南京大學

研 究 生 毕 业 论 文

（申请硕士专业学位）

论 文 题 目 高考地理问答系统中的句子理解研究

作 者 姓 名 汤莲瑞

学科、专业名称 计算机技术

研 究 方 向 自然语言处理

指 导 教 师 戴新宇 副教授

2017 年 **5** 月 **20** 日

学　　号：**MF1433042**

论文答辩日期：**2017** 年 **5** 月 **30** 日

指 导 教 师：　　（签字）

Research on Interactive Phrase-based Machine Translation

by

Lianrui Tang

Supervised by

Vice Professor Xinyu Dai

A DISSERTATION SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY OF NANJING UNIVERSITY IN CANDIDACY FOR THE DEGREE OF MASTER

May 20, 2017

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：高考地理问答系统中的句子理解研究

计算机技术 专业 2014 级硕士生姓名：汤莲瑞

指导教师（姓名、职称）：戴新宇 副教授

摘　要

人工智能技术正在飞速改变这个世界。在自然语言领域，围绕着自动问答系统 (QA) 也开展了越来越多的研究。高效、智能的问答系统，致力于为使用者直接提供更直接和更优质的答案，可以自动从大量的知识储备中进行检索、推理，从而将使用者从海量信息的搜索、筛选、抽取答案的过程中解放出来。2011 年，IBM 的 Watson 问答机器人参加问答类综艺节目“Jeopardy!”，并战胜了人类顶尖选手赢得冠军，自动问答系统再一次吸引了世人的眼光。

从某种程度上来说，高考作为中国大多数中学生最重要的考试，可以看做高水平的问答过程。本文的项目背景是面向中国高考地理试题的问答系统，并专注于对选择题的解答。在解决高考自动问答的过程中，我们面临多项挑战：首先高考题的问答形式与传统自动问答系统的问题存在明显区别；其次，高考题的灵活性远高于传统问答系统处理的问题，这意味着我们很难从现成的文本中直接匹配、抽取得到答案，所以在解题过程中，我们面临着与传统问答系统不同的挑战。

作为问答系统的第一步处理，问题理解的作用举足轻重。本文重点关注对于地理选择题的试题理解过程。我们将题面和一个选项拼接成的完整句子作为分析的对象，主要从两个方面来研究对于地理试题的理解问题：一方面是句子分割的浅层处理，另一方面是尝试使用 AMR 对试题文本进行深层语义分析。前者可以可以将一些相对较长的复杂句子拆分转化成两个或多个较短的简单句，后者则希望能够从句子中提取概念的语义关系，完整描述句子中涉及的语义关系和表达逻辑。

我们针对地理选择题的特点，提出了利用逗号对选择题的选项进行拆分，将较长的原句转换成语义等价的多个简单句，从而简化后续的处理步骤的输入，提高后续步骤的处理能力。在这项工作中，我们使用了最大熵分类器和基

于规则的启发式方法，通过两个步骤来实现句子拆分：首先识别选项中的逗号是否可以作为一个分割点，然后在识别句子的从句或并列结构的公共前缀边界。

AMR（Abstract Meaning Representation）是一种具有较为强大的表达能力的的新型语义表示方法，它可以将一句话的语义用单根的、有向的连通图表示出来，更强调句子的抽象语义，而非具象的语法表达方式。但是由于围绕AMR 的研究才刚刚起步，目前已有的 AMR 自动分析效果仍然还有很大待提升的空间。中文 AMR 的标注语料仍未达到一定规模，尚在进展中，所以关于 AMR 的中文应用研究几乎还是空白。本文在 AMR 方面工作主要是对现有AMR 分析算法进行一些实验分析，并首次验证 AMR 标注体系及自动解析算法在中文上的性能。针对地理试题，我们标注了一个小样本的 AMR 语料，并用现有算法来验证 AMR 在特定领域文本上的处理能力。

为了支撑上述两项问题理解的研究工作，我们还构建了一个地理试题标注工具，并通过这个工具建立一个高质量的地理试题语料库。除了可以标注句子分割和 AMR 这两种信息，该工具同时支持标注分词、词性、命名实体、地理术语、试题模板表示、成分句法等各项数据。

关键词： 问题理解；句子拆分；语义分析；AMR；地理文本；标注工具；

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on Interactive Phrase-based Machine Translation

SPECIALIZATION: Computer Technology

POSTGRADUATE: Lianrui Tang

MENTOR: Vice Professor Xinyu Dai

Abstract

Artificial intelligence is changing the world rapidly. Recent years, more and more research on automatic question-answering is carried out in the field of natural language process. A highly efficient and intelligent QA system aims to provide more direct answers to users with high quality, which can retrieve information from large-scale knowledge and make deductions automatically. Therefore, it free users from searching, filtering texts from the large quantity of information, as well as finally extracting the answers by themselves. In 2011, the QA robot Watson from IBM took part in a quiz show named Jeopardy! on TV, beat the top human players and became the champion. Once again, QA system attracted the attention of the world.

To some extent, the college entrance examination is the most importance examination for almost all the Chinese middle school student, which can be seemed as a high-level question answering situation. The background of this paper is a question answering system focused on the geography part of the Chinese college entrance examination. And we paid more attention to answering the choice questions of test paper. In the process of accomplishing the QA system for college entrance examination, we are faced with many challenges. Firstly, the question form is different from those for traditional QA systems. Secondly, the questions are much more flexible, which means we can hardly match the question to the original texts in the knowledge base directly. Therefore, the answer extraction is harder. We need to rely on automatic inference to generate answer from the those texts.

As the first stage of automatic question-answering, question comprehension plays a key role for the whole system. This paper focuses mainly on some issues on the understanding of geography choice questions. AMR(Abstract Meaning Representation)

is a new and powerful semantic representation for sentences. In this paper, we parse the Chinese sentences into AMR. As far as we know, this is the first work about Chinese AMR parsing. Then we tried to apply AMR on the understanding of geography questions. Besides, in order to research the performance of a variety of natural language processing tasks on geography question data, we build a tool for tagging various data on question texts, including question splitting, Chinese word segmentation, part-of-speech, named entities, geographical terms, template representation for questions, syntactic tree, and AMR. According to the feature of choice questions, we present a question simplification approach, which splitting a composed sentence into multiple simple sentences by commas. In this way, we can simplify the input for the following processing stages.

The research on AMR has just started. The performance of state-of-the-art approaches for parsing English sentences into AMR is still not satisfactory. Besides, the size of Chinese AMR corpus is relatively small. Some AMR corpus annotation work is still in progress. So there is almost no work about Chinese AMR. The work in this paper about AMR is very preliminary. And we just do some exploration to apply AMR to question understanding.

keywords: Question Comprehension, Semantic Parsing, AMR, Geographic Text, Annotation Tool, Sentence simplification

目　　录

插图清单

附表清单

第一章　绪论

1.1　研究背景

在人工智能技术日新月异的今天，人们对人工智能技术寄予了越来越多的期待。在各个领域，人们都在不断试图突破人工智能目前的极限。在自动问答领域，已经有很多商业化的系统为企业提供高效的解决方案，为用户提供更加快捷、准确的服务。在自动问答出现以前，我们获取知识的方式通常是在搜索引擎中搜索关键字，在得到的网页文本中一个个去搜寻是否包含了我们想要的答案。问答系统的出现是对搜索引擎功能的一次升级。问答系统希望不但能够从海量数据中找到与用户问题相关的文本，还能够从文本中直接准确地找出答案，免去使用者自己去从搜索结果中进一步寻找答案的过程。

通常我们所说问答系统可以针对一个自然语言的问句，在知识库中找到相关的支持文本，然后可能涉及到一些简单的推理，接着抽取出可能的答案，再对所有答案进行综合打分，并将最终的答案返回给用户。waston 系统是这类问答系统的一个变种，输入是一个陈述句，但是可能其中的一个命名实体或者时间被代词替代，waston 所做的事情就是首先识别出哪个代词是需要消解出来的，然后进行上述的问答系统的流程[1]。

从某种角度来说，考试的过程就是一种问答过程，而高考作为中国学生进入高等教育的关键考试，其试题更具有难度和代表性，高考是对考试者的知识积累、推理能力、判断能力的一种综合考察。为了探索问答系统的潜力，基于 863 项目《开放域知识集成、推理与检索关键技术及系统》，我们对地理试题的自动解答进行了研究。在高考地理试题中，选择题是一类重要的题型。不同于传统的问答系统，选择题不是一个有明确疑问词的疑问句，也不像 waston 那样去消解一个句子中未知的代词。选择题更像是对四个选项的陈述做出判断的判断题。并且我们很难直接从课本或者其他文本中直接得到相关文本，通过匹配来判断一个句子是否正确，而是需要根据上下文的时间、地点、假设等等，综合相关的知识点，经过复杂的推理和计算才能够得到正确答案，因此和传统的问答系统存在很大区别。

在试题自动解答的过程中，对问题的理解是一个关键步骤。这一步包括对

问题做各种基础的自然语言处理，得到一些基本的分析结果。对每一项基础分析任务，我们需要针对地理领域试题的特点，做出一些针对性的调整，提高通用工具对地理试题的处理能力。在本文中，主要从句子拆分简化和 AMR 语义表示解析两方面来研究试题理解问题。对于地理选择题的特点，我们提出了基于选项中的逗号对句子进行拆分简化的方法。另外，针对句子的理解，除了目前比较常见的语义角色标注等语义分析方法，我们还尝试使用了近些年新提出的 AMR 方法，这个表示体系对于句子语义有更强的表示能力，也是一种比较值得探索的新方向，因此本文也基于目前已有的研究，对 AMR 在中英文语义表示和自动分析方面做了一些实验分析，探索 AMR 目前可以达到的水准，并在一个小的地理领域试题数据上进行了实验，希望能够为后面的问题理解工作探索一个新思路。

1.2 中文句子分割的研究现状

逗号是一种十分常见的标点符号，在中文文本中，逗号出现的频率比英文等语言更高，据统计每个英文句子中平均有 0.869 1.04 个逗号，而每个中文句子中平均有 1.79 个逗号[2]。在中文的长句子分析中，逗号可以起到十分重要的作用。在中文中，逗号不仅仅可以作为一个句子内部从句或者短语之间的停顿符号[3]，也可以作为两个句法独立的句子之间的分割符[4][5]。所以对于包含逗号的中文长句来说，利用逗号来将长句分解成更短的句子，可以对很多自然语言处理任务有比较好的提升作用，例如机器翻译[6]、句法分析[2][?][7][3][8]等等。

Mei 等人[2] 在文章中指出，中文的逗号中，约有 30% 是用来将从句与主句或相邻的从句之间分隔开。文中指出逗号是一个中文句子的自然分割点，可以将逗号分割和句法分析结合起来，首先在合适的逗号位置将句子切分成几个短句，然后对每个短句分别做依存句法分析，再将短句对之间用一个依存关系连接起来，得到原长句的句法分析结果。不是所有的逗号都可以作为这样的分割点，有些逗号如果做为分割点，会导致一些词在短句内找不到 head 词，还有会导致一些词找到错误的 head 词。作者提出的方法认为，如果逗号分割的两个短句之间只存在一条依存关系边，则认为这个逗号是合理的分割点，通常这样的逗号出现在一个从句结束的地方。文章对每个逗号抽取了一些特征，使用 SVM 分类器对从句内逗号和从句间逗号进行分类，获得了 87.1% 的准确率，并且显

示可以使依存句法分析的性能提升 9.6%。

Xing Li 等人[3] 将标点符号看为分割标点和普通标点两种，前者可以将一个句子分割成几个子句。文章中主要使用了基于规则的方法来处理长句的句法分析。将句法分析分解成一个两步句法分析方法：首先用所有冒号、逗号、分好，将句子分成多个子句；然后第一步先对分割出的子句进行句法分析；再使用一种基于句法分析结果的规则的方法，判断出每个逗号是否分割一个并列结构，而不是多个从句，如果出现这种情况再使用规则的方法将这几个子句的句法树合并起来；最后在将每个子树的根节点的词性标签序列作为句法分析的输入，再做第二次 parsing，结合前面的子树结果就可以得到原句完整的句法树。实验结果证明这种方法可以有效缩短长句句法分析的时间，并且可以将句法分析的性能提升 7%。

毛奇等人[?] 为了处理句法分析中的长句问题，提出了单独解析块的概念，指由特定的标点符号分割句子生成的自然次序列。单独解析块又分为可单独解析块和不可单独解析块，区别在于，前者的内部词序列在正确的句法树中只有一个根节点。文章思想类似 Xing Li 等人[3] 的工作，但是由于使用规则的方法能够处理的情况比较局限，他们提出了一种基于统计的方法。因此将逗号分类的任务形式化为了对可单独解析块的识别问题。文中考虑了包括逗号在内的五种标点符号作为单独解析块的划分边界，提出一个特征集合，并使用了 Id3 决策树分类算法进行分类，对于可单独解析块的识别 F 值可以达到 85.1%，对不可单独解析块的识别 F 值为 69.7%。然后将单独解析块的识别加入句法分析的过程，首先对所有的可单独解析块进行句法分析，得到一个子树结构，然后在这些子树中抽取出其中的中心词与词性，再将这些中心词与其他不可单独解析块合并成一个词与词性序列，在对这个组合序列进行句法分析得到一颗全局句法树，最后再将之前得到的子树结构整合进最终的句法树中。实验结果表明，这种方法可以使句法分析效果在长度大于 40 的长句中，准确率提高 1.59%，召回率提高 0.93%，并且该方法有效缩短了句法分析花费的时间。

Jinhui Li 等人[8] 提出了另一种可以处理中英文句子的层次化句法分析的方法。文章从句法树森林中递归地识别出简单的组成成分，然后逐步减小句法树森林中句法树片段的个数，直至全部合并为一棵句法树。总体上大致分为三个步骤：词性标注、组块分析、句法分析。算法从由所有组块的句法树组成的森林开始，并且设计了一个 BIESO 标签体系，每一次句法树合并之前，都会使用最大熵模型从左到右地预测出每一棵子树的标签，然后对能够合并的连续子树

进行合并，得到新的更小的句法树森林。这篇文章虽然并未直接涉及到利用逗号来作为句子的分割点，但是在组块分析中，实际上也用到了逗号的信息。

李艳翠[9] 在研究中文篇章关系的工作中，从基本篇章单位的角度，对标点符号进行分类，分成篇章单位间标点和篇章单位内标点，其中篇章间标点中，逗号就占到了 61.3%。Yang Y.Q. 等人[10] 对逗号的作用主要分为了表明并列关系和从属关系的两类。其中表明并列关系的逗号包括三种：起到句子边界作用的逗号、分割父节点为非根节点的并列 IP 结构的逗号、分割并列动宾短语的逗号。另一种表明从属关系的分为三类：分割附属从句与主句的逗号、分割句子谓语与宾语的逗号、分割句子主语和谓语的逗号。这里更多地从语法功能的角度对逗号的作用进行了划分，而不是像上面的一些工作，完全从句法树的角度来对逗号进行分类。

总的来说，对中文长句中的逗号进行句子切分，并没有一个十分明确的标准，有些研究是从句法树的角度出发，认为逗号是否能够切分句子，取决于人工标注的句法树是否将逗号前后两个部分表示成独立的子树；有些则是先在所有逗号处进行切分，再根据并列结构等的句法特点，识别出不应作为切分点的逗号并修正；还有些是从逗号在篇章切分中的作用出发，结合句法特点来对逗号进行分类。如何利用逗号来得到更简短的句子，要考虑到切分结果对某种应用场景（比如句法分析）的作用，也就是明确切分的目的是为了什么，并结合所处理的语料特点，选择合适的分类标准，从而用逗号将长句切分成短句。

1.3 AMR 的研究现状

句法树库对自然语言处理领域的发展具有巨大的影响力，比如宾州树库就是一个典型的例子。但是在语义标注方面，目前已有的标注语料还比较分散，比如有单独的命名实体、指代消解、语义关系、篇章关系等等，目前还缺少一个能够将整个句子的语义逻辑关系组合在一起的语义标注树库。AMR 就是在这样的背景下被提出的，这是一种能够将句子语义表示成一个简单有向图的的表示方法，在这个有向图中，节点表示一个概念，通常是句子中的一个词语或者词组，或者在词语或词组的基础上抽象出来的概念，有向边表示节点之间的关系，边具有指示概念间语义关系的标签。这种表示体系的提出，以及基于这个体系的语料库的建立，可能会给自然语言理解的任务带来新的发展空间。

AMR 表示是一种由单个根节点的、有标注图的表示方法[11]，对人来说，

AMR 标注是易读的，同时对于程序来说，也很容易获取到该表示中的所有信息。AMR 的目标是能够从句子的不同的句法表达方式中，抽象出句子的语义，也就是说对于不同表达方式的同一个语义的句子，希望能够得到相同的 AMR 表示结果。在 AMR 表示中，用到了大量 PropBank 框架的内容，对具有多个表示框架的谓词，会在概念节点中注明对应的是该谓词的哪种用法，在标注它的论元时，也会在边上标记出相应的论文序号。

L Banarescu 等人[12] 在 2012 年提出了第一个版本的 AMR 标注规范，明确了 AMR 应该如何标注，并给出了大量的标注实例，说明了怎么选择根节点、节点的内容应该怎样确定、关系标签的几种类型、常见句式中如何添加新的抽象节点和标注关系标签、如何标注命名实体和数字时间及其它各类型实体等等。在这个标注规范的基础上，L Banarescu 等人[11] 在 2013 年公布了一个英文的 AMR 标注图库，包含大约 5000 句标注文本。

在中文方面，李斌等[13] 在《小王子》文本上标注了一个 AMR 图库，总共包含 1562 个句子，并且据悉，一个规模超过 5000 句的 AMR 图库正在标注过程中。中文的语法表达比英文更加随意，例如有时会出现一个谓词在句子中不是连续的字序列，而是被别的词语分割开来（例如“帮了很大的忙”中的谓词“帮忙”）。文中对一些中文 AMR 标注中特殊之处进行了详细说明，成为 AMR 在中文中的应用的一个开创性的工作。此外，在英文的 AMR 标注中，没有标注图中的概念节点与原句中的词语的对齐关系，通常都是借助自动对齐算法来进行对齐，因此可能会损失一定的精度。在这项中文 AMR 标注工作中，还加入了对齐信息的标注，包括概念节点的对齐以及少部分边的对齐信息。

有了 AMR 语料库之后，陆续出现了一些 AMR 自动解析的算法。最早的公开工作是来自 Flanigan 等人[14] 在 2014 年提出的一种两阶段的图算法，将 AMR 解析分成概念节点识别和概念节点间的边预测两个步骤，对于边的预测，采用了一种类似最大生成树算法的方式，得到所有概念节点间相互连接形成的连通图。[15] 对于 AMR 子图的生成提出了一种鲁棒性更强的方法。随后在 2015 年，Wang Chuan[16] 等人发现 AMR 的表示方法有一部分与句法分析的结果比较相似，受到基于转换的句法分析算法的启发，他们提出了一种基于转换的 AMR 解析算法，并设计了一套使用于 AMR 图生成的转换方法：在句子的依存句法分析结果的基础上，每次预测一种转换动作，一步步将句法分析的结果转换成 AMR 的表示。Pust 等人[17] 于同年提出了使用基于语法的机器翻译的方法来做 AMR 的解析，这篇研究将英文句子到 AMR 表示的转换看成是一种 string

到 tree 的翻译过程，并设计了一种 AMR 表示下的语言模型。Lucy Vanderwende 等人[18] 的工作则支持对英文、法文、德文、西班牙文、日文的 AMR 解析，他们设计了一些逻辑形式（Logical Form）到 AMR 的转换规则，通过这样的方法来实现对上述语言的解析，这些语言还没有可用的 AMR 语料库，所以这篇文章对于 AMR 在其他语言中的扩展也有重要的借鉴意义。

AMR 作为一种语义表示方法，被寄希望于提高多项 nlp 任务的性能，目前已发表的研究中，有将 AMR 用于提高事件检测任务的性能[19]，Xiang Li 等人利用现有的性能最好的 AMR 自动分析算法，对待检测事件的文件进行 AMR 解析，然后将 AMR 结果中的一些数据作为特征加入，实验结果证明这样的做法可以在原来的基础上提高 2.1% 的 F 值。还有研究将 AMR 用于无监督的实体链接任务[20]，结果表明使用了 AMR 信息的无监督实体链接的性能可以和有监督的实体链接性能相当。

目前已知的 AMR 英文语料规模达到了 4 万多句，中文的大规模语料库预计也将于不久之后公开，随着更多可用语料的出现，对 AMR 自动解析和将 AMR 应用于其他自然语言处理任务的研究会越来越多，因此 AMR 是一种具有潜力的语义表示方法，尽管目前在中英文上的 AMR 自动解析效果还不尽如人意，但是 AMR 的出现为语义分析和自然语言理解提供了一个全新的研究方向。

1.4 论文的主要工作

本文工作主要是关于地理试题文本的问题理解，具体是从下面三个主要方面开展工作：

其一，针对选择题选项的特点，我们提出对部分含有逗号的选项进行句子简化拆分。我们发现地理选择题中有 14% 的选项中包含一个或一个以上的逗号，在这些包含逗号的选项中，有 71.7% 的选项，可以通过在某处寻找一个边界，将句子分割成公共部分和非公共部分，然后组合成多个可以分别判断正误的句子。虽然这部分可拆分的选项在我们标注的所有数据中只占 10.1%，但是根据我们在后续试题语义模板化处理的过程中发现，对于这类句子的处理难度较大，较严重影响地影响了自动模板化的性能，因此对这部分的简化拆分是一个重要的步骤。

其二，探索 AMR 在中英文语义分析上的效果，及其在地理试题上的应用

效果。AMR 是一种新型的语义表示方式，在此之前，我们在高考问答系统中使用的主要语义分析方法，是将试题文本通过其句法结构和词性等特征，转换为一个我们制定的地理试题模板体系，每个模板根据其定义包括模板类型和语义槽，每个选择题和问答题的核心问题部分都可以转换成一个语义模板的表达。经调研发现，AMR 的表示体系中有很多类似的语义结构，如果针对地理文本对某些关键的实体和关系进行 AMR 的解析，在理想的效果下可以方便地转换为我们所需要的模板表达方式。所以我们探索了 AMR 当前的表达和自动分析能力，并尝试将其应用在地理试题上。

其三，为了支撑上述两项围绕问题理解的研究，我们需要构建一个地理试题的语料库，为此开发了地理试题标注系统。由于研究开始时缺乏足够的有标注地理文本，我们针对地理试题的结构特点，包括文本外部组织结构和文本内部的语法表达方式，设计并开发了一个标注系统，该系统支持特定格式的试卷导入，同时支持选择题和主观题的试卷格式，能够保留试卷原有的结构信息（比如高考题常有一些几道连续的题目共享的背景知识，主观题大题小题的嵌套，地理试题的主选项和小选项等等），支持对包括分词、词性、命名实体、术语、语义模板表示、成分句法分析、选择题主次文本、AMR 的标注。可以通过关键字、试卷名、某一项数据的标注状态、模板类型等多种方式进行试题检索，可以对所有标注内容导出成文本文件，提高了地理试题的标注效率和标注数据的检索和使用效率。

通过上述工作，我们建立了一个包含多项标注结果的地理试题语料，并为试题理解提供了更加简单的输入文本。另外实验结果表明，目前已有的 AMR 自动解析方法在中文上的性能还明显低于英文，。。。。。。

1.5 论文的组织

本文内容的组织如下：

第一章主要介绍本文研究内容的背景，以及论文主要工作内容，并简述了本文围绕地理试题理解的两个主要研究工作的相关进展，一方面是基于逗号的句子拆分的研究现状，另一方面是关于 AMR 语义表示体系的研究现状。

第二章主要介绍本文提出的基于逗号对选择题选项进行拆分简化的工作。首先会介绍基于逗号的句子分解的相关工作，然后阐述本文工作的内容。阐明做选项拆分的动机，详细说明如何在问题理解分析中对选择题选项文本进行分

割，主要分为两个步骤，第一步是判断句子是否可拆分，第二步是寻找到拆分的边界，类似补全句子成分的步骤。本章同时描述了我们选取的文本特征和算法，给出当前的实验结果，以及错误分析。

第三章主要介绍对 AMR 在中英文语义理解上的相关内容。首先介绍目前 AMR 的研究进展，包括标注体系及规范，语料建设情况，自动解析算法等内容。对于这个较新的任务，我们使用了目前性能比较领先的一种图算法，对多种中英文语料设计了多种实验，对标注数据和算法流程进行了阐述，并在少量地理题标注数据上做了实验，为后续 AMR 相关的工作提供参考。并总结 JAMR 对于中文 AMR 解析的问题，以及目前存在的一些缺陷。

第四章主要介绍地理试题标注系统的相关内容。主要包括系统架构、功能设计、数据库设计、交互方式等各个方面，以及目前通过该系统完成标注的地理语料规模，以及语料的数据统计结果等。

第五章主要是对本文工作的总结以及对未来工作的展望。

第二章　背景知识

2.1　引言

本章是本文主要工作展开的基础。本文对于高考地理试题中的问题理解主要从两方面展开：一方面是逗号在句子分割简化中的作用，由此将长句拆解成多个等价的短句，降低问题理解的难度；另一方面是考虑从新型的语义表示方法 AMR 出发，对问题的深层语义进行解析。所以本文的背景知识由两方面组成，一方面是关于逗号在中文长句中的作用及相关研究进展，另一方面是 AMR 的知识体系及研究进展。

对于逗号相关的中文长句处理，我们主要介绍中文中逗号的功能分类体系；对于 AMR，我们将从 AMR 表示体系的定义、标注体系、语料建设、自动对齐、自动解析算法、自动评价等方面详细介绍。通过对以上内容的分析介绍，可以对本文的工作有更清晰的认识，也为我们的工作打下了基础。

2.2　本章小结

本章主要介绍了基于短语的统计机器翻译系统的各部分重点内容。首先介绍了短语机器翻译的整体框架与流程。在此基础上，分别介绍了短语机器翻译的建模方法，参数训练方法，解码方法等内容。

短语翻译系统通过最小错误率训练进行对数线性模型的参数调节，在开发集上，直接对翻译评价指标进行参数调节，每次调节一维参数，经过多轮迭代获得最优参数。统计机器翻译最常用的自动评价指标为 BLEU。

短语翻译系统采用了对数线性模型进行建模，可以自然地使用各种特征。短语翻译系统主要使用了翻译概率、词汇化调序概率、语言模型概率、词计数、短语计数等特征。为了减少搜索空间，降低搜索复杂度，短语翻译系统采用了一种启发式搜索算法：柱搜索解码算法。柱搜索算法通过源端词数维护搜索栈，即翻译了相同源端词数的假设置于同一个栈中。柱搜索算法的核心是假设扩展和假设剪枝。翻译系统通过将翻译选项的目标端连接到局部假设之后，并同时更新假设的各项特征值来进行假设扩展。翻译系统通过局部假设的特征

值得分和未来代价估计相加进行局部假设的得分估计，并使用该得分估计进行剪枝，提高了局部假设之间的可比较性，从而降低了搜索错误。在假设扩展过程中，为了进一步减少搜索空间，翻译系统通过假设重组，丢弃无效假设。

本文中提出的交互式机器翻译相关内容均建立在短语翻译系统的基础上，特别是短语翻译系统的短语表、解码方法、模型参数训练等内容上。对短语机器翻译系统中的关键部分的理解有助于对本文的理解。

第三章　地理选择题选项拆分

3.1　引言

中文长句通常是一个复句，《现代汉语》对复句的定义为：复句是由两个或几个意义上紧密相关、结构上互不包含的分句构成的句子[21]。这样的几个分句之间往往是由逗号分隔开的，因此根据上述定义，我们可以利用逗号在复句中的作用，将句子拆解成一些更简短的句子，从而降低句法分析、机器翻译、语义理解等等任务的难度。但是逗号除了可以分隔从句或是分句，在中文中也可以用来分隔并列的词语等等。

在地理试题中，尤其是选择题中，我们观察到一个现象：选择题的一个选项中（不考虑题面）经常包含一个或一个以上的逗号，而这个逗号常常隔开了两件或以上可以分别判断正误的部分，但也有一些逗号隔开的是具有因果、递进等关系的子句篇章。如果我们可以判断出哪些逗号隔开的是相对独立的两件事情，哪些逗号不能用来作为分割点将长句中两个句法上并列的部分隔开，则可以将较长的一个陈述句拆解成多个更短的、更简单的句子，能够大大减小例如句法分析及语义理解的分析难度。

所以为了更好地处理选择题中选项包含了逗号的长句，我们提出了一种对选项进行分类的方法，即根据逗号分开的两个句子部分的特点，判断该逗号隔开的是否是句法上可以视作并列的两个部分，进而将一些长句转换为几个短句。我们使用了最大熵模型进行分类，本章将介绍整个实验的方法及结果。

3.2　相关工作

在前期调研中，我们发现还没有一个已发表的研究十分类似于本节所提到的工作，一方面是因为逗号的处理在英文中的重要性要低于中文，因此对于逗号在句子分析和理解中的重视程度并不是很高，另一方面是因为本文所提出的方法是针对高考地理试题中的选择题这种特殊文本类型所提出的，并不十分适用于通用的中文文本。所以我们需要为我们研究的问题寻找一些理论基础，所以在本节中，我们根据 Yang Y.Q. 等人[10] 的研究工作，介绍他们在研究中文篇

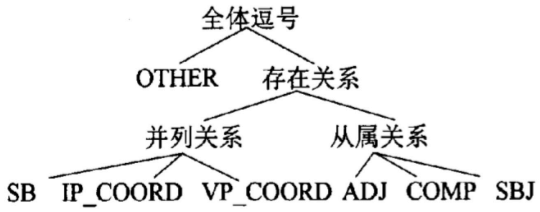

图 3-1: Yang 的逗号分类体系

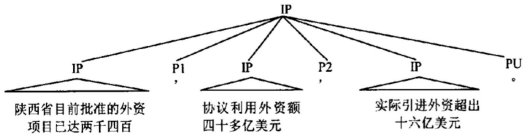

图 3-2: SB 类型的逗号

章关系时整理出的逗号在句子中的功能，这个分类与我们对逗号功能的判断和应用比较吻合。

上述工作将逗号的使用方法划分为 7 类，首先把逗号的使用方法在总体上分为两大类：一类是逗号连接的两个子句之间存在关系，即逗号是子句边界；另一类是两子句之间不存在关系，不能视作是子句或者是篇章的边界。两个子句之间存在的关系又可以分为并列关系和从属关系：并列关系可以分为三种类型（SB、IP_COORD、VP_COORD），从属关系也分为三种类型（ADJ、COMP、SBJ），图 3-1展示了这个分类体系。

下面我们详细介绍一下每种类型的逗号的具体作用。

1）SB（Sentence Boundary）：在某些语境下，可以起到分割句子边界作用的逗号。这类逗号要求逗号左右的子句都是 IP 结构，父节点为根节点，比如在流水句中，如图 3-2所示，该例中的两个逗号都是 SB 类型。

2）IP_COORD（IP Coordination）：分割父节点为非根节点的并列 IP 结构的逗号。如图 3-3所示，其中的 P4 和 P5 属于这个类型。

3）VP_COORD（VP Coordination）：分割并列动宾短语的逗号，这些动

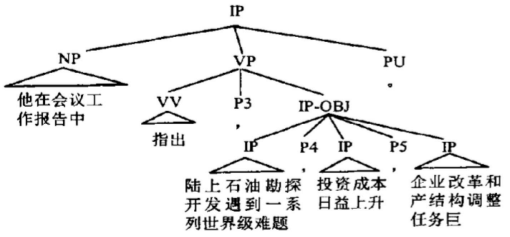

图 3-3: IP_COORD 类型的逗号

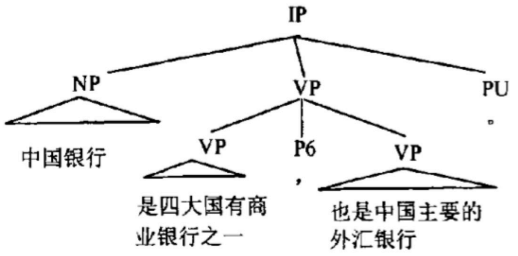

图 3-4: VP_COORD 类型的逗号

宾短语共享同一个主语。如图3-4所示，其中的 P6 属于这个类型。

4)ADJ（Adjunction）：分割附属从句与主句的逗号。附属从句指在句子中担当某种句子成分的主属结构，虽然从句部分的句子结构是完整的，但它不能脱离主句部分独立完成地表达语义。附属从句往往是状语从句，通常有条件状语、原因状语、目的状语、方式状语、伴随状语等等。例如“为了在运行机制上与保护区相配套，宁波保护区率先在中国实施了企业依法注册直接登记制的试行一站式管理。”中的逗号前是一个目的状语。

5）COMP（Complementation）分割句子谓语与宾语的逗号。通常对于宾语部分较长的复杂句子，会在谓语之后出现逗号，表示停顿，用于舒缓语气，通常在“表示”、“指出”、“认为”、“介绍”等提示性动词之后都会出现逗号。例如“钱其琛表示，我们对香港的前景始终是充满信心的。”中的逗号。

6）SBJ（Subjective）：分割句子主语和谓语的逗号。在句法结构上表示为逗号的左兄弟节点为 IP-SBJ 或者 NP-SBJ 结构，而右兄弟节点为 VP 结构。例如“出口快速增长，成为推动经济增长的重要力量。”中的逗号。

7）OTHER：其他用法，在分类体系图中则对应逗号前后两个部分不存在关系的情况。

在3.3中，我们会详细介绍地理试题文本中的逗号使用情况，并描述地理试题的分类的标准与上述的逗号分类体系之间的关联。

3.3 选项拆分方法

3.3.1 地理选择题的特点

选择题是高考地理试题中的一种重要的题型，占了大约一半的比重，所以在工作中，我们重点考虑了对选择题的理解的方法。首先我们介绍一下地理试题的基本特点，以及我们怎样将选择题处理成完整的句子。

一道选择题通常由四个选项组成，每个选项可能是一句文字描述，也可能是一些小选项的组合（例如“①②③”，小选项则属于题面的一部分）。由于我们重点是要理解选择题文本所表述的内容是否正确，所以通常是以“题面 + 选项文本”为一个单位作为理解对象。对于选项为小选项组合的情况，我们将小选项的内容找出来，和题面的其他部分拼接起来，得到一个完整的句子。

举个例子，一个不包含小选项的选择题如下:

（题面）该船即将进入

A 巴拿马运河

B 麦哲伦海峡

C 德雷克海峡

D 直布罗陀海峡

可以将题面与每一个选项进行拼接，得到四个完整的句子，然后判断这些句子的正误：

该船即将进入巴拿马运河

该船即将进入麦哲伦海峡

该船即将进入德雷克海峡

该船即将进入直布罗陀海峡

一个包含小选项的选择题如下：

（题面）岛内最大零售商业点位于甲村，主要形成原因是该村

① 地形平坦，交通便利

② 商业从业人口多

③ 商业组织形式复杂

④ 人口数量大

A ①②
B ①④
C ②③
D ③④

对于这类选择题，则不是将 ABCD 这些选项与原始句子直接拼接，而是将小选项内的文本与题面进行拼接，得到：
岛内最大零售商业点位于甲村，主要形成原因是该村地形平坦，交通便利
岛内最大零售商业点位于甲村，主要形成原因是该村商业从业人口多
岛内最大零售商业点位于甲村，主要形成原因是该村商业组织形式复杂
岛内最大零售商业点位于甲村，主要形成原因是该村人口数量大

我们提出的对地理选择题选项进行拆分的方法，就是将这样拼接出来的完整的一句话作为分类对象，在本文中我们称之为“试题文本”。拆分方法的提出主要基于选项中一个常见的现象，即选项经常是包含了由逗号隔开的几个短语或者动宾结构等等组成的部分，我们可以将这些部分拆分开来，分别与逗号两侧的 IP 或者 VP 等结构之前的公共部分拼接起来，使一个选项或者一个小选项对应的句子，拆分为多个更加简短的句子，这样得到的句子我们称为“拆分后试题文本”。举个直观的例子如下：

（题面）该海域沿岸
A 存在上升流，为热带雨林气候
B 有暖流经过，为热带草原气候
C 有寒流经过，为热带沙漠气候
D 盛行东南风，为热带季风气候

这题原本可以组成四个句子：
该海域沿岸存在上升流，为热带雨林气候
该海域沿岸有暖流经过，为热带草原气候
该海域沿岸有寒流经过，为热带沙漠气候
该海域沿岸盛行东南风，为热带季风气候

但是可以直观感觉出，每一句话实际上都可以分解成两句来单独判断正

误：

该海域沿岸存在上升流 & 该海域沿岸为热带雨林气候

该海域沿岸有暖流经过 & 该海域沿岸为热带草原气候

该海域沿岸有寒流经过 & 该海域沿岸为热带沙漠气候

该海域沿岸盛行东南风 & 该海域沿岸为热带季风气候

这样得到的每句话都更加简短，可以简化后续的句法分析、语义理解工作。在地理试题的理解中，我们的问答系统的项目为地理题制定了一套语义模板，需要在进行推理之前将自然语言的试题转换为模板表示。例如对于一个试题文本“该地区经济高速增长的根本原因是市场机制比较成熟”，对应的模板表示就是“原因 (该地区经济高速增长, 市场机制比较成熟)”。由于缺乏足够的标注数据，难以使用基于统计的机器学习方法来进行转换，为了提高转化的准确率和可信度，目前仍然是以基于规则的方法为主，这个规则考虑了词语、词性的特征，也考虑了句法分析结果中的特征。根据实验结果发现，这样的包含逗号的、含有句法并列成分的句子，是模板转换中错误率较高的一类。所以想办法将长的复杂句拆解成短的简单句，一方面可以提高句法分析的性能，另一方面也可以提高基于句法结果的模板转换的准确率。

根据上述的动机，我们对切分的问题描述如下，以选项中只要一个逗号的情况为例：

1、如果我们可以将试题文本分成三个部分：公共部分＋并列成分 1+ 并列成分 2，其中并列成分 1 和并列成分 2 之间的边界就是选项中的逗号，公共部分和并列成分的边界可能位于试题文本的选项逗号之前的任意位置，我们把这个边界叫做“公共部分右边界”。如图 ********

2、然后根据公共部分右边界的位置重新组合得到两个简短的句子：

-公共部分＋并列成分 1

-公共部分＋并列成分 2

并且这两个句子各自句法完整、语义通顺、两句合起来可以完整表达原试题文本的语义，则认为这样的试题文本是可拆分的，拆分方式即是如上。

但显然，不是所有包含逗号的选项都可以像上面这个例子这样，根据逗号分成几个部分然后分别和题面拼接成一个完整的句子。我们的目标是将所有的试题文本分类成“可拆分”和“不可拆分”两类。

例如“符合图中该城区实际情况的表述是 @ 北部地区的地租梯度，总体大于南部地区”（“@”标记题面和选项的边界），这个选项整体上是题面中

“是”的宾语从句，而逗号前面的部分是这个从句中的主语，这个逗号是 SBJ 类型。我们不能拆分得到“符合图中该城区实际情况的表述是北部地区的地租梯度”和“符合图中该城区实际情况的表述总体大于南部地区”这两个句子，这样得到的句子不再通顺。再比如“当火炬传递到 @④地时，当地正值多雨季节”，逗号前面是时间状语，逗号类型为 ADJ。

对于公共部分右边界的情况，根据定义也可以分成三种：与题面和选项的边界相同；位于题面内部；位于选项中逗号前的部分。但是位于题面内部这种情况在我们标注的数据中几乎没有出现过。所以这里给出一些试题文本的例子来说明另两种情况（“@”表示题面和选项的边界，“/”表示公共部分右边界）：

1. 位于题面选项边界：根据左图中等温线分布特点可知，例如“该海区 @/在北半球，A 处有暖流经过”。
2. 位于选项中逗号前：由 08 时到 20 时，例如“图中 @□ 地/风向偏北，风力逐渐减弱”。

此外，在问题描述中，我们提及拆分后得到的句子应该能够表达原试题文本的语义，但是有一种情况例外，即试题文本的选项中，逗号隔开的两个部分之间，存在潜在的因果关系，即没有明确的关系连词，但从逻辑上来说存在一定因果相关性。例如“2000 年到 2012 年，崇明县城镇化水平不断提高的主要原因是 @/第一产业效率提高，农村出现剩余劳力”这句话中，实际上“第一产业效率提高”是“农村出现剩余劳力”的一个原因，这里有篇章上的隐式因果关系[9]。这里如果拆开成两个句子，则两者之间的隐式因果关系就不能体现出来。但我们仍然认为这样的试题文本是可拆分的，原因有二：一是前文中提到，做选择题拆分的目的是为了简化后续句法分析、语义理解的难度，尤其为试题模板换提供更加简短的输入，而在模板化中，原本就不会体现这种隐式的因果关系，而是将前后两个部分作为两个独立的事实来考虑；二是从自然语言处理的角度来说，这样的因果关系几乎不能从语法层面获知，只有基于一定的背景知识和具备一定的推理能力的前提下，才能发现这层隐式因果关系，所以识别难度很大，因此不做识别。

基于以上描述，我们结合 Yang Y.Q. 等人[10] 对中文中逗号的功能分类，给出判断逗号所在试题文本是否可拆分的标准：

1. SB：地理题选项中几乎没有此类逗号，不考虑。
2. IP_COORD：所在选项可拆分，逗号之前的 IP 的左边界为公共部分右边

界，如“图中 @①地/风向偏北，风力逐渐减弱”。

3. VP_COORD：所在选项可拆分，逗号之前的 VP 的左边界为公共部分右边界，如“为防止艾比湖继续萎缩，在该湖流域应采取的措施是 @/修建水库，调节径流”。
4. ADJ/COMP/SUB：所在选项不可拆分，如一个 SBJ 类型的逗号的例子“符合图中该城区实际情况的表述是 @ 北部地区的地租梯度，总体大于南部地区”。

我们在实际的地理试题数据中，还发现了一些例外情况需要特别说明：

1. 在 Yang Y.Q. 等人的分类中没有提及这样的逗号，即逗号前面是一个 VP，后面是一个 IP，可能因为省略了某些成分导致前面的 VP 不能形成 IP，例如“英国 @ 多数河流/短，含沙少”，这里“短”只能形成一个 VP，而“含沙少”是一个 IP，这个 VP、逗号、IP 三者的父节点都是根节点，实际上“短”应该是“长度短”。对于这种情况，我们也认为是可拆分的。
2. 有时候选项中表述的两件事情是题面中明确提及的两个问题，题面中常包含关键字“分别是”“依次是”等等，这样的试题文本我们认为是不可拆分的，例如“与甲河段比较，乙河段的特点与主要成因分别是 @ 流量大，雨水补给量大”。

需要特别指出的是，ADJ/COMP/SUB 等非并列关系的逗号，也可能出现 IP_COORD 的句法形式，例如“玻利维亚 @/受寒流影响，多雾少雨”这句话的句法结构如图3-5所示，

所以，如何判断逗号的功能类型，不能仅仅从句法的角度判断，还是需要从语义本身出发，结合句法的特点来得到结论。综上我们给出了如何判断试题文本是否可拆分的直觉判断标准、从逗号功能角度出发的判断标准，基于上述论述，我们对地理试题库中的 859 道题进行了标注，在下一小节中将详细介绍标注数据的情况。

3.3.2 试题文本拆分数据

为了给试题文本拆分提供标注数据，我们对5中介绍的标注系统中的 55 套试卷、共 859 道选择题进行了拆分的标注，包括是否可拆分的二分类标注，以及公共部分右边界的标注。除了拆分标注本身，我们在训练中还用到了词性、分词、时间的实体识别等标注信息。包括拆分标注在内，这些数据的标注将

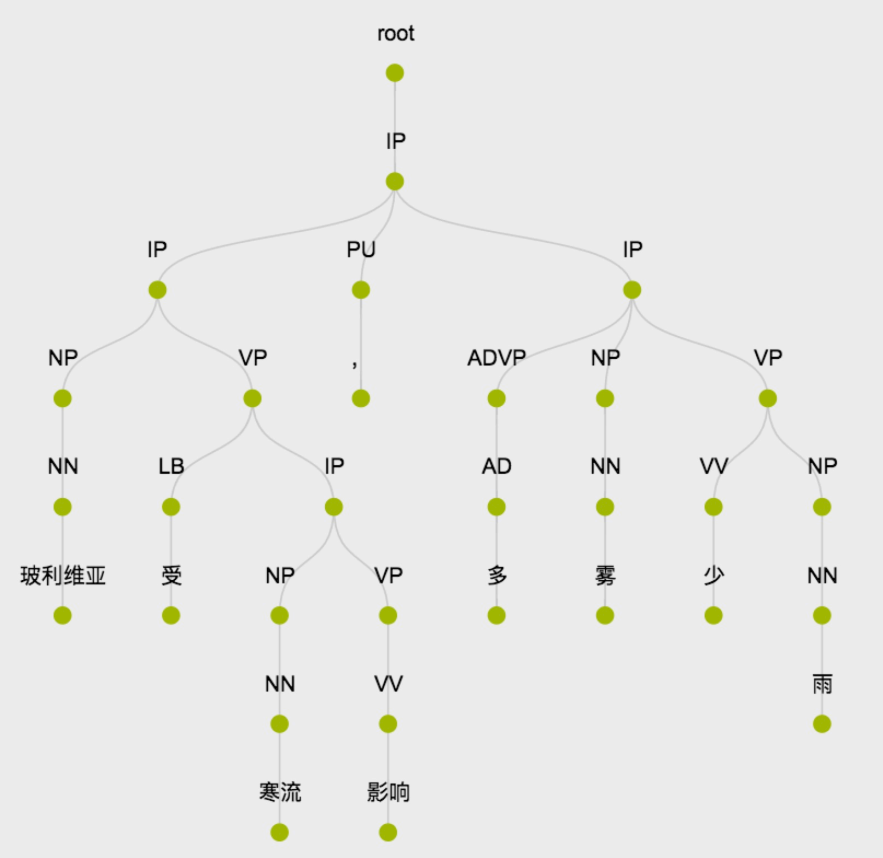

图 3-5: COMP 类型的逗号表现出 IP_COORD 类型逗号的句法特点

在5中详细介绍。这里提到的 55 套试卷是标注了拆分、分词、词性、命名实体识别的试卷。

标注数据的统计信息如图3-1：

可以看出，约有 14.1%（545/3853）的试题文本是含有逗号的，其中 71.7%（391/545）都是可以拆分的，在所有试题文本中占到 10.1%，所以从数据上来看，对地理试题做拆分，如果能够较好地利用拆分结果、尽可能多地找出可拆分的试题文本并拆分成短句，对句法分析和语义模板转换的性能将会有明显的影响。

但是我们的标注数据也存在两个问题：其一，尽管我们标注了 859 道题共 3853 个试题文本，但其中的 85.9% 都是选项不包含逗号的，也就是说跟拆分无关，所以在训练分类算法的时候，我们可以使用的训练数据十分有限，仅有 545 句；其二，在这 545 句中，存在严重的数据倾斜问题，可拆分的数据占 71.7% 而不可拆分的数据仅占到 28.3%，所以直接在这样的数据比例上进行训练，会导致不可拆分的数据的分类召回率较低，数据会倾向于被分类为可拆分

表 3-1: 拆分标注数据的统计信息

数据类型	数量
试卷	55
所有选择题	859
所有试题文本（仅拆分前）	3853
选项含逗号的试题文本	545
选项不含逗号的试题文本	3308
选项仅包含一个逗号的试题文本	509
选项包含一个以上逗号的试题文本	36
可拆分的选项含逗号试题文本	391
不可拆分的选项含逗号试题文本	154
公共部分右边界即题面选项边界的可拆分试题文本	303
公共部分右边界在选项中的可拆分试题文本	88

的，但是错误地将不可拆分的句子预测为可拆分，会导致后面的句法分析和模板转换的输入没有意义，因此我们需要尽可能避免这种情况的发生。第一个问题是因为标注工作量较大（包括整理试题上传、标注分词和词性等），虽然可以使用自动的分词、词性、命名实体识别标注结果，但我们为了排除自动分析错误的影响，只使用了有人工标注词法数据的试题文本来做训练；第二个问题是由地理试题本身的语料特点决定的。

3.3.3 选择题自动拆分

我们将自动拆分分成两步来进行：首先判断含有逗号的选项是否可拆分；如果可拆分，再判断公共部分右边界的位置在哪里。由于标注数据中只有 88 个公共部分右边界不在题面选项边界的试题文本，所以使用统计机器学习的方法来预测这个边界可能不太可行，但是这部分数据的比例仅有 22.5%。目前我们主要是判断是否可拆分，对于公共部分右边界的判断不是我们的工作重点，但会简单地介绍一些我们观察到的数据特点，并提出了一些初步的基于规则方法来解决这个问题。此外，语料中只有 36 个包含了一个以上逗号的选项，仅占所有含逗号选项的 6.6%，所以在实验中不考虑这类选项的分类，主要是针对仅包含一个逗号、两个选项部分的选项做分类，后面我们会使用一个简单的投票方

法来解决项目实际使用中遇到的多逗号选项的分类问题。

下面我们会分几个小节分别介绍第一步分类使用的最大熵模型、我们使用的上下文特征及实验方法、后处理方法、第二步基于规则的公共部分右边界识别、对于多逗号选项的分类方法。

3.3.3.1 最大熵模型

我们使用了最大熵模型（Maximum Entropy Model）[22] 作为分类器模型。最大熵模型是一种概率分布估计模型，可以方便地考虑待分类文本的上下文特征，曾被广泛应用于多项自然语言处理任务，包括词性标注[23]、语言模型[24]、文本分类[25][26]、组块分析[27] 等等。最大熵模型基于这样一个假设：如果没有任何其他的外部知识，对某个事件的分布概率一无所知的情况下，我们会倾向于选择一个模型使数据的分布尽可能均匀。直观地说，最大熵模型就是拟合所有已知事实，保持对未知事件的未知状态。换而言之，就是给定一些事实集，选择一种模型与现有事实一致，对于未知事件尽可能使其分布均匀。

从某种角度来说，我们对选项文本是否可拆分的分类，也是一种文本分类，但不是对于文本内容或者主题的分类，而是要考虑到文本的词、词性和句法等特征，来判断文本的结构类型。

最大熵模型是用来进行概率估计的，假设 a 是某个事件，b 是事件 a 发生的环境（也可以称为上下文），我们想知道 a 和 b 的联合概率是多少，记作 P(a, b)。在我们的拆分分类任务重，a 就是“可拆分”或者“不可拆分”，b 就是该试题文本中的一些特征。我们将所有 a 可能的取值组成的集合记作 A，所有 b 可能的取值组成的集合记作 B，概率估计问题就转换成，对于任意的 $a \in A$，$b \in B$，概率 P(a, b) 是多少？

以文本分类任务为例，假设我们有一个训练集，定义 $A = \{a_1, a_2, ..., a_m\}$，$B = \{b_1, b_2, ..., b_n\}$ 是文本的特征集合，num(a_i, b_j) 为训练集中二元组 (a_i, b_j) 出现的次数，根据训练集中出现的数据，可以这样估计 $P(a_i, b_j)$：

$$\tilde{p}(a_i, b_j) = \frac{num(a_i, b_j)}{\sum_{i=1}^{m} \sum_{j=1}^{n} num(a_i, b_j)} \tag{3-1}$$

但这种做法面临着严重的数据稀疏问题，对于很多 (a_i, b_j) 组合，在训练集中可能从未出现过，这种做法会武断地认为他们的概率为 0，是很不可取的。而最大熵模型可以使未知事件的概率分布总是尽可能均匀，即倾向于得到最大

熵。根据 Shannon 对熵的定义，熵的计算公式为：

$$H(p) = -\sum_{x} p(x) log_2 p(x) \tag{3-2}$$

所以求解满足最大熵原则的概率分布的公式如下：

$$p^* = \arg\max_{p \square P} H(p) \tag{3-3}$$

如果没有其他任何先验知识，因为 $\sum_{a \in A} p(a|b) = 1$，根据熵的性质，公式3-2得到最大值的条件是：

$$p(a|b) = \frac{1}{|A|} \tag{3-4}$$

我们可以定义很多这样的特征函数，它们之间可以互不相关，即特征函数可以灵活地将许多分散、零碎的知识组合起来完成同一个任务。

公式**??**中的 f_i 即是上述特征函数，λ_i 指示了特征 f_i 对于模型的重要程度。这个公式使模型由求概率值转化为求参数值 λ_i，一般的估计方式是 Darroch 和 Ratcliff 的通用迭代算法（Generalized Iterative Scaling, GIS）[28]，用来得到具有最大熵分布的所有参数值 λ_i。

3.3.3.2 特征的选择

在对包含逗号的地理选择题选项的特点进行观察分析后，我们选择了如下 18 个特征作为最大熵模型的特征模板，并以“据/P 图/NN 推测/VV ，/PU 2003-2013 年/NT 该/DT 市/NN@ 退耕还林/VV ，/PU 林地/NN 面积/NN 持续/VV 增加/VV”（以空格隔开每个词及其词性，‘@’表示题面和选项的边界）为例，说明每个特征在该例上的特征值：

1. wordNumDiff：选项中逗号前后两部分词的个数的差值（例子中为 3）
2. charNumDiff：选项中逗号前后两部分字的个数的差值（例子中为 4）
3. postagEditDistance：选项中逗号前后两部分的词性序列的编辑距离（例子中为 3）
4. lastPosComb：项中逗号前后两部分最后一个词的词性组合（例子中为“VV/VV”）
5. lastPosEqual：项中逗号前后两部分最后一个词性是否相等（例子中为

True）

6. firstPosComb：项中逗号前后两部分第一个词的词性组合（例子中为“VV/NN”）
7. firstPosEqual：项中逗号前后两部分第一个词的词性是否相等（例子中为False）
8. lastWordInTimian：题面的最后一个词（例子中为“市”）
9. lastTwoWordsInTimian：题面最后两个词的组合（例子中为“该/市”）
10. lastPostagInTimian：题面中最后一个词的词性（例子中为“NN”）
11. timeCombination：项中逗号前后两部分是否包含时间词的布尔值组合（例子中为“False/False”）
12. firstWordInSecondPart：项中逗号后部分的第一个词（例子中为“林地”）
13. firstPostagInSecondPart：项中逗号后部分的第一个词的词性（例子中为“NN”）
14. lastWordInFirstPart：逗号前部分的最后一个词（例子中为“退耕还林”）
15. lastCharInFirstPart：逗号前部分的最后一个词的词性（例子中为“VV”）
16. containCuewordsComb：项中逗号前后两部分是包含的线索词对应模板类型的组合（例子中为“None/None”）
17. containCuewordsMain：选项中是包含的主要线索词（例子中为“None”）
18. bothContainLonLat：项中逗号前后两部分是否包含经纬度的布尔值组合（例子中为“False/False”）

其中，词性数据采用的是宾州树库的词性体系。此外，在特征中提到了“线索词”这个概念，线索词是指在试题语义模板转化过程中，使用的是基于规则的转换方法，决定该试题文本对应哪个语义模板，是根据一个线索词列表决定的，比如出现“利于”“导致”“使”等等词语，就会触发“影响”这个模板。这个线索词列表我们是从地理题标注系统5中标注的试题模板及线索词数据中抽取出来的。但有一些线索词对于是否可拆分的判断影响不大，甚至可能产生一些干扰，比如“时间限定”“运动”“构成”“指示”等模板的线索词。所以我们只关注一些比较重要的、能够区分是否可拆分的模板类型的线索词，包括“影响”“因素关联”“原因”这三类，包括如下这些词（或者词的组合）：

1. 因素关联：表示、因、导致、需要、与/有关、因此、不利于、促进/了、通过、若/可、便于、造成、应、因为、若/则、由于、越/越、降低

2. 影响：控制、导致、拓展/了、受/影响、提高/了、提升/了、因素、借/优势、易/造成、宣传/了、利于、适宜、受/控制、受/制约、有利于、影响、便利/了、使
3. 原因：得益于、产生/原因、原因/是、缘于、目的、主要、形成/原因、的/原因/是、成因、主要/是、因为、由于、的/目的/是、原因

3.3.3.3 分类结果的后处理纠正

由于我们的训练数据比较有限，尽管有些我们考虑了某些文本特征并将其加入特征模板，有时候因为数据量有限，模型还是难以学出正确的分类。在我们观察到的结果中，大多数明显的可以用规则来解决的错误都是将不可拆分的误判成了可拆分的，有一部分原因也是因为数据倾斜现象的存在。所以我们针对一些明显的、具有规律性的错误分类样本，提出了一系列后处理纠错的规则，目前的规则均是用于将被预测为可拆分类的纠正为不可拆分类。包含下列9种情况：

1. 选项中都逗号后的部分的第一个词是某个特殊词，包括：利于，甚至，则，因此，便于，表示，但是，但，使，导致。这些词大多是线索词，或者是一些表示明确篇章关系的词语。例如：“符合图中该城区实际情况的表述是 @ 东南方向地租等值线密集，表示该方向空气质量较好”。
2. 选项中逗号前的部分的最后一个词是某个特殊词，包括：时。例如“下列对沿途地理现象的描述可信的是 @ 经红海时，可见沿岸大片森林”。
3. 选项被逗号隔开的两部分均包含表示经纬度的符号。例如“此时他可能位于 @24°N，120°E”。
4. 题面中包含某些特殊词，包括：分别，及。例如“图中河流的流向及河流与水渠的关系是 @ 河流自南向北流，水渠水汇入河流”
5. 选项被逗号隔开的两部分分别包含一组特殊的词中的一个，包括：越/越，若/则 (其中第一个词出现在第一部分中，第二个词出现在第二部分中）。例如：“图中 @ 大气中灰尘数量和颗粒越大，①越多”。
6. 选项中逗号前的部分包含某个特殊词，包括：因为，由于，因，借助。例如：“我国目前 @ 因消费量少于生产量，原油可以大量出口”。
7. 选项中逗号后的部分包含某个特殊词，包括：使、导致。例如：“大气中 @ 臭氧层遭到破坏，会导致①增加”。
8. 选项中逗号前的部分包括某组特殊词，包括：受/影响，受/控制。例

如：“玻利维亚 @ 受寒流影响，多雾少雨”。

9. 选项中逗号前的部分仅包含时间或者时间段，例如“仅考虑地球运动，图示窗户、屋檐的搭建对室内光热的影响有 @ 春分到夏至，正午屋檐的遮阳作用逐渐增强”。

对上述的后处理能够处理的场景，我们在特征模板中也尝试了将相关特征加入，例如逗号后部分的最后一个词、逗号前的部分的最后一个词、是否包含经纬度、线索词相关特征等等。但是可能由于我们的数据量比较有限，数据稀疏性比较大，所以学习出来的效果不太好。但后处理整理出来的这些规则，在地理高考试题的领域文本中均比较具有特色，是比较常见的现象。

但也存在后处理过程将正确分类为可拆分的样本错误纠正为不可拆分的样本。主要原因来自于标注数据基于的原则（能保持完整语法、不改变语义）

3.3.3.4 公共部分右边界的识别

标注数据中只有 88 个公共部分右边界与题面选项边界（下面简称“两个边界”）不一致的试题文本，占所有含逗号选项总数的 22.5%，没有足够的数据可以统计机器学习的方法来预测这个边界，因此我们通过对数据进行观察，总结了一些公共部分右边界位置的特点，提出了几点简单的基于规则的判定方法。

我们首先对公共部分右边界不在题面选项边界的 88 个试题文本进行了分析，发现在两个边界之间的词组类型分成下面积累：

1. 地点：有 52 个样本，占 59.1%。我们将方向位置（“东部”、“地点”）、指代地点（“甲地”、“乙河”、“两地”）、描述性地点（“板块交界处”）、术语性地点（“平原地区”、“中游”、“河口处”、“城市中心”）均视为地点。例如：“巴西 @ 亚马孙河/径流量大，流经经济发达地区”。
2. 名词性术语：有 24 个样本，占 27.3%。例如：“东北平原 @ 地势/中间高，南北低”、“在 7、8 月份，伦敦比北京 @ 气温/高，日较差大”。
3. 时间：有 4 个样本，占 4.5%。例如：“图 1 中，圣若阿金地区 @ 终年/受赤道低压带控制，降水多”、“地震发生时的防震措施错误的是 @ 在街上时/迅速离开电线杆和围墙，到开阔地躲避”。
4. 地点 + 术语：有 2 个样本，占 2.3%。例如：“图中城市 @ 上海市服务功能/强，辐射全国”，“上海市”为地点，“服务功能”为术语。

5. 术语 + 时间，有 1 个样本，占 1.1%。例如：“下列关于该地逆温特征的描述，正确的是 @ 逆温强度午夜/达到最大，后减弱”。
6. 其他，有 7 个样本，占 8.0%。例如：“京津冀协同发展利于 @ 城市间/的分工与协作，降低竞争力”、“此时，关于图中天气的正确叙述是 @ 北京比东京/气温低，气压高”、“关于两国大豆产区及生产条件的叙述，正确的是 @ 均/位于沿海地区，交通便利”、“读美国和巴西大豆生长周期表，下列叙述正确的是 @ 都在/春、夏之交播种，秋季收获”。

与上述提到的名词性术语相对的还有动词性术语，例如“退耕还林”“植树造林”等等，这种术语在所有术语中所占比例较小，并且在两个边界不一致的样本中没有出现过，所以我们暂不考虑。由上面的分析可以看出，如果两个边界不相同，那么两者之间的词组是比较有规律的。因此提出一个假设：如果选项中第一个词语是时间、地点、术语或者其中两者的组合，那么公共部分右边界很可能位于它们的右边。于是我们继续对两个边界一致的样本进行观察，发现选项第一个词是上述时间、地点、名词性术语的不在少数：

1. 名词性术语：例如“崇明县城镇化水平不断提高的主要原因是 @/第一产业效率提高，农村出现剩余劳力。”、“甲地位于喜马拉雅山东端，林线高于青藏高原其它地区。其主要原因是 @/纬度低，气温较高”。
2. 地点：例如“关于‘丝绸之路经济带’东、西部沿海地区差异的描述，正确的是 @/东部人口稠密，西部地广人稀。”
3. 时间：例如“乙地气候特点是 @/夏季高温多雨，冬季寒冷干燥”。

但我们综合上述两类发现，当选项的第一个词语是时间、地点、名词性术语（本段将这三类词语简称为“特殊词”）时：如果选项中逗号右边部分的第一个词语对应地也是同类特殊词，那么很有可能两个边界是一致的（此时边界在逗号左侧部分的该特殊词的左边）；如果选项中逗号右边部分的第一个词语不是对应类型的特殊词，那么公共部分右边界很可能位于选项开头的特殊词后面。对于两个边界之间的词组是特殊词组合的情况（如“术语 + 地点”），也可以在利用上述规则判断完选项开头的特殊词之后，再应用上述规则判断选项中第二个词语（例如“此次网购过程中 @ 成都/正午太阳高度变小，白昼变短”，“成都”为地点，“正午太阳高度”和“白昼”为术语）。此外，当选项中第一个词语是动词时，则两个边界很有可能是一致的（在标注数据中，没有出现选项第一个词语是动词，而两个边界不同的情况）；当选项中出现“均”“都在”时，公共部分右边界在这些词的后面。输入为被分类为可拆分

的试题文本的选项部分中，由逗号隔开的两个部分的分词及词性结果，注意，对于时间/地点/术语这三类词，我们使用特殊的词性标记 time/loc/term 来表示。我们将这个基于规则的识别方法用算法 3.1来给出。

算法 3.1 公共部分右边界识别伪代码

输入：选项中逗号左边部分的词与词性 $S1 = w_1, w_2, ..., w_m$，$P1 = p_1, p_2, ..., p_m$
选项中逗号右边部分的词与词性 $S2 = w_{m+2}, ..., w_n$，$P2 = p_{m+2}, ..., p_n$
输出：边界右边第一个词的下标 k
初始化：k = 0
for w_i in $S1$ **do**
  if w_i == “均” “都” **then**
    k = i + 1
    break
  else if p_i 是动词词性 **then**
    k = i
    break
  else if p_i in {time, loc, term} and p_i != p_{m+1+i} **then**
    continue
  else if p_i in {time, loc, term} and p_i == p_{m+1+i} **then**
    k = i + 1
    break
  else
    $k = i$
    break
  end if
end for

3.3.3.5 含多个逗号选项的处理

在我们有的标注语料中，只有 36 个试题文本的选项包含了一个以上逗号，占所有含逗号选项的 6.6%。我们没有将这部分语料特殊处理后加入训练语料中，如前所述，我们所有的特征模板、后处理、公共部分右边界识别算法，都是针对选项中只含有一个逗号的试题文本的。对于选项包含一个逗号以上的情况，我们只在预测的时候做一些特殊处理，使得我们训练的模型和后处理等算法能够运用其上。下文所述的算法只是解决试题文本自动拆分的第一步，即判断是否可拆分。对于公共部分右边界的识别，则直接扩展算法 3.1，在判断每个部分的第一个词是否属于同一类型特殊词时，将两个部分扩展为多个部分的第一个词之间的比较，只要有任意的另外一个部分与第一部分的开头有同类型的

特殊词，就终止判断。

对于含多个逗号选项的试题文本，判断其是否可拆分的算法思想类似于多分类问题在二分类模型下采用的 one-to-one 算法：首先用逗号将选项隔成多个部分 $parts=part_1, ..., part_n$；然后将任意两个 part 用逗号组合在一起，得到一个仅包含一个逗号的伪选项，并与题面组合在一起得到伪试题文本；再使用训练得到的二分类模型（仅能处理选项含一个逗号的情况）对每种组合下的伪试题文本进行是否可拆分的二分类；最后统计对所有伪试题文本的判断情况，将占多数的判断结果作为原试题文本是否可拆分的判断结果，如果“可拆分”和“不可拆分”的伪试题文本数量一致，则倾向于不拆分，判断原试题文本为“不可拆分”。算法 3.2给出了上述过程的伪代码，其中的 PREDICT() 函数，就是使用最大熵算法在训练数据上训练得到的模型对一个选项仅含一个逗号的试题文本判断是否可拆分的函数。

算法 3.2 选项含多个逗号的试题文本是否可拆分判断算法

```
输入：题面 T，选项被逗号隔开的多个部分 parts=part_1, ..., part_n
输出：原试题文本的是否可拆分判断结果
初始化：判断为可拆分的伪试题文本数量 y_num=0、判断为不可拆分的伪试题文本数量 n_num=0
for i in 1, ..., n-1 do
  for j in 2, ..., n do
    伪试题文本 = 题面 + part_i + ”，” + part_j
    predict_result = PREDICT(伪试题文本)
    if predict_result == y then
      y_num++
    else
      n_num++
    end if
  end for
end for
if y_num > n_num then
  return 可拆分
else
  return 不可拆分
end if
```

3.4 实验及结果分析

3.4.1 实验配置

本章实验使用的数据来自于第5章中介绍的标注系统，共涉及 55 套试卷、859 道选择题。具体数据的相关统计信息在第3.3.2小节中有详细介绍。由于高考问答系统的目标是解答北京高考地理试题，因此对标注试卷的选择尽量向北京高考试题靠近，避免不同试题风格的差异对系统性能带来干扰。这里的 55 套试卷包括：13 套北京高考真题、11 套其他地区高考真题、17 套北京高考模拟题、14 套北京期中/期末/会考/联考试题。

3.4.2 是否可拆分二分类实验

在是否可拆分的二分类实验中，我们使用 python 的 nltk 第三方工具包提供的最大熵模型工具作为分类模型。在每一种配置下的实验中，我们都使用十折交叉验证来得到平均的性能，在每次实验中，训练集的比例占原始数据的 90%，测试集占 10%。

3.4.2.1 特征的影响

我们一共提出了 18 个上下文特征用于最大熵模型的训练和预测，在第3-2小节中做过详细介绍。本节我们不改变训练数据中两类数据的比例，直接将原始数据集中的所有选项含逗号的试题文本按 9：1 划分成训练集和测试集。

我们依次添加每一个特征，观察实验结果的变化情况。实验结果如表3-2所示。

从上表可以看出，加了更多的特征之后总体的准确率反而出现了降低的趋势，我们认为是可能因为训练数据量比较小，每种特征都能提高某一类试题文本的辨识度，但是这些类的文本数量还比较小，受稀疏性影响比较大，容易产生过拟合，对预测产生了一定的干扰。同时发现，我们比较关注的 n_recall 指标出现了先上升后下降的趋势。实际上，n_recall 和 total_precision 是一对此消彼长的特征，因为在测试数据中，可拆分类型的数据也远多于不可拆分类型的数据，如果对占多数的可拆分数据识别正确，则更可能会使总体准确率提升，此时不可拆分类型数据的召回率就更低。

表 3-2: 增加特征对性能的影响

添加特征	y_precision	n_precision	y_recall	n_recall	total_precision
+wordNumDiff	89.6%	74.5%	90.3%	69.2%	84.8%
+charNumDiff	89.3%	71.8%	89.5%	68.5%	84.0%
+postagEditDistance	89.8%	73.2%	89.2%	70.0%	84.2%
+lastPosComb	90.4%	69.0%	86.8%	72.3%	83.0%
+lastPosEqual	90.4%	69.0%	86.8%	72.3%	83.0%
+firstPosComb	90.0%	62.1%	82.0%	73.1%	79.6%
+firstPosEqual	89.9%	62.1%	81.9%	73.1%	79.6%
+lastWordInTimian	90.3%	61.2%	80.8%	74.2%	79.2%
+lastTwoWordsInTimian	91.3%	56.7%	78.4%	78.5%	78.4%
+lastPostagInTimian	91.3%	58.1%	79.5%	78.5%	79.2%
+timeCombination	91.2%	59.1%	80.5%	77.7%	79.8%
+firstWordInSecondPart	91.1%	61.4%	82.2%	76.9%	80.8%
+firstPostagInSecondPart	91.0%	60.7%	81.6%	76.9%	80.4%
+lastWordInFirstPart	90.1%	61.3%	81.9%	76.2%	80.4%
+lastCharInFirstPart	91.0%	63.9%	83.8%	76.2%	81.8%
+containCuewordsComb	90.8%	64.6%	84.1%	75.4%	81.8%
+containCuewordsMain	90.6%	65.3%	84.9%	74.6%	82.2%
+bothContainLonLat	90.6%	65.3%	84.9%	74.6%	82.2%

表 3-3: 高 n_recall 特征下的性能

y_precision	n_precision	y_recall	n_recall	total_precision
90.8%	58.1%	79.2%	76.9%	78.6%

但不可拆分类的识别召回率比总体准确率更加重要，因此我们选出所有能使 n_recall 升高的特征（wordNumDiff、postagEditDistance、lastPosComb、firstPosComb、lastWordInTimian、lastTwoWordsInTimian），或者没有使 n_recall 发生变化但是不损失 total_precision 的特征（lastPosEqual、firstPosEqual、lastPostagInTimian、lastCharInFirstPart），来做一次所有数据上的封闭测试，得到性能如表 3-3所示。

不过发现这里的 n_recall 比表 3-2中的最好的 n_recall 还要低。所以我们在后面的实验中还是选择表 3-2中最高的 n_recall 对应的特征集来训练模型。

3.4.2.2 训练集数据类型比例的影响

在前文中我们曾提到过，对于试题文本的拆分我们倾向选择保守的拆分判断，即希望尽量能够提高不能拆分的试题文本的分类召回率。因为如果将不可拆分的试题文本进行拆分，会得到一些没有意义的句子，对后续任务比较不利；而将可拆分的试题文本判断为不可拆分，则只是保留了原句，下游任务的输入仍然是一个有意义句子。在实验数据中，选项含逗号的试题文本中有 71.7%（391/545）都是可以拆分的，存在比较明显的数据倾斜现象，模型也会倾向于将文本分类成可拆分的，这样就与我们希望提高不可拆分数据的识别召回率相矛盾。因此我们采用了重采样的方法，对训练数据中的不可拆分样本进行随机重采样，调整训练集中的两类数据的比例，测试集数据保持不变。

在每次实验调整训练集的数据比例之前，还是先按 9：1 的比例划分训练集和测试集；如果我们想让训练集中的可拆分数据的比例超过 71.7%，则随机从中抽取一个可拆分的数据，复制该样本再加入训练集，直至比例达到所需。如果想让训练集中的可拆分数据的比例低于 71.7%，则从其中不可拆分的数据中随机抽取后复制放回。

表 3-4中显示了不同的训练集可拆分数据比例 y_proportion 下的各项性能。

该实验的目的是为了在实际工程中应用，这些应用则需要 n_recall 尽可能高，即尽量不要将不可拆分的数据误识别为可拆分，同时 total_precision 也越高越好。但前文已经说过这两个指标是负相关的。从表 3-4中的结果来看，我们选择可拆分数据在训练集中的比例为 40% 应用在实际使用中，这个比例下，n_recall 接近 90%，比原始比例提高了 10.7%，total_precision 比原始比例降低了 6.2%。

3.4.2.3 后处理的影响

基于前两个实验，本实验对可拆分数据在训练集中的比例为原始比例和 40% 两种情况，给出使用后处理和不使用后处理的性能比较，如表 3-6所示，“正确纠正”指将应该是不可拆分的数据由原本预测为可拆分纠正为不可

表 3-4: 不同的训练集中可拆分数据比例下的性能

y_proportion	y_precision	n_precision	y_recall	n_recall	total_precision
10%	96.1%	38.3%	44.6%	94.6%	57.6%
20%	96.0%	42.0%	53.2%	93.1%	63.6%
30%	95.7%	46.7%	62.2%	91.5%	69.8%
40%	95.2%	50.1%	67.3%	89.2%	73.0%
50%	93.5%	51.7%	70.8%	85.4%	74.6%
60%	92.4%	52.5%	73.0%	82.3%	75.4%
70%	91.3%	53.7%	75.4%	79.2%	76.4%
原始比例	91.3%	58.1%	79.5%	78.5%	79.2%
80%	90.2%	59.0%	81.4%	74.6%	79.6%
90%	89.2%	69.5%	88.1%	69.2%	83.2%

表 3-5: 后处理的影响

可拆分数据比例	实验类型	y_precision	n_precision	y_recall	n_recall	total_precision	正确纠正	错误纠正
40%	无后处理	85.4%	44.2%	71.6%	62.3%	69.2%	-	-
40%	有后处理	95.8%	50.7%	67.6%	90.8%	73.6%	32	17
原始比例	无后处理	79.9%	47.8%	86.2%	37.7%	73.6%	-	-
原始比例	有后处理	92.2%	58.8%	79.5%	80.8%	79.8%	56	25

拆分，“错误纠正”指将应该是可拆分的数据由原本预测为可拆分纠正为不可拆分。

实验结果表明后处理过程对 n_recall 有提升作用，并且我们希望的就是能够在使 n_recall 尽可能高的前提下识别出更多的可拆分的数据，后处理过程对这个前提提供了更好的保证。在训练集 40% 的可拆分数据比例下，后处理过程能使 n_recall 提升 28.5%，总体识别准确率提升 4.4%。不过需要说明的是，由于我们的不可拆分数据较少，后处理规则是根据数据的特点提出来的，所以可能存在过拟合的现象，在更多的未知数据中的提升不一定有这么明显。

但该过程也同时会将很多可拆分的试题文本错误纠正为不可拆分的，我们观察了具体的结果后发现主要有两种情况：（1）有一部分错误是由于在标注错误引起的，例如“图中河流的流向及河流与水渠的关系是 @ 河流自南向北

表 3-6: 后处理的影响

边界类型	正确的个数/总数	正确率
与题面选项边界相同（A 类）	267/303	88.1%
在选项第一部分中（B 类）	71/88	80.7%
所有数据	338/391	86.4%

流，河流水补给水渠”被标注为可拆分，实际上仔细看了之后应该是不可拆分的，因为题面问的是两个问题，如果只和后面的一个部分拼接成句子，则题面中有一部分内容是无用的；（2）有一部分是由于相似的句式在不同句子在标注时出现了偏差，例如“由 08 时到 20 时，图中 @②地受高压脊控制，天气持续晴朗”被标注为可拆分，“大红门、动物园服装批发市场 @ 受历史因素影响，形成于传统商业区”、“孟加拉国的降水特征对农业生产的影响是 @ 旱季持续时间长，利于作物成熟”被标注为不可拆分。从我们对拆分的定义上，这些句子应该分类为可拆分（得到两个短句各自有意义且不改变原句意），但从直觉上来说又有一点模糊，前后两部分之间的因果关系较明显，所以对这类错误的纠正是可以接受的。

所以，这两种后处理中出现的主要错误都是可以接受的，不认为会对使用性能产生明显的负面影响。

3.4.3 公共部分右边界识别实验

本节使用第3.3.3.4中介绍的基于规则的方法，对数据集中所有可以拆分的数据进行公共部分右边界的识别。实验结果如表**??**所示。

实验结果表明该方法是可行的。对识别错误的样本给出如下分析：

1. 第二部分开头的术语是第一部分开头术语的一个属性，将公共部分右边界错误识别为题面选项边界：例如“图示两区域 @ 河流/以雨水补给为主，流量稳定”，“流量”是“河流”的属性，而不是与其地位相同的另一个术语。这类错误共有 6 例。
2. 多种边界均可接受：例如“1980　2000 年该城市人口密度的变化表现在 @ 甲区/人口密度最高，增长速度最快”，我们的算法会将边界识别在“人口密度”后面，但是因为题面中出现过这个词，所以认为第二部分可以不用补全这个概念，两种边界都是可以接受的。这类错误共有 16 例。

3. 第二部分中有省略的隐含术语：例如“由 08 时到 20 时，图中 @②地/气压升高，未来呈降温趋势”，第二部分隐含了术语“气温”。再例如“巴西 @ 亚马孙河/径流量大，流经经济发达地区”，第二部分没有一个明确可表述的隐含术语，但从语义上来说是描述了亚马逊河的另一个和“径流量”同等地位的属性。这类错误共有 12 例。
4. 第二部分中包含了时间词，干扰了同等地位术语的判断：例如“下列关于图中等温线与甲、乙、丙、丁四地气温的叙述，正确的是 @ 甲地/海拔高，1 月气温低于丙地”，“海拔”和“气温”是等地位的，但是第二部分中的“1 月”干扰了算法的判断。这类错误共有 2 例。
5. 类似术语但不是术语的词出现在开头：例如“企业总部留在北京主要因为 @/距离远，搬迁费用高”中的“搬迁费用”。这类错误共有 2 例。
6. 其他：例如“京津冀协同发展利于 @ 城市间/的分工与协作，降低竞争力”中，边界会被算法识别为“城市”之后、“间”之前，但这样得到的第二个句子就不再通顺。这类错误不具有典型性，出现地较少，上述其他类别不包括的错误都归于此类，这类错误共有 15 例。

3.5 本章小结

本章针对高考地理问答系统中的选择题的特点，提出了将复杂的试题文本（一个题面 + 选项组成的完整句子，选项中含有逗号）拆分成多个简单句的方法。逗号在中文中使用广泛，经常用来对短语或者从句进行并列，如果选项是一个包含逗号的较为复杂的句子，则如果能判断选项是否说的是两件或多件可以独立判断真伪的事情，然后将每个部分和题面分别组合成更简短的句子，对后续的语义模板转换的任务则有比较大的帮助。根据项目组相关人员的实验结果，在语义模板转换的错误分析中，由于选项较为复杂（含逗号，陈述多件事）导致转换失败或错误的比重较大。因此如果我们能够尽可能不错误识别不能拆分的句子，同时识别出一些可以拆分的句子并将其简化，则对语义模板的性能改进有较大意义。

在本章中，我们介绍了地里选择的特点，并详细阐述了试题文本拆分的直观定义，以及与中文逗号分类的关系，并介绍了我们标注的所有拆分相关的数据的情况。然后将拆分过程分成两个主要步骤：识别是否可拆分、识别公共部分右边界。在第一步中，我们使用最大熵模型，利用试题文本的上下文特征来

训练分类模型，并提出了提高不可拆分试题文本识别召回率的方法：一是提高训练集中不可拆分试题文本的比例，来抵消数据倾斜导致召回率低的问题；二是使用后处理过程，通过基于规则的方法，将一些样本的识别结果纠正为不可拆分。在第二步中，我们通过对地理试题数据的观察，提出了基于规则的方法来寻找公共部分右边界。

在第一步分类过程中，在训练使用 40% 的不可拆分数据比例，使用使召回率最高的 10 个特征，并使用 9 个后处理规则，使得不可拆分数据的召回率可以达到 90.8%，可拆分数据的召回率达到 50.7%，也就是说对不可拆分数据的识别错误很少的情况下，能找出一半的可拆分数据，这对后续的语义模板转换仍然是十分有意义的。而在第二部公共部分右边界的识别中，总体上可以达到了 86.4% 的准确率。

对于目前的性能，仍然可以有提升的空间，例如获取更多的数据，对本章提到的错误分析进行一些针对性的处理等等。

第四章　AMR 语义理解

4.1　引言

在 AMR 出现以前，自然语言理解在语义分析上的工作还比较零散，例如实体连接、命名实体识别、语义角色标注等等工作，从某种意义上来说都属于语义理解的研究范畴。这样一来，实际上每一项研究都只关注了语义理解的一个方面，没有一个公认的完整的语义表达体系，能够将句子中的语义完整地描述出来。AMR 的出现从某种程度上来说有希望能够解决这个问题，它是一种图表示方式，致力于将句子中的语义从表层的语法表示中抽象出来，将动作、实体、修饰等等各种语法要素都抽象成概念，然后以边来表示出概念之间的关系。

AMR 作为一种新型的语义表示方法，目前看来是一种比较强大的、有潜力的语义表示，可以给自然语言理解带来一种新的研究方向，并对一些相关领域的研究或者工程性的工作带来结果上的提升[19][20]。因此，本文针对高考地理问答系统中的地理试题的理解，考虑了 AMR 这种新的方法，但由于 AMR 在中文的处理上暂时还是空白，可使用的工具和语料还很悠闲，目前还未进入实际应用阶段。我们有幸和中文 AMR 语料标注工作小组合作，得到了一些可用的中文语料，因此本文针对 AMR 的主要工作是对比现有算法在不同语言（中英文）、不同语料规模、不同语料类型、人工对齐与否等方面的性能，并在少量的地理试题 AMR 标注数据上进行了一些实验，为将来的应用工作打下基础。

4.2　相关工作

AMR 是一种以有向的、简单的、类似树结构的（树出现环变成图的比例比较小）、边和节点有标记的图[11]，可以将句子中的语义概念和概念之间的语义关系以节点和边的形式表现出来。

AMR 有多种表现形式，根据定义可以以一个图的方式呈现出来，如图4.2.1.1；在图表示方法中，没有入边的节点就是根节点，在该图中就是“want-01”对应的节点。这张图是对“The boy wants the girl to believe him”这

图 4-1: AMR 图表示

句话的 AMR 表示。这里总共有四个节点，每个节点都对应一个 instance，实际上就是节点对应的概念。根节点是“want-01”对应的节点，后缀“-01”实际上是指定了这句中 want 这个谓词在 Propbank 体系中对应的语义框架，在这个语义框架下谓词有相应的多个论元，在这里就有 arg0 和 arg1 这两个论元。

这句话中有一个“want”动作，它的 arg0（谁在 want）是“boy”，它的 arg1（want 的是什么）是一个“believe”的动作；“believe”动作的 arg0（谁在 believe）是“girl”，它的 arg1（believe 什么）是“boy”。在这里“boy”有两个作用，一是作为“want-01”的 arg0，一是作为“belive-01”的 arg1，这个节点有两条入边，因此 AMR 图上出现了一个环，这在 AMR 中也叫做“重入”（reentrancy）。也正是因为“重入”现象的存在，使得 AMR 是一种图表示方法，而不是树表示方法。

上图这样的表示方法看起来稍有点复杂，AMR 还可以以一种文本方式表达出来：

```
(w / want-01
      :ARG0 (b / boy)
      :ARG1 (b2 / believe-01
            :ARG0 (g / girl)
            :ARG1 b))
```

这里的第一行就是 AMR 的根节点，冒号后面的标签表示节点之间边的标记，也就是节点之间的关系。每个节点由变量名（斜杠前的部分）和概念（斜杠后的部分）组成（除了多次出现的同一个节点），如果一个节点出现多次，则通过让该节点的变量名多次出现来表示，如该例中的节点 b。这里概念里的斜杠也可以看做是图表示方法中 instance 的简化表示。

AMR 也可以用逻辑三元组的来表示，上面的例子可以表示成：instance(w, want-01) /* w 是 want 的实例 */ instance(b, boy) /* b 是 boy 的实例 */ instance(b2, believe-01) /* b2 是 believing 的实例 */ instance(g, girl) /* g 是 girl 的实例 */ ARG0(w, b) /* b 是 want 的发起者 */ ARG1(w, b2) /* b2 是被 want 的东西 */ ARG0(b2, g) /* g 是 believe 的人 */ ARG1(b2, b) /* b 是被 believe 的人 */

上面例子中的这句话在英文中还可以有很多种不同的表示方法，例如：

1. The boy wants to be believed by the girl.
2. The boy has a desire to be believed by the girl.
3. The boy’s desire is for the girl to believe him.
4. The boy is desirous of the girl believing him.

概念 want-01 可以在句子中用动词 want、名词 desire 或者形容词 desirous 来表达。

再比如下面这个 AMR 表示：

```
(p / permit-01
        :ARG1 (g / go-02
                 :ARG0 (b / boy))
        :polarity ))
```

也有很多种英文表达方式：

1. The boy may not go.
2. The boy is not permitted to go.
3. It is not permissible for the boy to go.
4. The boy does not have permission to go.

由此可以看出，AMR 更加强调句子表达的逻辑语义，并将表层的语法表达方式抽象出去。

AMR 虽然是一种有向图表示，但是对于图的边和节点的关系仍然有一些限制条件[14]：

1. 简单性：应该是一个简单图，即任意两个概念节点之间最多有一条边
2. 连通性：应该是弱连通图
3. 确定性：对于某个节点，不能有标签相同的两条出边
4. 无环性：图中没有环（有向边形成的首尾相连的环）

4.2.1 标注体系

通过上文的介绍可知，AMR 主要有两个部分组成：概念（节点）、概念间的关系（边）。下面将从这两个方面介绍 AMR 的标注体系。

4.2.1.1 概念（concept）

AMR 的 concept 主要有三种来源：句中词语（例如4.2.1.1中的“boy”）、propbank 中的语义框架（例如中的“want-01”）、抽象的关键字。

从之前举的例子可以看出，AMR 试图对不同的表达方式甚至是相同语义的不同词语做出统一的标注方式。因此英文 AMR 对常见的 concept 制定了一个实体类型的标准列表，首先需要从这些实体名称中找出和想要描述的实体最相近的一个（例如有人会说“person”，有人会说“woman”，则可以通过这个列表统一起来）:

1. person, family, animal, language, nationality, ethnic-group, regional-group, religious-group
2. organization, company, government-organization, military, criminal-organization, political-party, school, university, research institute, team, league
3. location, city, city-district, county, local-region, state, province, country, country-region, world-region, continent, ocean, sea, lake, river, gulf, bay, strait, canal, peninsula, mountain, volcano, valley, canyon, island, desert, forest, moon, planet, star, constellation
4. facility, airport, station, port, tunnel, bridge, road, railway-line, canal, building, theater, museum, palace, hotel, worship-place, market, sports-facility, park, zoo, amusement-park
5. event, incident, natural-disaster, earthquake, war, conference, game, festival
6. product, vehicle, ship, aircraft, aircraft-type, spaceship, car-make, work-of-art, picture, music, show, broadcast-program
7. publication, book, newspaper, magazine, journal
8. natural-object
9. law, treaty, award, food-dish, disease

有时候句子中也会有单词指示该实体的类型，如果这个词比上述方式选择的类型更具体，则使用这个词代替，例如对于“the poet William Shake-

speare”，可以用：

```
(p / poet
        :name (n / name :op1 "William" :op2 "Shakespeare"))
```

如果上述列表中所有实体名都没有合适的，句子中也没有明确指明实体的类型名称，就可以用“thing”来表示这个实体。例如对于“Words are the source of misunderstandings”，AMR 中对“understanding”这个添加了“thing”这个 concept：

```
(s / source-01
        :ARG1 (t / thing
                :ARG0-of (m / misunderstand-01))
        :ARG2 (w / word))
```

如果一个实体有多个类型，例如著名诗人、画家 XXX，虽然同一种关系可以多次出现，但是:instance 是 AMR 中唯一一种只能出现一次的关系，这时候可以使用:mod 关系来修饰，例如对于“the poet Dr. Seuss”：

```
(d / doctor
        :name (n / name
                :op1 "Seuss")
        :mod (p / poet))
```

还有一类概念用来表示数量：monetary-quantity , distance-quantity , area-quantity , volume-quantity , temporal-quantity , frequency-quantity , speed-quantity , acceleration-quantity , mass-quantity , force-quantity , pressure-quantity , energy-quantity , power-quantity , voltage-quantity (zap!-, charge-quantity , potential-quantity , resistance-quantity , inductance-quantity , magnetic-field-quantity , magnetic-flux-quantity , radiation-quantity , concentration-quantity , temperature-quantity ,score-quantity ,fuel-consumption-quantity , seismic-quantity。通常这类概念后面都会有一个 quant 关系指向表示具体数值的节点。

此外还有一些常见的 concept 列举如下：

1. relative-position（相对位置）
2. product-of 和 sum-of（数学运算）
3. date-entity（时间）
4. date-interval（时间区间）

5. percentage-entity（百分比）
6. phone-number-entity（电话号码）
7. email-address-entity（电子邮箱）
8. url-entity（url）

对于特殊疑问句，AMR 还对被提问的部分设计了一个 amr-unknown 概念。例如“What did the girl find?”：

```
(f / find-01
        :arg0 (g / girl)
        :arg1 (a / amr-unknown))
```

4.2.1.2 关系（relation）

AMR 在英文标注中大约有 100 种关系，比较常见的可以大致分成以下 5 个类型：

1. 框架论元:arg0,:arg1,:arg2,:arg3,:arg4,:arg5.
2. 通用语义关系:accompanier, :age, :bene ciary, :cause, :compared-to, :concession, :condition, :consist-of, :degree, :destination, :direction, :domain, :duration,:employed-by, :example, :extent, :frequency, :instrument, :li, :location, :manner, :medium, :mod, :mode, :name, :part, :path, :polarity, :poss, :purpose, :source, :subevent, :subset, :time, :topic, :value.
3. 数量相关的关系:quant, :unit, :scale
4. 时间实体相关的关系:day, :month, :year, :weekday, :time, :timezone, :quarter, :dayperiod, :season, :year2, :decade, :century, :calendar, :era.
5. 列举关系:op1, :op2, :op3, :op4, :op5, :op6, :op7, :op8, :op9, :op10.

对于所有的关系，AMR 还允许这些关系的反转形式也作为关系标签，例如:arg0 的反转形式为:arg0-of，:location 的反转形式为:location-of。除了上面列举的，还有一些其他的关系标签，不再一一列举。

4.2.2 语料对齐与自动对齐

AMR 的语料对齐，即是将原句中的词语、词组与 AMR 图中的一个概念节点或者一个概念节点子图（例如一个时间概念子图，包括一个 date-entity 概念及其指向的表示年、月、日等信息的节点）进行对齐，这样的对齐信息对于自

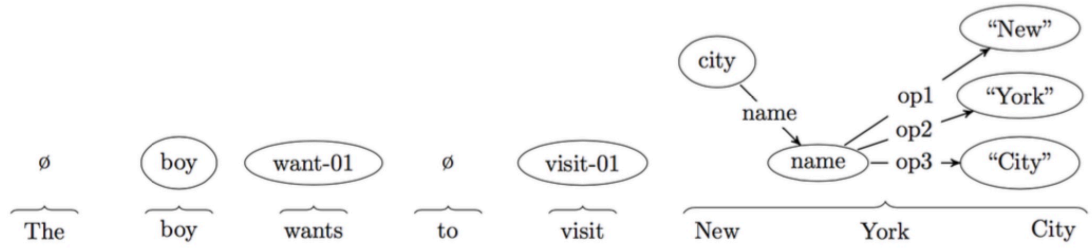

图 4-2: AMR 对齐

动解析算法有着重要的作用，例如在概念识别等阶段。图4-2给出了一个 AMR 节点与句子词语之间的对齐示意图。

在语料建设工作中，不同的语料可能存在一个差异，即是否有对齐信息。目前英文的语料都是不含有对齐信息的，上文的例子都是这一类标注；中文已公布的 1500 句小王子语料也是不含有对齐信息的，但是即将公开的数据将对齐信息考虑了进去，嵌入了 AMR 的标注中，因本文的 AMR 部分工作是与含有对齐信息的语料标注团队合作，所以在我们的实验中也使用到了这一部分数据。

下面给出同一个中文句子的有对齐标注和无对齐标注两个版本，并进一步说明对齐信息是如何标注的。这个例子是对来自《小王子》语料中的一句话的两种标注，这句话分词的结果是“画的是一条蟒蛇正在吞食一只大野兽。”：

无对齐信息：

```
(x12 / 吞食-01
        :arg0 (x14 / 蟒蛇
                :quant (x15 / 1)
                :cunit (x5 / 条))
        :arg1 (x16 / 野兽
                :mod (x17 / 大-01)
                :quant (x18 / 1)
                :cunit (x10 / 只))
        :domain (x19 / thing
                :arg1-of (x20 / 画-01))
        :time (x21 / 正在))
```

有对齐信息：

```
(x8 / 吞食-01
```

```
:arg0   (x6 / 蟒蛇
         :quant  (x4 / 1)
         :cunit  (x5 / 条))
:arg1   (x12 / 野兽
         :mod  (x11 / 大)
         :quant  (x9 / 1)
         :cunit  (x10 / 只))
:domain  (x24 / thing
         :arg1-of  (x1 / 画-01))
:aspect  (x7 / 正在))
```

可以看到，在 AMR 的整体结构上两者基本一致，最主要的差异在于变量（variable）名的选取。

对于无对齐信息的标注版本，在英文语料中，变量名通常是概念的首字母，如果出现了重复的首字母，则在后面添加一个数字来区分，例如“b2”等等。在中文语料中则通常是一个字母“x”加上一个数字，这个数字没有特别含义，只要能够将不同的概念区分开来即可。上述只是约定俗成的一种标注方式，并非标注规范要求，也可以采取别的方式来标注。

对于有对齐信息的标注版本，每个变量名中的数字都是有含义的。如果概念对应原文的某个词或者短语，呢么变量名的数字后缀就是对应的词或短语的下标（第一个词语的下标为 1 而不是 0）；如果概念是新增的抽象概念，例如对于时间实体加了“date-entity”概念节点或者上面增加的“thing”这个概念，他们的数字下标是任意一个超过句子长度的数字，表示该概念不直接对应于句子中的词语，以免混淆。

在中文的对齐标注中，根据概念与句子分词结果中的对齐方式分类，有这几种情况：

1. 对应一个词语：例如上面的“(x11 / 大)”，表示“大”这个概念对齐到句子的第 11 个词
2. 对应多个词语：例如“那么刺有什么用呢？”的 AMR 标注中有一个概念“x4_x6 / 有用-01”，表示“有用-01”这个概念对齐到句子的第 4 和第 6 个词上。对于连续的词，也需要逐个写出对应的区间内的每一个词的下标，以免和上面这种跳词现象混淆。需要注意的是，在英文的 AMR 标注中，几乎没有跳词现象。

3. 对应某个词语中的某几个字：例如“我把锤子、螺钉、饥渴、死亡，全都抛在脑后。”的 AMR 标注中有一个概念“x12_1 / 抛-01”，表示“抛-01”这个概念对齐到句子的第 12 个词的第一个字上（如果是多个字则依次写出每个字在词中的下标，例如“x5_1_2_3”）。

关于自动对齐，Flanigan 等人[14] 在提出 JAMR 解析算法的同时，为了训练概念节点的识别，提出了一种自动对齐的算法，主要是根据一系列的规则，用贪心算法来进行对齐，达到了 92% 的准确率、89% 的召回率和 90% 的 F 值，但是这种对齐方式只支持对概念节点和词语之间进行对齐，而没有考虑到有时边的 label 也是来源于某个词语，也存在对齐关系。Nima Pourdamghani 等人[29] 提出了另一种自动对齐方式，该方法主要分为预处理、训练、后处理三步，在预处理阶段将 AMR 转换成一个字符串（保留关系的标签），然后进行小写化、去除停用词、去除词缀、去除等预处理，然后利用 IBM 的对齐模型进行训练，在后处理再重建出对齐好的 AMR 图。在对齐过程中，这种方式包含了边的标签信息，所以可以对边也进行对齐。

4.2.3 自动解析算法

Flanigan 等人[14] 在 2014 年提出了一种两阶段的图算法 JAMR，这是第一个公开的 AMR 自动解析算法，为后面的研究提供了一个较强的 baseline。在 2015 年 Wang Chuan[16] 等人观察到 AMR 表示与句法分析结果之间的相似性，提出了一种基于转换的 AMR 解析算法。Pust 等人[17] 也在同年提出了一种使用基于语法的机器翻译的方法进行 AMR 的解析。Lucy Vanderwende 等人[18] 还提出了一种基于逻辑形式（Logical Form）到 AMR 的转换规则的解析方法，来处理没有大规模 AMR 标注语料的语言的 AMR 解析问题。本节重点介绍本文实验基于的 JAMR 图算法以及基于转换的 AMR 解析算法。

4.2.3.1 基于图的算法

Flanigan 等人[14] 将 AMR 解析分成两个阶段：第一个阶段是从句子中识别出概念节点；第二个阶段是预测这些节点之间的边，也就是概念之间的关系。

在概念识别阶段，句子将会被切分成多个连续的片段（span），通过序列标注算法来实现，并将每一个片段对应于一个概念子图（AMR 的一个片段，可能包含多个概念节点，例如一个时间实体包括一个 date-entity 概念和具体的

年月日概念等等），这些概念子图都来自于训练语料中这些片段曾经对齐到的概念子图，或者是一个空图（即放弃这个片段到 AMR 的对应）。通过对一系列连续的句子子串 b 和一系列的概念图片段 c 进行打分（两者的个数均为 k），使用公式4-1这个线性参数化函数进行优化，其中 f 是句子子串（span）和它可以对应的一个概念子图片段在上下文中的特征向量表示，包括词、长度、命名实体等等这些特征。

$$score(\mathbf{b}, \mathbf{c}; \theta) = \sum_{i=1}^{k} \theta^{\mathrm{T}} f(\mathbf{w}_{b_{i-1}\, b_i}, b_{i-1}, b_i, c_i) \tag{4-1}$$

在关系识别阶段，则是基于第一个阶段识别出来的概念子图片段，在子图之间添加关系得到最终的 AMR 结果，这个结果需要满足五个条件，其中四个是在本章开头介绍过的 AMR 的有向图需要满足的四个条件：连通性、简单性、确定性、无环性，另一个就是保持性，即在第一阶段识别出所有概念子图片段应该是最终的 AMR 结果的子图。首先会对训练语料中的边训练一个线性参数的函数，特征为这条边的一些上下文，在解码的时候根据训练出的参数和函数对候选边进行打分。这个阶段的算法描述如下：边的候选集合为所有概念节点之间两两连接的有向边（两个方向，所有可能的边类型）；首先对于第一阶段识别出来的概念子图中已经存在的边，相应的节点对之间的其它边都不再作为候选边，这些存在的边成为 AMR 最终结果的组成部分；根据训练出的边打分函数，对未确定关系的节点对之间的所有候选边打分，保留分数最高的一条边，其他的不再作为候选；将所有得分大于 0 的边选出作为 AMR 最终结果的组成部分；对剩余的候选边，根据得分从高到底依次判断每条边是否进入最终结果，如果这条边可以将之前未相连的两个子图连接起来，则保留这条边；直至所有的节点之间形成一个弱连通图。根据实验结果，虽然算法未保证无环性，但是在所有测试数据上均未得到有环的结果。此外，上述算法过程可以保证连通性、简单性，但不能保证确定性，因此作者通过应用拉格朗日松弛法来保证确定性。

4.2.3.2 基于转换的算法

Chuan Wang 等人[16] 提出了一种基于转换的 AMR 解析算法。这个算法也是两阶段算法：第一步先使用依存句法分析器得到句子的依存句法结果；第二步则是根据他们提出的基于转换的算法，将依存句法结果一步步转换到 AMR

图的结果。从语言学的角度来看，在一个句子的 AMR 和依存结构之间存在很多相似点：两者都有有向的节点间关系；AMR 的概念和关系虽然是从居中实际使用的词语中抽象出来的，但是通常存在一些规律性的映射关系；句中的实词通常会保留在 AMR 结果中，虚词则常被忽略。从这些现象中可以感觉出，可以使用有限次转换动作，将一个依存句法树转换为 AMR 结果。

为了完成从句法分析到 AMR 的转换，作者设计了一系列转换的动作，并学习一个模型去决定在每次转换中采取哪个动作。包括如下这些动作：

1. NEXT-EDGE-l_r：赋予一条边一个标签
2. SWAP-l_r：将两个节点之间的边的方向反转，并给这条边赋予一个标签。这条边的新的 head 取代原 head 节点的位置，成为别的节点指向的节点。这个动作用于解决 AMR 和依存句法树对 head 的选择的差异。
3. REATTACK_k-l_r：为一个节点 A 重新选择一个 head 节点，即去除该节点原有的一条入边，将其与另一个节点建立一条新的边，这条新边仍是 A 的入边，同时给这个新边赋予一个标签。这个动作的出发点是，当反转了边之后，有些关联到旧的 head 节点上的节点应该重新关联到新的 head 节点上。
4. REPLACE-HEAD：将某个节点 A 删除，然后让 A 指向的节点 B 取代被 A 的位置，也就是使 A 的 head 节点成为 B 的 head 节点。这个动作用于删除一些 AMR 中不包含但是依存句法树中包含的节点，例如介词等虚词节点。
5. REENTRANCE_k-l_r：添加重入边，给某个节点和任意一个可能的其他节点之间添加一条新的边，在 AMR 结果中形成一个环。
6. MERGE：将两个节点合并为同一个节点，让新节点继承旧的节点的所有出边和入边。这个动作通常用于为句子中连续的一个词语序列生成一个节点，通常可能对应一个命名实体，比如将一个名字的两个单词合并起来。
7. NEXT-NODE-l_c：给某个节点赋予一个概念标签，然后将该节点从待处理节点队列中移除，开始考虑在下一个节点上采取什么动作。
8. DELETE-NODE：删除一个节点及与其相关的所有边，这个动作只用于删除没有出边的节点，作用是删除一些功能词节点。

该转换算法的状态是一个三元组 $(\sigma,\beta,\mathrm{G})$，$\sigma$ 存储所有待处理的节点，初始化成依存句法树中的所有节点，β 存储 σ 中第一个节点的所有出边指向的节点，G 存储当前的解析结果，初始化为依存句法树，算法结束后为得到的 AMR 结果。算法会训练一个在某个状态下对所有动作进行打分的函数，在每次转换前，找出所有动作中得分最高的那个，应用到当前的解析结果 G 上进行一次转

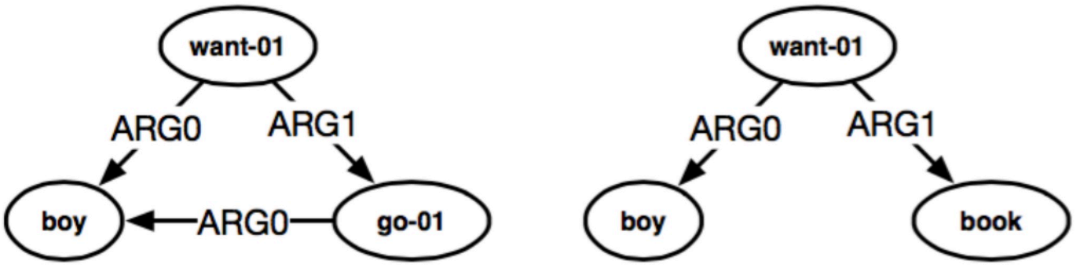

图 4–3: 两个待比较的 AMR 结果

换。打分函数的训练特征包括节点特征、节点对特征、路径特征、动作相关特征四类，涉及到词语、命名实体标签、依存句法等等句法和词法特征。

4.2.4 自动评价

目前，AMR 最常用的评价方式是计算自动生成的 AMR 和测试集中的参考 AMR 的 smatch 得分[14][30]。

在本章开头曾经介绍过，AMR 可以以逻辑三元组集合的形式表示出来，例如对于图中左边的 AMR，可以表示成：

instance(a, want-01)
^instance(b, boy)
^instance(c, go-01)
^ARG0(a, b)
^ARG1(a, c)
^ARG0(c, b)

三元组有两种形式，一种是 relation(variable, concept)，另一种是 relation(variable1, variable2)，例如上述前三行就是第一种三元组，后三行为第二种三元组。

对于图中右边的 AMR，可以表示成：

instance(x, want-01)
^instance(y, boy)
^instance(z, football)
^ARG0(x, y)
^ARG1(x, z)

smatch 就是计算两个 AMR 表示的三元组之间匹配的准确率、召回率和 F

表 4-1: 不同变量映射方式

映射方式	M	P	R	F
x=a, y=b, z=c	4	4/5	4/6	0.73
x=a, y=c, z=b	1	1/5	1/6	0.18
x=b, y=a, z=c	0	0/5	0/6	0.00
x=b, y=c, z=a	0	0/5	0/6	0.00
x=c, y=a, z=b	0	0/5	0/6	0.00
x=c, y=b, z=a	2	2/5	2/6	0.36
smatch score				0.73

值（通过将两个 AMR 中的变量名映射起来，然后计算两组三元组之间的相同的个数，再根据总的三元组个数计算准确率和召回率）。困难在于两个 AMR 所使用的变量名并不相同，所以两个 AMR 之间基于不同的变量映射方式，可以计算得到多种重合结果。smatch 则定义为：将两个 AMR 之间的节点一一对应起来，所能得到的最大的 F 值。例如针对上面的例子，有六种变量匹配的方式，如表 4-1。

所以上面这个例子的 smatch 得分就是所有映射方式中所能得到的最大的 F 值 0.73。虽然 smatch 的概念很简单明了，但是计算起来实际上没有这么简单，对于包含大量节点的 AMR，枚举所有的可能的映射方式的时间复杂度很高，所以有一些高效的算法被用于计算近似的 smatch 值，例如爬山法等等，这超过了本文的介绍范围，所以不再展开。

4.3 AMR 在中文上的应用

关于 AMR 在中文上的发展，我们在第 4.2节中已经介绍过。目前，在中文上已发表的 AMR 语料，我们已知的只有 1500 左右的小王子[13]；据我们所知，目前还没有针对 AMR 中文语料进行自动解析的算法及实验结果。

有幸与李斌老师带领的中文 AMR 标注小组合作，我们获得了除了小王子语料外的 5004 句基于 CTB 语料的 AMR 标注数据，以及 729 地理选择题试题文本（题面 + 一个选项组成的完整句子）的 AMR 标注数据。

在本节中我们将介绍如何将 AMR 自动解析算法应用到中文数据上。我们

的工作是基于 Flanigan 等人[14] 提出的两阶段的图算法 JAMR，对中文处理的需求进行模块的替换，对第4.3.2小节中提到的对齐语料进行对齐信息的提取，以满足 JAMR 的需求。

用 JAMR 训练以及分析新句子时，首先需要进行预处理，一是调用一些自然语言处理基础任务的工具包对语料进行分析，包括得到语料的分词（英文不需要分词，就是得到一个个 token）、命名实体识别结果、依存句法分析结果等，二是使用基于规则的算法来对齐 AMR 标注图中的 concept 和原句中的字符串，用于训练阶段得到 AMR 概念和触发该概念的字符串的对应关系。

我们没有改动 JAMR 的主体算法，只对上述的两种预处理针对中文做了一些工作，以使后续的工作可以在中文数据上奏效。在下面两个小节中，我们详细介绍每一项工作内容。

4.3.1 基本自然语言处理任务

上文中已经提到，JAMR 的预处理中需要得到分词、命名实体识别、依存句法分析这三项基本自然语言处理任务的结果，以用于后续的训练和预测。

对于分词，英文不需要进行分词，而中文的分词结果在标注语料中已经给出，所以我们只需关注另外两项处理。

对于命名实体识别，JAMR 原本使用的是 IllinoisNER，该工具能够识别出 PER（人名）、LOC（地名）、ORG（机构名）三种命名实体；在中文处理中，我们使用哈工大 LTP 的命名实体识别，对应地，该工具可以识别 Nh（人名）、Ns（地名）、Ni（机构名）。

对于句法分析，JAMR 原本针对英文的处理，和我们对中文的处理，都是使用了 StanfordParser，但是我们需要使用中文的模型，我们使用的是 xinhuaFactored.ser.gz 这个模型。

4.3.2 语料对齐

在第小节，我们详细介绍了对齐的含义，以及有对齐的标注语料是如何标注这个信息的。对齐信息将会用于 AMR 解析的第一步，即 concept 识别阶段；此外，实际上这个对齐信息就是 JAMR 的 AMR 解析算法第一阶段得到的结果，也可以直接将这个对齐结果作为第二阶段预测的输入，以此来观察算法在两个阶段的性能表现。

英文语料中没有标注对齐，JAMR 中使用了一种基于规则的自动对齐方法来对齐，可以达到约 90% 的 F 值[14]。例如对于这样一个 AMR 标注：

```
# ::snt Why should any one be frightened by a hat ?”
(f / frighten-01
        :ARG0 (h / hat)
        :ARG1 (o / one
                :mod (a / any))
        :ARG1-of (c / cause-01
                :ARG0 (a2 / amr-unknown)))
```

JAMR 会得到下面这样的对齐信息：

```
# ::alignments 5-6|0 3-4|0.1 2-3|0.1.0 8-9|0.0
# ::node 0.0 hat 8-9
# ::node 0.1 one 3-4
# ::node 0.1.0 any 2-3
# ::node 0.2 cause-01
# ::node 0.2.0 amr-unknown
# ::root 0 frighten-01
# ::edge cause-01 ARG0 amr-unknown 0.2 0.2.0
# ::edge frighten-01 ARG0 hat 0 0.0
# ::edge frighten-01 ARG1 one 0 0.1
# ::edge frighten-01 ARG1-of cause-01 0 0.2
# ::edge one mod any 0.1 0.1.0
```

对齐信息主要分成三类：（1）以“# ::node”开头的几行：节点编号、该节点的 concept、该 concept 在原句中对应的子串的下标范围（包括位置，不包括终止位置，下标从 0 开始，例如 5-6 表示句子下标为 5 的词，即“frightened”。这个范围称为“span”）；（2）以“# ::edge”开头的几行：边的出发节点的 concept、边的指向节点的 concept、边的出发节点的编号、边的指向节点的编号；（3）以“# ::alignments”开头的一行：所有在“node”行出现的 span 对应的节点的编号。

有时，一个 span 可以对应多个节点，例如对某个概念抽象出另一个概念，如“(q / question-01)”这个概念上抽象了一层得到“(t2 / thing:ARG1-

of (q / question-01))”，此时“thing”和“question-01”这两个概念节点对应到同一个 span 上。这种情况下，用“+”来表示对应多个节点，如“14-15|0.3.1+0.3.1.0”。

还有更特别的情况，例如“”The flowers have been growing thorns for millions of years .”中的“years”，可以对应到 AMR 标注中的两个节点，即“t2 / temporal-quantity”和“y / year”，而“temporal-quantity”概念除了“year”这个子节点，还有别的子节点。该例的 AMR 标注为：

```
(g / grow-03\\
    :ARG0 (f / flower)\\
    :ARG1 (t / thorn)\\
    :duration (m / multiple\\
        :op1 (t2 / temporal-quantity \\
            :quant 1000000\\
            :unit (y / year))))
```

对于一个 span 对应多个节点的情况，人工对齐的数据中只会标注出原词直接对应的概念与原词的对应信息（如“year”这样的节点），而对于抽象出的节点如“temporal-quantity”、“thing”等等，则没有对齐标注。例如原句有 10 个 token，这样的抽象概念的变量名可能是“x15”，这样就不能对齐到原句的词语上，也就是无对齐标注。在我们从人工标注的对齐语料中转换对齐信息格式的时候，对这种情况也只对齐出有对齐标注的节点，对于抽象的节点则不产生对齐结果。这一点也会对对齐的结果产生负面的影响。

对于有对齐标注的语料，我们将其转换成这种形式，但是在转换过程中我们发现中文对齐与英文有所不同，对于中文的对齐有一些更复杂的场景：

1. 一个概念对齐到不连续的几个词语之间的情况，例如“有什么用”中可以提取出了个概念“有用-01”。
2. 一个概念对齐到一个词的一部分，例如“市场分析家”中的“分析”可以作为一个概念。
3. “关系”有时候也与原句的一些词语存在对齐关系，例如“他大胆地向国王提出了一个请求”中，“大胆”会被对齐到一个概念上，该概念与其父节点的关系是“:manner”，而“地”字表明了“大胆”表示一种方式，所以“地”可以被对齐到“:manner”这个关系上。

由于 JAMR 现有模型还不能处理这些对齐情况，我们对这几种情况在工程

表 4-2: AMR 中英文语料说明

数据集	总句数	训练集	开发集	测试集
LDC2013E117（英文）	8252	3988	2132	2132
小王子（英文）	1562	1274	145	143
小王子（中文）	1562	1274	145	143
CTB（中文）	5003	4097	414	492
地理选择题（中文）	729	583	73	73

实现上作了一些兼容的处理，这里提出这三个问题，主要是对未来工作指出一些可改进之处。由于目前中文 CTB 的 AMR 标注数据还不够完善，尤其是对于前两种对齐情况还有较多的标注错误，所以目前还无法提出有意义的方法来解决这些问题。对于第一种情况，我们将概念视为对齐到多个词语的第一个词语上；对第二种情况，视为对齐到原词；对第三种情况，忽略所有边对齐信息。这样的让步几乎会丧失这些对齐信息的全部作用，但好在这样的对齐在语料中所占的比重很小，所以对算法的性能不会产生太明显的影响。

综上所述，我们主要的工作包括两个方面：一是通过替换处理模块，将 JAMR 应用于中文 AMR 的解析；而是利用中文 AMR 语料中标注的对齐信息，使其可以用于 JAMR 的后续处理，并指出其中存在的问题。

4.4 实验及结果分析

在本节中，我们会给出原始 JAMR 在英文上的结果，以及修改后的 JAMR 在中文上的结果。还会给出在中文上对使用 JAMR 自动对齐方法和人工标注对齐数据的效果对比。并且会在中英文语料上做封闭测试，以验证该算法模型的有效性。

4.4.1 实验数据简介

对于英文测试，我们会使用 Flanigan 等人[14] 在发表 JAMR 的论文中使用的数据、小王子语料，具体数据集情况如表4-2所示。对于中文测试，我们会使用小王子语料、CTB 语料、地理数据语料。

其中，LDC2013E117 是一个新闻语料，CTB 中文语料则是来源于论坛、

表 4-3: AMR 中英文封闭测试性能

数据集	Smatch（两阶段）			Smatch（gold 概念）			Span 识别		
	P	R	F-score	P	R	F-score	P	R	F-score
LDC2013E117（英文）	81.5%	82.2%	81.8%	95.9%	87.9%	91.7%	82.1%	90.2%	85.9%
小王子（英文）	74.9%	71.2%	73.0%	92.3%	76.0%	83.4%	79.5%	94.1%	86.2%
小王子（中文）	56.3%	43.8%	49.2%	66.4%	45.6%	54.0%	83.5%	92.4%	87.8%
CTB（中文）	57.8%	39.7%	47.1%	68.4%	49.8%	57.6%	85.4%	80.2%	82.7%
地理选择题（中文）	70.1%	58.7%	63.9%	74.3%	62.6%	67.9%	94.5%	94.0%	94.3%

博客。

4.4.2 封闭测试

我们对表4-2提到的五个中英文数据集都做了封闭测试，将表格中的测试集和训练集的数据合并起来，作为封闭测试的训练和测试数据，开发集保持不变。实验结果如表4-3所示。该测试对于中文，是使用 JAMR 自带的自动对齐算法做概念的对齐。

这里我们主要比对中英文小王子语料的实验结果，因为这两个语料规模一致，并且是平行语料，可以排除句子长度、句子类型（口语化还是正式）带来的影响。从实验结果来看，对于英文的 AMR 解析，仅使用小王子 1500 句的语料，JAMR 的两阶段测试总体 F 值可以达到 73.0%，可以说明该模型对于 AMR 的结构具有一定的处理能力；如果直接使用语料中的对齐结果作为概念识别的结果，只进行第二阶段的预测，F 值可以达到 84.4%。而对中文 AMR 的解析，两阶段测试的 F 值只有 49.2%，比英文低 23.8%；仅预测第二阶段的 F 值为 54.0%，比英文低 29.6%。

smatch（gold）概念指标可以反映第二阶段预测的情况，Span 识别则反映第一阶段识别的情况。从实验结果可以看出，中英文的第一阶段识别性能差不多，而第二阶段（关系）识别则拉开了很大的差距，第二阶段的识别与句子内部的结构关系紧密，与句法特征等等比较有关。

在 Levy 等人的工作[31] 中，论述了对中英文进行句法分析的难度差异。在 Durrett 等人在 2015 年的工作[32] 中，对比了多种句法分析方法在英文句法分析上的性能，其中最高的 F 值可以达到 92.4%，是使用集成方法达到的。而在 Zhang 等人 2011 年的工作[33] 给出的多个方法的性能比较中，在 CTB 语料上进

表 4-4: AMR 中英文开放测试性能

数据集	Smatch（两阶段）			Smatch（gold 概念）			Span 识别		
	P	R	F-score	P	R	F-score	P	R	F-score
LDC2013E117（英文）	67.6%	60.8%	64.1%	84.6%	77.6%	81.0%	75.0%	74.3%	74.7%
小王子（英文）	57.6%	39.8%	47.1%	75.1%	59.0%	66.1%	74.2%	66.9%	70.4%
小王子（中文）	44.6%	31.3%	36.8%	58.9%	42.4%	49.3%	72.7%	71.0%	71.9%
CTB（中文）	42.7%	31.2%	36.1%	57.5%	41.0%	47.8%	74.1%	68.5%	71.2%
地理选择题（中文）	51.9%	24.6%	33.3%	63.5%	51.4%	56.8%	84.4%	50.5%	63.2%

行句法分析的最好 F 值为 80.4%。中文句法分析在性能上与英文仍有明显的落后，而句法特征是 AMR 训练和预测使用的重要特征，因此句法分析上的错误是中文 AMR 解析性能明显差于英文的一个重要原因。

此外，可能 JAMR 用来识别英文概念之间的关系时，使用的一套特征模板不能很好适用于中文，需要重新针对中文数据调试出一套性能更好的模板。中文是一种意合的语言，而英文是形合的语言，中文的语义表达可能更加多变。

这里仅对中文性能较差的原因给出了一些推测，我们的实验结果仅仅是在中文上进行 AMR 解析的一次初步尝试，性能差异的具体解释和探求已经超出了本文的讨论范围。

4.4.3 开放测试

该测试对于中文，也是使用 JAMR 自带的自动对齐算法去做概念的对齐。JAMR 的概念对齐对于时间使用了只适用于英文的一套正则表达式，并不能直接在中文中使用。但是在语料中，时间所占的比例还是非常小的，所以暂时可以忽略这个原因给性能带来的影响。

值得注意的是，三个中文语料的两阶段测试 F 值都在 30% 至 40% 之间，而他们的语料规模有明显差异。如表 **??**所示，CTB 语料共有 5003 句，小王子语料有 1562 句，地理题语料有 729 句。我们猜测这样的结果是由于三个语料的句子长度差异导致的，句长统计结果如表 4-5所示。

CTB 语料的平均句长为 21.8，小王子的平均句长为 12.9，地理试题的平均句长为 12.4。统计结果为之前的猜测提供了一些证据。我们可以初步得到一个结论，平均句长较大的语料，AMR 的解析会更难做。

表 4-5: 中文语料句长分布统计

句子长度范围	len<10	10<=len<20	20<=len<30	30<=len<40	len>=40
CTB（中文）	10.0%	37.8%	31.6%	12.5%	7.2%
小王子（中文）	42.3%	42.1%	10.9%	2.8%	2.0%
地理选择题（中文）	32.0%	57.6%	9.9%	0.5%	0.0%

表 4-6: 中文 AMR 人工对齐数据的影响

数据集	Smatch（两阶段）			Smatch（gold 概念）			Span 识别		
	P	R	F-score	P	R	F-score	P	R	F-score
小王子（自动对齐）	44.6%	31.3%	36.8%	58.9%	42.4%	49.3%	72.7%	71.0%	71.9%
小王子（人工对齐）	40.7%	35.3%	37.8%	59.0%	50.7%	54.5%	63.8%	63.2%	63.5%
CTB（自动对齐）	42.7%	31.2%	36.1%	57.5%	41.0%	47.8%	74.1%	68.5%	71.2%
CTB（人工对齐）	44.9%	34.3%	38.9%	57.1%	45.9%	50.9%	75.7%	71.2%	73.4%
地理选择题（自动对齐）	51.9%	24.6%	33.3%	63.5%	51.4%	56.8%	84.4%	50.5%	63.2%
地理选择题（人工对齐）	48.9%	26.6%	34.4%	62.1%	50.1%	55.5%	72.5%	46.2%	56.4%

4.4.4 在中文上使用人工对齐数据

对齐信息是新旧 AMR 标注之间最主要的区别，因此我们想验证一下目前这种对齐标注的有效性。在本实验中我们对比了三个中文数据集使用自动对齐和人工对齐的实验结果，如表 4-6所示。

从实验结果来看，使用人工对齐不一定能够提高实验结果，例如小王子语料的 Span 识别，和地理选择题的 Span 识别以及 gold 概念下的 smatch 得分，都有不同程度的下降。性能下降可能有这几方面因素：一是自动对齐本身就有 90% 以上的准确率，二是人工对齐忽略了对新增抽象概念的对齐。另外还有一些在第 4.3.2小节中提到的一些中文对齐的问题。

4.5 本章小结

AMR 是一种新型的语义表示方法，目前在英文上开展了一些研究，包括语料标注、自动对齐、自动分析等等，但是在中文上仅有一个 1500 句左右的小王子语料是公开的，另外还有一些 CTB 语料、本文研究背景相关的地理选

择题语料正在标注中。本章工作基于英文 AMR 的第一个公开的自动解析工具 JAMR，将其中对英文处理的组件替换成中文，使其可以处理中文语料。

本章工作主要介绍了 JAMR 针对中文的改动、对中文有对齐标注的处理，并对中英文共 5 个语料进行了实验，实验包括封闭测试、非封闭测试、对齐数据对比测试。实验结果发现，JAMR 对英文 AMR 的描述能力比较强，封闭测试可以达到 81.8% 的总体性能，而对中文 AMR 的封闭测试只有不到 50%，可能是由于中文本身句法分析的准确性就比英文有明显区别，错误累积使得中文的 AMR 效果差得更多。在正常测试中，中文的实验结果平均比英文也要低 20% 左右。在中英文小王子的实验结果对比中发现，主要是在第二阶段关系识别时，中文数据的表现远差于英文，而概念识别的性能则相差不多。另外通过中文三个语料实验结果的对比，发现句子长度越长的语料，即使语料规模更大，也会比较难以得到更好的结果。在语料对齐方面，我们比较了在中文数据上使用 JAMR 的自动对齐和直接使用标注好的对齐数据两种方式的实验结果，发现人工标注的对齐不总是能提高实验结果。

本章工作主要是对 AMR 在中文上的应用做一次探索，尝试现有的算法在中文数据上的应用，并测试了在高考地理选择题数据上的应用效果。可以看出高考地理题仅用较小的语料规模就可以得到与 5000 句以上 CTB 语料差不多的实验效果。实际上，700 多句高考选择题文本中有很多是非常类似的，因为一道选择题有四个选项，在很多情况下，各个选项之间可能仅有几个词语发生了变化。所以实际上地理题的标注规模是很小的。因此可以猜测，AMR 在地理题这类句子较短、语义清晰、书面化的语料上，有比较好的应用前景，是一个值得探索的应用领域。

在未来工作中，除了可以尝试不同的 AMR 分析算法在数据上的预测效果，还可以针对有对齐信息的标注数据进行进一步利用，结合对齐信息和自动对齐，将抽象出的概念也进行对齐，使对齐信息的作用不会被丢失的抽象概念对齐抵消掉。在 JAMR 的算法中，对概念的识别比较依赖训练集中的出现过的概念，几乎不能识别未在语料中出现的概念，可以针对这一点对概念的识别做一些优化，利于对实词，即使未在训练语料中出现过，也将其转换为一个概念节点。对于中文 AMR 分析中，关系识别的效果与英文远差于中文的现象，也应该引起注意，思考一些针对中文的改进方法，比如寻找适用于中文的特征模板等。

第五章　地理试题标注系统

5.1　引言

为了给选择题试题文本拆分和地理试题的 AMR 分析提供标注语料，也为了给整个高考问答系统的各项涉及自然语言处理和理解的任务提供高质量的地理试题标注语料，我们开发了一个地理试题标注系统。除了标注拆分和 AMR 的数据，这个系统还支持对分词、词性标注、命名实体、术语、成分句法分析等传统自然语言处理任务在地理试题上的标注，以及试题语义模板、套话、上下文、题干核心成分分类、选项类别、题干前导部分分类等地理试题领域相关的、用于解题的各项数据的标注。

在此之前，我们没有统一的数据标注平台和规范，导致得到的数据比较零散、错误率高，我们开发这个系统主要是为了提供下列各项服务：

1. 由于各个任务所需要的数据之间互有关联（如词性标注依赖分词的结果），所以需要对同一套试卷进行各个方面的标注，但通常不同的标注内容由不同的人员完成，例如分词是一个同学负责，语义模板是另一些同学负责，我们这个基于 B/S 的系统可以让所有标注人员在同一个平台上看到别的同学的标注结果，实时性相对较高。
2. 由于各类标注数据自身的特点、不同类数据之间的相互关系，标注数据之间存在较多的约束关系，例如同一个句子的分词和词性个数需要一样，词性应该是有效的。该系统为各类数据的约束条件提供了检查，保证大多数容易出现的错误能够被自动检查出来，提醒标注人员进行修改，防止引入标注错误。
3. 有些标注内容有自动的标注工具，例如分词工具、词性标注工具、句法成分分析器等等，对这类数据的标注通常是基于自动分析的结果进行人工修正，而不是完全由标注人员去标注，这样可以大大降低标注的工作量，提高标注效率。另一方面，自动分析也可以用于对比人工修正结果与自动分析结果之间的差异。该系统对分词、命名实体识别、术语识别、词性标注、成分句法分析、试题语义模板提供了生成自动分析结果的功能。
4. 提供标注快捷键、标注提示等等，尤其在试题语义模板的标注功能上，避

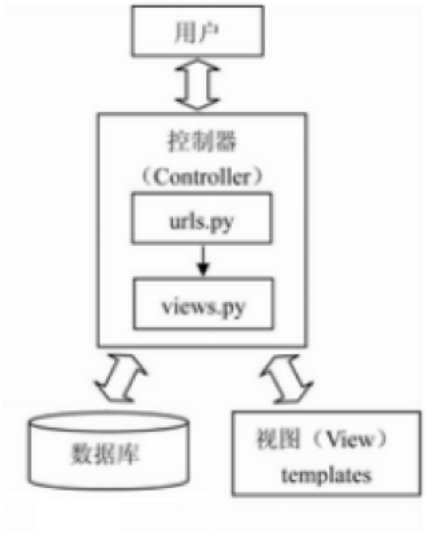

图 5-1: 试题标注系统基于 django 的架构

免了标注人员还要另外参考模板体系规范。可以快速进行标注

5. 对所有的标注数据记录标注人信息，数据出错或者有异议时，可以方便找到原始标注人进行讨论和修改。
6. 支持试题检索，可以在系统内所有标注数据中快速找到包含相关关键字的试题。
7. 支持按所属试卷、按试题语义模板类型两种方式导出数据，并且可以选择导出哪些标注内容，方便各项任务的研究人员获取数据。
8. 在添加试题时支持查重功能，对系统中已有的试题和试卷进行提示，避免重复劳动。

5.2 系统架构

该系统采用了 B/S 架构，使用 python 的 django 框架进行 WEB 环境的搭建，在后台使用了 mongodb 数据库来存储数据。系统架构如图 5-1所示。

系统主要包含四个核心功能：试题上传、试题浏览、试题标注、试题导出。对选择题和主观题分开管理，但是提供基本相同的标注流程，并对两种试题都提供了上述四种核心功能，同时对两者的差异进行了处理，例如在试题上传时对两种试题的格式特点分别解析。

在试题标注中，由于各项标注内容之间存在相互依赖关系，系统中为了这种依赖关系，要求在标注某一项数据之前，需要完成其依赖的标注内容在该试

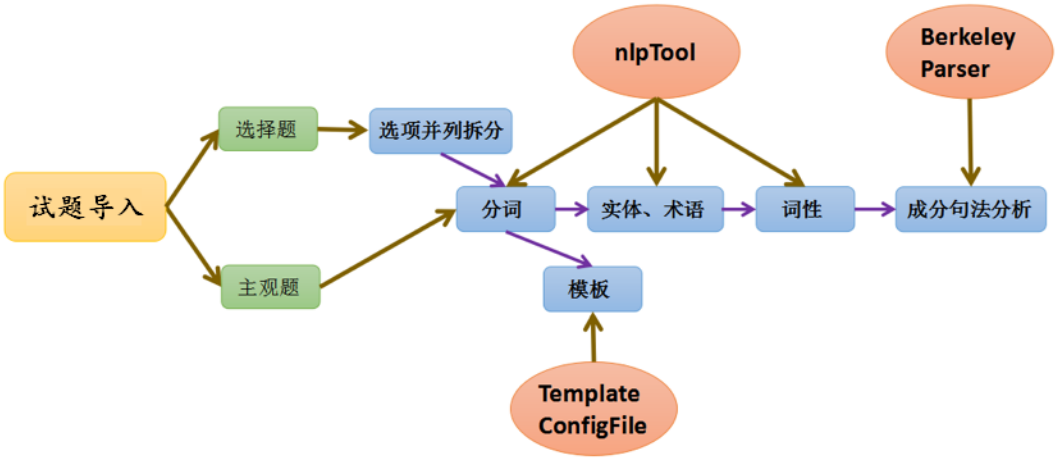

图 5-2: 标注流程示意图

题上的标注。选择题和主观题的标注流程如图 5-2所示。

首先是试题导入，试题为特定格式的 txt 文件，选择题和主观题分别在通过不同的文件上传。系统会判断 txt 文件内部格式是否正确，例如题目的层次、选项的个数等等，经过解析的试题数据会以层次化的结构，以一个 document 的形式存储一个文件中的试题。对于选择题，会特别进行选项并列拆分的标注，用于为第 3章中的地理选择题选项拆分任务提供标注数据。

图中标注任务之间的有向箭头表示数据之间的依赖，只有完成某一项任务的标注后，从该项任务指出的箭头所指向的任务才可以进行标注。具体来说，试卷会首先进行分词标注，在分词完成后可以进行试题语义模板（图中简称“模板”）、实体和术语的标注，完成实体和术语的标注会可以进行词性标注，完成词性标注后可以进行成分句法分析的标注。此外还有一些解题相关任务的数据标注未在该图中提现，例如套话、上下文、题干核心成分分类、选项类别、题干前导部分分类等，这些标注仅依赖分词标注。

图中上方指向标注任务的箭头，连接了标注任务和该任务在本系统中对应使用的自动标注工具的名字。nlpTool 是项目组其他成员基于地理试题领域文本进行性能调优的自动分析工具，支持包括分词、实体（时间、地点、数量词）、术语、词性等四项标注，例如在分词任务中加入了地理领域词典，在术语识别中使用地理术语表等资源。Berkeley Parser 用来给试题文本生成成分句法分析结果。

图中下方的 Template ConfigFile 记录了目前系统支持的模板类型及其填槽规范和示例，在标注的过程中，可能发现某个模板指定地不合理，或是想要增

图 5-3: 词性单项标注页面

加或删除一些模板，甚至更改整个模板体系，通过这个配置文件，可以灵活地修改系统支持的模板定义。

5.3 功能说明及使用方法

系统提供的标注是基于试卷的，通过试卷名找到对应试卷，然后开始标注。支持两种标注方式，一种是按照标注内容一项一项进行标注，另一种是对某一道题进行所有标注。通常在使用过程中，由于标注人员个有分工，基本上是采用按内容标注的方式，在系统中我们称之为“单项标注”，词性的单项标注如图**??**所示。以题目为单位的标注通常用于标注完成后对某道题的标注进行修改，我们称之为“单句标注”，如图**??**所示。两者之间的主要区别在于进入的入口，以及有单项标注页面的数据的标注，成分句法、语义模板等的标注都是使用同样的页面。

图 5-4: 单句标注页面

5.3.1 基本使用流程

本节以选择题为例，简要说明一张试卷的主要标注流程使用方法。首先我们提供了浏览试卷的页面，这个页面会显示系统中当前所有的选择题试卷，并显示每一项内容的标注完成情况及标注人信息，提供进入标注页面的超链接。此外，该页面还提供检索功能：对试卷名关键字的检索、对各项数据标注状态的检索。该页面如图5-5所示。

如果我们打算标注某个试卷的某项数据，需要点开图5-5所示页面中对应试卷所在行的最后一列中“单项标注”链接（如果点开单句标注），系统会根据该试卷当前的标注情况，为所有可以进行标注的内容提供超链接。可以标注的内容为：其所依赖的标注数据已经全部标注完成、依赖该标注的标注数据还未提交。不满足这两个条件的标注内容不能再通过“单项标注”进行修改，只

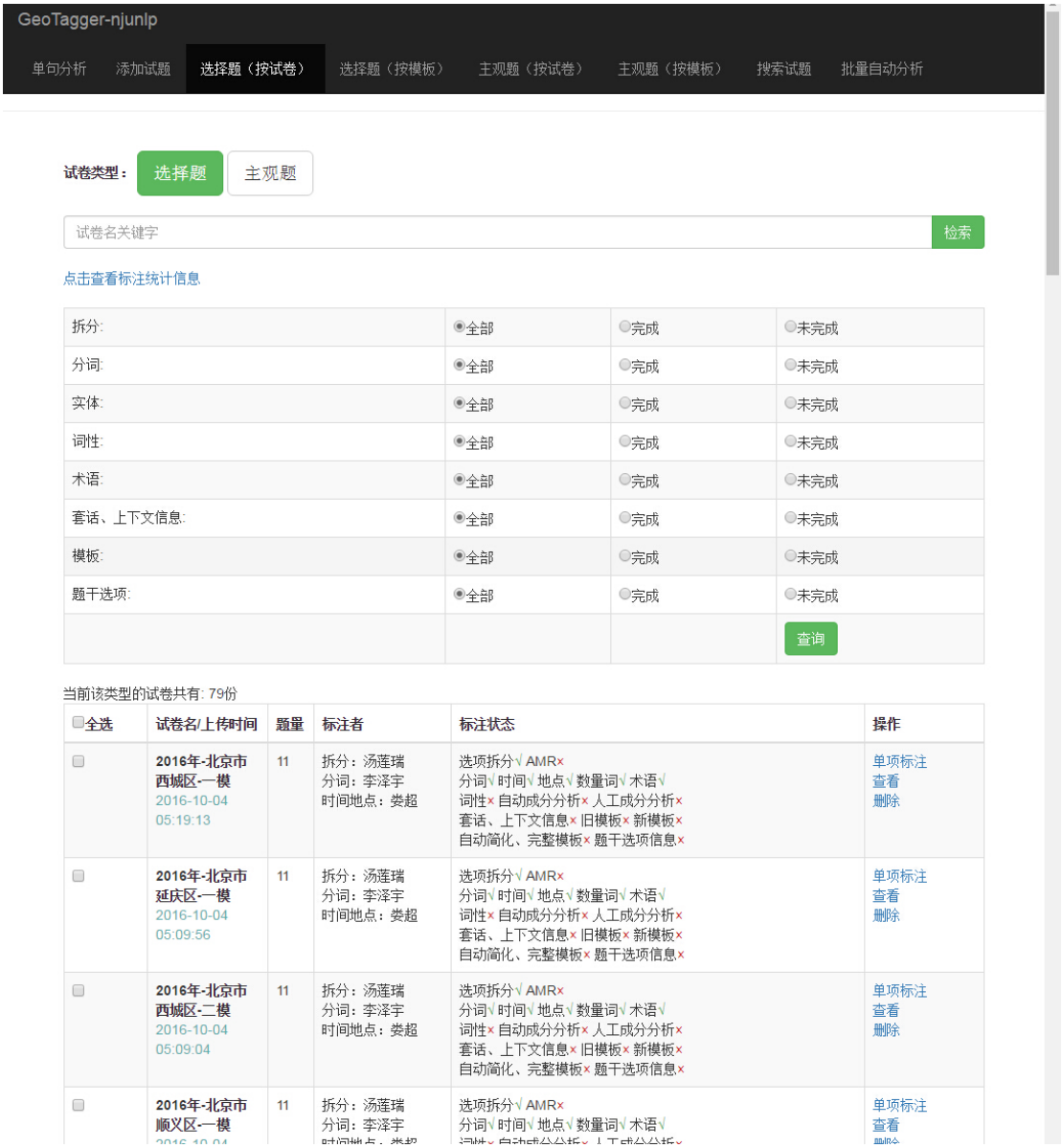

图 5-5: 试题浏览页面

能通过“单句标注”修改该数据及与之相关的依赖数据。例如图5-5中的第一份试卷“2016 年-北京市西城区-一模”，在它的“单项标注”页面中，提供如图5-6的标注超链接。因为分词是时间地点标注的依赖数据，当时间地点标注完成后，分词不能再整体标注。时间地点是词性标注的依赖数据，所以这里时间地点不能再整体标注。后五项数据都没有被依赖数据，而它们的依赖数据分词已经完成标注，所以这些内容是当前可整体标注的。

严格来说，整体标注的数据仅有分词、时间地点（即实体，实际上还包括数量词的标注）、词性、术语、套话/上下文标注这 5 项内容。因为这些标注

图 5-6: 单项标注页面

对于每一个试题文本来说都很简短，两行文本或者几个填空就可以完成，并且同一道选择题的四个（或以上）试题文本之间常常互有关联，例如对于分词来说，多个选项之间的题面分词结果是一致的。所以系统提供了在同一个页面中对该试卷所有试题文本进行某项数据的标注功能，例如词性单项标注页面如图**??**所示。

对于不方便进行整体标注的数据，我们为每个试题文本单独显示一个标注页面，例如下面两个小节介绍的试题文本拆分标注、AMR 标注都是每次标注一个试题文本。

5.3.2 试题文本拆分标注

标注系统会对所有选项中含有逗号的试题文本提供一个标注界面，并调过不含逗号的试题文本，如图 5-7所示。我们需要通过图 5-5中试卷所在行的最后一类中的“标注拆分”超链接进入这个界面，这个超链接只会在还未完成标注拆分或已完成标注拆分但没有提交分词结果（提交分词结果后不能再进入标注拆分）的试卷中显示，所以该图中没有这种试卷也就没有这个超链接。

首先我们可以选择该试题文本是否可拆分，如果选择“是”，则会显示下面的拆分后的几个部分，如果公共部分右边界位于选项中而不是题面和选项的

图 5-7: 拆分标注页面

边界处，则第二个部分及后面的部分，会在最前面补上缺失的公共部分。例如在例图中，选项文本的拆分情况为“下列关于太阳方向的叙述，不正确的为 @ 冬至日某地/太阳升起的方位是东偏南，落下的方位是西偏南”，原本两个逗号隔开的部分是“冬至日某地太阳升起的方位是东偏南”及“落下的方位是西偏南”，但是由于公共部分右边界位于“冬至日某地”后面，所以会将‘@’和‘/’之间的部分补全到第二部分的最前面，这样得到的每一个部分和题面进行拼接，就得到了拆分后的多个句子。如果选择“否”，则不会显示拆分后的部分。这样就标注了是否可拆分，以及可拆分的情况下公共部分右边界的位置。

5.3.3 AMR 标注

中文 AMR 的标注已经有一个更加专业和完整的工具，李斌老师开发了这个工具并用于中文语料的标注[13]，但不能直接接入我们的系统。为了避免重复劳动，我们没有重新开发一套完整的 AMR 标注工具，包括对中文 Propbank 的支持等功能。在地理试题标注系统中加入 AMR 标注的目的主要是为了保持数据的一致性和标注的完整性，将 AMR 标注结果和其他标注结果保存在一起，便于检索和使用。

系统为 AMR 标注提供了基本的功能，提供了一个简单的输入框，可以将使用专业工具标注得到的结果添加进来保存入系统，同时这个标注结果是随时

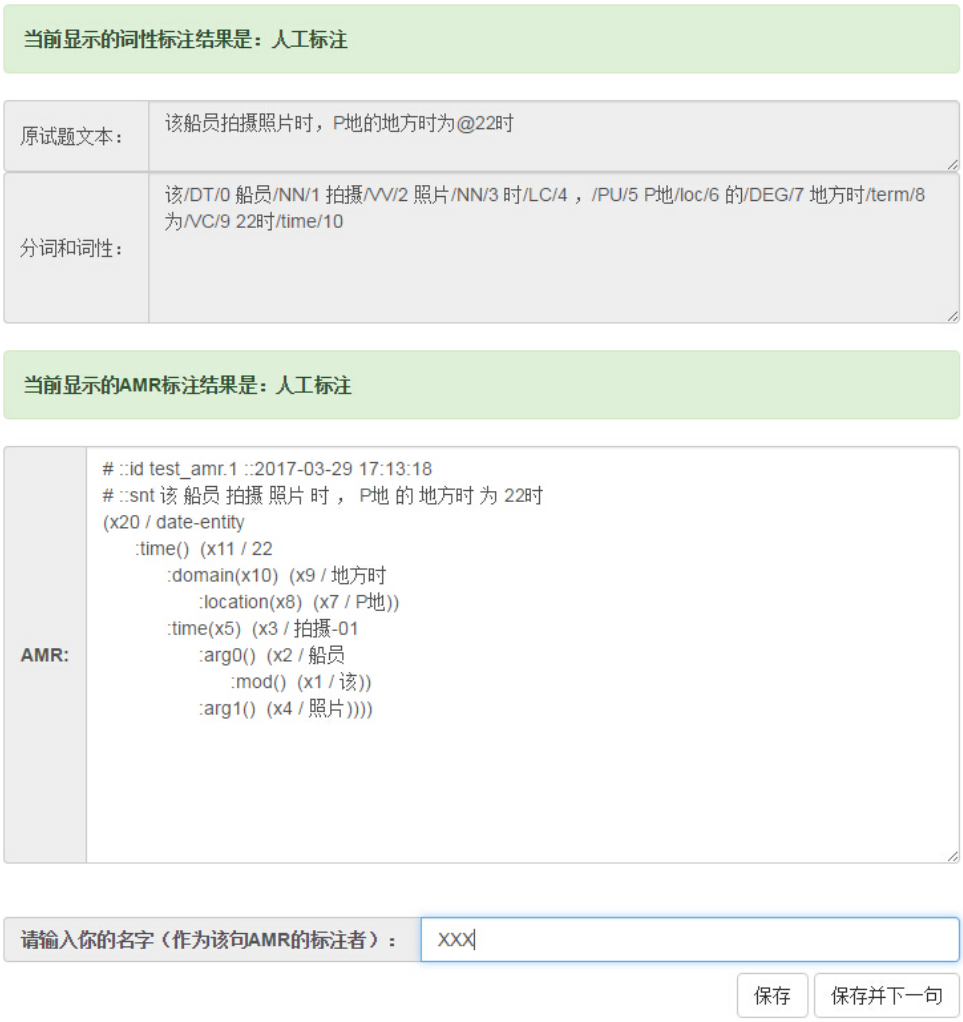

图 5-8: AMR 标注页面

可以修改的。标注界面如图5-8所示。

5.3.4 标注数据导出

我们提供两种数据导出方式：一种是以试卷为单位，每份试卷的数据生成一个文件；一种是以试题语义模板的类型为单位，找出所有对应类型的试题文本，为同一类型的试题文本及其标注数据生成一个文件。在两种导出方式下，都可以根据使用者的需要，选择需要下载哪些标注数据。例如按模板类型导出的页面如图5-9所示。

图 5-9: 按模板类型导出数据的页面

5.4 本章小结

本章介绍了地理试题标注系统的主要功能，包括整个系统的架构，基本的使用流程等等。该系统可以为多项地理试题问答系统的自然语言处理相关任务提供高质量的语料，包括本文对试题理解所做的复杂试题拆分、AMR 语义分析等功能。这个系统使标注人员之间的合作更加便捷，对标注效率的提升也很明显，并且可以检测出很多人工标注中可能出现的错误。

系统主要分为试题上传、检索、标注、导出等四个功能，已经投入实际使用，为十几位标注人员提供自然语言处理各项任务及地理试题理解相关的各项任务的标注环境，目前已经对 79 份试卷进行了标注。在使用过程中，也接受了

多位标注人员的反馈，对系统的便捷性、实用性、鲁棒性进行了提升。

第六章　总结与展望

6.1　工作总结

本文工作的背景是高考地理自动问答系统，本文工作的切入点主要是对高考地理试题的理解。试题理解是多方面的，包括基本的 NLP 各项任务，例如命名实体识别、时间识别、句法分析等等，也包括对句子语义的理解，在项目中提出了试题语义模板的概念，每一个题目的文本都需要转换成这样的模板格式，以便于后续的推理。地理试题的理解具有领域特点，本文的工作从几个角度出发，从不同方面来加强对地理试题的理解。

本文从以下几方面展开了具体工作：

1. 针对地理选择题的特点，提出了复杂选择题试题文本的拆分方法，即对选项中含有逗号的较长选项，判断其是否描述了两件或以上可以独立判断的事情，如果是，再去寻找这两个子句或者短语的公共句子前缀，将句子前缀与这几个部分分别拼接，得到几个较短的简单句子。在第一步判断是否可拆分中，使用了最大熵模型作为分类器，并对一系列上下文特征的效果进行验证；在第二步中，在对数据特点进行观察后，提出了一种基于规则的启发式方法，来寻找公共部分右边界。这个方法侧重于在尽可能高的不可拆分数据的召回率下，使可拆分数据的召回率达到一定水平。应用后，可以使一部分长句简化为短句进行后续处理，进而简化了试题语义模板的转换等工作在这些试题文本上的处理过程。这项工作从句子结构出发，间接地对试题语义理解的工作做出提升。
2. 将 AMR 的自动解析算法 JAMR 修改应用于中文，并在中英文共五个语料（包括一个小规模的地理选择题语料）上进行了测试。实验结果表明，JAMR 对中文的 AMR 解析效果明显差于英文，主要是在关系识别一步落后明显，而在概念识别阶段则无明显性能差异；此外，AMR 对较短的、书面化的文本，例如地理选择题文本上，通过较少的标注数据就能够得到和大数据集差不多的实验效果，说明在地理试题领域的 AMR 应用还是比较有前景的。这项工作探索了用深层语义理解方法来理解地理试题，尽管目前效果不尽如人意，但是为未来进一步的工作打下了基础。

3. 设计开发了一个地理试题标注系统，通过 B/S 架构可以使标注团队在同一份数据上进行合作标注，提供了包括各项基本自然语言处理任务、地理试题领域相关任务的标注，这其中就有另两项工作需要的拆分数据、AMR 数据的标注。对地理领域标注数据的需求催生了本系统的出现，这个标注系统也让标注人员可以更加高效、高质量地完成数据标注，为整个高考地理试题自动问答系统的多项任务提供了数据基础，促进了对地理试题理解的各项工作的开展。

6.2 未来工作

现有工作中，仍然存在一些不足之处，未来工作可以基于这些点进行更深入的研究：

1. 在地理试题拆分任务上，目前的数据量还比较少，仅有 500 多条选项含逗号的标注数据，未来可以标注更多的数据，减小数据过拟合的影响；还可以尝试更加有效的分类模型、寻找更好的特征等等，来提高分类精度。
2. 在 AMR 方面，还有许多值得研究之处，例如如何更好利用人工对齐的数据，结合对新增概念的自动对齐方法，使人工对齐的数据真正发挥作用；对 JAMR 的概念识别阶段，通过一些方法对训练语料中未出现的词，根据词性等信息，规则地或用模型学习出他们可以对应的概念。
3. 针对地理试题的特点，制订一套 AMR 概念、关系体系，对试题理解需要重点关注的信息增加学习权重，选择性地忽略一些短语内部关系，即考虑加大 AMR 概念的粒度，减小 AMR 图的复杂度，提高关键概念和关系的识别性能。

参考文献

[1] FERRUCCI D, BROWN E, CHU-CARROLL J, et al. Building Watson : An Overview of the DeepQA Project[J]. AI magazine, 2010 : 59 – 79.

[2] JIN M, KIM M-Y, KIM D, et al. Segmentation of Chinese Long Sentences Using Commas[J]. ACL SIGHAN Workshop 2004, 2004 : 1 – 8.

[3] LI X, ZONG C, HU R. A hierarchical parsing approach with punctuation processing for long Chinese sentences[J/OL]. Proceedings of the Second International Joint Conference on Natural Language Processing, 2004 : 7 – 12. http://acl.ldc.upenn.edu/I/I05/I05-2002.pdf.

[4] XU S, KONG F, LI P, et al. A Chinese sentence segmentation approach based on comma[J/OL]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, 7717 LNAI : 809 – 817. http://dx.doi.org/10.1007/978-3-642-36337-5_82.

[5] XUE N, YANG Y. Chinese sentence segmentation as comma classification[J/OL]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011 : 631 – 635. http://www.aclweb.org/anthology/P11-2111.

[6] WANG J, ZHU Y, JIN Y. A rule-based method for Chinese punctuations processing in sentences segmentation[J/OL]. Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014, 2014 : 195 – 198. http://dx.doi.org/10.1109/IALP.2014.6973504.

[7] KONG F, ZHOU G. Chinese Comma Disambiguation on K-best Parse Trees[J]. Natural Language Processing and Chinese Computing, 2014 : 13 – 22.

[8] LI J, ZHOU G, ZHU Q, et al. Syntactic Parsing with Hierarchical Modeling[J]. Asia Information Retrieval Symposium, 2008(Ratnaparkhi 1999) : 561 – 566.

[9] 李艳翠. 汉语篇章结构表示体系及资源构建研究 [D]. [S.l.] : 苏州: 苏州大学博士论文, 2015.

[10] YANG Y, XUE N. Chinese comma disambiguation for discourse analysis[C] // Meeting of the Association for Computational Linguistics: Long Papers. 2012 : 786 – 794.

[11] BANARESCU L, BONIAL C, CAI S, et al. Abstract Meaning Representation for Sembanking[C] // Linguistic Annotation Workshop and Interoperability with Discourse. 2013 : 178 – 186.

[12] BANARESCU L, BONIAL C, CAI S, et al. Abstract meaning representation (AMR) 1.0 specification[C] // Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL. 2012 : 1533 – 1544.

[13] LI B, WEN Y, WEIGUANG Q U, et al. Annotating the Little Prince with Chinese AMRs[C] // Linguistic Annotation Workshop Held in Conjunction with ACL. 2016 : 7 – 15.

[14] FLANIGAN J, THOMSON S, CARBONELL J, et al. A Discriminative Graph-Based Parser for the Abstract Meaning Representation[J]. Acl, 2014 : 1426 – 1436.

[15] ANGELI G, MANNING C D. Robust Subgraph Generation Improves Abstract Meaning Representation Parsing[J], 2014.

[16] WANG C, XUE N, PRADHAN S. A Transition-based Algorithm for AMR Parsing[J]. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015 : 366 – 375.

[17] PUST M, HERMJAKOB U, KNIGHT K, et al. Using Syntax-Based Machine Translation to Parse English into Abstract Meaning Representation[J/OL]. Arxiv, 2015(September) : 1143 – 1154.
http://arxiv.org/abs/1504.06665.

[18] VANDERWENDE L, MENEZES A, QUIRK C. An AMR parser for English , French , German , Spanish and Japanese and a new AMR-annotated corpus[J]. Naacl2015, 2015 : 26 – 30.

[19] KAI X L T H N, GRISHMAN C R. Improving event detection with abstract meaning representation[J]. ACL-IJCNLP 2015, 2015 : 11.

[20] PAN X, CASSIDY T, HERMJAKOB U, et al. Unsupervised Entity Linking with Abstract Meaning Representation[J]. Naacl2015, 2015 : 1130 – 1139.

[21] 周文翠, 袁春风. 并列复句的自动识别初探 [J]. 计算机应用研究, 2008, 25(3) : 764 – 766.

[22] A. Berger and S. D. Pietra and V. D. Pietra. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1) : 39 – 71.

[23] RATNAPARKHI A. A Maximum Entropy Model for Part-Of-Speech Tagging[C] // . 1996 : 133 – 142.

[24] LAU R, ROSENFELD R, ROUKOS S. Adaptive language modeling using the maximum entropy principle[C] // The Workshop on Human Language Technology. 1993 : 108 – 113.

[25] NIGAM K. Using maximum entropy for text classification[C] // IJCAI-99 Workshop on Machine Learning for Information filtering. 1999 : 61 – 67.

[26] 李荣陆, 王建会, 陈晓云, et al. 使用最大熵模型进行中文文本分类 [J]. 计算机研究与发展, 2005, 42(1) : 94 – 101.

[27] 李素建, 刘群, 杨志峰. 基于最大熵模型的组块分析 [J]. 计算机学报, 2003, 26(12) : 1722 – 1727.

[28] DARROCH J N, RATCLIFF D. Generalized Iterative Scaling for Log-Linear Models[J]. Annals of Mathematical Statistics, 1972, 43(5) : 1470 – 1480.

[29] POURDAMGHANI N, GAO Y, HERMJAKOB U, et al. Aligning English Strings with Abstract Meaning Representation Graphs[J]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, 1 : 425 – 429.

[30] CAI S, KNIGHT K. Smatch: an Evaluation Metric for Semantic Feature Structures[C] // Meeting of the Association for Computational Linguistics. 2013 : 748 – 752.

[31] LEVY R, MANNING C. Is it harder to parse Chinese, or the Chinese Treebank?[C] // IN PROCEEDINGS OF THE 41ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. 2003 : 439 – 446.

[32] DURRETT G, DAN K. Neural CRF Parsing[J]. Computer Science, 2015.

[33] ZHANG Y, CLARK S. Syntactic processing using the generalized perceptron and beam search[J]. Computational Linguistics, 2011, 37(1) : 105 – 151.

[34] FLANIGAN J, DYER C, SMITH A N, et al. CMU at SemEval-2016 Task 8: Graph-based AMR Parsing with Infinite Ramp Loss[J/OL]. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016 : 1202 – 1206.
http://aclweb.org/anthology/S16-1186.

[35] PRADHAN S, WARD W, HACIOGLU K, et al. Shallow Semantic Parsing using Support Vector Machines[J], 2004.

[36] CARRERAS X, MÀRQUEZ L. Introduction to the CoNLL-2005 shared task: Semantic role labeling[C] // Proceedings of the Ninth Conference on Computational Natural Language Learning. 2005 : 152 – 164.

[37] FRAZIER L. On Comprehending Sentences: Syntactic Parsing Strategies[J]. Dissertations Collection for University of Connecticut, 1979.

[38] BUCHHOLZ S, MARSI E. CoNLL-X shared task on multilingual dependency parsing[C] // Tenth Conference on Computational Natural Language Learning. 2006 : 149 – 164.

致　　谢

时光飞逝，转眼间在南京大学的三年硕士生活即将结束，我也将迎来新的工作和生活。在南京大学计算机科学与技术系以及自然语言处理研究组的三年，我认识了很多优秀的师长和同学，他们给我的工作和生活提供了许多直接帮助，让我受益匪浅，也通过他们自身的勤勉和严谨，给我留下了深刻的印象并树立了良好的榜样，让我间接地感受到了充满活力的学习氛围。

首先要感谢实验室的陈家俊老师，三年前接收我进入自然语言处理实验室，让我有机会接触到这么多优秀的同学。并且陈老师和蔼可亲、治学严谨，给我在为人处世和学习工作上都带来了深刻的影响。

其次，要感谢我的导师戴新宇老师，戴老师在我的三年研究生期间，一直指导我的工作，给予了我很大的帮助，在生活上也给予了很多关心。此次硕士毕业论文，也是在戴老师的指导下完成的。戴老师提出了很多宝贵的意见，使我能够顺利完成论文的写作。十分感谢戴老师的付出，让我能够走进自然语言处理的世界，使我受益良多。

同时，也要感谢黄书剑，尹存燕，沈思，李斌，张建兵等实验室的老师们。他们在我的工作中给予了很大的帮助和有价值的建议，使我能够在学习工作中走得更好。特别是感谢李斌老师对本文工作的支持，为我们的实验提供了宝贵的数据，以及在实验过程中提出了很多宝贵的建议。此外，还感谢程川，黄家君，胡光能，牛力强，周逸初，程善伯等几位师兄，给我在学习和工作上提供了很多帮助和建议，让我感受到了实验室作为一个集体，大家相互关心和帮助的氛围，希望你们在今后的人生中越走越好！还要感谢同一级的尚迪，郁振庭，周启元，季红洁，李小婉，王韶杰几位同学，为我提供了很多工作上的帮助。另外，还特别感谢李泽宇、娄超两位师弟，在高考地理试题解答项目中，我们共同合作，互相讨论学习，两位师弟工作认真，十分感谢他们的付出。

最后，我要感谢我的父母，他们给予了对我生活、学习、人生规划上无条件的支持和理解，让我能够选择我喜欢的工作和生活方式，并且在我人生的一路上给予鼓励和关心，让我感觉十分幸运。希望你们可以永远开心、幸福、健康。

附录

攻读硕士学位期间完成的学术成果

1. CHENG S, HUANG S, CHEN H, et al. PRIMT: A Pick-Revise Framework for Interactive Machine Translation[C]//The 15th Annual Conference of the North American Chapter of Association for Computational Linguistics: Human Language Technologies. 2016.

攻读硕士学位期间申请的专利

1. 黄书剑, 程善伯, 戴新宇, 陈家骏, 张建兵. 一种计算机中限定翻译片段的交互式翻译方法. 国家发明专利 (已公开). 申请/专利号: 201510330285.X.

攻读硕士学位期间获得的奖项

1. 2015 –荣获南京大学二〇一五年“计算机科学与技术系研究生优秀奖学金”
2. 2015 –荣获南京大学二〇一五年“计算机科学与技术系优秀研究生”

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

_____年_____月_____日

论文题名	高考地理问答系统中的句子理解研究				
研究生学号	MF1433042	所在院系	计算机科学与技术系	学位年度	2016
论文级别	☐ 硕士　☑ 硕士专业学位 ☐ 博士　☐ 博士专业学位　（请在方框内画勾）				
作者 Email	tanglr@nlp.nju.edu.cn				
导师姓名	戴新宇 副教授				

论文涉密情况：

☑ 不保密

☐ 保密，保密期：_____年_____月_____日　至　_____年_____月_____日

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。