论文编码：TP39

首都师范大学
Capital Normal University

首都师范大学本科生毕业设计论文

基于《自然语言处理》课程的在线学习系统
--基于本体模型的自动判卷系统设计与实现

Online Learning System Based on the Natural Language Processing Course
——Design and Implementation of Automatic checking System Based on the Ontology Model

论文作者_____
院　　系　信息工程学院
专　　业　智能科学与技术
学　　号_____
指导教师_____

完成日期　2014 年 4 月 20 日

首都师范大学信息工程学院

摘　　要

在中文信息处理中，文本的相似度计算是一项基础而又重要的工作，它在自然语言处理的诸多领域起着至关重要的作用。例如，自动问答、机器翻译、信息检索、自动文摘等领域，文本的相似度计算都是必须解决的关键问题，在智能信息处理领域一直以来都是人们研究和讨论的热点及难题。本体在自然语言处理上有着广泛的应用，其结构上独特的概念模型关系，给词语间相似度计算带来了新的处理方法。本体内部概念直接的关系，能有效的表示词语间的互相关系，从而体现句子内词语间在结构上的相互联系。结合同义词词林和知网在词语相似度计算上的应用，有效的解决了文本在词、句子、段落层次上的相似度计算问题。本文利用本体结构的特性，提出了一种基于本体内部概念之间的相似度计算方法。实验结果表明，该方法能充分利用本体特点来计算相关文本概念之间的相似度，证明了算法的有效性。并结合自动判卷系统，将此方法应用到具体的系统中。

关键词：文本相似度；本体；知网；同义词词林；自动判卷

首都师范大学信息工程学院

Abstract

In Chinese information processing, the text similarity calculation is a basic and important work, it plays a vital role in many fields of natural language processing. For example, automatic question answering, machine translation, information retrieval, automatic summarization, and so on, the text similarity calculation is the key problem to solve. It's a hot and difficult problem for people to study and discuss in the field of intelligent information processing. Ontology has been widely used in natural language processing, its unique concept model on the relationship between structure, take a new processing method to solve word similarity computation. Direct relationship to the internal concept of ontology, can effective mutual relations between words, which reflect between words in sentences on the structure of contact each other. Combination of Synonym word Lin and applications of HowNet words in similarity calculation, effectively solve the text in the word, sentence and paragraph level of similarity calculation problem. Based on the characteristics of the structure of ontology, this paper proposes a method based on ontology concept similarity calculation between internal. The experimental results show that the method can make full use of the characteristics of ontology to calculate related concept similarity between texts, prove the effectiveness of the algorithm. Combined with the automatic checking system, the method is applied to the specific system.

Keywords: Text Similarity; Ontology; HowNet; Synonym Word Lin; Automatic Checking

首都师范大学信息工程学院

目　录

第一章 绪论 1
1.1 论文研究背景及意义 1
1.2 国内外研究现状 3
1.3 本文的主要工作及内容安排 4
第二章 词语相似度计算 6
2.1 词语相似度的定义和特点 6
2.2 词语相似度计算方法概述 7
2.3 基于汉字字面的相似度计算 8
2.4 基于《知网》的词语相似度计算 9
2.5 基于同义词词林的相似度计算方法 12
2.6 本文的词语相似度计算方法 16
第三章 句子相似度计算 17
3.1 句子相似度的定义和特点 17
3.2 句子相似度计算方法概述 17
3.3 结合句子结构信息的相似度计算方法 20
第四章 基于本体模型的相似度计算算法 22
4.1 本体简介 22
4.2 本体模型的建立 23
4.3 基于本体模型的相似度计算 24
第五章 自动判卷系统的实现 29
5.1 系统背景介绍 29
5.2 相关技术 29
5.3 功能展示 30
5.4 核心代码 34
第六章 结论与设想 39
6.1 结论 39
6.2 收获与总结 39
6.3 设想与展望 39
参考文献 40
致　谢 42

首都师范大学信息工程学院

第一章 绪论

1.1 论文研究背景及意义

1.1.1 研究背景

基于《自然语言处理》课程的在线学习系统是一个大课题，主要是设计和实现一个关于《自然语言处理》课程的在线学习系统网站，目的是帮助那些学习《自然语言处理》课程的学生能自主的学习。为学生的学习提供便利，增加学生自主学习能力，同时减轻教师的教学负担。

在这一大课题下分了四个子课题，分别有我们小组的成员来完成。主要有

1) 基于 SSH2 的网站设计与实现；
2) 基于领域本体的自动答疑系统设计与实现；
3) 基于本体模型的自动判卷系统设计与实现；
4) 基于 Flash 的多媒体课件设计与实现

基于 SSH2 的网站设计与实现主要完成了在线学习系统的网站的设计，是课题的基础，由刘影同学负责。主要是基于 SSH2 的《自然语言处理》课程自主学习平台网站设计。网站在技术方面，主要是基于 SSH2 框架，数据库是采用 Mysql 数据库。网站主要分三种身份，分别是教师，学生和管理员。对每种身份，都有自己的管理模块。网站主要实现的功能有用户管理模块、考务管理模块、在线学习模块、交互平台模块等等。在用户管理模块中，主要是对各种身份用户信息的管理。考务管理模块中有对测试题的管理，考试管理，及成绩管理等。在线学习模块中，主要是对《自然语言处理》课程的课程学习。交互平台模块主要有通知管理，消息管理及论坛管理。

基于领域本体的自动答疑系统设计与实现是对在线学习系统中自动答疑系统的实现，由谢兴波同学负责。结合自然语言处理技术，利用本体知识，采用七步法构建《自然语言处理》课程的领域本体。并利用构建的领域本体，使用 Jena 进行推理，结合中科院分词软件，设计并实现了自动答疑系统。能对和《自然语言处理》课程相关的问题进行自动解答。《自然语言处理》课程有它的专业知识，要将其转换成领域本体知识库，就需要对它有深入的了解。因为其知识的零散性，使得如何构建领域本体变得更加困难。经过一番研究讨论，最终我们采用七步法来构建《自然语言处理》课程的领域本体。在构建了本体后，如何利用本体知识库，用来进行自动解答也是一大难点。在这里，我们小组使用了 Jena 推理机制，结合中科院的分词软件来进行设计。首先用中科院分词软件对问句进行分析，结合

用户词典及问句模板，来理解问句所要表达的信息。然后，对得到的问句信息进行处理，使用 Jena 推理在本体中得到相应的答案，最终展示到界面中。

基于本体模型的自动判卷系统设计与实现是对在线学习系统中自动判卷系统的实现，由我来负责。主要目的就是为了对两个短文本（标准答案及学生给定答案）进行相似度计算，判定学生给定答案的正确性。提出一种本体模型，并在此基础上提出一种基于本体的相似度计算方法。文本的相似度计算在自然语言处理领域一直是一大难题，也是一个关键的课题。如何解决两个短文本之间的相似度是自动判卷系统的核心问题。经过深入的研究及刘杰老师的指导，我设计出了一种基于本体模型的文本相似度计算。将文本转换成两个本体，最终把文本的相似度计算转换成对本体的相似度计算。利用本体的概念间的关系来记录文本中词与词之间的关系，同时利用概念来存储文本中所有关键信息。在自动判卷系统中，对学生答案和标准答案分别制作两个本体，然后对两个本体进行相似度计算。在计算相似度的时候，主要是计算本体概念间的相互联系。结合同义词词林及知网来处理词语的相似度。最终得到两个答案之间的相似程度。如果相似程度超过了一定的要求，则认为学生的答案具有一定的正确性。

基于 Flash 的多媒体课件制作是对在线学习系统中多媒体课件的制作与实现，由唐科技同学负责。利用 Flash 软件，结合《自然语言处理》课程，设计实现了一套多媒体课件。有对课程的学习，有动画展示，有答题互动，有趣味游戏。全方面体现了多媒体课件在课程学习上的优势。多媒体课件的制作主要考虑到增加学生在学习时的乐趣，因此在制作 Flash 多媒体课件的时候，着重表现了课件的趣味性及教学理念设计。设计了一系列的互动模块。包括课件开场动画展示、幽默的自然语言小故事、趣味“僵尸动画”答题游戏、课程知识点学习等模块。

下面是对我负责的区域，基于本体模型的自动答疑系统设计与实现的具体说明。

两个对象之间的相似度计算早已成为了自然语言处理理解、数据挖掘等领域中的基本问题，具体的说，它是对文本处理的核心问题。文本相似度计算被应用在自然语言处理中的诸多领域。例如自动问答、基于实例的机器翻译、信息检索、自动文摘等领域。本文就针对相似度计算的特点，将其应用到试卷的自动判定上，设计了一个自动判卷系统，用以解决试卷的自动判定。

自动判卷系统（Automatic Checking System）是指计算机能够根据学生给出的待判定答案和数据库中存在的标准答案进行比对，自动判定学生给出的待判定答案的正确性。自动判卷系统是对文本相似度计算的应用，实质上就是对两个文本之间的相似度进行比较。文本的相似度计算方法有许多，有简单的基于表面意义的字重叠相似度计算方法、基于语

料库统计的相似度计算方法，基于语言学相似度计算方法等。各种方法都有自己的优劣之处。本文主要提出一种基于本体的文本相似度计算方法，结合同义词词林和知网，解决文本在词语、句子、段落层次上的相似度，最终判定两个文本之间的相似度。

1.1.2 研究意义

文本相似度计算是自然语言处理中的关键问题，对其的研究有着重要的意义。本体在人工智能领域有着广泛的应用，目前对 Ontology 概念的统一看法认为“Ontology 是共享概念模型的明确的形式化规范说明”。因为本体在表示自然语言上有着其特有的优势，其概念模型能很好的表述出文本中词语间的相互关系。基于本体的文本相似度计算充分利用了本体的概念模型特点，表述了文本中词与词之间的关系，再利用相似度计算方法，结合同义词词林和知网，求出两个文本之间在语义和结构上的相似度。基于本体的相似度计算方法给计算两个对象之间的相似度提供了一个新的方法，是一种新的创新，新的尝试。

1.2 国内外研究现状

国外对英语的研究已经有了许多的研究成果，在相似度计算方面也有了一些突破。汉语的研究较英语来说有着一定的困难。英语是一种形合(Merplotactic)语言，造句是要求词的形态变化符合规则，注重句法平面。而汉语是一种意合(Semotactic)的语言，造句是要求词的意义搭配符合情理，注重语义平面。许多句子没有严格的按照语法规范，给汉语的研究带来了一定的难度。不过，国内对汉语的研究也有了一定的成果。分别体现在词语、句子的层面上。

1.2.1 词语相似度计算的研究进展

国外的词语相似度计算方法主要包括：基于单词构成字符的相似度计算方法、基于 WordNet FrameNet MindNet 等语义词典的计算方法、基于大规模的语料库统计的方法和基于搜索引擎查询的方法。在这些方法中，基于单词构成字符的计算方法和基于语义词典方法与汉语词语相似度计算中的字面相似度算法和语义词典方法类同，而后三种方法属于基于统计的方法。

在国内，对于汉语词语的相似度计算，人们也提出了许多较有代表性的方法。如从最初简单的字面匹配相似度计算方法，到后来的基于词素的相似度计算方法，以及最近具有代表性的基于汉语语义词典的计算方法。同时，对于基于概率统计的相似度计算，也有学者做了相关研究。基于字面匹配的相似度计算方法实现比较简单，不需要如语义词典一样的其他资源的支撑，但是该方法的弊端也是显而易见的。它无法体现词语在词性上的不同，

首都师范大学信息工程学院

及词在同义、近义词之间的相似性。基于汉语语义词典的相似度计算方法从词所表达的概念出发，从词的语义角度来计算词语相似度，计算结果与人的主观判断较为接近。然而该方法对语义词典的依赖性较大，语义词典的词汇量、组织方式、概念的表达方式、更新速度等直接决定了计算的效果。此外，单纯依靠语义词典无法计算词典中不包含的词（未登陆词）的相似度，所以必须考虑语义词典的扩展问题。不过，总的来说，目前基于统计的方法与基于语义词典的方法相比效果还不够理想。

1.2.2 句子相似度计算的研究进展

国内对于汉语句子的相似度计算研究方面取得了一定的成果。陈利人[1]等提出了句子相似度包括结构相似度和语义相似度的概念，认为句子相似度计算包括两个步骤：首先，经过词的结构相似度计算，得到句子的结构相似度；然后，在句子结构相似的基础上，再进行句子语义相似度的计算。张民等设计了一种基于词的汉语句子相似度计算方法[2]。由于进行相似度计算的两个句子往往长度相差较大，单词的个数不一致，因此，两个句子中每个单词都可能是相互相关的，这种相关性可通过相对位置加权处理。该方法考虑了词一级的相关性、词（同义词）的同现以及评价值的向后叠加，并设计了一个递归定义的算法，不仅可以给出句子级相似，而且也可以给出“亚句子”级相似。穗志方、俞士汶[3]根据语句相似度定义和计算的特殊要求，为被比句子设计了一种折衷的句子分析方法—骨架依存分析法，进而设计了一种基于骨架依存树的语句相似度计算模型。

在国外例如哥伦比亚大学的 Goldsdein 等人通过最大边缘相关的方法（Maximal Marginal Relevance）进行相似度计算。Lambros 等提出同时依据句子的表层结构和内容计算相似度的方法[4]。学者 Chris H.Q.Ding.等采用了隐含语义索引(Latent Semantic Indexing)的方法。在计算相似度时，系统使用了两级动态规划技术，应用动态规划算法允许在两个长度不同的句子之间计算语句相似度。Nirenburg[5]等提出了两种串匹配的方法，即更规范的“切块+匹配+重组”方法和整句级匹配的方法。这两种方法所采用的相似度衡量机制都是词组合法。该系统的相似度计算采用罚分制，两个句子匹配所得到的总罚分值由句子中每个对应单词对的比较所得的罚分组合而成。

1.3 本文的主要工作及内容安排

本文主要提出一种基于本体的文本相似度计算方法，一种多层次的文本相似度计算放。多层次分别指词语层、句子层和段落层。在词语层，结合同义词词林和知网，解决词语在同义词和语义上的相似度。在句子层，以句子为单位，构建一个本体模型，记录每个

首都师范大学信息工程学院

句子内所有词与词之间结构的关系。在段落层，提炼出段落中所有句子的信息，记录到本体模型中，得到一个本体模型。该本体模型包含了这段文本中所有的关键词的信息，并且记录了他们之间的关系。最后对得到的本体模型进行相似度计算。该方法解决了词语在语义上的相似度问题，及句子在结构上的相似度问题。

本文的主要结构有以下四部分

1) 词语的相似度计算方法与研究（第二章）
2) 句子的相似度计算方法与研究（第三章）
3) 基于本体模型的相似度计算方法与研究（第四章）
4) 自动判卷系统的设计与实现（第五章）

第一部分是对词语相似度计算方法的介绍与研究，分别介绍了基于汉字字面的词语相似度计算、基于同义词词林的词语相似度计算和基于知网的词语相似度计算。最终提出本文的词语相似度计算方法，对以上所提到的三种方法进行了有效的整合。

第二部分是对句子相似度计算方法的介绍与研究。研究汉语句子相似度的定义与计算方法，以及汉语句子相似度计算的特点；分析了基于相同词汇的句子相似度计算方法、基于相似词语的句子相似度计算方法、基于向量空间的句子相似度计算方法、结合句子结构信息的相似度计算方法的基本原理。

第三部分提出了基于本体模型的相似度计算方法，提出一种本体模型及其构建方法。并结合词语相似度计算，得到两个文本的相似度。

第四部分是对基于本体模型的相似度计算方法的具体应用，设计了一个自动判卷系统，并将之切入到在线学习系统中。

首都师范大学信息工程学院

第二章 词语相似度计算

2.1 词语相似度的定义和特点

词语是汉语最基本的语义和语法的单位，汉语词语的语义相似度计算是计算句子相似度的基础。Dekang Lin 曾从信息论的角度给出了句子相似度的一个统一的、与应用领域无关的非形式化定义。他认为，句子 A 与句子 B 之间的相似度一方面与它们的共性相关，共性越多，相似度越高；另一方面相似度也与它们的区别相关，区别越大，相似度越低；当 A 和 B 完全不同时相似度为 0，当 A 和 B 完全相同时相似度达到最大值[6]。但是上述定义的实际操作性并不强，在实际应用中常常根据应用领域的特点，从实用的角度对句子相似度的概念进行重新定义。词语相似度是一个主观性较强的概念，没有非常明确的客观标准可以衡量。因为词语之间的关系非常复杂，其相似或差异之处很难用一个简单的数值来进行度量。

脱离具体的应用背景谈论词语相似度，很难得到一个统一的定义,在具体的应用中，词语相似度的含义才会比较明确。词语本身包含的属性很广，涉及到词语的词法、句法、语义甚至语用等方方面面的特点[7]。其中，对词语相似性影响最大的应该是词语的语义。从语义的角度出发不考虑词法、语用等方面，我们定义词语的相似度为词语所表达概念的重合程度，也可以认为是两个词语所表示的语义在概念空间上距离的大小。两个词语所表达语义距离越小，所代表概念的重合程度越大，则词语的相似度越大，反之越小。词语的相似度是一个主观性相当强的概念，不同的人对于两个词语的语义相似度的感受程度往往不同。

在实际应用中往往根据实际的需要对这一概念进行调整和扩展。例如，在基于实例的机器翻译中，词语的相似度被定义为词语替换而不改变原来句意的程度，如果替换后，句意改变的程度越小则替换词与被替换词的相似度越大，反之越小；在信息检索中词语的相似度是指词语代表用户检索目标的程度，如果词语与用户的检索概念越相近则相似度越大，反之越小。本文中的文本相似度计算是基于受限领域自动问答系统的，所以将词语相似度的概念细化为词语在领域知识空间上所代表概念的重合程度。

在本文中，我们对词语的相似度进行定量计算，取值空间为[0，1]。两个词越相近取值越大，两个词语义距离越大，相似度值越小。一个词与其本身的相似度为 1。两个词在概念上完全不重合，其相似度为 0。

首都师范大学信息工程学院

2.2 词语相似度计算方法概述

从国内外研究情况看词语间语义相似度的计算方法大体上可以分成两类：一类是依赖于概念间结构层次关系组织的语义词典的方法，主要是根据概念之间的上下位关系和同义关系，利用概念距离方法(Conceptual Distance) 来计算；另一类是基于大规模语料库统计的方法，将词汇的上下文信息的概率分布作为词汇间语义相似度计算的参照。

2.2.1 基于语义词典的方法

基于按照概念间结构层次关系组织的语义词典的方法，主要是根据这类语言学资源中概念之间的上下位关系和同义关系来计算。这种方法建立在两个词语具有一定的语义相似性当且仅当它们在概念间的结构层次网络中存在一条通路（上下位关系）这一假设基础之上。基于语义词典的方法通常依赖于比较完备的概念间结构层次关系组织的大型语义词典，比如 Wordnet[8]等等，在汉语处理领域中，可以使用知网、同义词词林等语义词典。

国外有通过计算在 Wordnet 中词节点之间上下位关系构成的最短路径来计算词语之间的语义相似度[9]，还有提出根据两个词的公共祖先节点的最大信息量来衡量两个词的语义相关度等等。国内如王斌利用《同义词词林》来计算汉语词语之间的相似度。这种类型的方法一般都要利用概念距离。

概念距离[10]的方法其实就是一种简化的传播激活方法。传播激活理论的一个假设就是语义网络是沿着语义相似度的传播趋势建立的。两个概念之间的联接越多，它们的关系也就越密切，也就是说语义关系是基于两个概念间的联接集合的。而语义距离等价于两个概念之间的最短路径长度，这是它们的不同之处。当语义网络中只有上下位这一种关系时(因为上下位是计算相似度时的主要关系)，这两种距离是等价的，也就是说可以用语义距离来代替概念距离的计算。概念距离应该是语义相似度一种递降函数，也就是说两个概念越相似，这两个概念间的距离就越短。概念层次树中的最短路径长度只存在上下位关系时才可以被用来作为概念距离的度量。

基于概念间结构层次关系的语义词典方法比较直观而且简单有效，但是构造汉语语义词典却是一件规模浩大的系统工程。

2.2.2 基于语料统计的方法

该种计算方法用大规模的语料来统计，利用词汇的上下文信息的概率分布来计算词汇间语义相似度。例如，利用词语的相关性来计算词语的相似度[11]。事先选择一组特征词，然后计算这一组特征词与每一个词的相关性，一般用这组词在实际的大规模语料中在该词的上下文中出现的频率来度量。于是对于每一个词都可以得到一个相关性的特征词向量，

首都师范大学信息工程学院

然后利用这些向量之间的相似度，一般用向量的夹角余弦来计算来作为这两个词的相似度。国外如 Brown[12]的基于平均互信息的方法，Lillian Lee[13]的基于相关熵的方法等等，这类方法建立在两个词汇具有某种程度的语义相似当且仅当它们出现在相同的上下文中这一假设的基础上。也就是说凡是语义相近的词，它们的上下文也应该相似。国内关毅[14]等提出基于统计的汉语词汇间语义相似度计算，主要研究了汉语中的实词（名词、动词、形容词）。在汉语中，具有比较强的上下文约束关系的词类包括：形容词-名词，动词-名词，副词-动词，名词-动词等等。采用模糊集合中的隶属函数定义语义相似度的数学模型，然后以同义词词林的词汇分类体系为基础，提出了基于相关熵的汉语词汇间语义相似度的计算方法。李涓子[15]利用基于统计的思想来实现语义的自动排歧。

基于统计的定量分析方法能够对词汇间的语义相似性进行比较精确和有效的度量。但是，这种方法比较依赖于训练所用的语料库，计算量大，计算方法复杂，另外，受数据稀疏和数据噪声的干扰较大，有时会出现明显的错误。

2.3 基于汉字字面的相似度计算

由于汉语中绝大多数同义词都含有相同的汉字，在字面上具有较强的相似性，利用构成词语的汉字出现情况计算词语之间的相似度是可行的。在参考文献[16，17，18，19]中对该类算法进行了论述。基于单汉字的字面相似度计算的主要思想是：汉语词语的构成往往具有“重心后移”的特点，在表达某一个具体概念的词语时，其中心部分大都在词语的后半部分。在字面上，语素越靠后，它在表达主题概念中所起的作用越大。基于这一考虑，对词语中各个语素表达主题概念的作用进行加权量化处理：位置靠前的权值较小；位置越靠后，对应的权值越大。

基于汉字的字面相似度计算算法的描述如下：

A、B 分别表示进行相似度计算的两个词语，A 与 B 的相似度为 Sim(A,B)；|A|与|B|分别表示 A 与 B 所包含的字符个数，A(i)表示 A 中的第 i 个语素；SameHZ(A,B)表示 A 与 B 中共同含有的相同语素的集合，| SameHZ(A,B)|表示相同语素的个数；Weight(A,i)表示 A 中的第 i 个语素的权值，如果 A(i) ∈ SameHZ(A,B) 则 Weight(A,i)=i，否则 Weight(A,i)=0；同理，Weight(B,j)表示 B 中第 j 个语素的权值；位置系数 d 为被匹配词与待匹配词的语素总数之比。

$$d = \min\{\frac{|A|}{|B|}, \frac{|B|}{|A|}\} \tag{2.1}$$

词语之间的相似度受两个词之间含有的相同语素的个数和相同语素在各个词中的位置权重两方面因素影响。由此词语之间的相似度可按以下公式计算：

首都师范大学信息工程学院

$$SimZM(A,B) = \alpha \times (\frac{|SameHZ(A,B)|}{|A|} + \frac{|SameHZ(A,B)|}{|B|}) / 2 + \beta \times d \times (\sum_{i=1}^{|A|} Weight(A,i) + \sum_{j=1}^{|B|} Weight(B,j)) / 2 \quad (2.2)$$

式 2.2 中，α 和 β 分别表示相同语素相似度和相同语素位置关系相似度的权重系数，并且满足 α+β=1。

以单汉字为基本处理单位的字面相似度算法无需语义词典等资源的支持，运行速度快，具有直观、简单的特点。但是通过字面相似性来计算词语的相似度也存在着明显的缺点。该方法无法识别字面完全不同的同义词和近义词，而且不适合于非纯汉字组成的词语的相似度计算。

2.4 基于《知网》的词语相似度计算

语义分析是自然语言处理技术的一个重要的研究方面，需要一定的语义知识资源作为基础，一部能够表达概念关系的语义词典是自然语言处理工作一项不可或缺的基础性资源。在英语中，人们通常用 WordNet。本文利用董振东和董强先生创建的《知网》(HowNet)作为系统的语义知识资源。《知网》是一个以汉语和英语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。它是一个网状的知识系统，作为面向汉语计算需求的知识库，提供了设计真正的智能软件所需的知识，为自然语言处理提供了一个新的研究资源。《知网》是一座宝库，它包含了丰富的词汇语义知识和世界知识，是目前基于语义的中文信息处理所使用的主要知识库之一[7]。

2.4.1 《知网》系统的概述

《知网》作为一个知识系统，它着力反映的是概念的共性和个性。在知网中概念是对词语语义的一种描述，每一个词语可能有不同的语义，那么每一个词可以表达为几个概念。概念是用一种知识表示语言来描述的，这种知识表示语言所用的“词汇”叫做“义原”。例如：对于“医生”和“患者”，“人”是它们的共性。知网在主要特性文件中描述了“人”所具有的共性，那么“医生”的个性是他是“医治”的施事，而“患者”的个性是他是“患病”的经验者。对于“富翁”和“穷人”，“美女”和“丑八怪”而言，“人”是它们的共性。而它们的个性，即：“贫”、“富”与“美”、“丑”等不同的属性值，则是它们的个性。

同时知网还着力要反映概念之间和概念的属性之间的各种关系。总的来说，《知网》描述了下列各种关系：

(a) 上下位关系 (由概念的主要特征体现)

(b) 同义关系

首都师范大学信息工程学院

(c) 反义关系

(d) 对义关系

(e) 部件-整体关系（由在整体前标注 % 体现，如“心”，“CPU”等）

(f) 属性-宿主关系（由在宿主前标注 & 体现，如“颜色”，“速度”等）

(g) 材料-成品关系（由在成品前标注 ？体现，如“布”，“面粉”等）

(h) 施事/经验者/关系主体-事件关系（由在事件前标注 * 体现，如“医生”，“雇主”等）

(i) 受事/内容/领属物等-事件关系（由在事件前标注 $ 体现，如“患者”，“雇员”等）

(j) 工具-事件关系（由在事件前标注 * 体现，如“手表”，“计算机”等）

(k) 场所-事件关系（由在事件前标注 @ 体现，如“银行”，“医院”等）

(l) 时间-事件关系（由在事件前标注 @ 体现，如“假日”，“孕期”等）

(m) 值-属性关系（直接标注无须借助标识符，如“蓝”，“慢”等）

(n) 实体-值关系（直接标注无须借助标识符，如“矮子”，“傻瓜”等）

(o) 事件-角色关系（由加角色名体现，如“购物”，“盗墓”等）

(p) 相关关系（由在相关概念前标注 # 体现，如“谷物”，“煤田”等）

《知网》系统的哲学贯穿于整个知识系统的各个方面，是系统的灵魂。首先体现在它对概念的主要属性的确定上，概念的主要属性也是概念分类的类别。《知网》系统的最高层有三类：N 范畴，V 范畴，A 范畴。N 范畴包含实体，属性和单位。其中实体的直接下位有万物、时间、空间、部分。事件的直接下位有静态事件和行为动作。万物包括物质、精神和事情三类，它们通常是运动和变化的主体。而这种运动和变化总是发生在一定的时空之中。运动和变化体现于属性，并由属性值显示。

2.4.2 知网的结构

《知网》对概念的描述是比较复杂的。在《知网》中，每一个概念用一个记录来表示，如下所示：

NO.=017144

W_C=打

G_C=V

E_C=~网球，~牌，~秋千，~太极，球~得很棒

W_E=play

G_E=V

E_E=

DEF=exercise|锻练,sport|体育

首都师范大学信息工程学院

其中 NO.为概念编号，W_C，G_C，E_C 分别是汉语的词语、词性和例子，W_E、G_E、E_E 分别是英语的词语、词性和例子，DEF 是《知网》对于该概念的定义，称为一个语义表达式。其中 DEF 是《知网》的核心。这里所说的知识描述语言也就是 DEF 的描述语言。

在《知网》的文档中，对知识描述语言做了详尽的介绍。不过，由于该文档过于偏重细节，不利于进行词语相似度计算。可以将《知网》的知识描述语言具体归纳为以下几条[7]：

1) 《知网》收入的词语主要归为两类，一类是实词，一类是虚词；
2) 虚词的描述比较简单，用“{句法义原}”或“{关系义原}”进行描述；
3) 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下三种形式：
   a) 独立义原描述式：用“基本义原”，或者“(具体词)”进行描述；
   b) 关系义原描述式：用“关系义原=基本义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”来描述；
   c) 符号义原描述式：用“关系符号 基本义原”或者“关系符号(具体词)”加以描述；
4) 在实词的描述中，第一个描述式总是一个基本义原，这也是对该实词最重要的一个描述式，这个基本义原描述了该实词的最基本的语义特征。

2.4.3 基于知网的相似度计算方法

已经有不少学者对如何利用《知网》进行汉语词语相似度的计算进行了研究。刘群等对《知网》的知识描述语言进行了研究，分析其描述一个词义所用的多个义原之间的关系，区分这些义原在对于表达语义所起的不同作用，提出了计算词语相似度的方法[7]。孔翔勇研究了知网知识库系统的基本结构，并针对知网关系的网状表示结构进行了关系型数据库的转化，研究了语义层次网络的量化问题，通过构建知网的义原概念层次图，使知网的量化问题变得更加方便，并通过概念距离计算公式和统计信息方法对知网义原之间的语义相似度进行了计算[20]。夏天对刘群的方法进行了改进，重新定义了相似度与语义距离之间的关系式，并采用将未登陆词切分为知网的登陆词，通过对切分得到的词的定义项合并来计算未登陆词的相似度[21]。Li Sujian, et al.提出了一种词语语义相似度的计算方法，计算过程综合利用了《知网》和《同义词词林》。在义原相似度的计算过程中，不仅考虑了义原之间的上下文关系，还考虑了义原之间的其他关系。在计算词语相似度时，加权合并了《同义词词林》的词义相似度、《知网》语义表达式的义原相似度和义原关联度。其中，刘群提出的方法较有代表性，下面对该方法进行简单介绍。

首都师范大学信息工程学院

汉语词语可以分为实词和虚词。每一个虚词在知网中都对应一组句法义原和关系义原，通过计算两个虚词所对应的句法义原和关系义原之间的相似度，来得到虚词之间的相似性。由于虚词不表达实际概念，虚词与实词差别较大，所以规定实词和虚词的相似度为0。实词之间的相似度转化为实词定义项中的义原之间的相似度。每一个实词在《知网》中有一组定义项与之对应，在实际的环境中应根据上下文进行消岐来确定实词所对应的概念。如果是两个独立的实词进行相似度计算则将两个词所对应概念的最大相似度作为词语的相似度。实词的每一个概念是由一组义原描述式来定义的，这些义原描述式可以分为四个部分：第一独立义原描述式、其他独立义原描述式、关系义原描述式、符号义原描述式。第一独立义原之间的相似度通过义原在知网中义原根据上下位关系所形成的树中的距离来计算，公式为：

$$Sim(p1, p2) = \frac{\alpha}{d + \alpha} \tag{2.3}$$

其中 p1 和 p2 表示两个义原，d 是 p1 和 p2 在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数。其他独立义原对应一组义原，采用将整体分为部分的策略，将两组的义原按相似度由大到小的顺序进行配对，将义原对相似度的平均值作为整体的相似度。对于关系义原和符号义原，将对应关系义原和符号义原相同的作为一组，用各组义原的平均值来作为整体的相似度。

计算第一独立义原描述式、其他独立义原描述式、关系义原描述式、符号义原描述式四部分相似度的加权平均值作为词语的相似度。如下式所示：

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i Sim(S_1, S_2) \tag{2.4}$$

其中，β_i（1≤i≤4）是可调节的参数，代表相应部分的权重。且有：$\beta_1+\beta_2+\beta_3+\beta_4$=1，$\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$。后者反映了 Sim1 到 Sim4 对于总体相似度所起到的作用依次递减。由于第一独立义原描述式反映了一个概念最主要的特征，所以我们应该将其权值定义得比较大，一般应在 0.5 以上。考虑到主要部分对次要部分的抑制作用将上式改写为：

$$SimZW(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(S_1, S_2) \tag{2.5}$$

该方法从词所代表的概念的角度来计算词语的相似度，结果较符合人们的主观感受。但是该方法的计算效果对于知网有很强的依赖性。该方法规定《知网》未登陆词之间、未登陆词与登陆词之间的相似度为一个很小的常数。

2.5 基于同义词词林的相似度计算方法

2.5.1 同义词词林的介绍

首都师范大学信息工程学院

《同义词词林》是梅家驹等人[22]于 1983 年编纂而成,这本词典中不仅包括了一个词语的同义词,也包含了一定数量的同类词,即广义的相关词。由于《同义词词林》著作时间较为久远,且之后没有更新,所以哈尔滨工业大学信息检索实验室利用众多词语相关资源,完成了一部具有汉语大词表的“哈工大信息检索研究室同义词词林扩展版”。《同义词词林扩展版》收录词语近 7 万条,全部按意义进行编排,是一部同义类词典。哈工大信息检索研究室参照多部电子词典资源,并按照人民日报语料库中词语的出现频度,只保留频度不低于 3 的(小规模语料的统计结果)部分词语,剔除 14706 个罕用词和非常用词后,词表共包含 77343 条词语。

2.5.2 同义词词林的分类方法

同义词词林按照树状的层次结构把所有收录的词条组织到一起,把词汇分成大、中、小 3 类,大类有 12 个,中类有 97 个,小类有 1 400 个。每个小类里都有很多的词,这些词又根据词义的远近和相关性分成了若干个词群(段落)。每个段落中的词语又进一步分成了若干个行,同一行的词语要么词义相同(有的词义十分接近),要么词义有很强的相关性。例如,“大豆”、“毛豆”和“黄豆”在同一行;

“西红柿”和“番茄”在同一行;“大家”、“大伙儿”、“大家伙儿”在同一行。另外,“将官”、“校官”、“尉官”在同一行,“雇农”、“贫农”、“下中农”、“中农”、“上中农”、“富农”在同一行,“外商”、“官商”、“坐商”、“私商”也在同一行,这些词不同义,但很相关。

同义词词林词典分类采用层级体系,具备 5 层结构,如图 1 所示。随着级别的递增,词义刻画越来越细,到了第 5 层,每个分类里词语数量已经不大,很多只有一个词语,已经不可再分,可以称为原子词群、原子类或原子节点。不同级别的分类结果可以为自然语言处理提供不同的服务,例如第 4 层的分类和第 5 层的分类在信息检索、文本分类、自动问答等研究领域得到应用。研究证明,对词义进行有效扩展,或对关键词做同义词替换可以明显改善信息检索、文本分类和自动问答系统的性能。

2.5.3 同义词词林编码方法

同义词词林共提供了 5 层编码,第 1 级用大写英文字母表示;第 2 级用小写英文字母表示;第 3 级用二位十进制整数表示;第 4 级用大写英文字母表示;第 5 级用二位十进制整数表示。例如:“Ae07C01=渔民 渔翁 渔家 渔夫 渔父 打鱼郎”,“Ae07C01=”是编码,“渔民 渔翁 渔家渔夫 渔父 打鱼郎”是该类的词语。例如:

Ba01A02=物质 质素;

Cb02A01=东南 西北 四方;

首都师范大学信息工程学院

Ba01A03@ 万物;

Cb06E09@ 民间;

Ba01B08# 固体 液体 气体 流体 半流体;

Ba01B10# 导体 半导体 超导体。

具体编码如表 2.1 所示。

表 2.1 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	=\#\@
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

由于第 5 级有的行是同义词,有的行是相关词,有的行只有一个词,分类结果需要特别说明,可以分出具体的 3 种情况。在使用上,有时需要对这 3 种情况进行区别对待,所以有必要再增加标记分别代表几种情形。表 1 中的编码位是按照从左到右的顺序排列。第 8 位的标记有 3 种,分别是“=”、“#”、“@”。“=”代表“相等”、“同义”;“#”代表“不等”、“同类”,属于相关词语;。“@”代表“自我封闭”、“独立”,它在词典中既没有同义词,也没有相关词。

2.5.4 基于同义词词林的词语相似度算法

该算法根据同义词词林编排特点,提出基于同义词词林的义项相似度的主要思想是:基于同义词词林结构,利用词语中义项的编号,根据两个义项的语义距离,计算出义项相似度。算法如下。

首先判断在同义词林中作为叶子节点的两个义项在哪一层分支,即两个义项的编号在哪一层不同。例如: Aa01A01 与 Aa01B01,即在第 4 层分支。从第 1 层开始判断,相同则乘 1,否则在分支层乘以相应的系数,然后乘以调节参数$\cos(n \times \frac{\pi}{180})$,其中 n 是分支层的节点总数,该调节参数的功能是把义项相似度控制在[0,1]之间。

词语所在树的密度,分支的多少直接影响到义项的相似度,密度较大的义项相似度的值相比密度小的相似度的值精确。再乘以一个控制参数(n -k+1) /n,其中 n 是分支层的节点总数, k 是两个分支间的距离。这样把原本计算出的只对应在几点的值细化,精确计算结果。

若两个义项的相似度用 Sim 表示

1) 若两个义项不在同一棵树上:

$$SimCL(A, B) = f \tag{2.6}$$

2) 若两个义项在同一棵树上:

① 若在第 2 层分支,系数为 a

$$SimCL(A, B) = 1 \times a \times \cos(n \times \frac{\pi}{180}) (\frac{n - k + 1}{n}) \tag{2.7}$$

② 若在第 3 层分支,系数为 b

$$SimCL(A,B) = 1 \times 1 \times b \times \cos(n \times \frac{\pi}{180}) (\frac{n-k+1}{n}) \quad (2.8)$$

③ 在第 4 层分支,系数为 c

$$SimCL(A,B) = 1 \times 1 \times c \times \cos(n \times \frac{\pi}{180}) (\frac{n-k+1}{n}) \quad (2.9)$$

④ 在第 5 层分支,系数为 d

$$SimCL(A,B) = 1 \times 1 \times 1 \times d \times \cos(n \times \frac{\pi}{180}) (\frac{n-k+1}{n}) \quad (2.10)$$

举例说明如下:

Ae05A01=邮递员 邮差 信差 信使 绿衣使者 通信员 投递员

Ae05A02=交通员 交通 通讯员

Ae05A03=联络员 联络官 联系人

义项邮递员和联络员的相似度为

$$SimCL(A,B) = d \times \cos(n \times \frac{\pi}{180}) (\frac{n-k+1}{n})$$

其中 n=3, k=2(邮递员在 01 分支,联络员在 03 分支)。

该算法只考虑了分支处的处理,然后计算相似度的值,分层之后的层次编号不予考虑。

如计算的两个义项的编号相同,即在同一行内,则考虑用编号计算义项的相似度:当编号相同且末尾号为=时,相似度为 1;当编号相同而只有末尾号为#时,直接把定义的系数 e 赋给结果。例如:

Ad02B04#非洲人 亚洲人

则义项“非洲人”和“亚洲人”的相似度

SimCL(A,B) =e

当编号的尾号为@时,则代表这个词既没有同义词也没有相关词,在一个编号中只有一个词，所以不予考虑。

在经过多次试验,人工评定后将层数初值设置为 a=0.65, b=0.8, c=0.9, d=0.96, e=0.5, f=0.1。

在计算词语相似度时,把两个词语的义项分别两两计算,取最大值作为两个词语的相似度值。例如:词语“骄傲”的义项编号有“Da13A01=”、“Ee34D01=”;词语“仔细”的义项编号有“Ee26A01=”、“Ee28A01=”。分别计算义项的相似度为: 0.1, 0.1, 0.483920, 0.510077。可以得出“骄傲”和“仔细”的词语相似度为 0.510077,即 4 个义项相似度的最大值。

首都师范大学信息工程学院

2.6 本文的词语相似度计算方法

在本文中，对词语相似度计算分别从基于汉字字面、基于知网和基于同义词词林的三方面考虑。如上三节所述，这三种方法在计算词语相似度的时候都各有优势。

基于字面的词语相似度计算能解决词语在字面上的相似程度，是词语表面内容相似性的体现。基于知网的词语相似度则充分利用了知网的语义结构特点，解决了词语除表面内容的相似性外，对语义相似度的完善。基于同义词词林的相似度计算是对词语同义词方面的应用，解决了词语和词语之间在意义上的相似性。可以说，以上三种相似度计算方法都有各自的特点。因此本文提出一种结合以上三种相似度计算的方法。分别求出相应的相似度计算结果，最后取出它们之间最大的值作为两个词语之间的相似度。具体体现的含义是："两个词语之间的相似度可能体现在它们之间字面上的相似度，也可能体现在它们在语义上的相似度，还需要考虑它们在意义上的相似度。其中最相似的部分可以体现两个词语之间整体的相似度。"具体的公式如下公式（2.11）

$$Sim(A,B) = \max\{SimZM(A,B), SimZW(A,B), SimCL(A,B)\} \quad (2.11)$$

其中 $SimZW(A,B)$ 代表 A,B 通过基于汉字字面的相似度计算后得到的值，其公式见本章的第三节公式（2.2）；$SimZW(A,B)$ 代表 A,B 通过基于《知网》的相似度计算后得到的值，其公式见本章第四节公式（2.5）；$SimCL(A,B)$ 代表 A,B 通过基于同义词词林的相似度计算后得到的值，其公式见本章第五节公式（2.6）至（2.10）。

首都师范大学信息工程学院

第三章 句子相似度计算

3.1 句子相似度的定义和特点

句子相似度计算是中文信息处理中一个基础而核心的研究课题，在实际中有着广泛的应用，例如，在基于实例的机器翻译中利用句子相似原理进行源语言检索；在信息过滤技术中利用句子模糊匹配来过滤敏感信息；在自动问答系统中，对于相似问句的检索以及问句与答案的匹配；句子相似度有着这样广泛的应用，然而要给它下一个准确的、公认的定义确是一件十分困难的事情。Dekang Lin 曾从信息论的角度给出了句子相似度的一个统一的、与应用领域无关的非形式化定义。他认为，句子 A 与句子 B 之间的相似度一方面与它们的共性相关，共性越多，相似度越高；另一方面相似度也与它们的区别相关，区别越大，相似度越低；当 A 和 B 完全不同时相似度为 0，当 A 和 B 完全相同时相似度达到最大值[6]。但是上述定义的实际操作性并不强，在实际应用中常常根据应用领域的特点，从实用的角度对句子相似度的概念进行重新定义。例如，在多文档文摘系统中，相似度代表句子与局部主题的相近程度；在自动问答系统 FAQ 库的问句匹配中，相似度的值是用户问句与候选问句在问题内容上的接近程度。在基于实例的机器翻译中，相似度表现为待翻译句子与双语语料库中的句子在结构和语义上的相近程度。

句子的相似度计算一般分为三个层次[23]：语法相似度、语义相似度和语用相似度。目前，人们在句子的语法和语义相似度计算方面研究较多，语用相似度的计算有相当的难度，而且在目前的应用中语用相似度的计算不是十分必要，所以对语用相似度的计算所作的研究较少。

本文将句子通过词性标注后提取出关键词进行讨论，主要讨论句子的语义结构层面上的相似度。但是，句子的语义是受句子结构等语法信息影响的，因此如何有效的运用语法信息是提高句子相似度计算效果的关键。在定量计算中我们选定句子相似度的取值区间为[0，1]，数值越大代表句子越相似，1 代表两个句子完全相同，0 代表两个句子完全不同。

3.2 句子相似度计算方法概述

很多学者对句子相似度的计算进行了大量的研究，取得了一定的成果。目前对句子语义相似度计算的方法主要有：基于向量空间模型的 TF-IDF 方法、基于语义依存的相似度计算方法、使用编辑距离的方法等。其中，基于相同词汇的方法有很明显的局限性，对于同义词之间的替换无能为力。而使用语义词典的方法，可以很好地解决这一问题，但是单

首都师范大学信息工程学院

纯的使用语义词典的方法，并没有考虑到句子内部的结构和词语之间的相互作用关系，准确率不高。编辑距离通常被用于句子的快速模糊匹配领域，但是其规定的编辑操作不够灵活，也没有考虑词语的同义替换。最后基于统计的方法，需要构造大量的训练语料，工作量是十分巨大的，而且还存在着数据稀疏、计算结果对训练语料依赖性较大等问题。

目前，已经有一些成熟的句子相似度计算方法，这些方法从不同的角度出发，具有不同的特点和适用条件，下面分别进行介绍。

3.2.1 基于向量空间模型的 TF-IDF 方法

向量空间模型(VSM：Vector Space Model)是近年来使用较多且效果较好的一种信息检索模型。在 VSM 中,将文档看作是由相互独立的词条组(T1，T2，…，Tn)构成，对于每一词条 Ti，据其在文档中的重要程度赋以一定的权值 Wi，并将 T1，T2，…，Tn 看成一个 n 维坐标系中的坐标轴，W1，W2，…，Wn 为对应的坐标值。这样由(T1，T2，…Tn)分解而得的正交词条矢量组就构成了一个文档向量空间，文档则映射成为空间中的一个点。对于所有文档和用户查询都可映射到此文本向量空间，用词条矢量(T1，W1，T2，W2，…Tn，Wn)来表示，从而将文档信息的匹配问题转化为向量空间中的矢量匹配问题。假设用户查询为 Q，被检索文档为 D，两者的相似程度可用向量之间的夹角来度量，夹角越小，说明相似度越高。

向量空间模型的常用方法有 TF-IDF(Term Frequency Inverted Document Frequency)，TF-IDF 方法综合考虑了不同的词在所有句子中的出现频率(TF 值)和这个词对不同句子的分辨能力(IDF 值)。基于向量空间模型的 TF-IDF 方法被广泛地用来计算文本之间的相似度。假设所有句子中包含的词为 w1， w2，…，wn,则每一个句子都可以用一个 n 维的向量 T=<T1，T2，…，Tn>来表示。其中，Ti(1≤ i ≤n)的计算方法为：设 n 为 wi 在这个句子中出现的个数，m 为其它所有句子中含有 wi 的句子的个数，M 为句子的总数，那么 Ti=n×log(M/m)。从这个式子中可以看出，出现次数多的词将被赋予较高的 n 值，但这样的词并不一定具有较高的 T 值。例如，在汉语中“的”出现的频率非常高，即 TF 值(n 值)很大，但由于“的”在很多句子中都出现，它对于我们分辨各个句子并没有太大的帮助，它的 IDF 值（log(M/m)）将是一个很小的数。因此，这种方法综合地考虑了一个词的出现频率和这个词对不同句子的分辨能力。

用同样的方法[24]，可以计算目标句子的 n 维向量 T’＝＜T1’, T2’, … Tn’>。得到 T 和 T’后，它们所对应的两个句子之间相似度就可以 T 和 T’这两个向量之间夹角的余弦值来表示。

首都师范大学信息工程学院

$$Similarity(T,T') = \frac{\sum_{i=1}^{n} T_i * T_i'}{\sqrt{\sum_{i=1}^{n} T_i^2 \sum_{i=1}^{n} T_i'^2}} \quad (3.1)$$

但是基于向量空间模型的 TF-IDF 也有不足之处。首先，只有当句子所包含的词语足够多时采用该方法效果才会比较好。因为它是一种基于统计的方法，只有当句子包含的词数多时，相关的词才会重复出现，这种统计的效果才会体现出来。其次，TF-IDF 方法只考虑了词在上下文中的统计特性,而没有考虑词本身的语义信息，因此具有一定的局限性。

3.2.2 基于向量空间的句子相似度计算方法

依存句法是由法国语言学家 L.Tesniere 在其著作《结构句法基础》（1959 年）中提出，对语言学的发展产生了深远的影响，特别是在计算语言学界备受推崇。依存语法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，主张句子中动词是支配其他成分的中心成分，而它本身却不受其他任何成分的支配，所有受支配成分都以某种依存关系从属于支配者。二十世纪七十年代，Robinson 提出依存语法中关于依存关系的四条公理，在处理中文信息的研究中，中国学者提出了依存关系的第五条公理[25]：

1) 一个句子中只有一个成分是独立的；
2) 其它成分直接依存于某一成分；
3) 任何一个成分都不能依存于两个或两个以上的成分；
4) 如果 A 成分直接依存于 B 成分，而 C 成分在句中位于 A 和 B 之间，那么 C 或者直接依存于 B，或者直接依存处于 A 和 B 之间的某一成分；
5) 中心成分左右两边的其它成分相互不发生关系。

句子成份间相互支配与被支配、依存与被依存的现象普遍存在于汉语的词汇（合成词）、短语、单句、复句直到句群的各级能够独立运用的语言单位之中，这一特点称之为依存关系的普遍性。依存句法分析可以反映出句子中各成分之间的语义修饰关系，它可以获得长距离的搭配，并跟句子成分的物理位置无关[26]。

利用依存结构计算句子间的相似度，关键的一步是如何获得句子各成分间的依存关系信息。目前国内，哈尔滨工业大学计算机科学与技术学院智能内容管理实验室开发了依存句法分析器，该分析器的准确率能达到 86%以上。可以通过该依存句法分析器的分析，获得句子各成分之间的依存关系。

在利用依存结构进行相似度计算时，只考虑那些有效搭配对之间的相似程度。所谓有效搭配对是指全句核心词和直接依存于它的有效词组成的搭配对，这里有效词定义为动词、名词以及形容词，它是由分词后的词性标注决定的。

首都师范大学信息工程学院

相似度计算公式如下：

$$Sim(Sen1, Sen2) = \frac{\sum_{i=1}^{n} W_i}{\max\{PairCount1, PairCount2\}} \quad (3.2)$$

$\sum_{i=1}^{n} W_i$ 为句子 1 和句子 2 有效搭配对匹配的总权重，PairCount1 为句子 1 的有效搭配对数，PairCount2 为句子 2 的有效搭配对数。

3.2.3 编辑距离方法

辑距离指从一个以字符为单位的字符串转换为另一个字符串所需要的最小编辑操作的代价数。标准的编辑操作有“插入”、“删除”和“替换”三种。编辑距离最初面向的是不考虑语义的字符处理，在字符串相似度计算、自动文摘、拼写检查、图象识别、音乐识别、语音识别、数据清理等众多领域都有着广泛地应用，在句子的相似度计算方面，也取得了一些成果，例如 G.Leusch[21]利用编辑距离计算句子相似度并用于机器翻译评价。在编辑距离方法的基础上，又提出了一些新的方法。车万翔[27]利用改进编辑距离方法用于中文相似句子的检索。该方法的主要思想是：以普通编辑距离算法为基础，采用词语取代单个的汉字或字符作为基本的编辑单元参与运算。同时使用了 Hownet 和同义词词林两种语义资源，计算词汇之间的语义距离，加入词语的语义相似信息确定词语之间的替换代价，并且赋予不同编辑操作不同的权重。在不用经过词义消歧和句法分析的情况下，兼顾了词汇的顺序和语义等信息，使之更加符合中文句子相似度计算的要求。

由以上句子相似度计算方法的分析比较可以看出，对汉语句子相似度计算的定量研究目前还处于一个起始阶段，综合考虑句子的目标层、语义层、结构层等多种因素，计算句子的相似度是今后研究的一个重要方向。本文综合考虑影响汉语句子相似度的目标层、结构层、语义层的多种因素，提出多层次多种特征融合的句子相似度计算方法。

3.3 结合句子结构信息的相似度计算方法

在汉语的表述中，一个句子中能表达实际含义的词语往往会集中在句子中的某些关键部分，如句子的名词、动词、形容词等等。这些词语是整个句子所要表达句子意思的关键词。因此对于一个句子的信息提取，只需摘取出其中重要的关键词信息就能表示整个句子。在相似度计算中，对句子的虚词、助词等对整个句子的影响比较小。相反，形容词，副词，动词，名词，数词，字符串等词语在语句的意思上表示有着关键的作用。针对这个特点，结合本体结构，设计了一个本体模型，用来表示每个句子的关键词之间相关的关系。

例如句子“汉语可以分为五级的语法单位：语素、词、短语、句子和句群。”通过中科院分词软件的词性标注后，其结果如下“汉语/n 可以/v 分为/v 五/m 级/q 的/u 语法/n 单

首都师范大学信息工程学院

位/n ：/w 语素/n 、/w 词/n 、/w 短语/n 、/w 句子/n 和/c 句/q 群/n 。/w”。从结果上可以看出，主要体现句子含义的词主要集中在标注为 n（即名词）、v(即动词)、m（即数词）等的上面。而像“、”，“：”等标点符号及“和”，“的”等连接词和虚词在句子含义是没有关键作用。由此提出一个假设。

假设：影响句子含义的关键信息集中在句子中的形容词、副词、动词、名词、数词、和字符串上。

那么在这个假设的基础上，对句子信息的收集，可以只需提取出句子中的形容词、副词、动词、名词、数词、和字符串等词性，而后进行相似度计算。

通过提取句子的关键信息解决了句子对虚词，连词等非关键信息的过滤。不过，一个句子的含义除了关键信息的提取外，还有这些关键信息之间的相互联系，即句子的结构信息。句子内的关键信息相互存在联系，如名词后面可能有动词或是形容词，每个句子内如果关键信息之间的位置不同，也会影响整个句子所要表达的含义。因此，对于句子内关键信息的相互联系，也需要记录下来。那么如何记录两个关键信息之间的相互联系呢？在此本文提出了一种基于三元组的方法来记录关键信息之间的联系。即给定一个三元组[word1,relation,word2]。Word1、Word2 表示句子中一对相邻的关键信息，这些关键信息之间存在相互联系，可以称之为关系（relation）。

那么两个关键信息之间存在哪些关系呢？下面就是对两个关键信息之间的存在的关系进行分析。通过以上的假设，我们可以知道一个句子的关键信息主要包括形容词、副词、动词、名词、数词、和字符串等词性的词语。在一个句子中，关键信息是连续出现的，例如：“语法单位”这一个短语，它由两个词“语法”，和“单位”组成。它们组合在一起有着新的含义，因此需要记录下它们之间的相互联系。这里，语法是名词，单位也是名词，它们之间的关系可以记录成“语法”之后有名词“单位”。如果通过三元组来表示的话就是[“语法”，之后有名词，“单位”]。像这样的三元组就能准确的记录的句子中关键信息之间的相互联系。通过对关键信息的分析，最后可以提炼出六种关系，分别是“之后有名词”，“之后有形容词”，“之后有副词”，“之后有动词”，“之后有数词”，“之后有字符串”等这六种关系。

通过以上的叙述，可以得出一个对句子结构信息记录的新的形式，通过三元组的方式来表示出来。这种表示方式有效的记录了句子内的关键信息，及句子内关键信息之间的相互联系，在处理句子的相似度程度上有着极大的帮助。

首都师范大学信息工程学院

第四章 基于本体模型的相似度计算算法

4.1 本体简介

4.1.1 本体定义

所谓本体,最著名并被广泛引用的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”[28]。通俗地讲,本体是用来描述某个领域甚至更广范围内的概念以及概念之间的关系,使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义,这样,人机之间以及机器之间就可以进行交流。

4.1.2 本体的组成元素

从对本体的分析中知，概念和概念之间的关系是本体的两个非常重要的组成元素。此外，还包括概念的属性和性质，以及用于知识推理的公理。

(1) 概念

概念原本是思维科学的一个术语，是“思维的基本形式之一，反映客观事物的一般的、本质的特征。”在知识工程中，将概念理解为是物体或事件的模型知识（Modular knowledge of things and events）。OKBC[29]中类（Class）和本体中的概念非常相近。例如，猫（cats）和暗杀（assassinate）分别是物体和事件的模型知识，它们都是概念。在本体工程中，概念扮演着非常重要的角色，是人与机器交互的桥梁：

1) 概念是人类对现实世界理解的表意符号；

2) 概念是机器操作的主要对象；

3) 在人类和机器之间，需要建立一个数学模型使得人类能理解并控制机器的运作，而概念又是数学模型主要的构成元素。

(2) 关系

本体中的关系表示概念之间的一类关联，典型的二元关系如概念之间的 IS-A 关系，它形成了概念之间的逻辑层次分类结构。一般情况下用 R: C1 × C2 × … × Cn 表示概念类 C1, C2, …, Cn 之间存在 n 元关系 R。

(3) 概念的属性

概念的属性是指概念的一些描述方面。例如，概念“河流”有属性“年径流量”。属性可分为基本属性和角色属性，在本体中也要作相应处理：

1) 基本属性 (basic attribute) 是指本质的属性，例如“高度”；

2) 角色属性 (role attribute) 是指基本属性在不同领域中的应用，例如与基本属性“高

首都师范大学信息工程学院

度”相对的角色属性有“海拔、深度、高程等”。

(4) 概念的性质

概念的性质是所有实例所共同具有的一般抽象的特征，是用于区分概念集中不同的概念。用形式化的方法概念的性质 p(C)可表示为 a(C) = v 其中，a 为概念 C 的属性，v 是与属性 a 相对应的属性值。例如“矩形有 4 条边、4 个直角”，那么任意给定的矩形必有这样的性质。

(5) 公理

本体中公理是公认的事实（或推理规则），是用来知识推理的。

4.2 本体模型的建立

从上节对本体的介绍中可以得知在结构上可以看成是一个五元组[30]，即 $O:=\{C,R,H^c,Rel,A^o\}$。这里的 C 和 R 是两个不相交的集合,其中:C 中的元素称为概念(concept)；R 中的元素称为关系(relation)；H^c 表示概念层次,即概念间的分类关系(taxonomy relation)；Rel 表示概念间的非分类关系(non-taxonomy relation)；A^o 表示本体公理(axiom)。

本文第三章第三节中提到应用三元组来记录句子的结构信息，从而进行相似度计算。结合本体的特点，可以发现本体在设计上完全能满足存储句子信息的要求。因此，本文应用本体的特点，提出一种本体模型，用来存储句子的关键信息及其之间的关系。

句子中各类关键信息可以抽象的理解成本体元素中的概念，关键信息之间的关系可以理解成本体元素中概念之间的关系。一个句子的信息可以通过这种方式转换成本体的概念及概念之间的关系来存储，并进行计算。下面就是对本体概念的具体的设计。

通过之前的分析，可知句子的关键信息主要集中在形容词、副词、动词、名词、数词、和字符串等词性，可以将这些词性作为本体的概念建立。每个概念下的个体就是对应的相应的词语。如：“语法”一词，它是名词，属于本体的名词概念。本体内概念之间的关系则可以用来表示句子的关键信息之间的关系。由此可以总结出六种关系，分别是：“之后有名词”，“之后有形容词”，“之后有副词”，“之后有动词”，“之后有数词”，“之后有字符串”等这六种关系。再以短语“语法单位”为例，“语法”和“单位”都是句子中的关键信息，它们都是名词概念的个体。它们之间存在相互联系，这个联系在本体中可以用本体内概念之间的关系来表示。即“语法”，“之后有名词”，“单位”。

为了能形象化的描述本体模型，下面提供了一种示意图作为参考，如图 4.1。

首都师范大学信息工程学院

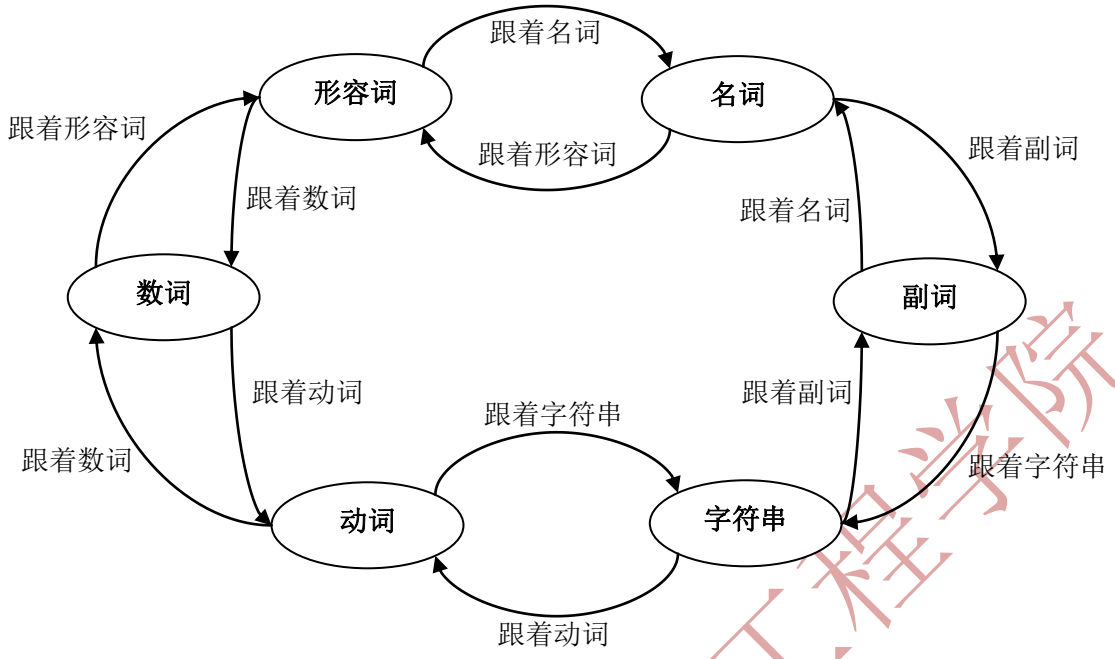


图 4.1 本体模型示意图

图中椭圆图形代表本体的概念元素，带箭头的曲线代表本体概念之间的相互关系，在曲线边上的文字是关系的具体内容。由图 4-1 可以清晰的看出本体是如何通过建立本体模型来表示句子的关键信息及关键信息之间的关系的。

4.3 基于本体模型的相似度计算

通过上节的本体模型可以将文本的句子转换成本体，进而进行相应的处理。本文主要是借助这种转换后得到的本体进行相似度计算，从而推断出两个文本之间的相似度。

先从总体分析，要计算两个文本之间的相似度，首先必须知道句子之间的相似度。要知道句子之间的相似度，就必需知道句子内词语之间的相似度和句子内句子结构之间的相似度。

对于句子内词语之间的相似度，本文在第二章里提出了一种结合词语字面，同义词词林，及《知网》的词语相似度计算方法。两个词语之间的相似度，有时体现在词语之间在字面上特别相近，有时体现在词语之间在语义上的同义性，有时还体现在词语之间在语意上的相近。这些层面上的相近都一定程度的反应了词与词之间的相似性，所以不能单独的考虑某种特定的相似。因此，本文提出一种同一的方式来处理词与词之间的相似度。取两个词之间最大的相似层面作为两个词之间的相似度，如第二章的公式 2-11 所示。

对于句子内句子结构之间的相似度，本文通过建立本体模型来进行处理。在本体模型中，句子的结构可以理解成句子内关键信息之间的相互关系，及本体内概念之间的关系。

首都师范大学信息工程学院

只要通过查找推理出句子内一个词语及与这个词语存在关系的所有其他词语，进行相似度的计算，最终得到整个文本中，围绕这个词的所有信息的相似度。将文本中所有词都做这样的处理，最终可以得到整个文本之间的相似度。

具体的算法流程如下：

1) 读取两个文本
2) 将两个文本进行分词、词性标注
3) 分别建立六个本体概念（如动词，名词等）
4) 分别建立六种本体概念之间的关系（如之后有动词，之后有名词等）
5) 对标注后的两个文本进行信息提取，建立两个本体（本体 1，本体 2）
6) 取本体 1 中一个概念下的个体 1
7) 取出本体 2 中对应的概念中的一个个体（个体 2）
8) 计算个体 1 和个体 2 的相似度（应用公式 2-11）
9) 如果个体 1 和个体 2 的相似度超过了一定的阈值，则执行下一步，否则执行第 7 步。
10) 在本体 2 中比对找出与个体 1 最相似的个体（个体 3）。
11) 在本体 1 和本体 2 中分别找出所有和个体 1 及个体 3 相关的所有信息。
12) 计算这些信息的相似度，记录下来
13) 在去取本体 1 中的下一个个体，重复 6-12 步的操作，直至遍历所有本体 1 中的个体。
14) 将每个个体得到的相似度进行汇总，加权求均值，得到最后的文本相似度。

通过以上的算法流程，可以将两个文本转换成两个本体，并且进行相似度计算。得到的最后的结果是一个 0 到 1 之间的小数，并且这个值越大表示这两个文本之间越相似。在自动判卷上，可以通过将这个值与预先预定的阈值进行比较，如果大于给定的阈值，则可以判定这两个文本相似，所表述的内容是一样的。即学生的答案和标准答案很接近，学生的答案正确。

下面通过一个例子来说明算法流程

给定两段文本：

① 隐马尔可夫模型的 5 个参数为状态数目、每个状态对应的可能观察值数目、初始状态概率分布、状态转移概率矩阵、观察值概率分布矩阵。

② 隐马尔可夫模型的 5 个参数为初始状态概率分布、每个状态对应的可能观察值数目、状态转移概率矩阵、状态数目。

首都师范大学信息工程学院

可以看出，两段文本表述的含义大部分是相同的，只是表达的时候语句的顺序不同，并且第二段文本少了一个参数。

按照算法的流程进行分析。首先读取这两段文本，之后对两段文本进行词性标注。其标注后的结果如下：

① 隐马尔可夫模型/bt 的/ude1 5/b 个/q 参数/cs 为/p 状态/n 数目/n 、/wn 每个/r 状态/n 对应/vi 的/ude1 可能/n 观察值/n 数目/n 、/wn 初始状态/n 概率分布/n 、/wn 状态/n 转移/vn 概率矩阵/n 、/wn 观察值/n 概率分布/n 矩阵/n

② 隐马尔可夫模型/bt 的/ude1 5/b 个/q 参数/cs 为/p 初始状态/n 概率分布/n 、/wn 每个/r 状态/n 对应/vi 的/ude1 可能/n 观察值/n 数目/n 、/wn 状态/n 转换/vn 概率矩阵/n 、/wn 状态/n 数目/n

之后根据词性标注后的两段文本，生成两个本体

本体 1：<ModelCom {http://www.sentiment.org/ontologies/example/#followA @rdf:type owl:ObjectProperty; http://www.sentiment.org/ontologies/example/#followM @rdf:type owl:ObjectProperty; http://www.sentiment.org/ontologies/example/#状态

@http://www.sentiment.org/ontologies/example/#followV

http://www.sentiment.org/ontologies/example/#转移;

http://www.sentiment.org/ontologies/example/#状态

@http://www.sentiment.org/ontologies/example/#followV

http://www.sentiment.org/ontologies/example/#对应;

http://www.sentiment.org/ontologies/example/#状态

@http://www.sentiment.org/ontologies/example/#followN

http://www.sentiment.org/ontologies/example/#数目;……}>

本体 2：<ModelCom {http://www.sentiment.org/ontologies/example/#followA @rdf:type owl:ObjectProperty; http://www.sentiment.org/ontologies/example/#followM @rdf:type owl:ObjectProperty; http://www.sentiment.org/ontologies/example/# 状 态 @http://www.sentiment.org/ontologies/example/#followN

http://www.sentiment.org/ontologies/example/#数目;

http://www.sentiment.org/ontologies/example/#状态

@http://www.sentiment.org/ontologies/example/#followV

http://www.sentiment.org/ontologies/example/#转换;

http://www.sentiment.org/ontologies/example/#状态

首都师范大学信息工程学院

@http://www.sentiment.org/ontologies/example/#followV

http://www.sentiment.org/ontologies/example/#对应; ……}>

因为本体比较大，这里给出两个本体的部分信息。从这两个片段可以看出，本体主要以类似于三元组的形式来存储。如[“状态”，“followN”，数目]，在句子中的表示是“状态数目”。

为了方便之后的运算，下面罗列两个本体中各个概念内的个体。

	形容词	动词	名词	数词	副词	字符串
本体1	无	转移，对应	矩阵，概率矩阵，概率分布，初始状态，观察值，可能，数目，状态	无	无	无
本体2	无	转换，对应	概率矩阵，数目，观察值，可能，状态，概率分布，初始状态	无	无	无

接下来是对两个本体的相似度计算，首先取本体 1 中动词概念下的一个个体“转移”。接下来需要在本体 2 的动词概念下找出与个体“转移”最相似的词语。查找的方式是分别计算两个词语在字面、同义词词林及《知网》上的相似度。这里在本体 2 中“转换”和“转移”是最相似的，经过公式 2-11 计算后得到的结果最大，为 1.0。

找出了最相似的词语后，接下来就是计算两个词语的相似度，及两个词语在本体内其他属性上的相似度。这里应用本体的推理机制，在两个本体中分别推理出和这个词语相关的所有信息。

“转移”在本体 1 中的所有信息有：{Name=转移，属于=动词, followN=概率矩阵}

“转换”在本体 2 中的所有信息有：{Name=转换，属于=动词, followN=概率矩阵}

结合原句分析，可以看出，转移和转换的后面都跟有名词概率矩阵。和本体推理得到的结果相符。在得到了两个词语在本体内的所有信息后，在对这些信息进行相似度计算。最终得到两个词在本体内其他属性上的相似度为 1.0。至此解决了本体 1 中“转移”个体的相似度计算。在从本体 1 中取出下一个个体“对应”，进行同样的操作。直至将本体 1 中所有个体的相似度都计算出来。

将所计算出来的所有个体相似度进行汇总，可以得到两个文本在形容词、动词、名词、数词、副词和字符串上的相似度。将这六个相似度进行加权平均，最终得到两个文本的相似度。在这个例子中，两个文本大部分是相同的，只是本体 2 少了个参数，而本体 2 中名词概念下也少了个个体“矩阵”。两个本体对“概率分布”个体的推理结果也有所区别。分别是：{Name=概率分布，属于=名词, followN=矩阵}和{Name=概率分布，属于=名词}。由此导致两个本体在名词概念上得到的相似度为 0.62。经过最终的加权平均，得到两个本体的相似的为 0.8433。

首都师范大学信息工程学院

可以看出，这样的结果与预期的结果很相近，符合实际情况。

首都师范大学信息工程学院

第五章 自动判卷系统的实现

5.1 系统背景介绍

自动判卷系统是对基于本体模型的相似度计算方法的一个应用。本文主要是对基于《自然语言处理》课程的在线学习系统的实现。这个在线学习系统强调提高学生的自主学习能力，为学生对《自然语言处理》课程的学习提供帮助。在在线学习系统中，设有学生的测试单元，学生可以在其中进行习题的练习，也可以进行在线的测试。

学生在进行习题练习时，就需要一个能自动判卷的系统来帮忙，为其评判自己在这个练习上是否正确的回答了问题，并且，希望能给自己在学习上提供相应的复习意见，方便自己的学习。

鉴于学生在这上面的需求，结合本文对文本相似度计算的方法，设计出一个能自动判定学生答案的自动判卷系统。实时的对学生的答案进行自动判定，并且给出相应的复习意见。

5.2 相关技术

“基于《自然语言处理》课程的在线学习系统”是一个完整的网站。它应用了 Java Web 网站开发技术，基于 SSH 框架，结合各种网站开发技术进行设计开发的。

5.2.1 SSH 框架

SSH 是一种多层 Web 应用程序架构，将 Struts2、Hibernate 和 Spring 这 3 个当前流行的 Java Web 开源框架进行整合，从而构建出灵活、易于扩展的多层 Web 应用程序。

1）Strut2 框架技术

Struts 是最早的 Java 开源框架之一，该框架是基于 Model2 的架构，也就是基于 MVC 模式的框架技术。它是一个免费的开源的 Web 层的应用框架，具有很高的可配置性，和有一个不断增长的特性列表。一个前端控制组件，一系列动作类，动作映射，处理 XML 的实用工具类，服务器端 Java Bean 的自动填充，支持验证的 Web 表单，国际化支持，生成 HTML，实现表现逻辑和模板组成了 Struts 的灵魂

2）Spring 框架技术

Spring 为系统提供 IOC 容器，将对象之间的依赖关系交由 Spring 进行控制，避免硬编码所造成的过度程序耦合。Spring 框架是轻量级的，能实现控制反转和面向切面编程的

首都师范大学信息工程学院

开源框架，由 7 个定义良好的模块组成．Spring 真正实现了逻辑层和应用层的分离，从而有助于我们更好地降低程序之间的耦合度，降低系统维护的难度。

3）Hibernate 框架技术

Hibernate 是一个优秀的持久化框架，负责简化将对象数据保存到数据库中或从数据库中读取数据并封装到对象的工作。通过简单的配置和编码即可替代 JDBC 繁锁的程序代码。是一个开放源代码的对象关系映射(ORM) 框架，它对 JDBC 进行了非常轻量级的对象封装，使得 Java 程序员可以随心所欲地使用对象编程思维来操纵数据库。

5.2.2 MyEclipse 简介

在线学习系统是在 MyEclipse 开发平台上进行开发设计的。MyEclipse 是一种企业级工作平台（MyEclipse Enterprise Workbench ，简称 MyEclipse）。是 Eclipse 集成开发环境的扩展，通过它可以在不同数据库和 JavaEE 上进行开发并且发布，整合应用程序服务器方面，它能极大提高工作效率。功能齐全的 JavaEE 集成开发环境，附带了优良的编码、测试和解析功能。

MyEclipse 平台具有很多优点，主要分为以下几类：

1) JavaEE 模型
2) WEB 开发工具
3) EJB 开发工具
4) 应用程序服务器的连接器
5) JavaEE 项目部署服务
6) 数据库服务
7) MyEclipse 整合帮助

5.2.3 其他技术

除了 SSH 框架和 MyEclipse 开发平台外，在线学习系统还应用了许多技术，如 Ajax，Jsp 等网站开发技术，也应用了许多平台，如 Flash，Dreamweaver。应用 Flash 平台能制作出 Flash 动画，建立多媒体学习平台，为在线学习系统增添乐趣，Dreamweaver 能对网站的界面进行设计修改，改善在线学习系统的界面外形。

5.3 功能展示

通过以上的技术，我们具体实现了《自然语言处理》课程的在线学习系统。其中，我主要负责对自动判卷系统的实现。它主要包括对学生习题的自动判定，及对学生考试后的

首都师范大学信息工程学院

测试卷进行自动判定。同时将判定结果进行汇总，给出相应的结果，包括正确与否，标准答案和学生答案，及最终的复习意见。为了形象的显示自动判卷系统的处理流程，下面我首先给出自动判卷系统的演示程序，演示自动判卷系统如何根据学生答案及标准答案，求出它们之间的相似度，最终给出相应结果。如图 5.1。

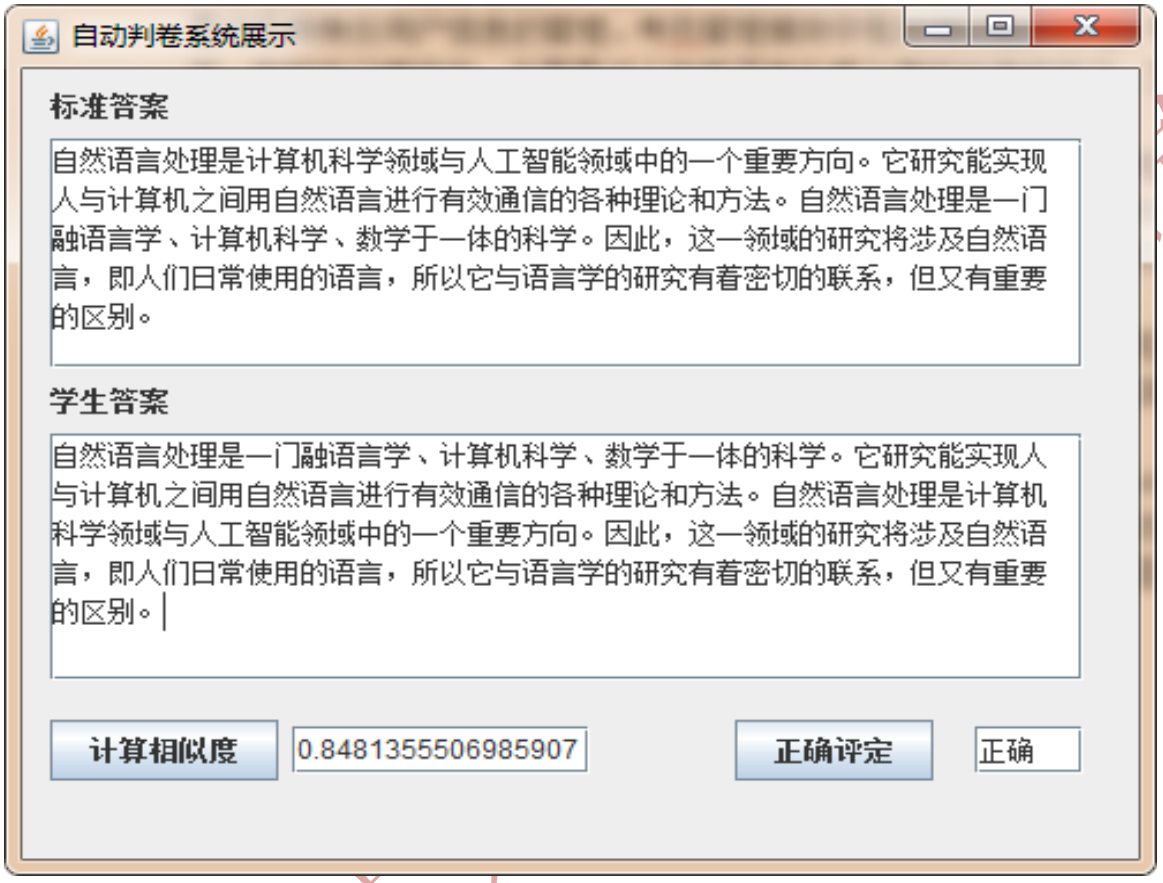


图 5.1 自动判卷系统演示程序

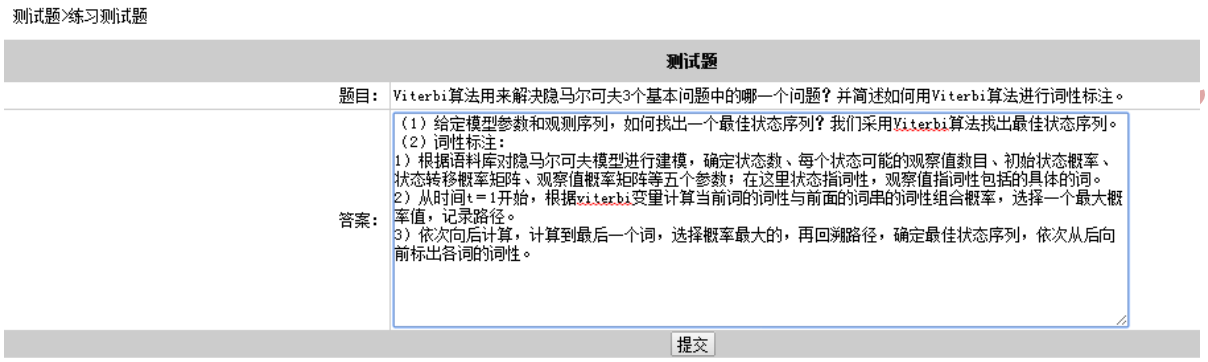
从图 5.1 可以看出，总共有两个文本，一个是标准答案，一个是学生答案。而学生答案和标准答案大体上是一样的，只是在句子的顺序上有着区别。以我们人的感官来说，这两个答案是很相似的，学生的答案应该是正确的。而从图中给出的计算结果可以看出，两个文本之间的相似度达到了 0.848，是个比较高的数值，和预期的结果相似。最终判定学生的答案是正确的。

通过这个演示程序，也可以证明基于本体模型的相似度计算方法在处理两个文本之间的相似度上有着其突出的优点。对两段相似的文本，虽然在语句的顺序上可能存在偏差。但是只要两段文本描述的内容是一样的，那么它们之间的相似度也就会比较高。

结合实际情况分析，学生给定的答案和标准答案往往有着很大的不同。它们在语句的顺序和语句内非关键信息的应用上存在很大的差异。但是，只要它们描述的关键信息是相同的，那么，就可以认为学生对题目有着正确的认识，给出了正确的答案。基于本体模型

的相似度计算方法忽略了句子中的非关键信息，并且记录了关键信息及其之间所有的相关关系。语句顺序的不同不会影响最终的结果。因此，基本本体模型的相似度计算方法能有效的解决自动判定方法，为自动判卷系统的实现提供理论基础。

下面是对自动判卷系统的具体实现后的功能展示，主要是对习题和测试两个单元的功能展示。如图 5.2，5.3，5.4，5.5：


测试题>练习测试题

测试题	
题目：	Viterbi算法用来解决隐马尔可夫3个基本问题中的哪一个问题？并简述如何用Viterbi算法进行词性标注。
答案：	（1）给定模型参数和观测序列，如何找出一个最佳状态序列？我们采用Viterbi算法找出最佳状态序列。 （2）词性标注： 1）根据语料库对隐马尔可夫模型进行建模，确定状态数、每个状态可能的观察值数目、初始状态概率、状态转移概率矩阵、观察值概率矩阵等五个参数；在这里状态指词性，观察值指词性包括的具体的词。 2）从时间t＝1开始，根据viterbi变量计算当前词的词性与前面的词串的词性组合概率，选择一个最大概率值，记录路径。 3）依次向后计算，计算到最后一个词，选择概率最大的，再回溯路径，确定最佳状态序列，依次从后向前标出各词的词性。
提交	

图 5.2 简答题练习题目

测试题练习结果	
题目：	Viterbi算法用来解决隐马尔可夫3个基本问题中的哪一个问题？并简述如何用Viterbi算法进行词性标注。
结果：	回答正确
正确答案：	（1）给定模型参数和观测序列，如何找出一个最佳状态序列？我们采用Viterbi算法找出最佳状态序列。（2）词性标注：1）根据语料库对隐马尔可夫模型进行建模，确定状态数、每个状态可能的观察值数目、初始状态概率、状态转移概率矩阵、观察值概率矩阵等五个参数；在这里状态指词性，观察值指词性包括的具体的词。2）从时间t＝1开始，根据viterbi变量计算当前词的词性与前面的词串的词性组合概率，选择一个最大概率值，记录路径。3）依次向后计算，计算到最后一个词，选择概率最大的，再回溯路径，确定最佳状态序列，依次从后向前标出各词的词性。
您的答案：	（1）给定模型参数和观测序列，如何找出一个最佳状态序列？我们采用Viterbi算法找出最佳状态序列。（2）词性标注：1）根据语料库对隐马尔可夫模型进行建模，确定状态数、每个状态可能的观察值数目、初始状态概率、状态转移概率矩阵、观察值概率矩阵等五个参数；在这里状态指词性，观察值指词性包括的具体的词。2）从时间t＝1开始，根据viterbi变量计算当前词的词性与前面的词串的词性组合概率，选择一个最大概率值，记录路径。3）依次向后计算，计算到最后一个词，选择概率最大的，再回溯路径，确定最佳状态序列，依次从后向前标出各词的词性。
涉及知识点：	隐马尔可夫模型
返回	

图 5.3 简答题练习结果

图 5.2 和图 5.3 是对习题中简答题类型的题目自动判定的展示。由图中可以看出，学生首先通过选择相应的测试题，进入到测试题练习界面进行测试题的练习。在给定的答案输入框内输入学生答案，之后将结果提交上去。在后台会得到学生给定的答案，同时会在数据库中提取出已有的标准答案。将两个答案通过基于本体模型的相似度计算方法计算得到结果。如图 5.1 中演示程序所展示的一样，将得到的结果与阈值进行比较。最终判定学生的答案是否正确。从图中可以看出，学生答案和标准答案基本一致，因此最终的结果也是显示为回答正确。

从图 5.3 里可以看出，在最终的结果显示上，除了会给出答案的结果外，还会给出学生的答案和标准答案，方便学生对两个答案进行比较，总结出自己的出错的地方。同时，还会对这个题目中所涉及的知识点进行罗列，如这道题目中涉及到的知识点是隐马尔科夫模型。那么最终结果中会将这个知识点罗列出来，学生可以通过点击这个知识点，跳转关于这个知识点的相关简绍中。学习所有和这个知识点相关的内容。

测试题的自动判定是自动判卷系统的一个具体实现，而测试卷的自动判卷则是在测试题的自动判定的基础上设计完成的。由于一张测试卷存在许多的测试题，因此，对于一张测试卷的判定，需要同时对多个测试题进行判定。在在线学习系统中，测试题的题型被设定为四种题型，分别是“单选题”，“多选题”，“填空题”，和“简答题”。每种题型都有自己的判定方式。最简单的是单选题，可以通过直接的字符匹配就能完成，多选题需要将所有学生的选项进行汇总，将汇总的结果和标准答案进行比对，得到最终的结果。填空题可以通过本文在第二章中基于字面的相似度计算方法进行计算，判定学生答案是否正确。简答题则是通过基于本体模型的相似度计算方法进行计算。下面给出一个具体的例子，通过图片的形式来展出。如图 5.4，5.5



图 5.4 学生开始考试



图 5.5 考试结果

从图 5.4 和图 5.5 可以看出，学生在开始考试的时候，系统会列出这个试卷内所有的测试题。学生对所有的测试题进行了简答，之后将它提交了上去。后台在得到学生的测试结果后，会对学生的每个题目的答案进行分析，和数据库中的标准答案进行比对，最终得到最后的结果。

5.4 核心代码

自动判卷系统最核心的部分就是对两个文本进行基于本体模型的相似度计算。基于本体模型的相似度计算的具体流程在本文的第四章中有所简绍。主要是通过将两个文本转换成两个本体，再进行本体之间的相似度计算，得到的结果与阈值进行比较。最后来判断答案的正确性。下面给出自动判卷系统在实现是的核心代码。

基于本体模型的相似度计算主要由 Similarity 类来处理。其中包括各种相似度计算函数。其类图如下图 5.6

首都师范大学信息工程学院



图 5.6 Similarity 类图

从图中可以看出，有两个本体，分别对应两个文本。还有许多的处理函数，如 getCiLinSim 是为了获得两个词语之间在同义词词林上的相似度，相应的还有 getLiuQunSim 和

getZiMianSim，分别表示两个词在《知网》语意上及字面上的相似度。基于词林和基于《知网》的词语相似度计算方法比较复杂，在本文的第二章有所介绍。对于它们的实现也是本文在处理词语相似度计算上的关键。下面给出它们的主要代码。

```
/**
 * 返回word1和wrod2经过同义词林方法计算相似度后得到的值
 * @param word1
 * @param word2
 * @return
 */
public double getCiLinSim(String word1,String word2){
    return Cilin.getInstance().getSimilarity(word1, word2);
}

/**
 * 返回word1和wrod2经过刘群基于知网方法计算相似度后得到的值
 * @param word1
 * @param word2
 * @return
 */
public double getLiuQunSim(String word1,String word2){
    return LiuConceptParser.getInstance().getSimilarity(word1, word2);
}

/**
 * 返回word1和wrod2经过字面上字符匹配方法计算相似度后得到的值
 * @param word1
 * @param word2
 * @return
 */
public double getZiMianSim(String word1,String word2){
    return new CharBasedSimilarity().getSimilarity(word1, word2);
```

在 Similarity 类图中，除了这三个函数外，还有像 getBestSimWord 函数，是为了获得本体中与词最相近的个体。getASim、getDSim、getMSim 等分别是为了获得本体在形容词，副词，数词等上面的相似度，getOwlSim 获得两个个体在本体关系上的相似度。

在处理相似度上，首先需要读取文本，之后对文本进行分词标注，然后构建本体。以上的操作都是通过 OwlBuild 类来处理的。OwlBuild 类的类图如下图 5-7 所示，其中分词标注使用的是中科院的分词软件。

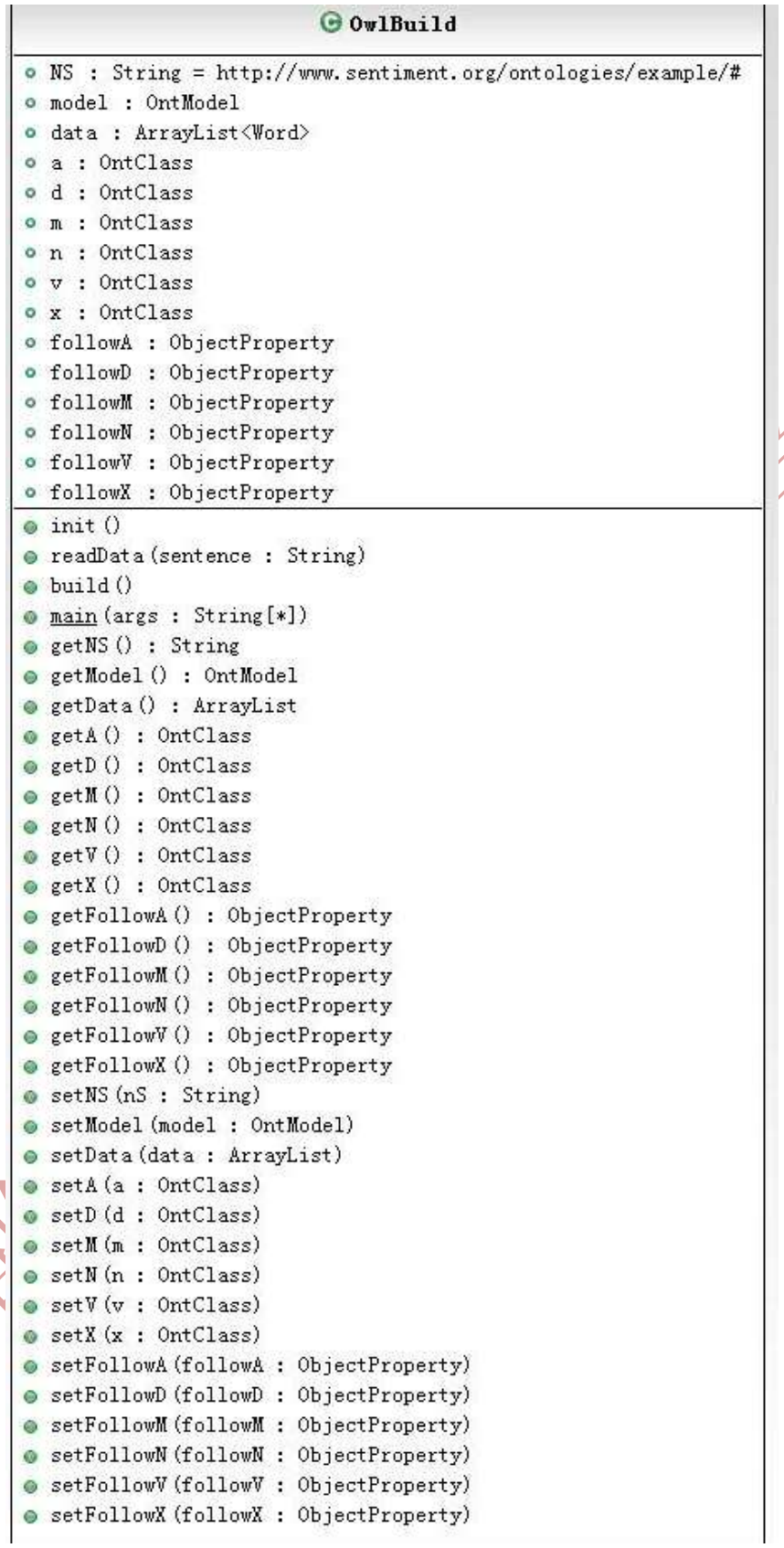


图 5.7 OwlBuild 类图

OwlBuild 中，A，D，M，N，V，X 分别代表本体的概念形容词，副词，数词，名词，

动词和字符串。FollowA，FollowD 等分别代码本体内概念之间的关系。Data 是用来存储读入的文本数据的。Model 为最后生成的本体。

建立本体的过程如下，首先通过 readData 函数来读取文本数据，并且进行词性标注。之后通过 build 函数来建立本体。

Similarity 类里的 BuildOwl 函数中完成了文本的读取和本体的建立。主要是通过调用 OwlBuild 类中的函数来实现的。建立了本体后，分别计算得到本体内各个概念下的个体的相似度，最终得到概念的相似度。以形容词概念的相似度获取为例，其具体代码如下：

```
/**
 * 获得两个本体在形容词上的相似度
 * @return
 */
public double getASim(){
    double asim=0.0;//形容词相似度
    double tempsim=0.0;//辅助相似度
    for(int i=0;i<astr1.size();i++){
        String word1=astr1.get(i);//读取一个形容词
        String bestsimword=this.getBestSimWord(word1, astr2);//在另一个本体中读取与这个形容词最相近的词
        double bestsim=this.getBestSim(word1, bestsimword);//计算出两个词的相似度
        double owlsim=this.getOwlSim(word1, bestsimword);//计算两个词在本体内所有关系的相似度
        tempsim=tempsim+bestsim*owlsim;//对所有的形容词的相似度进行汇总
    }
    if(astr1.size()>0){
        tempsim=tempsim/astr1.size();//得到最终形容词的相似度
    }
    asim=asim+tempsim;
    return asim;
    }
```

以上是形容词相似度获得的方法，动词、名词等相似度计算方法和这个类似，经过这些函数的调用，能分别算出本体在形容词、动词、名词等词性上的相似度。最后通过 getSim 函数对各类词性的相似度进行加权平均，最终获得两个本体的相似度。

第六章 结论与设想

6.1 结论

通过几个月的工作，基本完成了自动判卷系统的设计与实现了。自动判卷系统的实现主要依据的是基于本体模型的相似度计算。这也是本文的核心。

基于本体模型的相似度计算为文本的相似度计算方法提供了新的处理方式。通过将文本转换成本体，然后再进行本体之间的相似度计算，进而得到文本之间的相似度。从中可以看出本体是一种工具，用来存储文本的信息。在计算词语的相似度的时候，则是应用了公式 2-11，将字面，同义词词林，《知网》三种词语相似度计算方法进行有机的结合。基于本体模型的相似度计算注重文本在句子关键信息上的相似度，及关键信息之间联系的相似度。与句子在文本的位置无关，与句子中非关键信息无关。由此最大程度的降低了无关信息对文本相似度结果的影响，也忽略了句子在文本位置不同对计算结果的影响。

6.2 收获与总结

本课题是对自动判卷系统的设计与实现。经过这个课题的实践，我对文本的相似度计算有了深入的了解。了解了本体，及各种相似度计算方法。同时，还了解了许多网站开发的计算，熟悉了各种网站开发平台。是对课堂所学知识的综合实践。

基于《自然语言处理》课程的在线学习系统是一个大课题，包含多个子课题，是由一个小组完成的。在完成这个大课题的时候，我们小组互相合作，一起探讨，最终完成了整个课题的设计与实现。在期间我们深深的感受到了小组合作的好处，相互学习，共同进步。

6.3 设想与展望

基于本体模型的相似度计算除了在自动判卷系统上可以应用外，还可以扩展到其他的领域。如论文的查重，作弊的检测都能有所应用。此外，也可以将这个方法应用到基因序列的相似度比对上。只需将本体模型中的概念及概念间的关系做相应的变换，就能将之转换成对基因序列的相似度计算上。不过，基于本体模型的相似度计算还存在一些不足。

1) 本体模型内的概念需要一定的领域知识作为基础。不是统一的模型。

2) 不能区分出句子内并列词语之间位置的变换。

这些不足目前还在探讨解决的过程中。

首都师范大学信息工程学院

参考文献

[1] 陈利人，陈群秀. 基于实例的日汉机器翻译中的句子相似度计算研究,计算语言学进展与应用.
[2] 张民，李生，赵铁军，周明. 一种汉语句子间相似度的度量算法和实现. 计算语言学进展与应用，1995：152-158.
[3] 穗志方，俞士汶. 基于骨架依存树的语句相似度计算模型. 中文信息处理国际会议（ICCIP'98），1998.
[4] Lambros Cranias, Harris Papageorgiou, and Stelios Piperdis. 1994.
[5] Nirenburg, S., C. Domashnev and D. J. Grannes. "Two Approaches to Matching in Example-Based Machine Translation." Proceedings of TMI-93, Kyoto, 1993.
[6] Lin, Dekang. An Information-Theoretic Definition of Similarity[C]. In Proceedings of the 15th International Conference on Machine Learning table of contents, San Francisco, 1998
[7] 刘群，李素健. 基于《知网》的词汇语义相似度计算[C]. 第三届中文词汇语义学研讨会，中国台北，2002
[8] R. Richardson, A. F. Smeaton, J. Murphy. Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words 1994.
[9] William D. Lewis. Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity.
[10] Richardson, R., Smeaton, A. F., and Murphy, J. Using WordNet for conceptual distance measurement. Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group.
[11] 鲁松，白硕. 词距离的计算方法.自然语言理解与机器翻译，2001：401-409.
[12] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer（1991），Word-Sense Disambiguation Using Statistical Methods，ACL-91.
[13] Lillian Lee. Similarity-Based Approaches to Natural Language Processing. PhD thesis, Harvard University, Cambridge, MA, 1997.
[14] 关毅，王晓龙. 基于统计的汉语词汇间语义相似度计算. 语言计算与基于内容的文本处理，2003：221-227，清华大学出版社.
[15] 李涓子，黄昌宁. 一种自组织的汉语词义排歧方法.中文信息学报，1999，13（3）：1-8.
[16] 吴志强. 经济信息检索后控词表的研究[D]. 南京：南京农业大学硕士学位论文，1999
[17] 朱毅华，侯汉清，沙印亭. 计算机识别同义词的两种算法比较与测评[J]. 中国图书馆学报，2002，28(140):82-85
[18] 王源. 后控规范的计算机处理[J]. 现代图书情报技术，1993，(02)
[19] 宋明亮. 报纸文献机助自由标引研究及对后控制词表动态维护的思维[D]. 上海：空军政治学院硕士学位论文，1994
[20] 孔翔勇. 基于知网的汉语词语相似度计算[D]. 哈尔滨：哈尔滨工业大学硕士学位论文，2002
[21] G.Leusch, N.Ueffing, H.Ney. A novel string-to-string distance measure with applications to machine translation evaluation. Machine Translation Summit Ⅸ,New Orleans, Louisiana, USA, 2003.
[22] 梅家驹,竺一鸣,高蕴琦,等.同义词词林[M].上海:上海辞书出版社, 1993: 106-108.
[23] Chatterjee, Niladri. A Statistical Approach for Similarity Measurement Between Sentences for EBMT[C]. In Proceedings of Sysposium on Translation Support Systems(STRANS-2001), 2001: 278-294
[24] 秦兵，刘挺，王洋，郑实福，李生. 基于常问问题集的中文问答系统研究.哈尔滨工业大学学报，

首都师范大学信息工程学院

2003，35（10）：1179-1182.
[25] 郭艳华，周昌乐. 一种汉语语句依存关系网协动生成方法研究. 杭州电子工业学院学报，2000，20(4)：24-32.
[26] 李彬，刘挺，秦兵，李生. 基于语义依存的汉语句子相似度计算.计算机应用研究，2003，12：15-17.
[27] 车万翔, 刘挺, 秦兵,李生. 基于改进编辑距离的中文相似句子检索. 高技术通讯, 2004,(7):15-19.
[28] Gruber TR. A translation approach to portable ontology specifications. Technical Report, KSL 92-71, Knowledge System Laboratory, 1993.
[29] Chaudhri V K, Farquhar A, Fikes R, et al. Open Knowledge Base Connectivity 2.0. Knowledge Systems Laboratory, 1998
[30] Maedche A. Ontology Learning for the Semantic Web. Boston: Kluwer Academic Publishers, 2002.

首都师范大学信息工程学院

致　谢

这次的毕业设计项目对我来说是非常有意义的，刚刚选定题目的时候，我就想作为即将毕业的最后一个作品，一定要好好完成。

在制作毕业设计的这几个月过程中，使我受益颇多。首先，我学会了如何开展工作，如何做准备工作，接触新的事物并自学如何使用一个新的软件。其次，在遇到问题的时候要冷静，要全方位的思考，要寻求解决办法，多查阅资料，多在网上查找好的方法，多和他人探讨互相学习。另外，由于基于《自然语言处理》课程的在线学习系统是一项比较庞大的系统项目，工作态度要严谨、要细心，对各项任务要有条理的进行，小组成员要分工合作。养成良好的工作习惯和团队精神不仅是完成毕设所需要的，更是我们以后工作学习中不可缺少的。最后，理论源于实践，只有将理论应用于实践才能有更好的价值，要将理论和实践结合起来，不断探索，不断求知，才能更好的完成需要的工作。

在毕业设计的过程中，我的导师刘杰老师给了我巨大的帮助，不仅给我提供大量的参考资料，为我们的毕设提供了指导思想，在我们遇到问题时悉心指导并给出了很多宝贵意见，使得我的毕业设计任务能够顺利完成，在此我衷心感谢刘老师的指导与帮助。

身为毕业设计小组的一员，我对小组的其他成员表示深深的感谢。在毕业设计的过程中，他们让我学到了很多，也帮助了我很多。我们一起工作，一起探讨，一起解决遇到的问题，让我从中受益匪浅。

首都师范大学信息工程学院