分考虑到检索系统的点击因素，统计发现，绝大多数用户在使用检索体统进行信息检索时会去点击那些靠前的检索结果，而对于那些位于后面的检索结果，则只有极少数的用户会去查看并且点击它们。这就说明从实用性角度来考虑，在检索结果的前面几个结果中提高准确率会使得系统的性能提高的更多。

P@N的计算方法很简单，从其定义就能推测出来，即求前N条检索结果的准确率，假设前N条检索结果中有m条是与查询相关的检索结果，那么P@N的值就为m/N。

准确的来说，这三种评价指标各有优劣，对于我们系统性能的评价，在实验中我们一律使用的是P@N。选择P@N作为评价指标的原因有一下几点：第一，P@N更加符合用户对于检索系统的需求，这一点是P@N本身的特性；第二，计算方便，这一点充分考虑到P@N的计算快速有效，对实验帮助较大；第三，P@N更加注重单一的查询性能，这一点很符合我们的系统需求。

在后续所有实验中，P@N的N值取10。

4.3 实验结果

本节将会详细给出参数训练的实验结果，包括初始化，中间结果以及最终的参数训练结果。

首先是参数的初始化情况，实验中参数的个数为7，分别是模型中7个特征函数的权重，命名分别为f1, f2, f3, f4, f5, f6, f7。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 参数名 | f1 | f2 | f3 | f4 | f5 | f6 | f7 |
| 初始化 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 调参范围 | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ | ∞ |

表4.5 模型参数初始化示例

表4.5给出的是训练开始时模型参数初始化的情况，为了实验方便，将所有模型参数初始化为1.0，并且将每个参数的调参范围为设定为整个实数范围（计算机能表示的实数范围），这样能够有利于找到最优解。

接下来是mert算法的参数初始化，主要有一下几个参数：

1、-m：评价指标

2、-maxIt：最大迭代轮数

3、-ipi：每轮迭代的中间参数点数

4、-N：n-best的n值

5、-seed：随机数种子

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| -m | -maxIt | -ipi | -N | -seed |
| PN | 15 | 20 | 10 | 12341234 |

表4.6 mert参数初始化示例

表4.6给出了mert参数初始化情况，评价指标为P@N；最大迭代轮数为15；每轮迭代中间参数点为20个；n-best取10，与P@N相对应；随机种子设为12341234。

至此，实验初始化已经完成，接下来进行参数优化实验，我们给出每轮参数结果以及评价指标的变化。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | f1 | f2 | f3 | f4 | f5 | f6 | f7 |
| It0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| It1 | -2019.9 | 842.2 | 1.0 | 5.4 | 1.0 | 1.0 | 1.0 |
| It2 | -2019.9 | 462.7 | 1.0 | 2.6 | 1.0 | 1.0 | 1.8 |
| It3 | -165.7 | 52.8 | 0.8 | -2.6 | 0.9 | -0.8 | 0.1 |
| It4 | -165.7 | 52.8 | 0.8 | -2.6 | 0.9 | -0.8 | 0.1 |

表4.6 参数随迭代轮数的变化

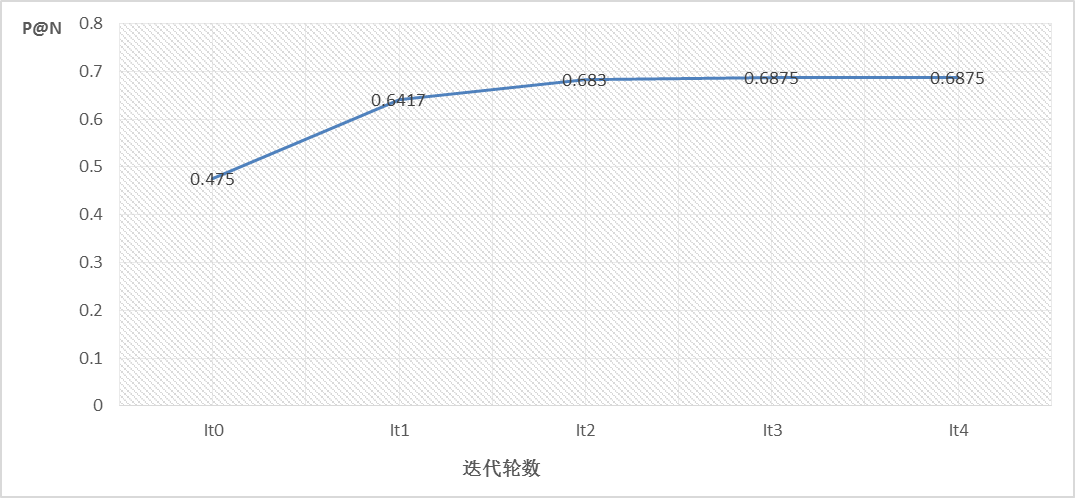
表4.6所示的是参数在迭代训练过程中的变化，从中我们可以看出，总共进行了4轮迭代，其中It0表示的是初始的参数设定，It1, It2, It3, It4分别表示第1、2、3、4轮迭代的参数结果。

图4.7评价指标随迭代轮数变化示意图

图4.7所示的是评价指标P@N在参数训练过程中的变化曲线，从整体趋势看，P@N的值随着迭代轮数的增加而增长，第3轮迭代时达到最大并且第4轮迭代没有变化，说明此时处于最优解状态。

我们将训练的参数代入模型，并且利用测试数据进行测试，并将测试结果与训练结果做比较。

表4.8 训练集与测试集结果对比

表4.8所示的就是经过训练的参数在训练集和测试集上的评测对比示例，从中我们可以看出，在测试集上的P@N比训练集上要低，这也是符合常理的。

接下来，我们分析算法参数中的-ipi的值对于最后结果的影响，我们将-ipi的值分别设为10、20、30、40进行对比试验，得到如下结果：

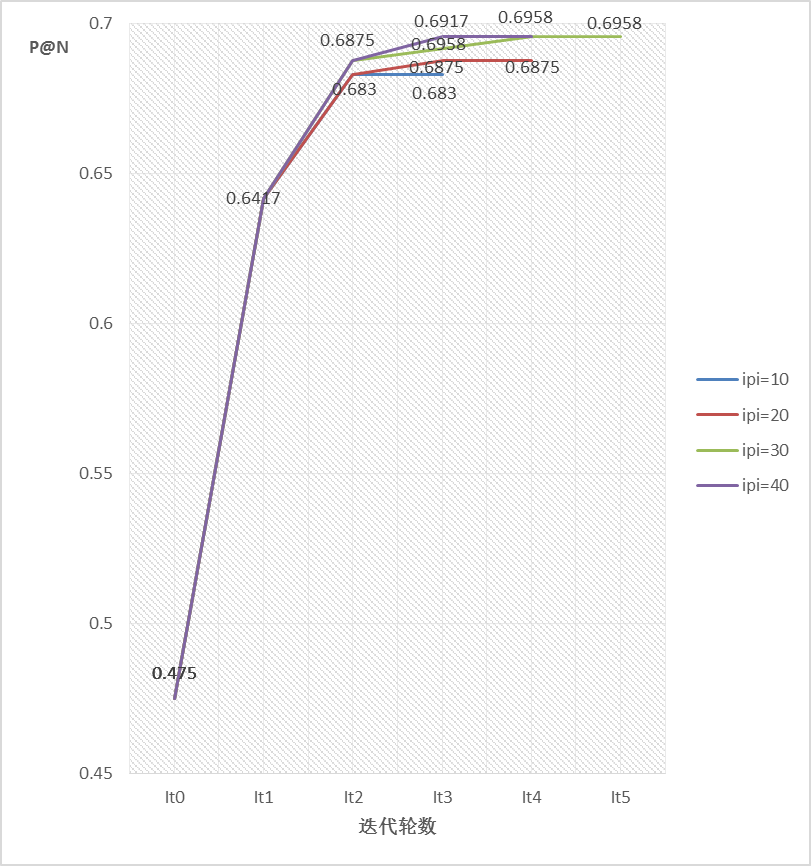
图4.9显示的是ipi分别为10、20、30、40时训练曲线的对比情况，从中可以直观的看出，当ipi=30和ipi=40时，训练得到的最终评价指标最高，不同的是党ipi=40时找到最优解所需要的迭代轮数更加少。

图4.9 不同参数下的实验对比曲线

4.4 实验分析

对于上一节的实验结果，我们发现一些值得思考的问题，首先，为什么n-best的n值需要跟P@N取相同的值？为什么参数ipi的值会对实验结果产生重要影响？

对于第一个问题，我们的解释很简单，因为mert训练的优点在于直接对评价目标进行优化，既然我们的P@N取N=10，那么我们将n-best取n=10正好满足优化的目标就是我们最终对系统性能评测的指标，这样使得优化的目标与结果相互统一。

对于第二个问题，我们需要利用图3.3进行解释。在3.3.2小节中对mert算法原理进行解释时就已经说明过，mert算法是通过对直线交点将横坐标划分为若干个区域。那么显然的，在每轮迭代中考虑点的个数会对实验有很大影响，根据原理，当每轮迭代考虑到所有直线的交点时，算法的优化应该出现最优解，所以参数ipi的值对实验结果有很大影响，并且当ipi值越大时，实验效果会越好，这一点也符合实际的实验结果。

通过对比实验，我们选取了参数ipi的值为40，此时模型参数训练具有最优解，参数具体情况如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| f1 | f2 | f3 | f4 | f5 | f6 | f7 |
| -71.69 | 24.42 | 4.52 | -1.53 | 0.23 | 0.22 | -0.05 |

对这些参数取值的合理性我们也需要进行一些分析，这里以f1, f2, f3为例，它们对应的三个特征函数分别是：

1. 中文查询与英文问题的长度关系：值越大，相似度越低；
2. 关键词在英文问题中出现的次数：值越大，出现的次数越多；
3. 关键词在英文答案中出现的次数：值越大，出现的次数越多。

按照逻辑关系，f1的值应该是一个负数，实验结果为-71.69，符合理论分析；f2与f3的值应该都是正数，并且关键词在英文问题中出现的次数对相关性影响应该更大，所以有f2>f3，实验结果为f2=24.42, f3=4.52，符合理论分析，由此可见，上述实验的结果基本符合理论分析，没有出现违背逻辑关系的情况出现，基本认定实验结果的有效性和合理性。

1. 总结

本章主要对这篇论文的工作进行总结，并且提出工作中的一些不足和需要改进的地方，最后是对未来工作的一些展望。

5.1 工作总结

随着社会经济的发展，人们对于健康问题的关注度越来越高，而且伴随着互联网技术的高速发展，人们的生活节奏加快，在线医疗渐渐成为人们生活中看病治疗的一种重要方式。语言沟通的障碍导致现在的主流在线医疗平台都是单语种的，这与越来越多的国人倾向于去国外就医的趋势产生了矛盾，建立一个跨语言的在线医疗平台成为一种广泛的需求。

相对于传统的单语在线医疗平台，跨语言的在线医疗平台有着更为广泛的信息来源，更加专业的医疗手段，并且能够为更多的人提供服务帮助。研究如何构建一个面向医疗的跨语言信息检索系统对于构建跨语言的在线医疗平台具有重大的意义。

本文以医疗问答信息的跨语言检索为目标，构建了一个跨语言信息检索系统，主要研究了基于特征匹配的对数线性检索模型在跨语言信息检索问题上的效果以及基于词典的跨语言信息检索方法的有效性。

在这里，总结整篇论文工作的主要创新点以及贡献如下：

第一，我们对于面向医疗问答信息的跨语言检索问题进行了形式化的定义。通过结合对跨语言信息检索的理解，提出了自己的检索框架，并将问题缩小到针对问答形式的信息检索；

第二，结合基于词典翻译的跨语言信息检索方式提出了基于特征匹配的对数线性检索模型，该模型在参考最大熵翻译模型的基础之上，利用词典翻译的方式获取查询与文档之间对应的特征信息，利用求最大检索概率的方式进行检索，最后结合由粗到精的检索策略使得检索效率得到提高；

第三，利用最小错误率训练算法对检索模型的参数进行训练优化。通过分析最小错误率训练算法在对数线性模型参数优化过程中的原理和优势，合理的将其运用到检索模型的参数优化中，提供了快速有效的参数训练方法；

最后，通过对面向医疗的跨语言信息检索问题的研究，为构建跨语言的在线医疗平台提供一个理论的基础。

5.2 工作展望

在面向医疗的跨语言信息检索系统构建中，虽然有一些初步的成果，但是也存在一些问题，这些问题需要进一步的工作来逐渐完善它们，由于各种原因的限制，在这里只能就这些问题进行一些讨论，具体有一下几个方面：

1. 数据的更新与完善
2. 检索模型的完善
3. 参数训练的优化
4. 检索对象的扩充