

硕士学位论文

面向问答的问句关键词提取技术研究

RESEARCH ON QUESTION KEYWORDS EXTRACTION TECHNIQUES FOR QUESTION ANSWERING

王煦祥

哈尔滨工业大学

2016 年 6 月

国内图书分类号：TP391.2
国际图书分类号：681.37

学校代码：10213
密级：公开

工程硕士学位论文

面向问答的问句关键词提取技术研究

硕 士 研 究 生：王煦祥

导 师：张宇 教授

申 请 学 位：工程硕士

学 科：计算机科学与技术

所 在 单 位：计算机科学与技术学院

答 辩 日 期：2016 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.2

U.D.C: 681.37

Dissertation for the Master Degree in Engineering

**RESEARCH ON QUESTION KEYWORDS
EXTRACTION TECHNIQUES FOR QUESTION
ANSWERING**

Candidate:	Wang Xuxiang
Supervisor:	Prof. Zhang Yu
Academic Degree Applied for:	Master of Engineering
Speciality:	Computer Science and Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June, 2016
Degree-Conferring-Institution:	Harbin Institute of Technology

摘 要

问答系统是目前自然语言处理领域中的研究热点之一，它以精准的答案直接回答用户以自然语言方式表达的问题。在问题分析时，提取问题中的关键词对于理解其语义至关重要；在问题检索时，关键词的提取的效果直接影响到信息检索的结果和答案的相似度计算与排序。因此，关键词提取是问答系统的基础，面向问答的问句关键词提取技术研究对提升问答系统的性能有着积极作用，能够为问答系统带来更好的用户体验。

本文重点研究了两类问句关键词提取技术：无监督的关键词提取方法和有监督的关键词提取方法。有监督的关键词提取方法又分为：基于特征选择的机器学习方法和自动学习特征的深度学习方法。

基于图模型的关键词提取算法发展较为迅速。本文提出了基于依存分析排序的无监督方法提取关键词，引入词向量，从语义的角度衡量词语的相似度，引入依存句法分析，从句法结构的角度来表示两个词语之间的关联度，利用基于图的排序算法，更加准确地对候选词语进行排序，提高关键词抽取的效果。

基于特征选择的机器学习方法提取关键词，将依存句法特征应用到关键词提取技术中，通过特征分析，选取最有效的特征，利用最大熵模型训练分类器，来判断候选词是否为关键词。实验结果表明，依存句法特征有助于提高关键词提取的效果。

自动学习特征的深度学习方法提取关键词，能够让机器自动学习关键词的特征，并将特征学习融入到模型建立的过程中，避免了特征工程。在我们的研究中，利用 LSTM 模型构建神经网络层次，将目标词语的上下文信息都输入到模型中，更好地利用了词语的语义信息。同时，为了解决人工标注训练数据不足，无法满足模型训练需求的问题，我们提出了两段式的训练方法。实验证明了深度学习的关键词提取方法的有效性。

关键词： 问答；关键词提取；依存句法分析；机器学习；深度学习

Abstract

Question answering system is currently a hot point in natural language processing research. It provides precise answer directly to the users who ask queries in natural language form. For question analysis, keywords help understand the semantic for given question; for information retrieval, the results of keywords extraction can influence retrieval results, answers similarity calculation and ranking. In sum, keywords extraction is the foundation of the question answering system. And research on techniques of question keywords extraction for question answering can improve the performance of the question answering system and user experience.

In this thesis, we focus on two kinds of techniques of question keywords extraction: i.e. unsupervised approach and supervised approach. For supervised approaches, we mainly research on two techniques: machine learning-based approach and deep learning-based approach.

In recent years, there have been lots of work on graph-based keywords extraction. In this thesis, we propose an unsupervised dependency-based ranking approach for question keywords extraction. We utilize word embeddings to better evaluate similarities between two words at the semantic level, and employ the dependency relationship between two words to evaluate their relevance at the syntactic level. A graph-based ranking model is utilized to more precisely rank candidate keywords, thus it improves the performance of question keywords extraction.

For machine learning-based approach, we integrate dependency features into our model. By feature analysis, we choose the most effective features. A max entropy machine learning algorithm is employed to train the classifier that classifies whether the given word is a keyword. Experimental results show that dependency features can help improve the performance of question keywords extraction.

By utilizing the deep learning technique to extract keywords, we can jointly integrate the feature selection and model building processes together, to automatically learn effective features for question keywords extraction and avoiding the feature engineering. In this thesis, to better utilize the semantic information for contextual words, a LSTM architecture is employed to build the neural network. Meanwhile, we propose a two phases training method to solve the problem of the lack of enough manually annotated training corpus. Experimental results evaluate the effectiveness of the deep learning-based approach.

Keywords: question answering, keywords extraction, dependency parsing, machine learning, deep learning

目录

摘 要	I
ABSTRACT.....	II
第 1 章 绪 论	1
1.1 课题背景及研究目的和意义	1
1.1.1 课题背景	1
1.1.2 课题研究的目的是和意义.....	1
1.2 国内外研究现状	2
1.2.1 国外研究现状.....	2
1.2.2 国内研究现状.....	3
1.2.3 国内外研究现状简析.....	4
1.3 本文研究内容及章节安排	5
1.3.1 本文研究内容.....	5
1.3.2 本文章节安排.....	6
第 2 章 基于依存分析排序的无监督方法提取关键词	8
2.1 引言	8
2.2 应用词向量的通用关键词提取方法	8
2.2.1 TextRank 算法	8
2.2.2 词向量	10
2.2.3 基于词引力值排序的关键词提取方法.....	12
2.3 基于依存分析排序的关键词提取方法	14
2.4 语料库的建设	16
2.4.1 语料收集	17
2.4.2 标注规范	17
2.5 实验与分析	18
2.5.1 实验方法与数据.....	18
2.5.2 实验评价指标.....	19
2.5.3 实验结果及分析.....	19
2.6 本章小结	20
第 3 章 基于特征选择的机器学习方法提取关键词	21
3.1 引言	21

3.2 MAUI 系统介绍	21
3.3 基于最大熵模型的关键词提取方法	22
3.3.1 最大熵模型介绍	22
3.3.2 提取关键词	23
3.4 实验结果与分析	26
3.4.1 实验设置与结果	26
3.4.2 实验分析	26
3.5 本章小结	28
第 4 章 自动学习特征的深度学习方法提取关键词	29
4.1 引言	29
4.2 LSTM 网络介绍	29
4.2.1 RNN	29
4.2.2 LSTM	32
4.3 基于 LSTM 模型的关键词提取方法	34
4.3.1 LSTM	34
4.3.2 以目标词为中心的 LSTM	35
4.4 两段式训练方法	37
4.4.1 生成训练数据	37
4.4.2 两段式训练	38
4.5 实验结果与分析	39
4.5.1 实验设置	39
4.5.2 实验结果与分析	39
4.6 本章小结	41
结 论	42
参考文献	43
哈尔滨工业大学学位论文原创性声明和使用权限	46
致 谢	47

第1章 绪 论

1.1 课题背景及研究目的和意义

1.1.1 课题背景

随着互联网的快速发展，网上信息越来越多。面对如此庞复的信息，搜索引擎很大程度上方便了用户对信息的检索与查询。但是随着网络信息的日渐增多，用户很难迅速从搜索引擎返回的大量信息中找到所需内容。因此，人们对网络信息的检索提出了更高的要求，用户更希望搜索引擎能够提供更为快速、准确且详尽的所需信息。自动问答系统（Question Answering System）正是为了满足人们的这种需求而发展起来的。它允许人们用自然语言的方式进行提问，将用户所需的答案直接返回而不是相关的网页，具有快捷、方便、高效等特点。

问答系统是目前自然语言处理领域中的研究热点之一，它以精准的答案直接回答用户以自然语言方式表达的问题。问答系统相较于目前广泛应用的搜索引擎，具有许多优势：搜索引擎只对用户输入的查询返回相关的列表，用户需要翻看列表进行筛选才有可能得到所需的信息，而问答系统可以直接精准地回答用户提出的问题，更加简洁；搜索引擎擅长处理用户以关键词形式呈现的查询，用户为了获取信息经常需要花费精力构造关键词，而问答系统处理的是用户以自然语言形式呈现的查询，更加贴近用户在实际生活中与人交流的方式，显得自然贴切。

问答系统一般包括三个主要部分：问题分析、信息检索以及答案抽取^[1]。其中，问题分析通常包括问题分类、关键词提取和关键词扩展。问句中关键词的提取结果主要用于信息检索中获得与查询问题相关的文档，因此关键词提取的效果将直接影响后续的信息检索的处理效果。对于用户输入的问题，我们需要提取出对后续信息检索有帮助的关键词，并不是问题中的每个词都能够提取出来作为检索系统的关键词。

1.1.2 课题研究的目的和意义

信息的表达方式随着信息时代的发展而日益多样，其中利用文本来表达信息的方式又是不可替代的。随着网络的发展，线上文本信息的数量程爆炸式增长，手工获取所需文本信息的难度日益增大，因此如何高效地获取信息成为一个十分重要的课题。

为了能够有效地处理海量的文本数据，研究人员在文本分类、文本聚类、自动文摘和信息检索等方向进行了大量的研究，而这些研究都涉及到一个关键而又基础的问题，即如何获取文本中的关键词。因此，在自然语言处理和信息检索等任务中，关键词提取技术已逐渐成为热点研究问题。现有的研究成果中，关键词提取技术已被广泛应用于新闻服务、查询服务等领域，并被证明能够在信息检索、自动摘要、文本分类等任务中发挥重要作用。与此同时，海量信息处理也对关键词提取技术提出了新的挑战。

关键词是对文本主题信息的精炼，高度概括了文本的主要内容，能帮助用户快速理解文本的主旨，易于使用户判断出文本是否是自己所需的内容，从而提高信息访问和信息搜索的效率。不仅如此，由于关键词精炼、简洁的特点，可以利用关键词以较低的复杂度进行文本相关性的计算，从而高效地进行文本分类、文本聚类和信息检索等处理。在这些应用中，使用最广泛的是信息检索，用户在搜索引擎或问答系统中输入关键词，系统将出现这些关键词的文本或问题答案返回给用户。

在查询问句中，关键词代表了用户问句的主体含义。在问题分析时，提取问题中的关键词对于理解问题的语义至关重要。在信息检索中，需要从用户输入的问句中提取出对检索有用的关键词，关键词的提取的效果直接影响到信息检索的结果和答案的相似度计算与排序。因此，关键词提取是问答系统的基础，如何快速准确地从问句中提取关键词对于提升问答系统的性能至关重要。

综上所述，关键词能够帮助人们高效便捷地管理和检索资源，关键词提取技术是信息时代人们在海量数据中遨游的重要依赖。问句中的关键词提取对于提高检索和问答系统的效果起着基础性的作用，因此问句关键词提取算法的研究有着很高的理论价值和实用价值。

1.2 国内外研究现状

1.2.1 国外研究现状

对于关键词自动提取工作，国外诸多研究者经过不断地探索，提出了许多有价值的研究成果。早在 1958 年，Luhn^[2] 就对自动标引进行了研究，首次将计算机引入到文本提取中，一直到现今，关键词提取研究经历了半个多世纪的发展历程。

基于统计的关键词提取方法得到广泛发展，包括词频，TF-IDF 和共现频率等统计信息。El-Beltagy^[3]提出的 KP-Miner 算法，首先选取被标点和停用词分隔

的词序列作为候选，通过频度以及规则对候选词进行过滤，然后计算候选短语的总体权重，包括 TF-IDF 以及位置和短语长度这两个辅助特征，并对权重值进行排名，来提取关键词和关键短语。

一些较常用的机器学习方法，也逐渐应用到关键词提取领域中。基于机器学习的关键词提取方法，首先需要选取候选词的特征，然后根据提取出的特征使用机器学习算法进行学习。一个实例的特征通常分为两类：数据集内部特征和基于外部资源特征。数据集内部特征是从训练集中计算得出的，如 TF-IDF、共现频率、词语第一次出现的位置（用文档词语个数进行归一化）、在训练集中被标记为关键词的次数等。基于外部资源特征是从除训练集之外的文档中计算得出，如基于维基百科的关键短语^[4]、是否是搜索引擎的搜索日志^[5]、关键词之间的语义相似度，PMI（Web）^[6]等。通过对有效的特征进行选择，并进行特征分析，来不断提升关键词提取的效果。Turney^[7]提出了基于 C4.5 决策树算法的 GenEx 关键词自动提取机，首次将关键词自动提取任务看作为有监督的机器学习问题，其操作步骤是根据训练集计算最优参数集，此过程引用了稳态基因算法，最后根据训练得到的参数自动进行关键词抽取，同时研发的基于 GenEx 的 Extractor 系统已取得专利，在商业上成功推广。Witten^[8]等人提出的 KEA 算法，把关键词自动提取作为一个有监督的学习过程，其主要步骤是首先对文本进行预处理，生成关键词候选词，然后根据候选词在文中首次出现的位置与 TF-IDF 值计算每个候选词的特征值，利用贝叶斯模型对特征值进行训练，最后根据训练得到的预测模型对文档进行关键词提取。

基于语义的研究工作也较多。Li^[9]等通过计算词语的语义相似度来构建词汇链，利用词汇链的长度作为特征来构造关键词提取模型。

近年来，基于图模型的关键词提取方法发展较为迅速。Mihalcea 和 Tarau 等人^[10]受 PageRank 的启发，提出 TextRank 算法，依据词与词之间的语义关系构建链接，利用 TextRank 算法迭代，计算得出关键词。Gollapalli^[11]等提出了基于 CiteTextRank 算法的关键词提取方法，通过构建词语网络，并结合引用网络中被引用句的上下文内容来提高关键词提取的准确率。

1.2.2 国内研究现状

在国内，中文文本的关键词自动提取也有不少的研究。由于中文文本没有明显的词边界，一定程度上影响关键词抽取的质量，增加了中文关键词提取的难度。简立峰^[12]提出了基于字的中文关键词提取算法，在字的基础上构建 PAT 树而获取词串，同时利用互信息进行短语识别，从而进行关键词的提取，有效

地回避了分词的过程。李素建^[13]等人提出基于最大熵模型的关键词的自动提取方法，将每一个连续出现的字串看作一个概率事件。王昊^[14]等将关键词提取问题转化为基于字的序列标注问题，采用条件随机场（Conditional Random Field, CRF）的方法进行关键词自动标注，在不需要中文分词的情况下基本解决中文关键词提取问题。

在基于语义的关键词提取方法中，李纲^[15]等提出利用基于《知网》的词语语义相关度算法对词汇链的构建算法进行了改进，然后结合词频与位置特征进行关键词提取。王立霞^[16]等提出利用《同义词词林》计算词语的语义相似度，构建词语的语义相似度网络，利用居间度密度来衡量词语的重要程度，将词语语义特征融入关键词提取过程中，取得了较好的效果。

基于图模型的关键词提取方法中，张敏^[17]等提出了基于 KeyGraph 算法的关键词提取方法，通过构建词语网络并利用网络节点中心度理论来提高关键词提取的准确率，可以提取出相对低频却具有重要意义的关键词。刘通^[18]提出基于复杂网络的文本关键词提取算法，构造基于词语共现关系的复杂网络，综合考虑词频和相邻节点的贡献度，来衡量词语的重要程度。

1.2.3 国内外研究现状简析

从关键词提取任务诞生到至今的 50 多年来，国内外诸多学者已经进行了许多颇有价值和成效的研究，取得了很多富有实践意义的成果。

基于统计的关键词提取方法得到广泛发展与应用，主要包括词频、逆文本频率和共现频率等统计信息。一些较为常用的机器学习方法，如支持向量机、最大熵模型、条件随机场和遗传算法等也逐渐应用到关键词提取任务中。关键词提取的研究和语言学也有着紧密的联系，因此基于语义的研究工作也较多，其中包括词性分析、语法分析、句法分析和语义依存分析等。近年来，基于图模型的关键词提取方法发展较为迅速，这类方法一般将候选关键词抽象成图的顶点，再根据统计信息或知识信息构建网络。

虽然关键词提取技术被国内外学者广泛研究，但目前国际上最好的方法^[19]的性能与其它核心自然语言处理任务相比，仍存在差距。关键词提取任务存在的难点主要是“关键”的度量和“词”的选取。对应于文本关键词中的“关键”是文本中最重要、最能代表文本含义，起决定作用的意思。如何度量文本中词语的关键性是关键词提取任务中一个至关重要的问题。虽然关键词能够在一定程度上概括文本的主要内容，但是绝大多数关键词都取自文本词语级别，并不能直接将关键性、有意义的短语提取出来。而短语有许多词语没有的优点，它

比词语概括能力强，所蕴含的信息更加丰富、完整。另一方面，语言的差异也为关键词提取任务带来不同的问题。不同于英语，汉语文本没有明显的词边界，词语之间无天然分隔符，关键词抽取技术很多都先依赖词典分词，因此所有词语级别的关键词都是词典词，这就可能造成无法发现一些未收录在词典中的词语——未登录词。因此通常意义上的关键词实际上包含关键短语和未登录词，而这部分关键词的提取是十分困难的。

1.3 本文研究内容及章节安排

1.3.1 本文研究内容

在问句关键词提取技术的研究工作中，主要包含以下两个方面的研究内容：无监督的关键词提取方法研究、有监督的关键词提取方法研究。图 1-1 为本文所研究的关键词提取技术特点及其之间的关系。

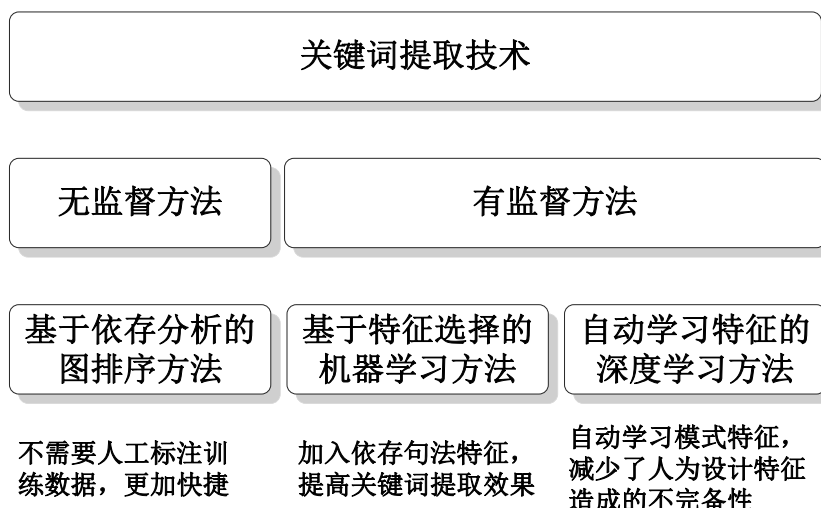


图 1-1 关键词提取技术特点及其之间的关系

信息爆炸的网络时代，标注训练集合非常耗时耗力，无监督方法不需要人工标注训练集合的过程，因此更加快捷。无监督的方法提取关键词，利用候选关键词的性质如统计性质，对它们排序，选取得分最高的若干个作为关键词。我们对基于图模型的关键词提取方法进行研究，在基于词引力值排序的关键词提取方法基础上进行改进，加入了依存句法信息，不仅从语义的角度来衡量词语之间的相似度，还从句法结构的角度来衡量词语之间的关联度，从而提高关键词提取的效果。基于依存分析排序的关键词提取方法，利用词语之间的关联度，构建无向有权图，利用经典的 TextRank 算法进行图排序，选取得分最高的若干个词语作为关键词。此外，为了更好地获取词级别的语义信息，我们还将

引入词向量^[20]来对词语进行表示。

虽然无监督的方法不需要人工标注的训练集，但效果远远达不到有监督的方法。有监督的方法提取关键词，将关键词抽取问题转换为判断每个候选关键词是否为关键词的二分类问题，它需要一个已经标注关键词的文档集合训练分类模型。任务的目标是使用人工标注的关键词训练出一个分类器，来判断候选词是否是关键词。有监督的关键词提取方法又分为：基于特征选择的机器学习方法和深度学习的方法。

基于特征选择的机器学习方法来提取关键词，首先需要选取候选词的特征，然后根据提取出的特征使用机器学习算法进行学习。通过对有效的特征进行选择，并进行特征分析，来不断提升关键词提取的效果。基于最大熵模型的关键词提取方法，使用最大熵模型训练分类器，来判断候选词是否为问句关键词，同时将依存句法特征应用到关键词提取技术中，提高关键词识别的效果。

深度学习^[21]提出了一种让计算机自动学习模式特征的方法，并将特征学习融入到了建立模型的过程中，从而减少了人为设计特征造成的不完备性。而目前以深度学习为核心的某些机器学习应用，在满足特定条件的应用场景下，已经达到了超越现有算法的识别或分类性能。深度学习的方法来提取关键词，能够让机器自动学习关键词的特征，而免去人工选取特征的过程。因此我们的研究中，在传统的基于特征选择的机器学习算法的基础上，引入深度学习的方法。

1.3.2 本文章节安排

本文主要从基于依存分析排序的无监督关键词提取方法、基于特征选择的机器学习方法和自动学习特征的深度学习方法三个方面来阐述我们的问句关键词提取技术的研究工作。

本文的内容结构安排如下：

第 1 章为绪论，从课题的背景和研究意义出发，主要介绍了本文的研究目的，以及国内外关键词提取技术的研究现状，并提出了本文的主要研究内容。

第 2 章介绍基于依存分析排序的无监督关键词提取方法。首先介绍了经典的 TextRank 算法和基于词引力值排序的关键词提取方法，在此基础上引入了我们的基于依存分析排序的关键词提取方法。同时介绍了词向量的生成方法并阐述了如何将词向量的词表示方法引入关键词提取技术。然后描述了问句关键词标注语料的构建工作。最后使用上述方法在人工标注的测试集上进行实验，验证了依存分析对于关键词提取的有效性。

第 3 章介绍基于特征选择的机器学习方法提取关键词。首先介绍了关键词

提取技术研究的基准系统——MAUI 系统。之后介绍了基于最大熵模型的关键词提取方法，并将依存句法特征应用到关键词提取技术中。最终，我们对所使用的特征进行了特征分析实验，通过实验结果的对比，证明了依存句法特征的有效性，同时说明了我们的方法相比较基准系统而言，准确率得到了大幅度的提升。

第 4 章介绍自动学习特征的深度学习方法提取关键词。提出基于 LSTM 模型的关键词提取方法和两段式训练方法。两段式训练方法用于解决深度学习模型的训练需要大规模训练数据，而人工标注的训练数据不足的问题。通过对比实验，我们验证了基于 LSTM 模型的关键词提取方法的有效性。

第2章 基于依存分析排序的无监督方法提取关键词

2.1 引言

信息爆炸的网络时代，标注训练数据耗时耗力，无监督方法不需要人工标注训练集的过程，因此更加快捷。无监督的方法提取关键词，利用候选关键词的性质如统计性质，对它们排序，选取得分最高的若干个作为关键词。而仅仅利用候选关键词的统计性质是不够的，因此我们考虑利用词语的语义信息和句子的依存句法信息，来帮助我们进行排序。

在关键词提取的研究领域，国内外学者提出了很多基于图模型的关键词提取方法，其中包括经典的 TextRank 算法和基于词引力值排序的方法。本章将基于以上两个关键词提取方法，提出改进的方法。

本章将介绍基于依存分析排序的无监督关键词提取方法。引入词向量来从语义的角度表示词语，同时介绍词向量的生成方法并阐述如何将词向量的词表示方法引入关键词提取技术。引入依存句法分析，介绍如何从句法结构的角度来表示两个词语之间的关联度，从而更加准确地对候选词进行排序，提高关键词提取的效果。

2.2 应用词向量的通用关键词提取方法

2.2.1 TextRank 算法

TextRank 算法^[10]是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的网页排序算法 PageRank，通过把文本分割成若干组成单元（词语、句子）并建立图模型，利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词提取、文摘。和 LDA、HMM 等模型不同，TextRank 不需要事先对多篇文档进行学习训练，因其简洁有效而得到广泛应用。

TextRank 的一般模型可以表示为一个有向图 $G = (V, E)$ ，由顶点集合 V 和边集合 E 组成， E 是 $V \times V$ 的子集。对于一个给定的顶点 v_i ， $in(v_i)$ 为指向该顶点的点集合， $out(v_i)$ 为顶点 v_i 指向的点集合。顶点 v_i 的得分定义如公式(2-1)所示。

$$S(v_i) = (1-d) + d \times \sum_{j \in in(v_i)} \frac{1}{|out(v_j)|} S(v_j) \quad (2-1)$$

其中， d 为阻尼系数，取值范围为 0 到 1，代表从图中某一特定顶点指向其他任意顶点的概率，一般取值为 0.85。

使用 TextRank 算法计算图中各顶点的得分时，需要给图中的顶点指定任意的初始值，并递归计算直到收敛为止。即图中任意顶点的变化率小于给定的阈值时就可以达到收敛，一般该极限值取 0.0001。由于图的连通性，通常经过较少次数的迭代就会达到收敛。

TextRank 算法还可以应用于无向图中。在无向图中，顶点的入度与出度相等，即 $in(v_i) = out(v_i)$ 。

传统的 PageRank 算法中，边是无权重的。然而，TextRank 模型是用于自然语言文本的，两个顶点之间的权重能够表示两个顶点之间的强弱关系。因此，我们引入一个新的排序算法，在计算顶点得分的时候能够将边的权重值考虑进去。带权重值的顶点 v_i 的得分定义如公式(2-2)所示，图中任两点 v_i 、 v_j 之间边的权重为 w_{ij} 。

$$WS(v_i) = (1-d) + d \times \sum_{v_j \in in(v_i)} \frac{w_{ji}}{\sum_{v_k \in out(v_j)} w_{jk}} WS(v_j) \quad (2-2)$$

利用 TextRank 算法进行关键词抽取，是利用局部词汇之间关系(共现窗口)对候选关键词进行排序，直接从文本本身抽关键词。其主要步骤如下：

- (1) 候选关键词选取。对于每个句子 S_i ，进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的词语，如名词、动词、形容词，即 $S_i = \{t_{i,1}, t_{i,2}, \dots, t_{i,n}\}$ ，其中 $t_{i,j} \in S_i$ 是保留后的候选关键词。
- (2) 构建候选关键词图 $G = (V, E)$ 。其中 V 为顶点集，由(1)生成的候选关键词组成。然后采用共现关系 (co-occurrence) 构造任两点之间的边，两个顶点之间存在边，当且仅当它们对应的词语在长度为 K 的窗口中共现， K 表示窗口大小，即最多共现 K 个词语。
- (3) 根据公式(2-1)，迭代计算各顶点的得分，直至收敛。
- (4) 对各顶点得分进行倒序排序，从而得到得分最高的 t 个词语，作为关键词。

图 2-1 是一个在问句上构建候选关键词图的例子，其中窗口大小 K 为 5，最后选取得分最高的三个词“特产”、“上海”、“世博会”作为关键词，而人工标注的关键词为：“宁波”、“特产”、“世博会”。

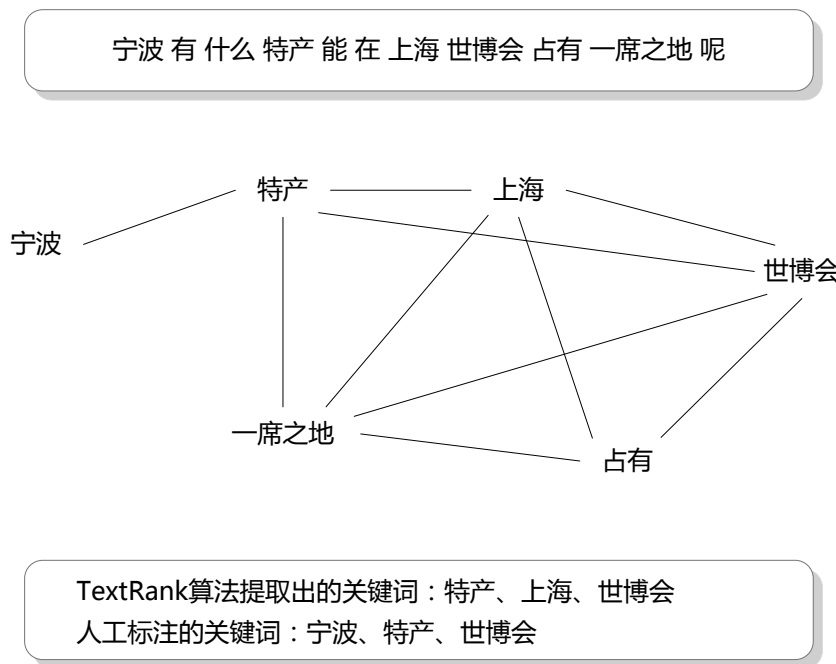


图 2-1 在问句上构建候选关键词图的例子

2.2.2 词向量

在早期基于规则和统计的自然语言处理问题中，词语被表示成符号，这些符号就是词语本身。例如，“话筒”、“麦克”等等。自然语言处理的问题要转化为机器学习的问题，通常需要首先将这些符号数学化，词向量就是用来将自然语言中的词语进行数学化的一种方式。一种最简单的词表示方法是 One-hot 表示（One-hot Representation），这种方法用一个维度很大的向量来表示一个词。向量的维度为词典的大小，向量的分量中只有一个 1，其位置对应该词在词典中的位置，其它位置都是 0。例如，“话筒”表示为 $[0,0,0,1,0,0,0,0,\dots,0]$ ，“麦克”表示为 $[0,0,0,0,0,0,0,1,\dots,0]$ 。很明显，One-hot 表示方法容易受维数灾难的困扰，尤其是将其用于深度学习的一些算法时。这种表示方法还存在一个重要的问题就是“词汇鸿沟”现象：任意两个词之间都是孤立的。仅仅从这两个向量中看不出两个词是否有关系，不能很好地刻画词与词之间的相似性，哪怕是“话筒”和“麦克”这样的同义词也不能幸免于难。

而在深度学习（Deep Learning）中，一般采用分布式表示（Distributed Representation）的方法表示词向量，它最早是 Hinton^[22]于 1986 年提出的，通常被称为“Word Representation”或“Word Embedding”。这种方法将词用一种低维实数向量表示，每一维代表词的一种特征，理论上来说，这种表示方法能够体现词的语法和语义特征。例如，“话筒”表示为 $[0.435,0.534,\dots,0.128]$ ，维度

以 50 维或 100 维较为常见。其优点在于相似的词在距离（如 cosine 相似度、欧氏距离等）上更接近，根据词与词之间的距离能判断出它们的语法、语义相似性，从而反映词之间的相关性。同时，较低的维度保证了在应用特征向量时有一个可接受的复杂度。因此，新近提出的许多语言模型，如潜在狄利克雷布（Latent Dirichlet Allocation, LDA）模型和潜在语义分析（Latent Semantic Analysis, LSA）模型，以及目前流行的神经网络模型等，都采用这种分布式表示的方法来表示词向量。

利用神经网络算法可以生成词向量。通常情况下，词向量和语言模型是捆绑在一起训练的，即训练完成后二者同时得到。利用神经网络来训练语言模型的思想最早由百度深度学习研究院的徐伟^[23]提出。Bengio^[24]于 2003 年提出用一个三层的神经网络来构建语言模型。其后有一系列的相关研究工作，其中包括谷歌 Tomas Mikolov 等^[20]提出的 CBOW 模型和 Skip-gram 模型。

在我们的研究中，选择 CBOW 模型和 Skip-gram 模型获取上下文相关的词向量。这两种方法旨在用较低的计算复杂度获得词语的分布式表示。在传统神经网络模型的基础上，采用对数线性模型结构，针对模型训练运算量过大的问题进行改进，去除了神经网络的非线性隐含层，从而降低了训练的复杂度。同时将词向量的计算与神经网络中 N-gram 模型的训练分开，提高训练效率，其模型结构如图 2-2 所示。

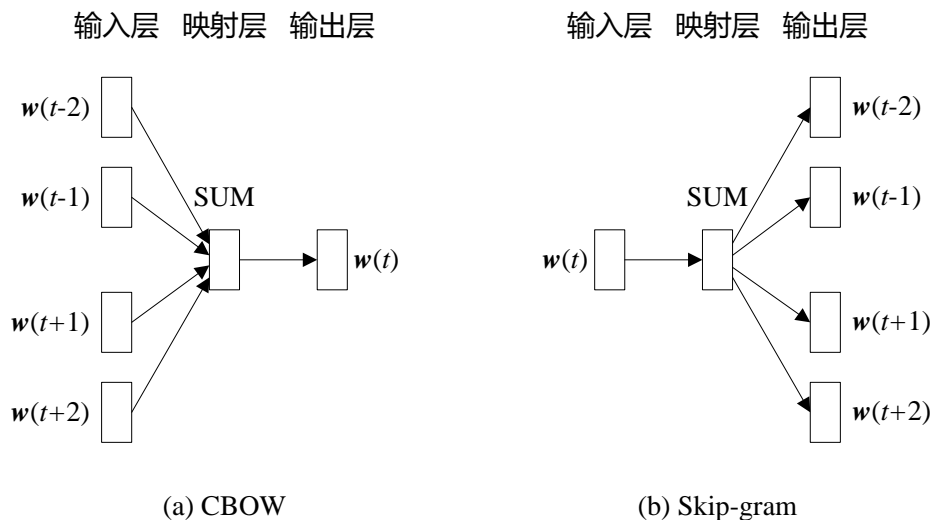


图 2-2 CBOW 和 Skip-gram 模型结构

CBOW 模型根据上下文预测目标词语 $w(t)$ 。从模型结构图中不难看出，对于 CBOW 模型，其整体结构类似普通的前馈神经网络模型，唯一的区别在于其去除了隐含层，只保留了输入层、映射层（Projection Layer）和输出层。输入层、

输出层表示每个词语的词向量，均采用分布式表示方法，维度一般为 50 维或 100 维。映射层的维度为 D ，窗口大小 C 表示上下文长度。CBOW 模型在训练时和前馈神经网络模型有如下区别：映射层不再是将输入词语的向量表示按顺序排列，而是将他们相加，采用均值表示单个词向量，达到减少计算量的目的。由于词语在历史信息中的顺序不影响其在映射层中的表示，这种结构被称为连续空间中的词袋模型。此外，由于这里的目的是寻找词语的向量表示，而不是语言模型，因此无需进行语言模型概率的计算，模型可以利用未来的信息 $w(t+1)$ 、 $w(t+2)$ 等训练当前词语 $w(t)$ ，真正实现利用上下文信息得到最优的词向量。

Skip-gram 模型的结构与 CBOW 模型相反，根据当前词语 $w(t)$ 预测上下文。

由于两种词向量模型的结构不同，它们在表达上各有优势。CBOW 模型在语法测试中准确率更高，表明其通过对上下文信息的学习，能够有效获取更多的语法信息；Skip-gram 模型在语义测试中有更好的效果，说明它产生的词向量能够更准确地从语义层面对词语进行描述，其区分性更为明显。二者的共同的优点在于，都能够从大规模语料中快速获得高质量的词向量。对大规模数据的有效利用，使模型能够产生更为精准的词向量，从而能够更好地描述不同词语之间的相关性。

2.2.3 基于词引力值排序的关键词提取方法

基于词引力值排序的关键词提取方法^[25]，提出利用文本中的词语构建无向有权图，使用 TextRank 算法对词语进行重要度排序，进行关键词提取。该方法不仅利用了文本内部的统计信息，还引入了外部语义信息词向量来度量词语之间的关系，从而提高了关键词提取的效果。基于词引力值排序的关键词提取方法是具有语料独立性的无监督方法，利用该方法在各类长度不等的公开数据集上进行实验，均能获得不错的效果。

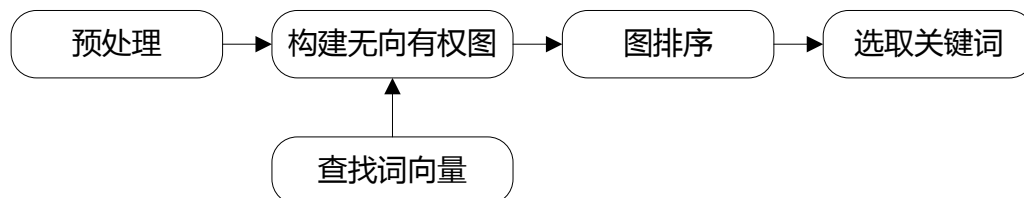


图 2-3 基于词引力值排序的问句关键词提取流程图

基于词引力值排序的问句关键词提取的主要流程如图 2-3 所示，主要包括以下几个步骤：

(1) 预处理过程。

预处理过程包括分词、去停用词，将剩下的词语作为关键词的候选词。如

果在英文语料上进行实验，不需要进行词干提取的步骤，因为具有相同词干的词语不一定表示相同的含义，并且词向量中可能不包含词语的词干形式。

(2) 构建无向有权图。

预处理之后，所有的候选词被表示为图的顶点，如果两个词语在一个句子内共现，则这两个顶点之间存在边。而边的权重值是两个词语之间的词引力值。下面将引出词引力值的概念。

我们认为，仅仅用两个词语的语义相似度是不足以衡量词语的重要程度的。直觉上，如果两个词在文本中出现的频率都特别低，那么即使这两个词语的语义相似度非常高，也并不能说明这两个词语很重要。相对而言，如果两个词语在文本中的重要程度很高，那么它们中至少有一个词语的出现频率要很高，并且两个词语相似度也很高。例如，在一篇关于旅行的文本中，如果“冲浪”和“海水”这两个词语都只出现了一次，很显然，这两个词不是这篇文本的核心词；如果“冲浪”出现了 100 次，那么即使“海水”只出现了两次，因为它们之间有着较高的相似度，因此“海水”也很可能成为关键词。受牛顿万有引力定律的启发，我们引入了一个公式来衡量两个词之间的引力。一般地，如果两个词都有很大可能性成为关键词，我们给予这两个词的引力一个较大的值。牛顿万有引力定律指出，宇宙中任意两个物体存在引力，引力大小与它们质量的乘积成正比，与它们距离的平方成反比。因此我们认为，两个词语之间也存在引力，物体的质量可以理解为词频，距离可以理解为两个词语的词向量之间的欧氏距离。

文本中词语 w_i 和 w_j 的引力 f 由公式(2-3)计算获得。其中 $freq(w)$ 是词语 w 在文本中出现的频率， d 是词语 w_i 和 w_j 的词向量之间的欧氏距离。

$$f(w_i, w_j) = \frac{freq(w_i) \times freq(w_j)}{d^2} \quad (2-3)$$

我们认为，经常以固定模式出现的多个词语是一个短语，即某个模式出现的频率越高，这个模式就越有可能是一个短语。两个样本间的相关性通常用 dice 系数来表示。将 dice 系数引入自然语言处理领域来衡量两个共同出现的词语是短语的可能性^[26]。dice 系数的定义如公式(2-4)所示。其中 $freq(w_i, w_j)$ 是词语 w_i 和 w_j 的共现频率。

$$dice(w_i, w_j) = \frac{2 \times freq(w_i, w_j)}{freq(w_i) + freq(w_j)} \quad (2-4)$$

词语间的词引力值是 dice 系数和两个词的引力的乘积，如公式(2-5)所示。其中，dice 系数用来衡量两个词语成为短语的可能性，两个词的引力用来衡量

两个词的相关度。

$$attr(w_i, w_j) = dice(w_i, w_j) \times f(w_i, w_j) \quad (2-5)$$

(3) 进行图排序。

使用 TextRank 算法进行图排序。在有权重的 TextRank 算法中，边的权重值越高，顶点的得分将会越高。在无向图 $G = (V, E)$ 中， V 是顶点的集合 E 是边的集合， $C(v_i)$ 是与顶点 v_i 有边连接的顶点集合。顶点 v_i 的得分由公式(2-6)计算得出，其中 $attr(v_i, v_j)$ 由公式(2-5)计算得出。

$$S(v_i) = (1-d) + d \times \sum_{v_j \in C(v_i)} \frac{attr(v_i, v_j)}{\sum_{v_k \in C(v_j)} attr(v_j, v_k)} S(v_j) \quad (2-6)$$

(4) 选取得分最高的 t 个词语，作为问句的关键词。

2.3 基于依存分析排序的关键词提取方法

尽管基于词引力值排序的关键词提取方法在各类公开数据集上都能取得不错的效果，但是由于问句的短文本特性，使得在计算两个词语之间的相关性时，共现频率这样的特征很难发挥作用。对于已有的基于图排序的关键词提取方法，一个主要的问题是在计算两个词语相关性的时候，忽略了词语之间的句法结构关系。为了解决这个问题，我们提出了一种新的利用词语间的依存句法关系作为线索的关键词提取方法。对于给定的问句，我们不仅利用统计信息和词向量信息，还构建依存关系图来计算词语之间的关联强度，进而根据依存关联度来构建有权图，利用 TextRank 算法迭代计算出词语的重要度得分。

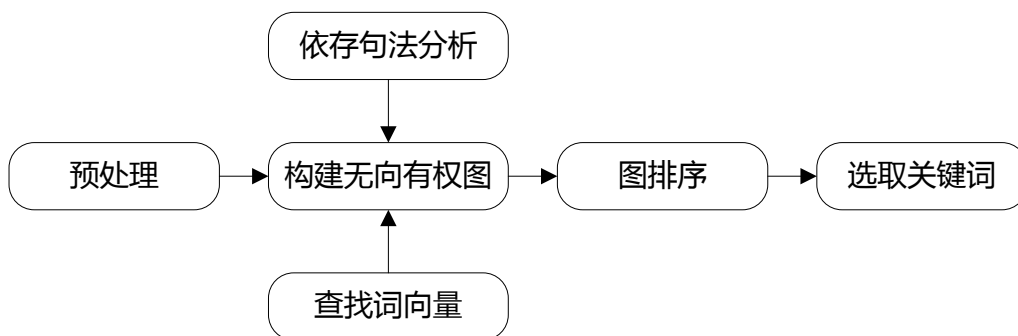


图 2-4 基于依存分析排序的问句关键词提取流程图

基于依存分析排序的问句关键词提取的主要流程如图 2-4 所示，主要步骤与基于词引力值排序的关键词提取方法基本相同，包括预处理过程、构建无向有权图、进行图的排序、选取得分最高的 t 个词语，作为问句的关键词。

(1) 预处理过程。

预处理过程包括分词、去停用词，将剩下的词语作为关键词的候选词。

(2) 构建无向有权图。

预处理之后，所有的候选词被表示为图的顶点，如果两个词语在一个句子内共现，则这两个顶点之间存在边。边的权重值我们需要利用词语之间的统计信息、词向量信息和依存句法分析信息计算求得。

现有可以用来计算两个词语之间的相关度的方法有：点互信息（Pointwise Mutual Information, PMI）、平均互信息（Average Mutual Information）等等^[27]。然而它们只考虑了词语之间的统计信息，并没有考虑句法的依存关系。词语之间的句法依存关系对于衡量词语的重要度有积极作用。“总理李克强在调研上海外高桥时提出了什么机制”这句话的依存句法分析结果如图 2-5 所示。蓝色的弧线表示两个词语之间存在句法关系，红色的标签表示两个词语之间具体的依存关系。有了依存句法分析结果，我们就可以比较容易的看出，“提出者”是“李克强”，而不是“上海”或“外高桥”，即使它们都是名词，而且距离“提出”更近。

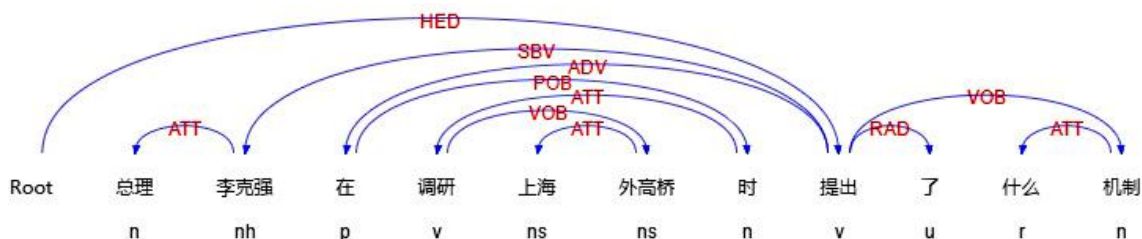


图 2-5 依存句法分析结果

依存句法分析的结果类似树结构，我们去掉它的根节点，并且忽略弧的指向，可以得到一个无向的依存关系图 $G' = (V', E')$ ， $V' = w_1, w_2, \dots, w_n$ ， $E' = e_1, e_2, \dots, e_m$ ，其中 w_i 表示词语， e_j 表示两个词语之间的无向关系。依存句法分析结果对应的无向的依存关系图如图 2-6 所示。

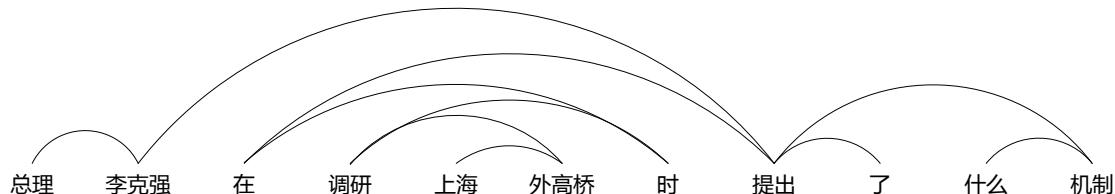


图 2-6 依存关系图

无向的依存关系图保证了问句中的任意两个词之间都有一条依存路径，而依存路径的长短反映了依存关系的强弱。图 2-7 显示了依存关系图中词语之间的依存路径的长度。很明显，从图 2-7 (a)中可以看出，“李克强”和“提出”是

直接相连的，因此它们之间的依存路径长度为 1。而从图 2-7 (c)中可以看出，“李克强”和“外交桥”相距较远，它们之间的依存路径长度为 5。这说明“李克强”与“提出”比“李克强”与“外高桥”有更紧密、更强烈的关系。

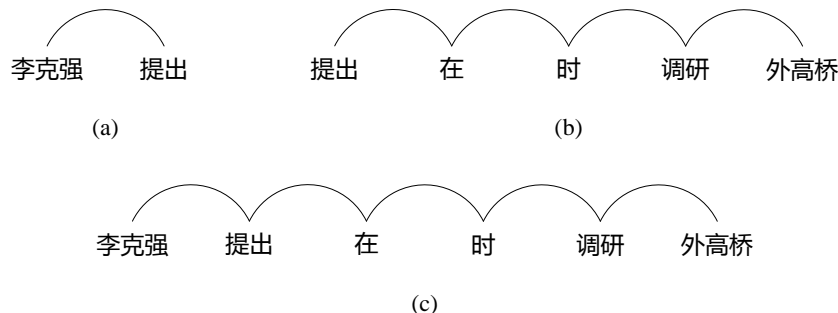


图 2-7 依存路径

因此，我们根据依存路径的长度，引入了依存关联度^[28]的概念，如公式(2-7)所示。其中， $dr_path_len(w_i, w_j)$ 表示词语 w_i 和 w_j 之间的依存路径长度。

$$Dep(w_i, w_j) = \frac{1}{b^{dr_path_len(w_i, w_j)}} \quad (2-7)$$

两个词语之间的关联度，即边的权重值是两个词的引力与依存路径长度的乘积，如公式(2-8)所示。其中，两个词的引力 $f(w_i, w_j)$ 由公式(2-3)得到。

$$weight(w_i, w_j) = Dep(w_i, w_j) \times f(w_i, w_j) \quad (2-8)$$

(3) 进行图排序。

使用有权重 TextRank 算法进行图排序。在无向图 $G = (V, E)$ 中， V 是顶点的集合， E 是边的集合， $C(v_i)$ 是与顶点 v_i 有边连接的顶点集合。顶点 v_i 的得分由公式(2-9)计算得出，其中 $weight(v_i, v_j)$ 由公式(2-8)计算得出。

$$WS(v_i) = (1 - d) + d \times \sum_{v_j \in C(v_i)} \frac{weight(v_i, v_j)}{\sum_{v_k \in C(v_j)} weight(v_j, v_k)} WS(v_j) \quad (2-9)$$

(4) 选取得分最高的 t 个词语，作为问句的关键词。

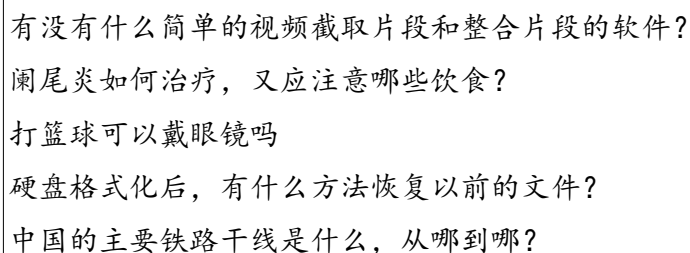
2.4 语料库的建设

在以往的研究工作中，关键词抽取的工作大多在论文、科技文摘要、新闻等语料上进行的，Liu^[29]在会议中抽取关键词的研究工作，也是通过人工标注获取一定数量的训练语料。因此，作为问句关键词提取技术研究的第一步，我们首先需要解决数据问题，应当收集相当规模的问句，并在此基础上制定详尽、规范、可执行的标注规则，以构建高质量的问句关键词抽取语料库。

2.4.1 语料收集

在语料收集方面，我们爬取了百度知道¹中的约 110 万条问句，作为我们研究的语料基础。百度知道是全球最大中文互动问答平台，用户可以根据自身的需求，有针对性地提出问题，因此问句都是真实的，具有研究意义。问句语料以纯文本保存，每一行是一条数据。

图 2-8 是我们所收集的百度知道问题中的一些数据。



有没有什么简单的视频截取片段和整合片段的软件？
阑尾炎如何治疗，又应注意哪些饮食？
打篮球可以戴眼镜吗
硬盘格式化后，有什么方法恢复以前的文件？
中国的主要铁路干线是什么，从哪到哪？

图 2-8 百度知道问题数据

2.4.2 标注规范

为了规范问句关键词的标注过程，缩小不同标注者主观意见所带来的标注差别，我们设定了详细的标注规则。

问句的关键词反映的是一个问句的话题（topic），即该问句在说些什么。在抽取或审定问句关键词时，须遵照以下原则。

- 关键词需从中文分词后的结果中产生。
- 全体关键词需覆盖问句的主要话题，以用于后续问题检索等。
- 关键词的数量不宜过多，应尽量覆盖问句所表达的主题。
- 一般来说，关键词不应是某些宽泛的、一般化的词语，如：“中国”、“平台”、“认识”等。特定场景除外，如讨论国际形势，“中国”、“美国”则可以被标注为关键词。
- 绝大部分关键词应是命名实体、专有名词或名词性词语，如：“天安门”、“马云”、“阑尾炎”、“硬盘格式化”、“铁路”等。动词酌情标注为关键词，数量尽量地少，如在某些金融经济类话题中，出现像“下滑”、“衰减”、“增长”等动词，由于它们反映了问句的话题，也可酌量收作关键词，但数量要有所控制。

一些问句的关键词标注如图 2-9 所示。

¹ <http://zhidao.baidu.com>

问句：有没有什么简单的视频截取片段和整合片段的软件？
 关键词：视频，截取，整合，软件
 问句：阑尾炎如何治疗，又应注意哪些饮食？
 关键词：阑尾炎，治疗，饮食
 问句：打篮球可以戴眼镜吗
 关键词：篮球，戴，眼镜
 问句：硬盘格式化后，有什么方法恢复以前的文件？
 关键词：硬盘格式化，恢复，文件
 问句：中国的主要铁路干线是什么，从哪到哪？
 关键词：中国，铁路，干线

图 2-9 一些问句的关键词标注

在实际标注过程中，我们共标注了 1000 个问句，并选取其中 800 个问句用作训练集，余下的 200 个问句作为测试集。

为了进一步了解问句关键词标注的情况，我们对测试集中的 200 个问句进行了关键词个数统计。在测试集的 200 个问句中，总词数共 2351 个，其中被人工标注为关键词的个数为 732 个，关键词约占总词数的 1/3。

2.5 实验与分析

2.5.1 实验方法与数据

在实验中，分词、依存句法分析由哈尔滨工业大学语言技术平台^[30]（Language Technology Platform, LTP）自动生成，两个词之间的依存关系的路径长度使用弗洛伊德算法^[31]算出。我们使用 google 的开源工具 word2vec^[32]，利用 2012 版搜狗新闻数据²（SogouCS）训练得到词向量。训练词向量时，使用 Skip-gram 模型，每个词向量的维度为 100 维。

实验中，我们使用人工标注的 200 个问句作为测试集。我们对测试集中的 200 个问句进行了关键词个数统计，其中，总词数共 2351 个，其中被人工标注为关键词的个数为 732 个，关键词约占总词数的 1/3。因此，实验的结果均是在 $t = \frac{1}{3}n$ 的时候得到的，这里 n 表示问句中的总词数， t 表示经过排序后选取得分最高的 t 个词语作为关键词。

² <http://www.sogou.com/labs/dl/cs.html>

2.5.2 实验评价指标

在问句关键词提取的实验结果中，对于提取的关键词，我们使用了准确率（Precision, P）、召回率（Recall, R）以及标准 F_1 值作为对关键词提取结果的评价指标。

关键词抽取实验的准确率、召回率以及 F_1 值的定义分别为公式(2-10)、公式(2-11)以及公式(2-12)所示。其中 $c_{correct}$ 是一个方法所有准确提取的关键词数目， $c_{extract}$ 是所有提取的关键词数目，而 $c_{standard}$ 是所有人工标注的标准答案数目。 F_1 值相对于准确率和召回率，可以更全面的衡量关键词提取的效果。

$$P = \frac{c_{correct}}{c_{extract}} \quad (2-10)$$

$$R = \frac{c_{correct}}{c_{standard}} \quad (2-11)$$

$$F_1 = \frac{2PR}{P+R} \quad (2-12)$$

实验中，采用严格的对照标准，即关键词必须完全匹配才算正确。

2.5.3 实验结果及分析

基于依存分析排序的无监督方法提取关键词实验的具体结果如表 2-1 所示。其中,TextRank 方法是 2.1.1 节中所介绍的关键词提取方法,WordAttractionRank 是 2.1.3 节中所介绍的基于词引力值排序的关键词提取方法, DependencyRank 是 2.2 节中介绍基于依存分析排序的关键词提取方法。TextRank 方法与 WordAttractionRank 方法均是基准方法, DependencyRank 是我们所提出的方法。

表 2-1 基于依存分析排序的无监督方法提取关键词实验的最佳结果

方法	P	R	F_1
TextRank	59.75%	58.61%	59.17%
WordAttractionRank	64.90%	63.66%	64.28%
DependencyRank	66.30%	65.03%	65.66%

观察表中的结果我们可以发现，基于依存分析排序的关键词提取方法在各项指标上均达到最高值，说明我们利用词语间的依存句法关系作为线索的关键词提取方法是有效的。因为我们综合利用统计信息、词向量信息和依存关系信息，不仅考虑了句子的语义信息，还考虑了句子的句法结构信息。

基于依存分析排序的关键词提取方法中，有一个可调节的参数 b ，图 2-10

显示了相同测试集，参数 b 的不同取值对关键词提取效果的影响。

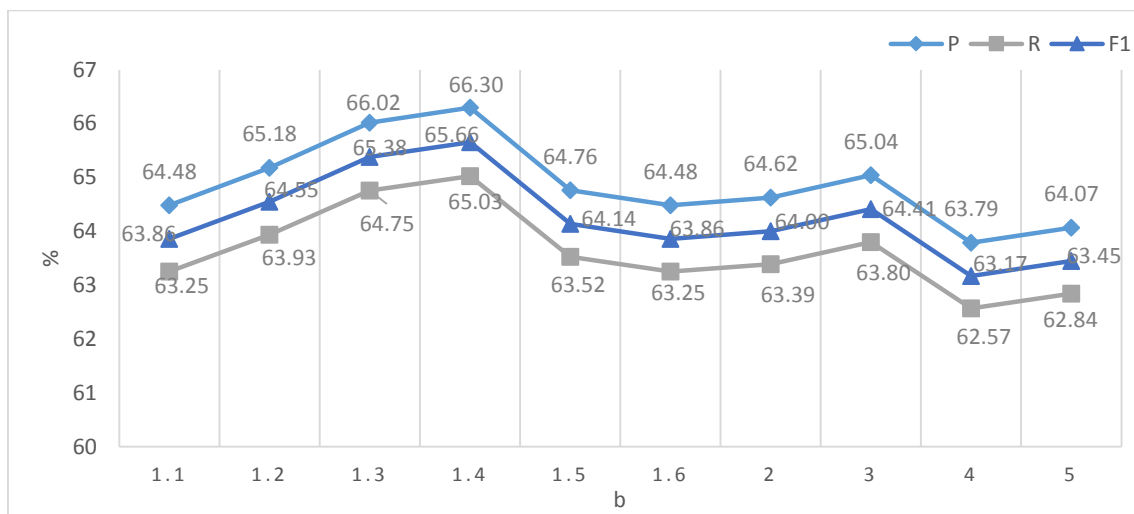


图 2-10 参数 b 的不同取值对关键词提取效果的影响

通过对比实验我们可以看出，当参数 b 的取值为 1.4 时，关键词提取的效果最好。

2.6 本章小结

本章首先介绍了以往的基于图排序的关键词提取工作，本文受 Wang^[25]和 Zhang^[28]工作的启发，提出了基于依存分析排序的无监督关键词提取方法。这个方法主要利用了词语间的依存句法关系作为线索，对于给定的问句，我们不仅利用统计信息和词向量信息，还构建依存关系图来计算词语之间的关联强度，进而根据依存关联度来构建有权图，利用 TextRank 算法迭代计算出词语的重要度得分，从而提取关键词。实验表明，利用词语间的依存句法关系作为线索的关键词提取方法是有效的，基于依存分析排序的关键词提取方法能够提高关键词提取的性能。

第3章 基于特征选择的机器学习方法提取关键词

3.1 引言

有监督学习已经被证明在很多任务上是有效的。由于在有监督学习的过程中，利用学习算法可以从训练语料中学习大量有用的信息，因此其效果往往要优于无监督算法。在我们的任务中，也同样采用了有监督学习算法对问句中的关键词进行识别。

有监督的方法中以基于特征的分类模型居多，Medelyan 等人提出的 Maui 算法，具有领域独立性和语言独立性的特点，仍是目前新闻领域的最有效的关键词提取工具之一。本章将该方法作为基准系统，在此基础上进行进一步研究。

本章将介绍基于特征选择的机器学习方法提取关键词，将关键词提取任务转换为是否为关键词的二元分类问题，利用人工标注的训练数据集，使用最大熵模型训练分类器，来判断候选词是否为问句关键词。同时还将依存句法特征应用到关键词提取技术中，提高关键词识别的效果。

3.2 MAUI 系统介绍

Maui (Multi-purpose automatic topic indexing) 是一种多功能自动主题标引算法，Medelyan 等人^[33]于 2010 年发布了 Maui 工具³。Maui 除了能进行关键词提取外，还能执行主题分配、术语提取等任务。Maui 算法在 KEA 算法基础上进行了改进，提取了新的特征并使用了更加有效的分类器。Maui 具有域独立性和语言独立性的特点，它已经成功在计算机科学、医学、物理学、生物学和生物信息学的文档上进行了实验，同时也在博客和新闻数据集上进行了实验，并取得了较好的效果。目前，在新闻领域的关键词提取中，Maui 仍然是最有效工具。因此，我们使用 Maui 作为问句关键词提取技术研究的基准系统。

Maui 算法提取关键词，首先选取一系列关键词的候选词，然后人工选取一些特征，最后在大规模人工标注的语料上训练决策树分类模型。训练完成后，Maui 为问句中的候选词打分，然后返回这些候选词的得分，得分最高的 k 个候选词作为问句的关键词。

Maui 提取的特征包括频率特征、位置特征、语义相关性特征、特异性特征（长度、词性、命名实体）等。Maui 使用决策树作为分类器，最终通过训练得

³ <https://code.google.com/archive/p/maui-indexer/>

到关键词的判别模型。

Maui 方法将作为我们关键词提取研究的基准系统。

3.3 基于最大熵模型的关键词提取方法

3.3.1 最大熵模型介绍

最大熵模型（Maximum Entropy Model, MaxEnt）是一种机器学习方法，在自然语言处理的许多任务中（如文本分类、词性标注、中文分词、浅层句法分析及句子边界识别等）都有很好的应用效果。

最大熵原理是由 Jaynes^[34]在 1957 年提出的，其主要思想是，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不做任何的主观假设。在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以我们称这种模型为“最大熵模型”。

最大熵原理是统计学习的一般原理，将它应用到分类中得到最大熵模型。假设分类模型的一个条件概率分布 $P(Y|X)$ ， X 表示输入， Y 表示输出，这个模型表示的是对于给定的输入 X ，在条件熵 $H(Y|X)$ 最大的条件下，以条件概率 $P(Y|X)$ 输出 Y 。最大熵模型的一般式如公式(3-1)所示。

$$\max_{p \in P} H(Y|X) = - \sum_{(x,y)} p(x,y) \log p(y|x) \quad (3-1)$$

为了进一步说明最大熵模型在分类中的应用，我们给出一些在分类中常用的定义，以此推出最大熵模型的约束条件和其具体表达式。

- 样本， (x, y) 。 x 表示输入，即特征信息， y 表示输出，即分类标签。
- 训练数据集， $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 。 (x_i, y_i) 为训练数据集中的样本。
- 特征函数， $f(x, y)$ 。描述输入 x 和输出 y 之间的某一个事实，其定义如公式(3-2)所示。它是一个二值函数，当 x 和 y 满足这个事实时取值为 1，否则取值为 0。

$$f(x, y) = \begin{cases} 1, & x = x_i \text{ and } y = y_i \\ 0, & \text{other} \end{cases} \quad (3-2)$$

- 样本特征函数期望值， $\tilde{E}(f)$ 。特征函数 $f(x, y)$ 关于样本分布的期望值，如公式(3-3)所示，其中 $\tilde{p}(x, y)$ 是 (x, y) 在训练数据集中出现的概率。

$$E(f) = \sum_{(x,y)} p(x,y) f(x,y) \quad (3-3)$$

- 模型特征函数期望值, $E(f)$ 。特征函数在模型中的期望值如公式(3-4)所示, 其中 $\tilde{p}(x)$ 是训练数据集中 x 出现的概率。

$$\begin{aligned} E(f) &= \sum_{(x,y)} p(x,y) f(x,y) \\ &= \sum_{(x,y)} p(x) p(y|x) f(x,y) \end{aligned} \quad (3-4)$$

如果模型能够获取训练数据集中的信息, 那么就可以假设 $E(f)$ 和 $\tilde{E}(f)$ 这两个期望相等, 即

$$E(f) = \tilde{E}(f) \quad (3-5)$$

或

$$\sum_{(x,y)} p(x) p(y|x) f(x,y) = \sum_{(x,y)} \tilde{p}(x,y) f(x,y) \quad (3-6)$$

对于给定的训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 以及其特征函数 $f_i(x, y), i = 1, 2, \dots, n$, 最大熵模型等价于约束最优化问题:

$$\begin{aligned} p^* &= \arg \max_{p \in P} H(Y|X) = - \sum_{(x,y)} p(x,y) \log p(y|x) = - \sum_{(x,y)} p(x) p(y|x) \log p(y|x) \\ \text{s.t. } &E(f_i) = \tilde{E}(f_i), i = 1, 2, \dots, n \\ &\sum_y p(y|x) = 1 \end{aligned} \quad (3-7)$$

最大熵问题的求解是将约束最优化问题转化为无约束最优化的对偶问题。对于最大熵模型对应的最优化问题, GIS^[35]、IIS^[36]等最优化算法都能解出, 这里不作详细描述。

3.3.2 提取关键词

首先我们需要对问句进行预处理, 预处理过程包括分词、词性标注、依存句法分析、去停用词。将去停用词后剩下的词语作为关键词的候选词, 词性标注、依存句法分析的结果是后续要使用的特征信息。

我们利用机器学习算法来进行关键词的提取。在这里, 关键词提取被视为一个二元分类任务。对于问句中的每个词, 我们都利用机器学习算法对其进行分类, 分为“是关键词”和“非关键词”两类。利用机器学习算法的步骤为: 首先, 我们对标注好的关键词数据进行观察, 然后根据数据的特点, 提取若干分类特征。一般来说, 我们根据数据统计分布情况尽可能提取出区分性大的特征, 认为这些特征可以对分类进行最好的指导。然后, 对于这些分类特征, 提取特征向量, 并利用机器学习算法(如支持向量机、最大熵等)对这些特征向量进行训练, 得到一个关键词判别模型, 该模型即可以对问句中的关键词进行识别。

对于测试问句中的每个候选词，提取相应的特征，将特征向量放入训练好的模型，即可得到一个词语是否是关键词的概率，我们可以利用这个概率进行分类。下面，对我们的基于最大熵模型的关键词提取方法进行详细的描述。

我们选取了机器学习常用的最大熵算法进行分类器的构建。根据观测到的数据特点，我们提取了如表 3-1 所示的，包括统计特征、词法特征、结构特征和依存句法特征共计四类、11 维特征。

表 3-1 特征表

种类	特征
统计特征	词频
	逆文本频率
词法特征	词性
	前一个词的词性
	后一个词的词性
	当前词与前一个词词性的组合
	当前词与后一个词词性的组合
结构特征	是否是命名实体
	位置信息
依存句法特征	依存关系
	依存弧所指向词的词性

在第一类特征，统计特征中，我们抽取了如下三维特征：

- 词频特征：对于问句中的每个词，我们要计算这个词在当前问句中出现的频率。我们认为，在一个问句中，重复出现的词可能是更有意义的词。因为问句中的关键词是能够表述问句需求的词语，应该会被适当地加重。如果一个词在一个问句中只出现了一次，相较于出现次数较多的词，其重要性可能会降低。
- 逆文本频率特征：该特征是预先通过大规模文本统计得到的。一个词的逆文本频率等于该词在所有文本中出现的次数除以总文本数所得商的对数。计算逆文本频率特征的意义在于，通过计算一个词的逆文本频率，对一个词的区分度进行度量。如果包含一个词的文档越少，其逆文本频率越大，说明词的区分度大。区分度大的词，往往更容易成为关键词。对于一些停用词，如“的”，虽然其词频高，但是其逆文本频率极低，接近于零。

在第二类特征，词法特征中，我们抽取了六维特征：

- 词性特征：即当前词的词性。通过观察数据，我们发现，名词（n）更有可能成为关键词，因此我们抽取了词性特征，从词法的角度对一个词的重要程度进行度量。
- 前一个词的词性特征&后一个词的词性特征：要判断一个词是否为关键词，该词上下文中的词性也同样重要。例如，如果一个词前边的词是动词，则这个词很有可能成为关键词。因此，我们抽取了当前词的前一个和后一个词的词性，根据它们的词性，对当前词是否为关键词做出判断。
- 词性组合特征：我们将一个词同其前一个或后一个词的词性组合在一起，形成这两个特征。这样做的目的在于，人工将一个词同其上下文（前后词）联系在一起。我们发现，如果连续两个词都是名词，则这两个词更有可能成为关键词。尤其是在问句中，因为句子本身不长，成分不多，所以这种组合特征显得尤为重要。
- 是否是命名实体特征：显而易见，如果一个词被判别为命名实体，那这个词在问句中很大可能扮演着重要的角色，因此我们提取了这样一维特征。在这里，我们并没有对命名实体进行区分（人、机构等），因为在数据中我们发现，一个词是否是关键词和其命名实体的种类关系不大，因此我们只对一个词是否是命名实体进行了区分。

第三类特征，结构特征中，我们对一个词的位置进行了区分。我们发现问句中句首和句尾的词一般来说不会很重要，因为句首的词很多都是问句的引出词，如“求助”，“为什么”这类，并不是关键词。但是如果句首的词是名词，是一个命名实体，则很大程度可能是关键词，因为很多时候，这种句首的词表示一个问句所提问的实体词。如“李白的出生年月日是什么”，在这里，李白就是整个问句的实体词，应该被识别为关键词。

第四类特征是依存句法特征，包括词语的依存关系和依存弧所指向词的词性特征。这一类特征是根据问句关键词的特点，由我们提出的一类特征。依存语法（Dependency Parsing, DP）通过分析语言单位内成分之间的依存关系揭示其句法结构。通过依存句法分析，我们可以识别出句子的核心谓词、主语和宾语。通常情况下，名词性宾语有很大可能性被标注为关键词。例如，问句“打篮球可以戴眼镜吗”的依存句法分析如图 3-1 所示。从分析结果中我们可以看到，句子的核心谓词是“打”和“戴”，打的宾语是“篮球”，戴的宾语是“眼镜”，而“篮球”和“眼镜”都是问句的关键词。



图 3-1 依存句法分析图

3.4 实验结果与分析

3.4.1 实验设置与结果

在实验中，分词、词性标注、依存句法分析由哈尔滨工业大学语言技术平台^[30]（Language Technology Platform, LTP）自动生成。我们使用最大熵作为分类器，模型训练过程中使用了 `mallet` 工具⁴。

实验数据方面，我们使用人工标注的 800 个问句作为训练集，200 个问句作为测试集。

我们使用 MAUI 算法和基于最大熵模型的关键词提取方法，在相同的测试集上进行实验，实验结果如表 3-2 所示。在基于最大熵模型的关键词提取实验中，我们使用了词频特征、逆文本频率特征、词性特征、当前词与前一个词词性的组合特征、当前词与后一个词词性的组合特征、是否是命名实体特征、位置信息特征、依存关系特征和依存弧所指向词的词性特征，共计 9 维特征。

表 3-2 实验结果对比

方法	P	R	F ₁
MAUI	61.42%	83.08%	70.63%
MaxEnt	78.00%	83.09%	80.46%

通过实验可以看出，我们的方法相比较基准系统而言达到了一个比较好的水平，准确率得到了大幅度的提升，而召回率并没有降低，整体 F 值提高了 10 个点，达到了 80%。

3.4.2 实验分析

为了鉴别基于最大熵模型的关键词提取方法中提出的 11 个特征是否对关键词具有识别作用，我们进行了特征分析实验。特征分析采用的方法是每次去

⁴ <http://mallet.cs.umass.edu/>

除其中一个特征后进行对比实验。特征分析的实验结果如表 3-3 所示。

表 3-3 特征分析实验结果

特征种类	去除特征	P	R	F ₁
	NULL（保留全部特征）	78.14%	82.82%	80.41%
统计特征	词频	77.99%	82.56%	80.21%
	逆文本频率	81.85%	33.02%	47.06%
	词性	77.03%	82.16%	79.51%
词法特征	前一个词的词性	78.14%	82.83%	80.41%
	后一个词的词性	78.24%	82.82%	80.47%
	当前词与前一个词词性的组合	77.59%	79.76%	78.66%
	当前词与后一个词词性的组合	76.97%	81.89%	79.35%
	是否是命名实体	78.16%	82.42%	80.23%
结构特征	位置信息	77.29%	82.02%	79.59%
依存句法特征	依存关系	77.85%	82.82%	80.26%
	依存弧所指向词的词性	81.15%	71.64%	76.10%

根据表中的结果，我们可以看出，在统计特征中，当去除逆文本频率特征后，相比使用全部特征 F₁ 值下降的程度最大，从 80.41% 下降到 47.06%，说明逆文本频率特征对于关键词识别有非常大的影响。而去除词频特征后，F₁ 值仅有很小幅度的下降，这是因为问句具有短文本的特性，问句中的每个词出现的次数基本为 1，因此频率没有很大的区分度。

在词法特征中，去除词性特征、当前词与前一个词词性的组合特征、当前词与后一个词词性的组合特征和是否是命名实体特征，F₁ 值均有下降，说明这些特征对关键词识别有影响。而去掉前一个词的词性特征和后一个词的词性特征，P、R 和 F₁ 值没有下降，其中一些指标反而提高了。这是因为，前一个词的词性特征和后一个词的词性特征的分类效果包含在当前词与前一个词词性的组合特征和当前词与后一个词词性的组合特征中，说明增加这两个特征对于关键词提取没有帮助，反而引入了更多的噪声。因此，我们去掉这两个特征。

在结构特征中，去除位置信息特征，F₁ 值稍有下降，从 80.41% 下降到 80.26%，说明位置关系对关键词识别有较小的影响。

在依存句法特征中，去掉依存弧所指向词的词性特征，F₁ 值有较大幅度的下降，从 80.41% 下降到 76.10%，而去掉依存关系特征，F₁ 值下降程度较小，说明这两个特征均对关键词提取的效果有影响，依存弧所指向词的词性特征对关键词的识别影响较大。这也能够证明我们提取的依存句法特征的有效性，它能

够通过词之间结构关系上的特点对关键词进行判断，从而提升整体的效果。

对于实验结果，我们观察到主要存在以下几类典型的错误。

- (1) 第一类错误主要存在于句子中一些无用的成分。用户在描述问题时，会加入一些无用的描述性的语句，目的是为了回答者对其提问有更多的了解。但是，很多时候，这些描述是不应该被识别为关键词的。比如这样一个例子：“cf 两周年活动抽奖问题，我的任务都做了”。在这样一个句子中，由于后半句中的“任务”是名词而且在句子中存在定中关系，因此我们的模型将“任务”识别为关键词。但是，后半句的描述都是无意义的，很明显不应该将其识别为关键词。这也从某种程度上表明了语义在问句关键词提取中的重要性，我们后续的一些工作就是要通过某些方法利用词的语义信息对关键词进行识别。
- (2) 第二类错误主要出现于有数字的问句中。用户在描述问句的时候，会加入一些数字，而对于其中大多数数字而言，我们的统计信息特征是无效的，比如“2006”，“2008”这类年份的词语或者“500”这类表示数目的词语，因为这些词并不会大量重复出现，而且出现的意义也不尽相同。因此，在处理存在数字的问句时，我们的模型总是会出现错误。比如下面一个句子“2006 北京高考分数线？”这个句子里，“2006”应该被识别成为一个关键词，因为它能够对“高考分数线”进行区分。而下边一个句子“2008 北京奥运会主题曲”中，“2008”则不应该被识别成一个关键词，因为“北京”就足够对“奥运会”进行区分了。这种情况下，我们的模型很难对这类问题进行正确地处理。

3.5 本章小结

本章首先介绍了有监督的关键词提取技术研究的基准系统——MAUI 系统。在此基础上，提出了基于最大熵模型的关键词提取方法，使用最大熵模型作为分类器，创新地将依存句法特征应用到关键词提取技术中。通过特征分析，我们选取了最有效的特征，并证明了依存句法特征的有效性。

我们还将我们的方法和 MAUI 方法在同样的训练集和测试集上做了对比实验，实验表明，我们的方法相比较基准系统而言，准确率得到了大幅度的提升，而召回率并没有降低，整体 F_1 值提高了 10 个百分点，证明了基于最大熵模型的关键词提取方法的有效性。

第4章 自动学习特征的深度学习方法提取关键词

4.1 引言

深度学习提出了一种让计算机自动学习模式特征的方法，并将特征学习融入到了建立模型的过程中，从而减少了人为设计特征造成的不完备性。而目前以深度学习为核心的某些机器学习应用，在满足特定条件的应用场景下，已经达到了超越现有算法的识别或分类性能。

深度学习的方法来提取关键词，能够让机器自动学习关键词的特征，而免去人工选取特征的过程。因此我们的研究中，在传统的基于特征选择的机器学习算法的基础上，引入深度学习的方法，充分利用问句中词语的上下文信息。我们认为，这种对上下文信息的利用能够更为充分地从语义的角度对问句建模。而在自然语言处理中，传统方法不能很好地获取语义信息，因此深度学习方法相对于传统方法来说，效果更好。此外，为了更好地获取词语级别的语义信息，我们使用词向量来对词进行表示。

4.2 LSTM 网络介绍

4.2.1 RNN

在深度学习领域，传统的前馈神经网络（Feed-Forward Neural Net, FNN）在手写数字识别、目标分类等很多任务上都具有出色的表现。目前为止，FNN 在分类任务上比其它方法要略胜一筹。

尽管如此，对自然语言处理任务而言，FNN 还是有很大的局限性。在自然语言处理任务中，信息之间有着复杂的时间关联性，即当前时刻状态受之前时刻状态影响较大。例如，要预测句子的下一个词语是什么，一般都会用到前文中出现的词语，因为一个句子中前后词语并不是独立的。此外，需要输入的序列通常是不定长的，因为无法规定句子中出现词语的个数。为了解决上述问题，我们使用递归神经网络（Recurrent Neural Net, RNN）。

递归神经网络相比传统的前馈神经网络，引入了定向循环的结构，能够处理前后关联的、不定长的序列输入问题。在传统神经网络模型中，层与层之间是全连接的，但是每层之间的节点是无连接的。在 RNN 中，因为其特有的定向循环结构，使得隐含层之间的节点不再是无连接的，而是有连接的，即一个隐含层的输入不仅同当前输入有关，还包含上一时刻隐含层的输出。图 4-1 是一

个典型的 RNN 结构与其展开结构。图上的每个节点表示每一时刻 RNN 网络中的一层。

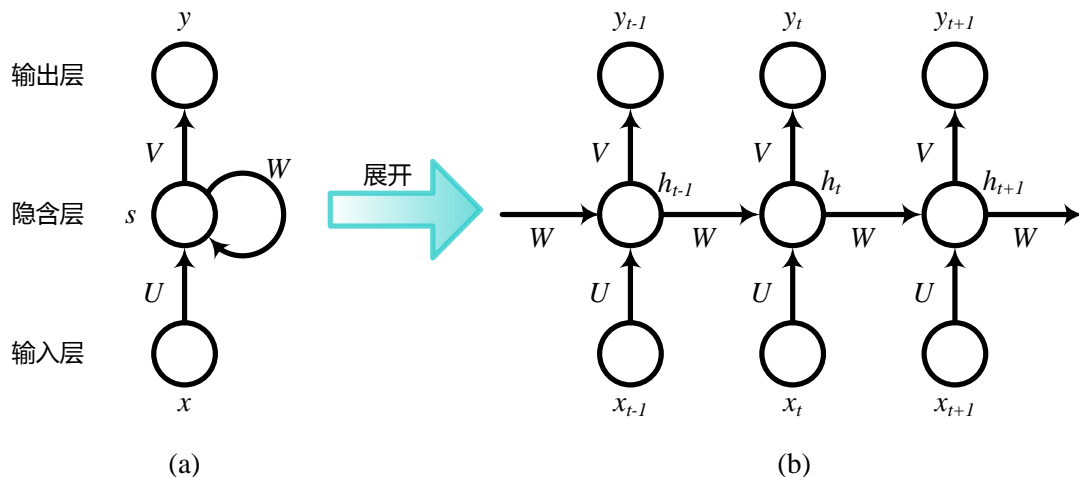


图 4-1 RNN 结构图

RNN 网络包含输入层 (Input Layer)、隐含层 (Hidden Layer) 和输出层 (Output Layer)，输入集被标记为 $\{x_0, x_1, \dots, x_t, x_{t+1}, \dots\}$ ，输出集则被标记为 $\{y_0, y_1, \dots, y_t, y_{t+1}, \dots\}$ ，隐含层的输出集标记为 $\{h_0, h_1, \dots, h_t, h_{t+1}, \dots\}$ 。U 表示输入层到隐含层之间的连接权值，V 表示隐含层到输出层之间的权值，W 表示上一时刻隐含层到当前时刻隐含层的连接权值。这些隐含层中的隐含单元完成了最为主要的工作。隐含层的反馈，不仅仅进入输出端，而且还进入了下一时刻的隐含层。

图 4-1(b) 将递归神经网络结构展开成一个全神经网络结构。例如，对于一个包含 5 个词语的句子，其展开的网络便是一个五层的神经网络，每一层代表一个词语。对于该网络的计算过程如下：

- (1) x_t 表示第 t ($t = 1, 2, 3 \dots$) 时刻的输入，例如， x_0 是第一个词的词向量， x_1 是第二个词的词向量。
- (2) h_t 为隐含层第 t 时刻的状态，它是网络的记忆单元。 h_t 根据当前输入层的输出与上一时刻隐含层的状态进行计算，如公式(4-1)。其中 f 一般是非线性的激活函数，如 \tanh 或 sigmoid 。在计算 h_0 时，即第一个词语的隐含层状态，需要用到 h_{-1} ，但是其并不存在，在实现中一般置为零向量。

$$h_t = f(Ux_t + Wh_{t-1}) \quad (4-1)$$

- (3) y_t 是第 t 时刻的输出，如下一个词语的向量表示，使用公式(4-2)进行计算。

$$y_t = \text{softmax}(Vh_t) \quad (4-2)$$

在传统神经网络中，每层的参数是不共享的。而在 RNN 中，每一层各自都共享参数 U 、 V 和 W ，也就是说 RNN 中的每一时刻都在做同样的事，只是输入不同，因此大大减少了需要学习的参数个数。

图 4-1 中每一时刻都会有输出，但是每一时刻都要有输出并不是必须的。例如，我们需要预测一条语句的情感倾向，我们仅仅需要关心最后一个词语输入后的输出，而不需要知道每个词语输入后的输出。同理，每一时刻都需要输入也不是必须的。RNN 的关键之处在于隐含层，因为隐含层能够捕捉到序列的信息。

由于传统的 RNN 模型展开后相当于多层的 FNN，层数对应历史输入数据的个数，层数过多会导致训练参数时梯度消失和历史信息损失的问题。也就是说，越远的序列输入对训练权值所能起到的“影响”越小，所以训练的结果往往偏向于新的信息，即不太能有较长的记忆功能，能够利用的历史信息非常有限。

我们可以利用图 4-2 对这些问题做一个直观地解释。图中节点的颜色深浅表示最初时刻输入信息对该节点造成的影响的大小。从前反馈的过程来说，最初时刻输入信息的影响随着时刻的变化，不断有新的信息输入而逐渐减小，在处理较后时刻（如图中的时刻 6）的信息时，已经无法利用最初时刻输入的信息。从训练时的反向传播过程来说，恰恰相反，例如时刻 6 输出层的误差在通过梯度向前传播的过程中，由于梯度逐渐变小，传播误差也逐渐变小，导致无法有效地更新较前时刻对应的权值。

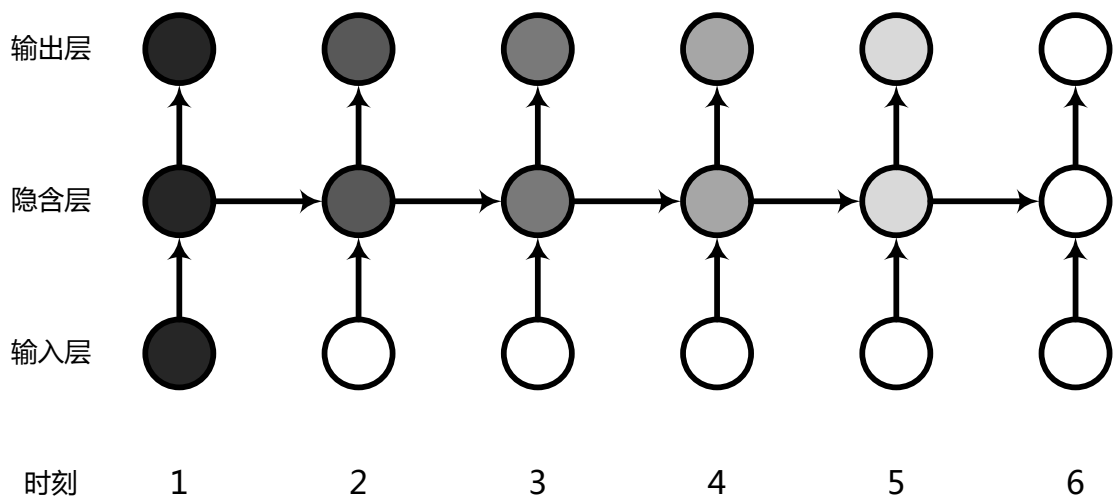


图 4-2 梯度消失示意图

4.2.2 LSTM

Hochreiter 等^[37]在 1997 年提出了长短时记忆模型(Long Short-Term Memory, LSTM)模型,用于解决 RNN 的梯度消失的问题。近些年,LSTM 被广泛应用于各种自然语言处理任务中,如情感分析^[38]、文本分类^[39]、语音识别^[40]和阅读理解^[41]等。

目前使用最广泛、最成功的 LSTM 网络的单元如图 4-3 所示。LSTM 单元代替了之前 RNN 网络中的隐含层。

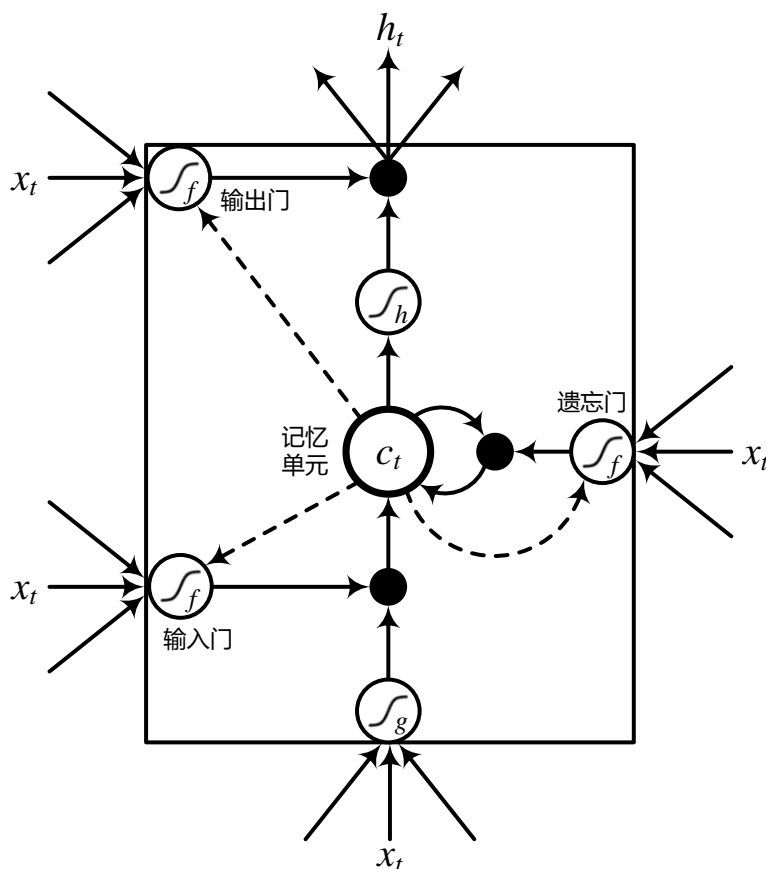


图 4-3 LSTM 单元结构图

在 LSTM 单元中,设计了专门的记忆单元(Memory Cell)用于储存历史信息。历史信息的更新和使用分别受三个门的控制——输入门(Input Gate)、遗忘门(Forget Gate)和输出门(Output Gate)。

其中, x_t 为 t 时刻 LSTM 单元的输入, c_t 为 t 时刻 LSTM 单元的值, h_t 为 t 时刻 LSTM 单元的输出。LSTM 单元的更新过程如下:

- (1) 按照传统 RNN 的公式计算当前时刻候选记忆单元 \tilde{c}_t ,如公式(4-3)所示。

W_{xc} 和 W_{hc} 分别对应输入数据和上一时刻 LSTM 单元输出的权值。

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1}) \quad (4-3)$$

- (2) 计算输入门的值 i_t ，如公式(4-4)所示。输入门是用于控制当前输入数据对记忆单元状态的影响。所有的门的计算除了受当前时刻输入数据 x_t 和上一时刻 LSTM 单元输出值 h_{t-1} 的影响之外，还受上一时刻 LSTM 记忆单元值 c_{t-1} 影响。

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1}) \quad (4-4)$$

- (3) 计算遗忘门的值 f_t ，如公式(4-5)所示。遗忘门是用于控制历史信息对当前时刻记忆单元状态值的影响。

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1}) \quad (4-5)$$

- (4) 计算当前时刻的记忆单元状态值 c_t ，如公式(4-6)所示。其中， \odot 表示逐点乘积。记忆单元状态值的更新取决于前一时刻自身记忆单元状态值 c_{t-1} 和当前时刻的候选记忆单元值 \tilde{c}_t ，并通过输入门和遗忘门分别对这两部分进行调节。

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4-6)$$

- (5) 计算输出门 o_t ，如公式(4-7)所示。输出门用于控制记忆单元状态值的输出。

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1}) \quad (4-7)$$

- (6) 计算 LSTM 单元的输出，如公式(4-8)所示。

$$h_t = o_t \odot \tanh(c_t) \quad (4-8)$$

上述公式中的函数 σ 一般为 sigmoid 函数，取值范围为(0, 1)。

输入门、输出门、遗忘门以及独立的记忆单元的设计结构，使得 LSTM 单元获得储存、读取、重置和更新较长距离历史信息的能力。如果输入门保持关闭状态（ $i_t = 0$ ），那么记忆单元不受后来输入信息数据的影响；与此同时，如果遗忘门不遗忘（ $f_t = 1$ ），那么记忆单元中的信息就会一直保存，并且能够通过输入门的打开状态（ $o_t = 1$ ）供模型使用。

图 4-4 是 LSTM 网络中，信息传递的一个例子。如上所述，同图中节点的颜色深浅表示最初时刻输入信息对该节点造成影响。LSTM 单元的输入门、遗忘门和输出门分别对应隐含层节点的下侧、左侧和上侧。为了简化表示，门只能取两个值：0 和 1，对应开和关，“○”表示开，“—”表示关。记忆图中，记忆单元在输入门关和遗忘门关的情况下，记住并保存了第一时刻输入的信息，并通过输出门的开关来输出信息。

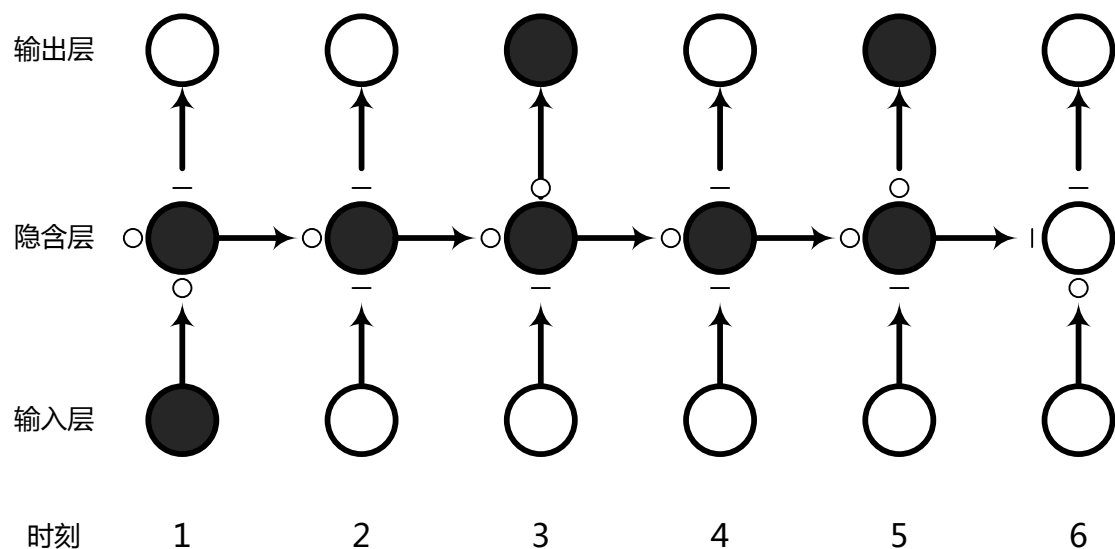


图 4-4 LSTM 信息传递示意图

LSTM 理论上可以使模型处理当前时刻数据的时候同时考虑历史信息。在实践中已经证明，LSTM 在解决自然语言处理的任务中是非常有效的，它可以学习长期依赖信息，因此得到了广泛地使用。LSTM 网络结构也是本章所用模型的重要组成部分。

4.3 基于 LSTM 模型的关键词提取方法

4.3.1 LSTM

基于 LSTM 模型的问句关键词提取方法研究利用 LSTM 构建神经网络层次，对问句建模，从而能够有效获取词级别语义信息。相比较传统的神经网络而言，LSTM 能够克服对不定长序列输入的不足，更好地存储历史信息，进而能够得到更好的效果。LSTM 将一个词序列的历史信息存储在一个实值的历史向量中，并用这个历史向量通过迭代将整个序列连结到一起。通常来说，我们认为这个包含了历史信息的向量即可表示当前的词序列。换句话说，这个向量包含了从问句开始到当前词所需要的全部语义信息。我们通过 LSTM 结构，将一个词语及其历史信息共同建模到一个向量中，并可以用这个向量来表示一个词序列所包含的信息。因此，我们可以利用 LSTM 构建一个能考虑上文信息的网络模型，将当前词和上文出现的词语共同建模到一个向量中，来预测其成为关键词的概率。

该方法主要通过 LSTM 构建一个分类器，对于句子中的每个词，预测其成为关键词的概率。对于一个词，我们抽取从其句首开始的所有词组成词序列，

放入 LSTM 中进行学习。LSTM 的输入是通过大规模语料预训练好的词向量序列，LSTM 的输出是这个词成为关键词的概率，根据设定的阈值，我们即可根据概率来选出句子中的关键词。利用深度学习的方法一方面能够有效避免冗杂的特征工程，另一方面使用预训练的词向量可以充分利用词的语义信息来从语义层面帮助判别关键词。

图 5 为我们利用 LSTM 构建的系统结构图。图中 $word_i$ 表示问句第 i 个词语的词向量， h_i 表示第 i 时刻隐含层的输出。例如问句“宁波有什么特产能在上海世博会占有一席之地呢”中的“世博会”一词，我们将从句首开始到当前词的所有词语的词向量都依次输入到模型中，最终输出为“世博会”是关键词的概率，根据设定的阈值，判断“世博会”一词是否为关键词。

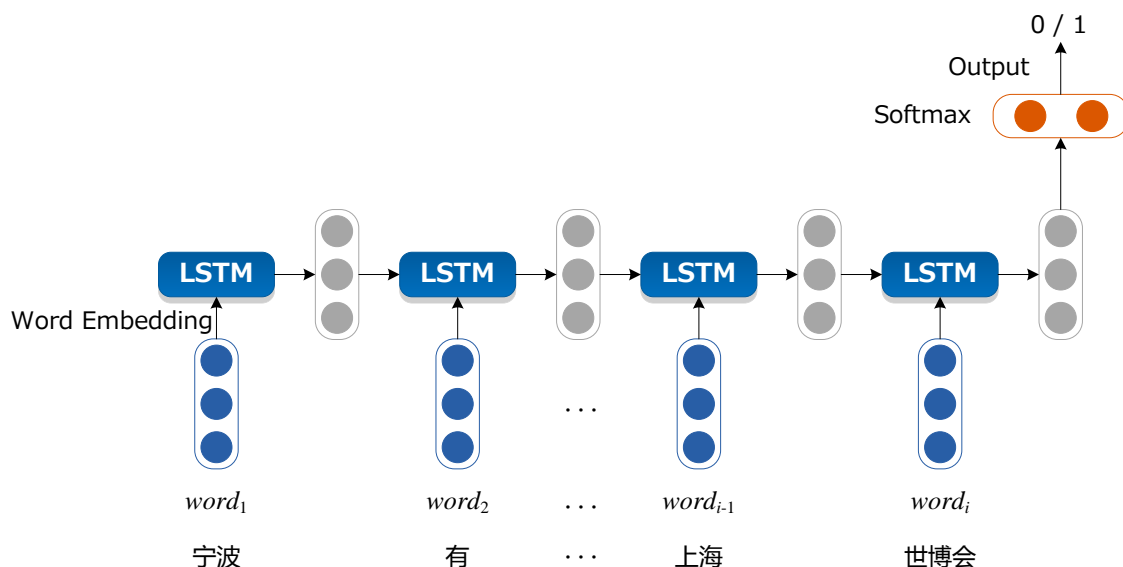


图 4-5 基于 LSTM 模型的关键词提取方法的系统结构图

4.3.2 以目标词为中心的 LSTM

在问句关键词提取任务中，要判断当前词语是否是关键词，当前词语的下文信息与上文信息同样重要。然而，由于网络是单向的，只能顺序将从句首到当前词的所有词语输入到网络中，而忽略了后文信息对判断当前词是否为关键词所产生的影响。

因此，为了将目标词的上下文均输入到模型中，我们在之前的 LSTM 模型上进行了一些改动，使用了以目标词为中心的 LSTM 结构（Target-Centered LSTM, TC-LSTM）。其主要思想是将目标词的上文信息和下文信息都输入到模型，两个方向上的信息共同建模到一个向量之中，能够更好地对目标词的重要

程度进行表示，从而来预测其是否是关键词的概率。我们相信，同时捕获目标词上下文信息的关键词提取方法，能够提高关键词提取的准确性。

具体来说，我们使用两个 LSTM 网络，左边的是 $LSTM_L$ ，右边的是 $LSTM_R$ ，分别用于接收目标词的上文信息和目标词的下文信息。图 4-6 是我们利用 LSTM 构建的以目标词为中心的关键词提取方法的系统结构图， $word_i$ 表示问句第 i 个词语的词向量， \bar{h}_i 表示第 i 时刻 $LSTM_L$ 隐含层的输出， \bar{h}_i 表示第 i 时刻 $LSTM_R$ 隐含层的输出。 $LSTM_L$ 的输入是从问句句首开始到目标词所有词语的词向量，从左至右，按顺序输入； $LSTM_R$ 的输入是从目标词到句尾所有词语的词向量从右至左，逆序输入。这样，目标词总是作为 LSTM 最后一个单元的输入。这是因为，将两个方向上的信息共同建模到向量中是时，将目标词分别作为 LSTM 最后一个单元的输入，能更充分地利用以目标词为中心的两个不同方向的信息，更好地对目标词语的重要程度进行向量表示。之后，我们将 $LSTM_L$ 的最后一个隐含向量 \bar{h}_i 和 $LSTM_R$ 的最后一个隐含向量 \bar{h}_i 连接到一起，反馈给 softmax 层，来生成目标词是关键词的概率。

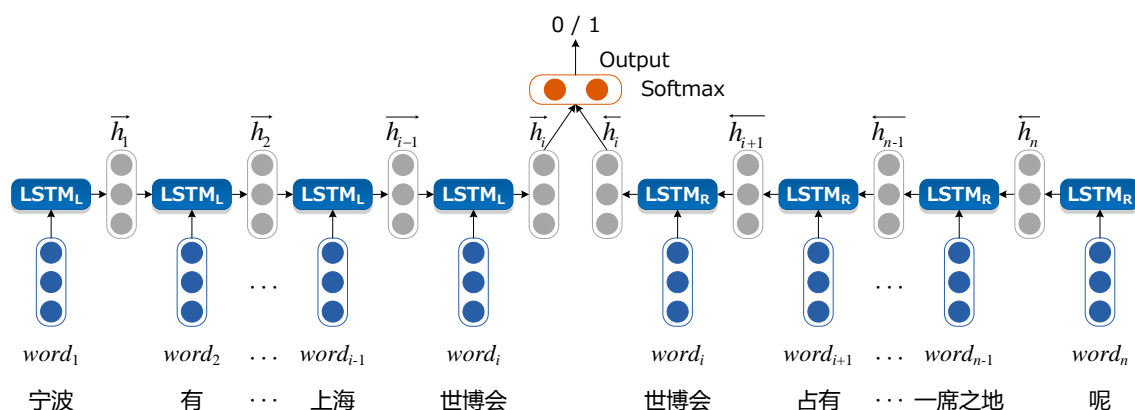


图 4-6 以目标词为中心的 LSTM 模型关键词提取方法的系统结构图

例如问句“宁波有什么特产能在上海世博会占有一席之地呢”，我们要判断“世博会”是否为关键词。首先，我们将目标词“世博会”和其上文信息，即“宁波”、“有”、“什么”、“特产”、“能”、“在”、“上海”和“世博会”这些词语的词向量从左到至右以此依次输入到 $LSTM_L$ 中，将目标词和其下文信息，即“世博会”和“占有”、“一席之地”、“呢”这些词语逆序输入到 $LSTM_R$ 中，模型最终输出“世博会”是关键词的概率，根据设定的阈值，判断“世博会”一词是否为关键词。

4.4 两段式训练方法

由于深度学习模型需要大规模训练数据，而人工标注的训练数据又十分有限，不能满足模型的训练需求。因此我们提出两段式的训练方法，首先用一个简单的方法自动标注问句中的关键词，生成不精确的标注数据，来对我们的模型进行预训练，然后使用人工标注的数据再训练我们的模型。

4.4.1 生成训练数据

为了获得大规模的问句关键词标注数据，我们使用一个简单有效的方法来自动生成大规模、不精确的关键词标注数据。

在收集问句语料时，我们从全球最大中文互动问答平台——百度知道中爬取数据。百度知道页面中不仅有问题数据，还有问题的描述、问题的最佳答案、和问题的相似问题等数据，而这些数据都可以帮助我们自动进行关键词的自动标注。直观上我们认为，如果一个词是某个问句的关键词，那么这个词很可能在问题描述中再次出现，用于更具体地描述问题；也很可能在答案中出现，用于回答问题；也有很大可能性在相似问题中出现。

另一个我们可以利用的数据资源是百度开放数据，它是用户搜索结果的日志文件，其中包含用户输入的 `query`、用户点击网页的标题和用户点击网页的 `URL`。百度开放数据记录了用户需要检索时输入的 `query` 关键词，以及输入这个关键词后在搜索引擎返回的结果中点击了某个网页的 `URL`。如果用户点击的网页是百度知道中的问题，那么说明用户想查询这个问题的信息，而用户输入的 `query` 关键词很大可能性是这个问句的关键词中的词语。

因此我们使用如下方案进行关键词的自动标注。

- (1) 对百度开放数据进行过滤，选取用户点击网页是百度知道中的问题的那些记录，记录用户输入的 `query` 和百度知道的 `URL`。
- (2) 利用过滤得到的百度知道 `URL`，使用爬虫爬取问题页面中的信息，包括问句、问题描述、最佳答案、和相似问题信息。
- (3) 对问句进行去分词、去停用词处理，剩下的词为候选关键词。
- (4) 对于搜索 `query`、相似问题、问题描述和最佳答案这些信息赋予不同的得分。对于问句中每个候选关键词，如果这个词语在搜索 `query`、相似问题、问题描述和最佳答案这些项目里出现，则获得对应项目的得分，简单地将各项得分相加，如果最终得分超过设定的阈值，则这个候选关键词被标注为关键词。各项的得分以及阈值如表 4-1 所示。其中，

$\#(similar_questions)$ 是要标注问句的相似问题的个数，一般为 4 或 5， $\#(word_in_similar_questions)$ 是相似问题中出现候选词的个数。

表 4-1 各项得分与阈值

Items	Score
搜索 query	0.75
相似问题	$\frac{\#(word_in_similar_questions)}{\#(similar_questions)}$
问题描述	0.5
最佳答案	0.25
阈值	1

我们认为，query 中出现的词就是用户搜索的词语，很大程度上就是关键词，因此赋予比较高的权重；相似问题的总条数一般都是 4 条或者 5 条（4 条是因为另一条是百度经验的链接），如果在 4 条相似问题中，有 3 条都出现当前词语，那么得分就是 0.75，也很大程度上说明是关键词，和 query 的权重相同，如果 5 条相似问题中有 4 条出现该词语，那么得分是 0.8，置信度也很高；问题描述和答案只作为辅助参考。最终权重值设置为 1，如果当前词的得分超过或等于阈值，那么它被标注为关键词。

经过过滤筛选，去掉标注为关键词的数量少于 1 的句子，并去掉了问题分类为“游戏”、“外语”、“主板”、“显卡”、“软件”、“电脑”、“相机”、“数学”的问题，最终自动标注 243882 个问句。

4.4.2 两段式训练

值得注意的是，虽然我们已经生成了大规模的问句关键词标注数据，但是这些大规模的标注数据和人工标注的数据还是有一定差距的，是不够准确的。因此，我们应该采取一些措施，使我们的模型能够在利用大规模数据训练后，适应精确的人工关键词标注语料。

在本章的关键词提取方法中，我们使用两段式的训练方法来处理大规模训练数据与人工标注的关键词数据不匹配的问题。首先，在第一阶段，我们使用上节所描述的自动标注的大规模训练数据，来训练一个基本的模型，根据验证集选择一个最好的模型。然后，第二阶段，我们在基本模型的基础上，使用人工标注的训练数据继续训练。

使用自动标注的大规模数据和精确的人工标注数据结合的训练方式，远比

单独使用人工标注的数据训练模型要有效。尽管自动标注的大规模数据不够精确，但是像上节所介绍的，大部分的关键词还是可以被正确标注的。因此，将大规模训练数据和精确的人工标注数据结合的训练方式对于关键词提取模型是有效的。

4.5 实验结果与分析

4.5.1 实验设置

本章中深度学习的网络结构使用深度学习框架 Keras⁵搭建。Keras 是基于 Theano 的一个深度学习框架，用 Python 语言编写，是一个高度模块化的神经网络库，支持 GPU 和 CPU。

本章从百度知道抓取了 1.70 GB 的数据，数据集其中包含了 375048 个问句。经过自动生成标注数据和过滤筛选，去掉标注为关键词的数量少于 1 的句子，并去掉了问题分类为“游戏”、“外语”、“主板”、“显卡”、“软件”、“电脑”、“相机”、“数学”的问题，最终自动标注 243882 个问句。使用这 243882 个问句标注结果作为第一阶段的训练集。

实验中，我们使用人工标注的 800 个问句作为第二阶段的训练集，200 个问句作为测试集。训练集中，随机抽取 10% 的数据作为验证集。

我们使用 google 的开源工具 word2vec^[32]，利用 2012 版搜狗新闻数据（SogouCS）训练得到词向量。训练词向量时，使用 Skip-gram 模型，每个词向量的维度为 100 维。

4.5.2 实验结果与分析

在人工标注的测试集上，我们对基于深度学习的问句关键词提取方法进行了实验。在实验结果中，我们对比了传统 LSTM 和 TC-LSTM 的效果，从而验证了 TC-LSTM 的有效性。实验结果如表 4-2 所示。

表 4-2 基于 LSTM 模型的关键词提取方法实验结果

方法	P	R	F ₁
MaxEnt	78.00%	83.09%	80.46%
LSTM	78.31%	86.90%	82.38%
TC-LSTM	79.76%	87.57%	83.48%

⁵ <http://keras.io>

通过实验结果对比,我们发现,相比较传统机器学习方法而言,使用 LSTM 方法能够获取更好的效果, F 值达到 83.48%, 说明了深度学习方法的有效性。当我们使用了 TC-LSTM 时, 效果相比较传统 LSTM 来说不论在准确率还是召回率上都有了一定的提升, 表现出后文信息对于关键词提取的重要性。即加入后文信息后, 同时考虑一个词的上下文能够对当前词是否为关键词做出更为准确的判断。例如, 在问句“谁知道 2010 世界杯 6 月 11 日北京时间几点开幕?” 中, 我们的 TC-LSTM 系统正确预测出“2010”是关键词, 而 LSTM 系统并没有得到正确的结果。这是因为, 当只考虑上文信息时, 很难对于“2010”是否是关键词进行预测, 而当我们考虑其后文信息时, 因为看到了“世界杯”, 因此“2010”被预测为了关键词。

通过观察实验结果, 对于深度学习方法而言, 我们发现了一个非常明显的错误, 即当关键词没有对应的词向量表示的时候, 我们的方法很难对关键词进行正确的判断。例如问句“2009 快乐女声曾轶可个人资料详细介绍?”, 在这个句子中, “曾轶可”应该被标注为关键词, 但是我们的方法将其预测为“非关键词”。这主要是因为, 在我们预训练的词向量中, 并没有“曾轶可”对应的词向量, 而在使用 LSTM 预测的时候, 我们把每个词, 包括当前词放入次序列中进行预测, 当一个词找不到对应词向量的时候, 会用零向量代替, 因此可能导致错误的预测结果。

为了验证我们提出的两段式训练方法的有效性, 我们做了对比实验。在实验中, 我们对比了未使用两段式训练和使用了两段式训练的实验结果, 并分别对 LSTM 和 TC-LSTM 进行了测试。实验结果如表 4-3 所示, 其中“人工标注训练数据训练”表示没有使用大规模自动生成的训练数据进行预训练, 只使用了人工标注的训练数据集进行训练得到的结果; “两段式训练”表示第一阶段使用自动生成的大规模数据进行训练, 第二阶段在预训练的基础上, 继续使用人工标注的训练数据集进行训练得到的结果。

表 4-3 两段式训练方法对关键词提取的影响

方法		P	R	F ₁
人工标注数据训练	LSTM	79.35%	77.78%	78.56%
	TC-LSTM	79.81%	79.50%	79.66%
两段式训练	LSTM	78.31%	86.90%	82.38%
	TC-LSTM	79.76%	87.57%	83.48%

通过对比试验, 我们能够发现, 如果只使用人工标注的训练数据, 相比较于两段式训练方法, LSTM 模型和 TC-LSTM 模型的效果均有所下降。这能够证

明，利用两段式训练方法能够提升整体的训练效果。

这是因为，深度学习训练的过程中，需要大量的数据来对参数进行更好地优化。也就是说，当数据量很大的时候，能够得到一个更优的模型。而我们的大规模训练数据由于是自动标注的，因此会出现一些错误。为了尽可能修复这些错误，我们在第一次大规模数据训练的基础上进行再次训练，利用人工标注的数据对参数进行调节，从而得到了一个更优的模型。实验结果证明了两段式训练方法的有效性。

4.6 本章小结

本章提出了一种基于 LSTM 模型的关键词提取方法，利用 LSTM 构建神经网络层次，将目标词的上文信息和下文信息都输入到模型，对问句进行建模，更好地利用了词语的语义信息。本文使用了两段式的训练方法，来训练我们的模型，利用一个简单有效的方法来自动生成大规模、不精确的关键词标注数据，解决了人工标注的训练数据不足，无法满足模型的训练需求的问题。

实验证明，深度学习的关键词提取方法比以往机器学习的关键词提取方法更为有效，系统性能得到了提高。

结 论

面向问答的问句关键词提取技术研究如何从用户输入的问句中提取出对检索有用的关键词，来提高问句检索的性能和答案相似度计算与排序的性能，为问答系统带来更好的用户体验。

本文主要探索了两类关键词提取技术：无监督的关键词提取方法和有监督的关键词提取方法。有监督的关键词提取方法又分为：基于特征选择的机器学习方法和深度学习的方法。

本文提出了基于依存分析排序的无监督方法提取关键词。该方法引入词向量，从语义的角度衡量词语的相似度，引入依存句法分析，从句法结构的角度来表示两个词语之间的关联度，从而利用基于图的排序算法，更加准确地对候选词语进行排序。实验表明，基于依存分析排序的无监督方法能够提高问句关键词提取的效果。

基于特征选择的机器学习方法提取关键词，将依存句法特征应用到关键词提取技术中，通过特征分析实验，选取最有效的特征，利用最大熵模型训练分类器，来判断候选词是否为关键词。实验结果表明，依存句法特征有助于提高关键词提取的效果。

自动学习特征的深度学习方法提取关键词，将特征学习同模型建立融合在一起，让机器自动对关键词的特征进行学习，有效避免了特征工程。在我们的研究中，为了更好地利用上下文词语的语义信息，利用 **LSTM** 模型构建了神经网络层次。同时使用了两段式的训练方法，解决了人工标注的训练数据不足，无法满足模型的训练需求的问题。实验证明，深度学习的关键词提取方法比以往机器学习的关键词提取方法更为有效。

在问句关键词提取任务中，仍有许多困难需要在未来的研究中解决。例如，缺少标准语料，虽然我们可以在互联网上收集到大量的问句，但是缺少对于这些问句的关键词标注结果，虽然我们已制定了相应的关键词标注规范并进行了语料库构建，但是人工标注的语料数量上还远远不够。又例如，如何将问句关键词提取得到的结果应用到问答系统中，提高问答系统回答问题的能力，也是未来的研究方向之一。

参考文献

- [1] 郑实福, 刘挺. 自动问答综述[J]. 中文信息学报, 2002, 16(6): 46-52.
- [2] Luhn H P. The automatic creation of literature abstracts[J]. IBM Journal of research and development. 1958, 2(2): 159-165.
- [3] El-Beltagy S R. KP-Miner: A Simple System for Effective Keyphrase Extraction[C]. Innovations in Information Technology, 2006. EEE, 2006: 1-5.
- [4] Medelyan O, Frank E, Witten I H. Human-competitive tagging using automatic keyphrase extraction[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, 2009: 1318-1327.
- [5] Yih W, Goodman J, Carvalho V R. Finding advertising keywords on web pages[C]. Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 213-222.
- [6] Peter Turney. Coherent keyphrase extraction via web mining [J]. In Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003: 434-439.
- [7] Turney P D. Learning to extract keyphrases from text[R]. NRC Technical Report ERB-1057, National Research Council, Canada. 1999: 1-43.
- [8] Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]. Proceedings of the 4th ACM Conference on Digital Libraries, Berkeley. California, US: ACM, 1999: 254-256.
- [9] Li Z, He B. Adding Lexical Chain to Keyphrase Extraction[C]. Web Information System and Application Conference (WISA), 2014 11th. IEEE, 2014: 254-257.
- [10] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Proceedings of EMNLP, 2004. 4(4): 275.
- [11] Gollapalli S D, Caragea C. Extracting Keyphrases from Research Papers Using Citation Networks[C]. AAAI. 2014: 1629-1635.
- [12] Chien L F. PAT-tree-based keyword extraction for Chinese information retrieval[C]. ACM SIGIR Forum. ACM, 1997, 31(SI): 50-58.
- [13] 李素建, 王厚峰, 俞士汶等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报. 2004, 27(9): 1192-1197.
- [14] 王昊, 邓三鸿, 苏新宁. 基于字序列标注的中文关键词抽取研究[J]. 现代图书情报技术, 2011, 27(12): 39-45.
- [15] 李纲, 戴强斌. 基于词汇链的关键词自动标引方法[J]. 图书情报知识, 2011 (3): 67-71.

-
- [16] 王立霞, 淮晓永. 基于语义的中文文本关键词提取算法[J]. 计算机工程, 2012, 38(01): 1-4.
- [17] 张敏, 耿焕同, 王煦法. 一种利用 BC 方法的关键词自动提取算法研究[J]. 小型微型计算机系统, 2007, 28(1): 189-192.
- [18] 刘通. 基于复杂网络的文本关键词提取算法研究[J]. 计算机应用研究, 2016, 33(2): 365-369.
- [19] Hasan K S, Ng V. Automatic keyphrase extraction: A survey of the state of the art [J]. Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland: Association for Computational Linguistics, 2014.
- [20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
- [21] Schmidhuber J. Deep learning in neural networks: An overview [J]. Neural Networks, 2015, 61: 85-117.
- [22] Hinton G E. Learning distributed representations of concepts[C]. Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.
- [23] Wei Xu, Alex Rudnicky. Can artificial neural networks learn language models? [C]. International Conference on Spoken Language Processing. 2000:202-205.
- [24] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
- [25] Wang R, Liu W, McDonald C. Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors[C]. Software Engineering Research Conference. 2014: 39.
- [26] Stubbs M. Two quantitative methods of studying phraseology in English [J]. International Journal of Corpus Linguistics, 2003, 7(2): 215-244.
- [27] Terra E, Clarke CLA. Frequency estimates for statistical word similarity measures[C]. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 165-172.
- [28] Zhang W, Ming Z, Zhang Y, et al. The Use of Dependency Relation Graph to Enhance the Term Weighting in Question Retrieval[C]. COLING. 2012: 3105-3120.
- [29] Liu F, Pennell D, Liu F, et al. Unsupervised approaches for automatic keyword extraction using meeting transcripts[C]. Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics. Association for Computational

-
- Linguistics, 2009: 620-628.
- [30] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: a Chinese Language Technology Platform[C]. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics: 2010: 13–16.
 - [31] Hougardy S. The Floyd-Warshall algorithm on graphs with negative cycles [J]. Information Processing Letters, 2010, 110(8): 279-281.
 - [32] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems. 2013: 3111-3119.
 - [33] Medelyan O, Perrone V, Witten I H. Subject metadata support powered by Maui[C]. Proceedings of the 10th annual joint conference on Digital libraries. ACM, 2010: 407-408.
 - [34] Jaynes E T. Information theory and statistical mechanics [J]. Physical review, 1957, 106(4): 620.
 - [35] Darroch J N, Ratcliff D. Generalized iterative scaling for log-linear models [J]. The annals of mathematical statistics, 1972: 1470-1480.
 - [36] Berger A. The improved iterative scaling algorithm: A gentle introduction [J]. Unpublished manuscript, 1997.
 - [37] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.
 - [38] Tai K S, Socher R, Manning C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[J]. Computer Science, 2015.
 - [39] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432.
 - [40] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 6645-6649.
 - [41] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[C]. Advances in Neural Information Processing Systems. 2015: 1684-1692.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向问答的问句关键词提取技术研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：王煦祥

日期：2016 年 6 月 29 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

(1) 学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；(2) 学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；(3) 研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：王煦祥

日期：2016 年 6 月 29 日

导师签名：张

日期：2016 年 6 月 29 日

致 谢

时光流逝，不知不觉已经在实验室度过了两年的时间。虽然时间不长，但在赛尔这个温暖的集体中，我得到了成长和进步，那些收获、幸福、感恩与回忆，从不会伴着时间而流逝。

在即将离开实验室之际，向所有在学习、生活和工作中给予我帮助、指导和关心的老师、同学们表示深深的感谢。

首先，感谢我的导师张宇教授，从我做本科毕业设计开始就为我选题和指导。感谢张老师在和我的研究生学习和科研工作中，对我悉心的指导与培养，并给予我足够的信任和自由发挥的空间，培养我独立完成研究任务的能力，使我得到成长和进步，并受益良多。除此之外，张老师治学严谨的态度和豁达的胸怀是我永远学习的榜样，并将积极影响我今后的学习和工作。

感谢刘挺老师，给予我宝贵的机会，让我有机会加入赛尔大家庭，并给予我学习生活中的指导与帮助。感谢张伟男老师，在科研中给予我指导，在生活中给予我建议。感谢秦兵老师、车万翔老师和赵妍妍老师，用辛勤工作和日夜操劳为我们编织了幸福温馨的赛尔大家庭。

感谢 QA 组的所有同学，两年的时间里我们一起学习、工作。感谢尹庆宇师兄在学习上的无私帮助和生活上的悉心照顾，让我倍感温暖。

感谢赛尔 14 级所有的小伙伴们，谢谢大家在我们共同的成长过程中的相互帮助、支持与鼓励，祝福大家在未来都能前程似锦。

感谢实验室的每一位成员，在这个温暖的大家庭中我得到了很多，希望在未来我也能对这个给予我帮助的大家庭有所回报。

感谢所有我在哈工大结交的好朋友、好姐妹，是你们充实了我在工大的生活，让我的青春更加丰富多彩。

最后，感谢我的父母，感谢你们一直以来的支持和鼓励，感谢你们在背后默默地付出，用爱与陪伴激励我不断前行。