学校编码：10384　　分类号_____密级_____
学号：200440020　　UDC _____

厦门大学

硕 士 学 位 论 文

全宋词语料库建设及其宋词风格与情感分析的计算方法研究

Research on the Establishment of Song Dynasty Poetry Corpus and the Computational Methods of Style Identification and Emotion Analysis

苏 劲 松

指导教师姓名：周 昌 乐 教授

专 业 名 称：计算机应用技术

论文提交日期：2007 年 5 月

论文答辩时间：2007 年 月

学位授予日期：2007 年 月

答辩委员会主席：_____

评 阅 人：_____

2007 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（√）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

目　录

Contents

摘　要

计算诗学是计算机自然语言处理技术的一个全新应用领域，其主要内容是建立诗词语料库，采用现代自然语言处理中的技术来挖掘语料库中所蕴含的信息，以此来辅助文学工作者们对诗词进行研究。本文以计算机辅助研究宋词为目的，建立全宋词语料库，并在此基础上开展了对宋词风格和情感分析的计算方法的初步研究。主要内容如下：

由于机器学习和古典文学数字化的需要，本文提出的方法和实验研究都基于语料库的数据驱动进行。语料库建设工作主要包括：基于统计抽词建立词表，结合格律特点对宋词进行切分，对宋词进行词性等方面的标注。同时，本文还建立了相关宋词知识库。该方面工作是整个课题的研究基础，具有十分重要的意义，主要内容集中在第二章到第五章。

针对高度抽象的艺术概念“风格”的辨析，本文将该问题转化为模式识别中的文本分类问题。在前面工作的基础上，分别建立了基于“字”特征和基于“词”特征的分类模型，并且通过遗传算法训练权值，建立两个模型的线型组合模型。在实验中，本文在 KNN 下比较了三个模型的优劣。这部分内容集中在第六章。

宋词中包含着诗词作者丰富的感情表达。针对该方面的研究，本文尝试将情感计算引入到宋词的机器理解中。文中采用了多重松弛迭代计算方法，对宋词词语的情感标注问题进行了研究，通过语境的利用，构建了一个实验性系统并取得了较为准确的词语情感标注，为以后的词句情感意义的理解提供了基础。该部分内容集中在第七章。

在文章的最后，对全文的研究工作进行了总结，并规划了今后进一步的研究方向。

关键词： 计算诗学；宋词辅助研究；语料库；风格评判；情感标注

Abstract

Computational Poetry is a new application field for computer natural language processing technology. The major work is to establish the poem corpus for the application of NLP knowledge mining in the assistant research on the poetry. To help us to know the Song Dynasty poetry, in the present paper, we set up the correlative annotated corpus and develop the primary research on computational methods of the style identification and emotion analysis. The main research outputs are as follows:

The methods and experimental study introduced in this paper are all driven based on corpus database because of the needs of machine learning and digitalization of classical poetry. The corpus work involves the establishment of word list with statistical word extraction, segmentation of poetry based on the foregoing word list and rules and forms, annotation on the segmented poetry such as part-of-speech tagging and so on. Meanwhile, we also set up the correlative knowledge database. It's very import to establish the corpus well, because the whole research project bases itself on it. This part is explained from Chapters2 to Chapters5 in this thesis.

Aiming at the differentiation of highly abstract artistic conceptual styles, we convert this problem into text categorization problem of Pattern Recognition. Basing on the above work, we found three categorization modules, including the module base on the "character" feature, the module base on the "word" feature, the linear combined module of fore modules which get the weight by the genetic algorithm. We compare these three modules with the KNN in this part, which is introduced in Chapters6.

About the assistant research on the authors' emotions which are expressed in the poetry, we try to import the emotions and their computation to the machine understanding of poetry. The problem of the emotional meaning tagging was studied by using the multiple relaxation alternate algorithm. We designed an experimental system and obtained accurate emotion tagging matching according to the context

information. It is the first step to the machine understanding of emotional meaning of poetry, thus the foundation for the future research is laid. The part of the content is covered in Chapters7 of this paper.

The summaries and conclusions of the research work, as well as the suggestion for the further researches come at the end of paper.

Keywords: Computational Poetry；Computer assistant research on Song Dynasty poetry；Corpus；Style Identification；Emotion Tagging

第一章 绪 论

1．1 前言

中国素来享有“诗国”之称，诗词是一种特殊文体的大众化文学形式，在汉语文化的成长、演变与传播中有着极重要的地位，她以独特的艺术形式，以恒久不衰的魅力成为中国文学的骄傲而流传千古，而其中的宋词作为宋代文学的典范，赢得了众多文人骚客的青睐，成为中国古代诗词中一颗璀璨的明珠。因此，通过对宋词进行研究进而了解宋代文化一直是文学工作者的一个研究热点。自古以来，对宋词的分析研究，往往都是具有丰富诗词写作的文学人士或者具有诗词美学研究功底的文学专家才能进行的。随着当代计算机技术的迅速发展，特别是在自然语言处理方面取得了巨大的进展，我们不禁想到“能否用计算机来辅助我们进行宋词研究，以此来加深我们对古典宋词的理解呢？”对此，本文采用现代自然语言处理中的一些技术，从计算机辅助研究宋词的角度出发，阐述了计算机辅助研究诗词的产生和应用背景、宋词语料库的建设、宋词风格的机器评判，宋词词语情感的机器标注等一系列研究工作。

1．2 论文中基本概念的界定

结合国内学者的相关研究成果[1][2]，本文对一些基本概念进行了界定：

1) 词义：词（词汇）的意思。是一个相对宽泛的概念，包含了人们通常所说的词义、隐喻义和引申义等。词义是决定如何界定词切分单位的主要标准。

2) 字：宋词研究的基本单位，在计算机系统中，“字”就是指给每一个汉字分配的唯一的机内码。与此相对应，本文中的“字”指的是字形相同的汉字符号。只要字型相同，即使读音、字义不同，也被认为是一个字。

3) 词：文中所指的“词”是一个广义的概念，可以称为“切分单位”，它既包括语义上独立的“词”，也包括语义上具有结构的“词组”。在具体语言处理中，词的定界应该是与应用相关的。因此可以说词是语言处理系统的基本单元。

4) 子句：宋词词句根据词的格律特征可以切分为子句，子句是词的连续连接组合。本文在这里提出子句的概念是为了方便宋词切分和语料库建设。

5) 宋词语料库建设：本文采用现代自然语言处理的若干技术，结合宋词本身的特点来进行宋词的计算机辅助研究，而其中宋词语料库的建设就是整个研究的基础，主要内容包括宋词“词”概念的基本界定、宋词数据库的整理、宋词的机器自动切分和人工校对、宋词生语料的标注等。

6) 宋词风格的评判：即“宋词的风格评判与分析”，本文将其视为关于风格的文本分类问题，尝试由计算机对宋词风格自动进行识别分类。文中主要以豪放派和婉约派宋词为研究对象。

7) 宋词词语情感的标注：即“标注出宋词词语在不同语境中的情感意义”，本文构建了一个实验性系统，利用上下文语境，采用多重松弛迭代计算方法，以切分后的词为单位进行情感标注，该方法为后续关于研究如何利用合一算法得出词句情感意义的奠定了基础。

1．3 相关领域已有的研究

20世纪80年代以来，随着计算机应用技术的不断发展，以语料库为基础的研究在语言学和计算机科学研究中都取得了丰硕的成果。无论是在语言学研究还是自然语言处理领域，语料库都已经成为重要的基础资源，发挥了越来越重要的作用。正是基于以上认识和技术条件的支持，人们开始建立了古代诗词语料库，运用自然语言处理技术结合古代诗词本身的特点来对诗词进行计算机辅助研究。在本文所涉及到的研究内容，我国学者已经开始了初步的探索，如厦门大学的周昌乐教授提出了“计算诗学”[3]的概念并开展了一系列相关研究工作；北大计算语言所与台湾元智大学古文献研究所合作开展了“古代诗词研究的计算机支持环境”的研究[4][5][6]；中科院自动化所的费越和重庆大学的易勇对计算机自动生成对联分别进行了探索[2][7]；重庆大学的李良炎提出了基于词连接的自然语言处理技术，并将其应用于诗词语言的理解[8]。在此，本文逐一作简单介绍：

1） 计算诗学概念的提出

厦门大学周昌乐教授在其著作《心脑计算举要》[3]中第一次提出了计算诗学的概念：使用计算思想、方法和技术等从事诗歌（推而广之，也可以包括其它文学形式）的研究工作，可以统称为计算诗学的研究。广义的计算诗学，可以包括许多方面的工作，主要是对诗歌文本的各种规律的研究，例如像诗歌机器分类、

诗歌风格的计算机辅助归纳、诗学知识的计算机辅助发现、诗歌创作的计算机辅助系统工具、诗歌用词用语的统计、诗学语料库、文献库等等。而狭义的计算诗学，则主要是指使计算机系统具备诗歌理解、欣赏和创作的能力，如诗歌作品的机器理解、计算机诗歌创作系统以及计算机歌曲创作系统等。以此为出发点，厦门大学艺术认知与计算实验室借助先进的人工智能理论与方法，开展汉语隐喻分析与理解研究、诗词计算分析与创作研究，以及诗歌机器翻译系统的开发等。

2） 唐宋诗的计算机辅助分析

北京大学计算语言学研究所运用计算语言学手段对中国古诗词进行研究，相关的研究成果能够对古诗词、古汉语领域的研究提供有益的帮助。从其研究成果来看，对古汉语计算语言学研究也为现代汉语的研究提供了一个新的视角，有利于从一个新的角度来观察现有的一些概念和问题。

在胡俊峰的博士论文“基于词汇语义分析的唐宋诗计算机辅助深层研究”[1]中，将一些现代计算语言学技术根据古诗词语言的特点加以改造，取得了一些有益的成果。其研究系统提取积累了有关中国古诗词的语料及语言信息知识库，为今后的研究奠定了良好的基础。总体而言，对古诗词的分析加工目前还只限于词汇与词汇共现一级，一些相关的应用如：词汇自动切分，相似词句检索技术等都是建立在这个基础上的。同时，由于古诗文体简短，大量使用隐喻，题材相对固定等，基于古诗词初步展开了有关篇章分析、意象分析、认知心理的计算语言学分析等研究。

3） 春联艺术的初探

中国科学院自动化研究所的费越在其博士论文“汉语语义的多层次集成研究——及春联艺术系统设计”[7]中采用神经网络的方法研究形象思维层次的“语义”，并用春联领域内的词语进行实验。在神经网络的学习过程中，语义的数值表现序列是从无序到有序的一个动态过程，在某种程度上类似于人类学习词语的过程。在采取格语法语义表示的基础上，文章提出了汉语处理的神经网络并行模型，在语义表示和并行模型的基础上，构造了六个汉字以内的计算机春联系统，例如上联“岁岁平安日”对得下联“年年如意春”。

重庆大学的易勇在其博士论文“计算机辅助诗词创作中的风格辨析及联语应对研究”[2]中，分析了传统对联的特点，将联语的应对生成问题抽象为有监督的

序列学习问题。将对联的上下联分别看作两个具有相同长度的语言单位的序列，采用机器学习方法对其进行学习。提出了不限字数的联语应对生成的计算模型，并分别用N元统计语言模型序列学习法、隐马尔可夫模型序列学习法和基于转换的驱动序列学习法对联语生成进行建模分析，构建了基于语料库不限字数的计算机联语应对实验系统，取得了较好的实验结果，如：针对庆祝神州五号载人飞船发射而出的上联“九天揽月，华夏英豪驰宇宙”对得下联“四海迎春，神州崛起舞天下”。

4) 基于词联接的自然语言处理技术

重庆大学的李良炎在其博士论文“基于词连接的自然语言处理技术及其应用研究”[8]中提出基于词联接的自然语言处理技术，并用于诗词语言的理解，提出了词联接最大语义符合度计算和最优句树搜索的初级语言分析算法，进行了诗词语料标注测试、诗词语言初级分析测试、诗词语言豪放与婉约风格的评价测试，取得了成功，在深入分析自然语言处理技术背景的基础上，提出并初步构建了基于词联接的自然语言处理技术（Term Connection Technique for NLP，简称TCT），并应用到诗词语言处理系统中。

1．4 课题的研究背景和主要内容

1．4．1 研究背景

本课题研究受国家自然科学基金项目“面向英汉机器翻译的汉语隐喻释义方法研究”的支持，主要开展了关于计算诗学方面的研究。

1．4．2 主要内容

图1中显示了本组课题研究的基本框架结构。可以看出，全宋词熟语料库建立、词表和其它相关知识库的建立是整个课题组研究的基础。