硕 士 学 位 论 文

基于深度学习的跨语言信息抽取研究

Research on Cross-language Information Extraction Based on Deep Learning

作 者 姓 名：陈 龙

学科、 专业：计算机应用技术

学 号：21309153

指 导 教 师：周惠巍 副教授

完 成 日 期：2016 年 06 月 11 日

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　要

基于机器学习的信息抽取方法性能依赖训练语料的质量和数量。然而标注数据在不同语言分布不均衡，阻碍了中文的信息抽取研究。针对这一问题，研究人员提出跨语言信息抽取方法，利用资源丰富语言（源语言）的标注数据来训练资源匮乏语言（目标语言）的信息抽取系统。然而，语言鸿沟和机器翻译错误影响了信息抽取的性能。本文研究基于深度学习的中英文跨语言信息抽取技术，主要包括以下内容：

（1） 基于降噪自动编码机（DAE）的双视图跨语言信息抽取

提出基于 DAE 的双视图跨语言信息抽取方法。在源语言和目标语言向量重构过程中，DAE 适当引入噪音，减少翻译错误对分类性能的影响。同时，在中英文两个视图，分别构建分类模型，融合两个模型的分类结果，减少语言鸿沟对分类性能的影响。在跨语言情感分类和跨语言模糊限制语识别两个任务进行实验，证明 DAE 模型和双视图方法均能有效提高跨语言信息抽取性能。

（2） 基于双语词表示的跨语言信息抽取

提出基于双语词表示的跨语言信息抽取方法，双语词表示的学习分为无监督和有监督两个学习阶段。无监督学习阶段利用 DAE 进行中英文双语词表示学习，捕获中英文双语语义信息。有监督学习阶段将训练语料的标注信息嵌入双语词表示，提高信息抽取性能。实验表明，双语词表示能够有效捕获双语语义信息和标注信息，克服双视图方法难以深入融合两种语言的不足。

（3） 基于联合表示学习的跨语言信息抽取

提出基于联合表示学习的跨语言信息抽取方法。采用长短时记忆递归网络（LSTM），学习中英文双语表示。在词语义表示基础上，引入上下文情感（模糊）信息表示，联合训练情感词（模糊限制语）在特定语境下的语义和情感（模糊）信息。实验表明 LSTM 能够有效实现跨语言信息抽取。同时，词语义表示与上下文情感（模糊）信息的联合表示学习能够进一步提高跨语言信息抽取的性能。

本文研究了基于深度学习的跨语言信息抽取方法，减少了翻译错误、语言鸿沟等对跨语言分类性能的影响，获得了有效的双语语义信息和标注信息，提高了跨语言信息抽取性能，为深度学习理论在跨语言的相关研究提供了有益借鉴。

关键词：跨语言信息抽取；双视图；深度学习；双语词表示；联合表示学习

Research on Cross-language Information Extraction Based on Deep Learning

Abstract

The performance of machine learning based information extraction system relies on the quality and quantity of training corpora. However, labeled data in different languages are very imbalanced. The lack of labeled data limits the research progress in Chinese and other resource-scarce languages. In order to solve this data imbalance problem, cross-language information extraction (CLIE) is proposed, which leverages resources in one language (source language) to improve the information extraction performance in another language (target language). However, the gap between the source language and target language limits CLIE performance. Besides, the errors introduced by machine translation systems inevitably affect the performance of CLIE system. This paper focuses on CLIE based on deep learning technique, the contents consist of the following three aspects:

(1) Two-view Cross-language Information Extraction Based on Denoising Autoencoders (Two-view DAE)

This paper adopts denoising autoencoders (DAE) for CLIE task. Noises are properly added to the training examples in DAE reconstruction, for enhancing the robustness to the translation errors. Meanwhile, we train the classifiers in English view and Chinese view respectively, and combine the outputs from the two views to obtain the final classification results. The two-view approach could make full use of the complementary advantages in English and Chinese, which bridges the language gap between English and Chinese. The experiments are conducted on cross-language sentiment classification (CLSC) and cross-language hedge cue detection (CLHCD) tasks. The experimental results show that both DAE and two-view approach are effective and could improve CLIE performance.

(2) Cross-language Information Extraction Based on Bilingual Word Representations (BWR)

This paper proposes an approach to learning bilingual word representations for CLIE. The learning process consists of two phases: unsupervised learning phase and supervised learning phase. In the unsupervised learning phase, DAE is used to learn bilingual word representations at the same time, capturing bilingual semantic information between the two languages. In the supervised learning phase, label information is integrated into bilingual word representations to improve CLIE performance. The experimental results on CLSC and

CLHCD tasks show that the learned bilingual word representations could effectively capture both bilingual semantic information and label information. This could overcome the problem that it is difficult for two-view approach to combine English and Chinese semantic information deeply.

(3) Cross-language Information Extraction based on Joint Representation Learning (JRL)

This paper adopts long short memory term RNN (LSTM) to learn word semantic and context information representations jointly for CLIE. In the learning process, we use word semantic representations and context sentiment (hedge) information representations to learn the semantic information and sentiment (hedge) information of sentiment words (hedge cue) in specific contexts. The experimental results show that LSTM could learn bilingual word semantic representations for CLIE effectively. Meanwhile, joint representation learning approach could further improve the performance of CLIE.

In this paper, we conduct the research on cross-language information extraction based on deep learning technique. Two-view DAE approach is proposed to enhance the robustness to the translation errors and bridge the language gap between English and Chinese. BWR approach is proposed to effectively capture both bilingual semantic information and label information. JRL approach is proposed to solve the data sparseness problem and to learn the latent semantic information. Those deep learning based approaches effectively improve CLIE performance, and could provide valuable references for future work on deep learning based CLIE research.

Key Words: Cross-language Information Extraction; Two-view Approach; Deep Learning; Bilingual Word Representation; Joint Representation Learning

目　　录

1 绪论

1.1 研究背景与意义

随着大数据时代的到来，每天数以亿计的文本数据在互联网上不断更新，这些数据对人们的生产生活起着至关重要的作用。单凭人力很难从海量数据中获取自己所需的信息。另一方面，计算机技术的迅猛发展给人们的生活带来诸多便利。人们迫切需要在计算机的协助下，从大量数据中挖掘知识，迅速获取相关信息。因此，信息抽取研究应运而生。

信息抽取（Information Extraction）研究的目的是从文本中抽取特定的事实信息[1]。例如，可以从产品评论、微博、微信中抽取信息，分析说话者的情感倾向；也可以在计算机的协助下，从科技文献中快速抽取相关知识，方便用户使用。自然语言处理（Natural Language Processing，NLP）研究利用计算机对人类的语言进行各种类型的处理。随着人工智能技术的发展和大规模真实数据的增长，基于机器学习方法的自然语言处理技术表现出无法比拟的优越性。

基于机器学习方法的信息抽取系统性能依赖于训练语料的质量和规模。目前，英文等一些语言的信息抽取研究较早，语料资源较为丰富，但中文等其他语言的研究起步较晚，缺乏足够高质量的语料资源，阻碍了这些语言的信息抽取研究。针对不同语言训练数据分布不均衡的问题，研究人员[2-4]提出利用英文等资源丰富语言（源语言，Source Language）的训练语料对训练资源匮乏的语言（目标语言，Target Language）进行信息抽取的方法，即跨语言信息抽取（Cross-language Information Extraction，CLIE）研究。

传统的跨语言信息抽取采用机器翻译系统将源语言的训练语料翻译成目标语言，作为目标语言的训练语料，训练获得目标语言分类模型，对目标语言的测试语料进行信息抽取[5]。这种方法简单易行，并且模型可以取得较好的分类性能。但中英文之间存在较大的数据差异，不可避免地影响了信息抽取的性能。而且，现有的机器翻译技术无法达到十分令人满意的效果。例句 1.1 源于亚马逊网站用户对某图书的评论，例句 1.2 是其对应的中文机器翻译结果。在例句 1.1 中，说话人觉得“bizarre”这个词用在这本书上是匪夷所思的，证明说话人对这本书的评价是消极的。而其翻译（例句 1.2）表达的意思是这本书很有趣，情感倾向变为积极的了。可见机器翻译错误，可能引起情感倾向的改变，影响跨语言信息抽取性能。

（例句 1.1）Bizarre is too interesting a word to use for this book.

（例句 1.2）奇怪的是，这本书很有趣。

近年来，深度学习技术（Deep Learning）[6]的兴起推动了图像识别[7]、语音识别[8]和自然语言处理[9]等多个人工智能领域研究的发展。深度学习基于一种深层非线性网络结构，实现复杂函数逼近，很好地实现高变函数等复杂高维函数的表示[10]。从仿生学角度来说，深度学习的原理与大脑皮层一样，分层对输入的数据进行处理，抽取其在不同层的信息，最终获得数据的本质特征[11]。

深度学习技术也被用于跨语言信息抽取任务，取得了较好的性能[12-15]。基于深度学习的跨语言信息抽取首先基于机器翻译或平行语料获得两种语言的句子对（Sentence Pair）或文档对（Document Pair），然后将对齐的文档对输入到深度学习模型中，学习文档的共享表示（Shared Representations）。共享表示可以将源语言和目标语言的数据映射到一个公共的数据空间，并在这一空间进行训练和测试。

跨语言信息抽取是 NLP 领域中一项具有挑战性的任务。源语言和目标语言之间存在语言鸿沟，利用源语言训练样例学习获得的分类模型很难适用于目标语言的测试语料，影响了分类性能。同时，多数跨语言信息抽取方法采用机器翻译系统获得目标语言的训练语料，翻译产生的错误也会降低信息抽取性能。如何减少语言鸿沟和翻译错误对信息抽取性能的影响，是跨语言信息抽取研究的难点。

本文以中英文跨语言情感分类（Cross-language Sentiment Classification，CLSC）和中英文跨语言模糊限制语识别（Cross-language Hedge Cue Detection，CLHCD）两个任务为例，研究和探索基于深度学习的跨语言信息抽取技术。重点研究和探索基于降噪自动编码机的双视图跨语言信息抽取，基于双语词表示的跨语言信息抽取和基于联合表示学习的跨语言信息抽取技术。减少语言鸿沟和机器翻译错误对跨语言信息抽取性能的影响，深层挖掘特征的双语语义信息和上下文信息，构建高质量的信息抽取系统。

1.2 跨语言信息抽取研究现状

传统的跨语言信息抽取采用机器翻译系统将源语言的训练语料翻译成目标语言，作为目标语言的训练语料，训练获得目标语言分类模型，对目标语言的测试语料进行信息抽取。然而，现有的机器翻译技术无法获得高质量的目标语言训练语料，影响了单语分类模型的性能。Li 等[5]基于交叉验证评价译文质量，选用高质量的翻译样例加入训练集，取得了比简单机器翻译更好的分类结果。为了克服语言鸿沟对分类性能的影响，Wan[2]将协同训练的方法（Co-training）用于中英文跨语言情感分类。首先利用 Google 翻译系统将英文训练语料翻译为中文，将中文未标注语料翻译为英文。然后在中英文两个视图中分别基于标注语料训练获得中、英文情感分类模型，再利用这两个分类模型对

中、英文未标注语料进行分类，选取高可信度样例加入训练集，迭代训练获得中英文分类模型。最后融合这两个分类模型的分类结果。Gui 等[16]提出了自训练（Self-training）和协同训练的混合模型，取得了比单独使用自训练或协同训练模型都好的分类性能。然而，上述方法虽然采用双视图的方法获得最终分类结果，但源语言和目标语言的分类模型是分别训练的，难以深入融合两种语言的语义信息。

此外，迁移学习（Transfer Learning）也用于跨语言信息抽取。迁移学习的目标是在不同数据分布的领域中，利用源领域（Source Domain）的知识辅助目标领域（Target Domain）学习模型，并对目标领域的数据进行分类[17-19]。研究人员将源语言看作源领域，将目标语言看作目标领域，利用迁移学习的方法实现跨语言信息抽取。Xu 等[17]基于 Transfer AdaBoost 和 Transfer Self-training 两种迁移学习算法，将高可信度的英译中翻译数据加入训练集中，迭代获得中文情感分类模型，最后对中文测试语料进行分类。Gui 等[18]提出一种 Transductive Transfer Learning 算法，在迁移过程中利用一种监测机制避免负迁移，降低翻译噪音对分类性能的影响。Chen 等[19]在迁移学习过程中，引入知识验证机制，过滤掉机器翻译噪音，在 NLPCC 2013 的“跨语言情感分类”共享任务数据集上取得 83.59%的分类准确率。基于迁移学习的跨语言信息抽取方法可以通过精细的算法过滤掉不可信的训练样例，提高分类性能，但这种方法容易产生过拟合的现象，很难推广应用于其他数据集。同时，当未标注数据与标注数据的分布差异较大时，迁移学习的方法难以取得令人满意的分类性能。

深度学习[6]在自然语言处理领域得到了广泛的应用。表示学习（Representation Learning）[9][11][20][21]是深度学习在自然语言处理领域的一种常用方法，即通过深度学习获得词、词性等特征的向量表示，用于 NLP 任务。这种方法不仅解决了传统“one-hot”表示方法带来的数据稀疏问题，而且可以挖掘词汇的语义等深层次信息。Zeng 等[20]基于卷积神经网络（Convolutional Neutral Network，CNN）[22]学习词汇的语义表示和位置信息表示。Xu 等[21]基于长短时记忆递归网络（Long Short Term Memory，LSTM）学习获得最短依存路径上的词序列表示。

近年来，研究人员将双语词表示（Bilingual Word Representations）用于跨语言信息抽取研究，取得了较好的效果。利用深度学习模型，基于双语平行语料，将源语言与目标语言数据映射到同一向量空间，训练获得双语词表示。这种方法已经被成功用于机器翻译[23]、跨语言文本分类[12-14]、跨语言序列标注[24]等 NLP 任务。Chandar A P 等[12]采用自动编码机（Autoencoder，AE）学习双语词表示，用于跨语言情感分类。首先分别学习源语言和目标语言的特定语言表示（Language Specific Representations）；然后将两种语言的特定词表示映射到同一向量空间，训练获得双语词表示。为了提高跨语言分类性

能，Chandar A P 等[13]将上述两个特定语言表示学习过程合并为一个双语词表示学习过程，即通过 AE 的双语重构（Bilingual Reconstruction）抽取双语语义信息。Rajendran 等[14]基于 AE 建立多视图模型，将多种语言的样例映射到同一空间，实现多语言分类任务。Zhou 等[4]基于堆叠式降噪自动编码机（Stacked Denosing Autoencoder，SDA）训练双语词表示，用于中英文跨语言情感分类，有效减小了语言鸿沟对分类性能的影响。然而，上述方法基于深度学习技术学习双语词表示，获得双语语义信息，但是没有将特定任务的标注信息加入双语词表示的学习过程。

Tang 和 Wan[3]提出一种双语情感表示（Bilingual Sentiment Representation）学习方法，用于跨语言情感分类。基于线性函数，学习双语语义信息。同时，将情感信息嵌入到双语词表示中，增强了双语词表示的情感表达能力。但他们的方法仅采用线性函数将源语言和目标语言的样例映射到同一语义情感空间，没有采用深度学习技术挖掘深层语义信息。

1.3 跨语言情感分类

情感分类（Sentiment Classification）是一种对含有情感色彩的文本进行分析，判断作者态度（支持或反对）的技术。英文等一些语言的情感分类研究较早，情感资源较为丰富，但中文的情感分类研究起步较晚，缺乏情感标注语料，阻碍了中文情感分类的研究。如何将英文情感资源应用于中文情感分类任务，即实现跨语言情感分类，引起了研究人员的关注。中英文之间的语言鸿沟会降低情感分类的性能，如何减小语言鸿沟对分类性能的影响是跨语言情感分类研究的重点和难点。

本文实验采用 NLPCC 2013“跨语言情感分类”评测任务语料，探索深度学习在中英文跨语言情感分类中的应用。该语料源自中英文亚马逊的产品评论数据，包括书籍、DVD 和音乐三个领域的篇章文档，以 XML 格式进行存储。图 1.1 为跨语言情感分类语料示例：图 1.1（a）为英文训练语料，图 1.1（b）为中文测试语料。语料库中各领域训练和测试数据数量如表 1.1 所示，其中英文训练语料正负样例比为 1:1。

```
<item>
        <summary>Possibly the worst book I have ever read</summary>
        <polarity>N</polarity>
        <text>While I realize that the majority of people thought this to be a stellar book,
                …
                It would be impossible to recommend this book to anyone</text>
        <category>books</category>
</item>
```

（a）英文语料

```
<item>
        <review_id>0056332</review_id>
        <summary>好难，看不懂</summary>
        <text>自认为C语言基础很好，结果根据评论买了C++primer打算学习
                …
                建议初学C++的童鞋不要买这本书了，真的不适合入门。</text>
        <category>book</category>
</item>
```

（b）中文语料

图 1.1　跨语言情感分类语料实例

Fig. 1.1　An example of CLSC corpora

表 1.1　NLPCC 2013 跨语言情感分类语料篇章数量统计信息

Tab. 1.1　The statistics of NLPCC 2013 CLSC corpora

领域 / 语料	书籍	DVD	音乐
英文训练语料	4000	4000	4000
中文测试语料	4000	4000	4000
中文未标注语料	47071	17814	29677

本文将英文训练语料及其中文翻译用于训练，将中文测试语料及其英文翻译用于测试。

首先采用准确率（*Accuracy*）分别评测书籍、DVD 和音乐三个领域的跨语言情感分类性能，然后采用平均准确率（*Average*）衡量三个领域的总体分类性能，计算方法如公式（1.1）所示。

$$\begin{aligned} Accuracy_C &= \frac{\#system\_correct_C}{\#system\_total_C} \\ Average &= \frac{1}{3}\sum_C Accuracy_C \end{aligned} \tag{1.1}$$

式中，$\#system\_correct_C$ 为被正确分类的样例数，$\#system\_total_C$ 为类别 C 的样例总数。

1.4 跨语言模糊限制语识别

模糊性是自然语言中的一种常见现象。Lakeoff[25]在 1972 年提出模糊限制语（Hedges）是“把事情弄得模模糊糊的词语”。由模糊限制语引导的信息为模糊限制信息（Hedge Information）。为了在信息抽取中获得真实信息，应将事实信息与模糊限制信息区分开来，因此，模糊限制语识别具有重要意义。

例句 1.3 源自于中文生物医学文献[26]，例句 1.4 和例句 1.5 均源自于英文生物医学文献[27]。其中，例句 1.3 和例句 1.4 是模糊限制性句子，而例句 1.5 是确定性句子。

（例句 1.3）在梯度幅值<*ccue*>较</*ccue*>大的区域使用各向异性扩散的 TV 约束方式，在梯度幅值<*ccue*>较</*ccue*>小的区域使用各向同性扩散的 Tikhonov 约束方式。

（例句 1.4）Interestingly, the MnlI-AluI fragment <*ccue*>could</*ccue*> suppress the basal-level activity of the conalbumin promoter in both Jurkat and HeLa cells.

（例句 1.5）No relationship could be established between glucocorticoid receptor binding and antidepressant medication.

中文例句 1.3 中“较”为模糊限制语，用于表示程度变动[26]。英文例句 1.4 中“could”是模糊限制语，用于表示可能性。英文例句 1.5 中同样包含“could”，但是却没有模糊性的含义。可见同一个词在不同的语境中，表达的模糊性不同。

本文将模糊限制语识别看作二值分类问题，首先基于模糊限制语语料库抽取模糊限制语词典（Candidate），然后基于词典匹配的方法筛选出训练语料中的候选词作为训练样例：标注为模糊限制语的候选词为正例，其余候选词为负例。最后，基于训练样例学习分类模型，对测试语料中的候选词进行分类，获得识别结果。

CoNLL 2010 评测任务[27]极大地促进了英文模糊限制语识别研究。然而，中文模糊限制语识别的研究十分匮乏，缺乏标注的模糊限制语语料库，阻碍了中文模糊限制语识别的研究。本文基于深度学习的方法，利用英文模糊限制语语料，进行跨语言中文模糊限制语识别。然而，中英文常用的模糊限制语及其使用频率不同，影响识别性能。一些中文（英文）模糊限制语在翻译成英文（中文）时，模糊性会发生改变。如例句 1.3 中的“较”是中文模糊限制语，表示程度变动，一般被翻译成英文形容词或副词的比较级。但这些词却不是英文模糊限制语，导致英文分类模型很难将中文样例中的“较”正确地识别出来。这种较大的数据差异给跨语言模糊限制语识别任务带来了困难和挑战。

实验采用 CoNLL 2010 的模糊限制语识别任务语料作为英文训练语料。该语料包括生物医学和维基百科两个领域，本文仅采用生物医学领域的摘要和全文语料作为训练数据集。

中文测试语料采用文献[26]构建的中文模糊限制信息语料库。为了使中文测试语料与英文训练语料属于同一领域，本文仅采用该语料库中生物医学领域的摘要和全文部分。跨语言模糊限制语识别实验语料统计信息如表 1.2 所示。

表 1.2　跨语言模糊限制语语料统计信息

Tab. 1.2　The statistics of CLHCD corpora

领域 / 语料信息	摘要		全文	
	英文训练语料	中文测试语料	英文训练语料	中文测试语料
词个数	277499	295513	59628	34042
候选词语个数	6230	13452	1533	1115
模糊限制语个数	2694	2323	682	512
模糊限制语候选占比（%）	43.24	17.27	44.49	45.92
句子数	11871	10724	2670	959
模糊限制性句子数	2101	1712	519	329

采用精确率、召回率和 *F* 值评价系统性能。

（1） 精确率（*Precision*， *P* ）表示在模糊限制语识别中，正确识别的模糊限制语占全部识别的模糊限制语的比例，计算方法如公式（1.2）所示：

$$P = \frac{TP}{TP + FP} \tag{1.2}$$

其中，*TP* 表示识别正确的模糊限制语个数， *FP* 表示识别错误的模糊限制语个数。

（2） 召回率（*Recall*， *R* ）表示在模糊限制语识别中，正确识别的模糊限制语占测试语料中所有模糊限制语的比例，计算方法如公式（1.3）所示：

$$R = \frac{TP}{TP + FN} \tag{1.3}$$

其中，*TP* 表示识别正确的模糊限制语个数，*FN* 表示系统未识别出的模糊限制语个数。

（3） *F* 值（*F-score*， *F* ）表示精确率与召回率的调和平均数，计算公式如（1.4）所示：

$$F = \frac{2 \times P \times R}{P + R} \tag{1.4}$$

1.5 本文研究内容

本文基于深度学习模型进行跨语言信息抽取研究。基于跨语言情感分类和跨语言模糊限制语识别两个任务验证本文提出方法的有效性。

具体工作分为以下三个部分：

（1） 基于降噪自动编码机的双视图跨语言信息抽取

提出一种基于降噪自动编码机（DAE）的双视图跨语言信息抽取方法。在源语言和目标语言向量重构过程中，降噪自动编码机适当引入噪音，减少翻译错误对分类性能的影响，提高信息抽取系统的鲁棒性和抗噪音能力。同时分别在中英文两个视图下构建分类模型，最后线性融合两个视图各自的分类结果，减少语言鸿沟对分类性能的影响。实验证明基于降噪自动编码机的双视图模型能有效提高跨语言信息抽取性能。

（2） 基于双语词表示的跨语言信息抽取

为了深入融合中英文两种语言的语义信息，提出基于双语词表示的跨语言信息抽取方法，学习过程分为无监督学习阶段和有监督学习阶段。在无监督学习阶段，利用降噪自动编码机同时进行中英文双语重构，捕获双语语义信息。在有监督学习阶段，将训练语料的标注信息嵌入双语词表示，提高信息抽取性能。实验表明，双语词表示能够有效地捕获双语语义信息和训练语料的标注信息，克服双视图方法难以深入融合两种语言的不足。

（3） 基于联合表示学习的跨语言信息抽取

为了解决基于降噪自动编码机方法产生的数据稀疏问题，提出一种基于联合表示学习的跨语言信息抽取方法。采用长短时记忆递归网络（LSTM），学习中英文词的语义表示。在此基础上，引入上下文情感（模糊）信息表示，联合训练获得情感词（模糊限制语）在特定语境下的语义信息和情感（模糊）信息，提高跨语言信息抽取性能。实验表明，通过 LSTM 学习获得中英文双语语义表示，能够有效挖掘特征深层次的语义信息。同时，词的语义表示与上下文情感（模糊）信息的联合表示学习，能够进一步提高跨语言信息抽取的性能。

1.6 本文组织结构

本文包括五章内容，各章研究内容如下：

第 1 章，介绍跨语言信息抽取的研究背景以及跨语言信息抽取研究现状，概括本文的主要研究内容。

第 2 章，介绍与本文相关的机器学习方法，包括支持向量机和逻辑斯蒂回归模型，以及常用的深度学习模型：降噪自动编码机和长短时记忆递归网络。

第 3 章，研究降噪自动编码机在跨语言信息抽取中的应用。在降噪自动编码机重构过程中适当引入噪音，减少翻译错误对分类性能的影响。同时，在中英文两个视图分别构建分类模型，最后融合两个分类模型的分类结果，减少语言鸿沟对分类性能的影响。实验证明降噪自动编码机和双视图的方法均能提高跨语言信息抽取性能。

第 4 章，研究基于双语词表示的跨语言信息抽取方法。双语词表示的学习分为无监督学习和有监督学习两个阶段。无监督学习阶段通过降噪自动编码机的双语重构，捕获中英文双语语义信息；有监督学习阶段将标注信息嵌入双语词表示。实验证明双语词表示能够捕获双语语义信息和训练语料的标注信息。

第 5 章，研究基于联合表示的跨语言信息抽取方法，引入上下文情感（模糊）信息表示，与词的语义表示一同作为 LSTM 的输入，用于联合表示学习。实验证明基于 LSTM 学习获得的中英文双语词表示，能够有效实现跨语言信息抽取。同时，上下文情感（模糊）信息表示可以进一步提高跨语言信息抽取性能。

2 统计机器学习方法

介绍本文使用的分类模型：支持向量机和逻辑斯蒂回归模型；常用的深度学习模型：降噪自动编码机和长短时记忆递归网络。

2.1 支持向量机

支持向量机（Support Vector Machines，SVM）是一种有监督的二值分类模型，是 Cortes 和 Vapnic 在 1995 年首次提出的[28]。它是定义在特征空间上间隔距离最大的分类模型。近年来，支持向量机在自然语言处理的各类任务中有着广泛的应用[29-31]。

支持向量机采用间隔最大化求分类超平面[32]。我们假定在特征空间有正负训练样例

$$(x_1, y_1)(x_2, y_2)...(x_n, y_n), x_i \in R^d, y_i \in \{-1, +1\} \quad (2.1)$$

其中，x_i 是一个 d 维向量，表示第 i 个训练样例，y_i 是 x_i 所属的类别，用-1 和+1 表示：-1 表示负样例，+1 表示正样例。

SVM 的学习目标是在 d 维空间中，利用法向量 w 和截距 b 构建一个分类超平面，满足 $w \cdot x_i + b = 0$。基于分类超平面，我们可以将空间中的样例分为两部分：法向量指向的那部分样例是正例，另一侧是负例。

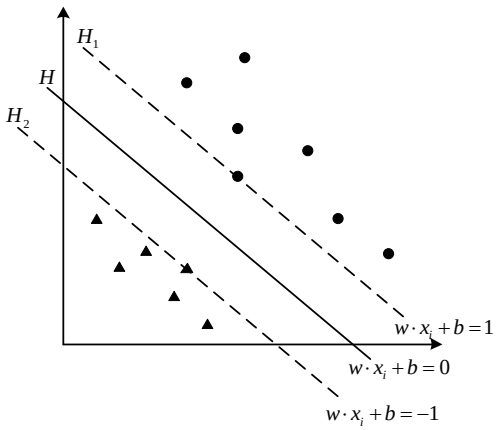


图 2.1 分类超平面示意图

Fig. 2.1 Classification hyperplane of SVM

图 2.1 为当 $d = 2$ 时的分类超平面示意图。H 为分类超平面，分类超平面两侧的“●”和“▲”分别表示正例和负例。H_1（$w \cdot x_i + b = 1$）和 H_2（$w \cdot x_i + b = -1$）分别是经过距

离分类超平面最近的样例的平面，均与 H 平行，它们与分类超平面的距离为间隔距离 $\frac{1}{\|w\|}$。H_1 和 H_2 上的样例称为支持向量。

SVM 的工作原理是优化分类超平面，使得间隔距离最大化。基于此，可以定义 SVM 的目标函数为：

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(w\cdot x_i+b)-1\geq 0, \quad i=1,2,...,n \end{aligned} \tag{2.2}$$

其中，w 和 b 分别为模型训练参数。经过训练后，最优解为 w^* 和 b^*，由此可以得到分类超平面 $w^*\cdot x+b^*=0$。

本文使用 SVM 作为基本分类器，仅介绍 SVM 的基本原理，对 SVM 的其他理论不再详细描述。

2.2 逻辑斯蒂回归

逻辑斯蒂回归（Logistic Regression），简称逻辑回归，是机器学习中经典的对数线型分类模型。设连续随机变量 X 服从以下分布函数（2.3）和密度函数（2.4）：

$$F(x)=P(X\leq x)=\frac{1}{1+e^{-(x-\mu)/\gamma}} \tag{2.3}$$

$$f(x)=F'(x)=\frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-(x-\mu)/\gamma})^2} \tag{2.4}$$

式中，μ 为位置参数，$\gamma>0$ 为形状参数，则称 X 服从逻辑斯蒂分布[32]。

给定样例 $x\in\mathbb{R}^d$，服从逻辑斯蒂分布。我们将分类结果表示为条件概率分布 $P(Y\mid X)$，所得分类概率如公式（2.5）所示：

$$\begin{aligned} P(Y=1\mid x)&=\frac{1}{1+e^{-w\cdot x+b}} \\ P(Y=0\mid x)&=1-P(Y=1\mid x)=\frac{e^{-w\cdot x+b}}{1+e^{-w\cdot x+b}} \end{aligned} \tag{2.5}$$

式中 $Y\in\{0,1\}$ 是输出的类别标签。其中模型参数 $\theta=\{w,b\}$，$w\in\mathbb{R}^d$，$b\in\mathbb{R}$。

对于训练集$T=\{(x_1,y_1),(x_2,y_2),...,(x_i,y_i),...,(x_n,y_n)\}$，我们在训练模型时，利用最大似然估计法学习参数，得到逻辑回归模型。在公式（2.6）中，$L(x,y;\theta)$为模型的对数似然函数，对$L(x,y;\theta)$求极大值，即可得到模型参数的估计值。

$$L(x,y;\theta)=L(x,y;w,b)=\sum_{i=1}^{n}[y_i(w\cdot x_i+b)-\log(1+e^{w\cdot x_i+b})] \quad (2.6)$$

通常在训练逻辑回归模型时，常采用梯度下降算法（Gradient Descent）对参数进行更新。如公式（2.7）所示：

$$\hat{\theta}=\theta-l\cdot\frac{\partial L}{\partial\theta} \quad (2.7)$$

式中$\hat{\theta}$为更新后的参数，l为学习率。

2.3 深度学习模型

2.3.1 降噪自动编码机

反向传播（Back Propagation，BP）是一种有效的训练神经网络的方法。但深层神经网络来说，BP 算法训练速度缓慢，而且模型参数容易收敛到局部最优点上[33]。为解决此问题，Hinton 等[6]在 2006 年提出自动编码机（Autoencoder，AE），首先利用无监督的重构的方法对模型进行预训练，然后利用 BP 算法进行优化。

自动编码机包括一个编码器（Encoder）和一个解码器（Decoder）。在编码阶段，对于给定输入向量$x\in\mathbb{R}^d$，编码器基于$h=f_\theta(x)=s(W\cdot x+b)$将输入向量$x$映射到隐藏层$h$，其中参数$\theta=\{W,b\}$，$s=\frac{1}{1+e^{-x}}$为sigmoid激活函数。在得到隐层表示 h 后，解码器将进行解码工作，基于 s 将 h 映射到$\hat{x}$，即$\hat{x}=f_{\theta'}(h)=s(W'\cdot h+b')$。其中参数$\theta'=\{W',b'\}$。一般规定$W'=W^T$，这样，编码器和解码器就具有关联的权重矩阵[34]。至此,自动编码机完成一个重构过程（Reconstruction）。

我们将训练集上的重构误差（Reconstruction Error）定义为交叉熵（Cross-entropy）的形式，作为模型的目标函数，如公式（2.8）所示，并利用梯度下降算法训练模型参数，使目标函数最小化。

$$L(\theta,\theta')=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}[x_{ij}\log\hat{x}_{ij}+(1-x_{ij})\log(1-\hat{x}_{ij})]$$
$$=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}[x_{ij}\log f_{\theta'}(f_{\theta}(x_{ij}))+(1-x_{ij})\log(1-f_{\theta'}(f_{\theta}(x_{ij})))] \quad (2.8)$$

式中，n 表示训练集部分数据的样例个数，x_{ij} 表示第 i 个输入向量的第 j 个元素。

降噪自动编码机（Denosing Autoencoder，DAE）是 Vincent 等[34]在 2008 年提出的一种新的“编码–解码”结构。它是在自动编码机的基础上，向训练样例加入噪音，增加模型的鲁棒性。降噪自动编码机的重构过程如图 2.2 所示。

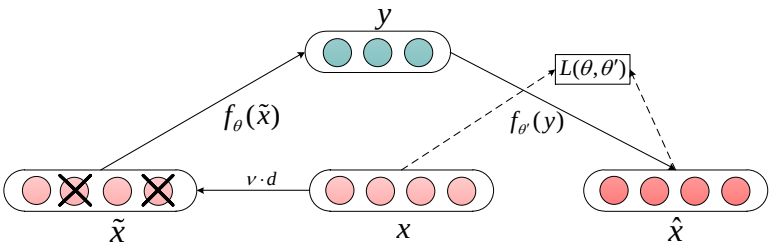


图 2.2　DAE 的重构过程

Fig. 2.2　Reconstruction of denosing autoencoder (DAE)

对于给定的输入向量 $x\in\mathbb{R}^d$，按一定比例 ν 从中随机选出 $\nu\cdot d$ 个元素，并将它们的值设置为 0，得到 $\tilde{x}$。ν 是损失率（Destruction Fraction），我们可以通过调整 ν 控制加入噪音的程度。之后，降噪自动编码机对 $\tilde{x}$ 进行重构：利用 $h=f_\theta(\tilde{x})=s(W\cdot x+b)$ 将其映射到隐藏层，再基于 $\hat{x}=f_{\theta'}(h)=s(W'\cdot h+b')$ 将 h 映射到 $\hat{x}$。重构误差仍然定义为训练集上的交叉熵，如公式（2.8）所示，并利用梯度下降算法训练模型参数。

2.3.2　长短时记忆递归网络

递归神经网络（Recurrent Neural Network，RNN）被证明可以充分利用上下文信息，有效处理序列数据[35]。它采用一种基于“覆写”形式的前馈网络逐步传播每个时刻的数据，即 $s_t=f(s_{t-1},x_t)$，其中 s_t 为时间序列 t 时刻的状态，x_t 是 t 时刻的输入，f 是非线性激活函数。根据链式求导法则，通过这种方式计算获得的梯度被表示为累积的形式。然而，随着序列长度的增加，后面时刻的梯度值趋近于零，造成梯度弥散现象[36]。

鉴于此，Hochreiter 和 Schmidhuber[37]在 1997 年提出长短时记忆递归网络（Long Short Term Memory，LSTM）解决这个问题。LSTM 的主要思想是在每一时刻，引入一个记

忆单元（Memory Cell c_t）和一系列控制门的结构，包括输入门（Input Gate i_t）、输出门（Output Gate o_t）和忘记门（Forget Gate f_t）来控制信息的输入和输出。

LSTM 中，在每个时刻 t，输入门 i_t、输出门 o_t 以及忘记门 f_t 都会接收到当前时刻的输入向量 x_t 和前一时刻的隐藏表示 h_{t-1}。通过控制值的变化，接收或者过滤前一时刻的信息：当值为 0 时，表示过滤相关信息；反之，表示接收当前信息。同时，LSTM 还会重新生成一个新的记忆单元，由输入门 i_t 控制，而 $t-1$ 时刻的记忆单元 c_{t-1} 由忘记门 f_t 控制。t 时刻最终的记忆单元 c_t 由 $t-1$ 时刻的记忆单元 c_{t-1} 和当前新生成的记忆单元共同决定，如公式（2.9）所示。可以证明，这种结构计算的梯度是累加的形式，不会造成由累积形式引起的梯度弥散现象。

然而，对于标准的 LSTM，i_t、o_t 和 f_t 的状态只与当前时刻的输入向量 x_t 和前一时刻的隐藏表示 h_{t-1} 有关，并没有包含前一个时刻记忆单元 c_{t-1} 的相关信息。这样，当前时刻记忆单元 c_t 的状态与前一时刻基于单元 c_{t-1} 的状态就失去了直接的联系。Gers 和 Schmidhuber 在标准的 LSTM 中引入一种“peephole”连接的机制，提出一种 peephole LSTM 模型，目的是在每个相邻时刻的记忆单元之间建立联系[38]。除了当前时刻的输入向量 x_t 和前一时刻隐藏表示 h_{t-1} 外，i_t 和 f_t 还接收前一时刻记忆单元 c_{t-1} 的状态，o_t 接收当前时刻记忆单元 c_t 的状态。隐藏表示 h_t 的计算过程如公式（2.9）所示，单元结构示意图如图 2.3 所示。

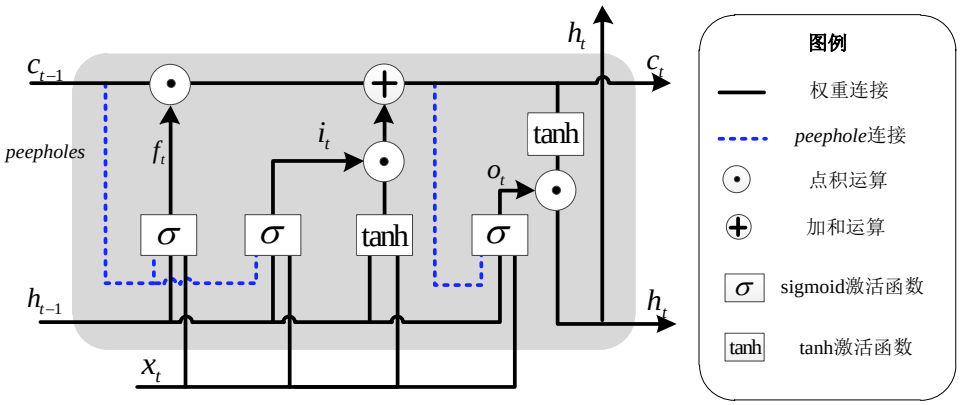


图 2.3　peephole LSTM 单元示意图

Fig. 2.3　Detailed architecture of a unit in peephole LSTM

$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + V^{(i)}c_{t-1} + b^{(i)}) \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + V^{(f)}c_{t-1} + b^{(f)}) \\
c_t &= f \odot c_{t-1} + i_t \odot \tanh(W^{(c)}x_t + U^{(c)}h_{t-1} + b^{(c)}) \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + V^{(o)}c_t + b^{(o)}) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{2.9}
$$

式中，W、U 和V 分别是 x_t、h_{t-1}和记忆单元c 的转换矩阵，b 为偏移量。σ 是sigmoid 激活函数， $\tanh(x)$为双曲正切函数，$\odot$ 表示点积乘法运算。

当前时刻的隐藏表示 h_t 作为下一时刻的输入，获得下一时刻的隐藏表示 h_{t+1}。LSTM 通过这种递归的方式，计算每个时刻的隐藏表示，而最后一个时刻的隐藏表示被认为是整个序列的表示。本文采用的 LSTM 均为 peephole LSTM 模型。同时，为了更好地学习词序列的上下文信息，我们将每个时刻输出的隐藏表示进行平均池化（Mean Pooling），获得序列的隐藏表示，如公式（2.10）所示。

$$
h_s = \text{average}(h_1, h_2, ..., h_t, ..., h_T) \tag{2.10}
$$

式中， h_s 为序列的隐藏表示，T 为时间序列的长度。

2.4 本章小结

本章介绍了本文用到的统计机器学习方法。首先介绍了支持向量机和逻辑斯蒂回归模型。然后介绍了常用的深度学习模型，包括降噪自动编码机和长短时记忆递归网络。降噪自动编码机在自动编码机的基础上加入噪音，增强了模型的鲁棒性；长短时记忆递归网络在递归神经网络的基础上引入了控制门的机制，解决了梯度弥散的问题。

3　基于降噪自动编码机的双视图跨语言信息抽取

为克服翻译错误和中英文语言鸿沟对跨语言信息抽取性能的影响，提出基于降噪自动编码机的双视图跨语言信息抽取方法（Two-view Cross-lingual Information Extraction based on Denoising Autoencoders，Two-view DAE）。在中英文两个视图，基于降噪自动编码机，分别训练获得中英文分类模型，并分别对测试语料进行分类，最后融合两个视图的分类结果。DAE 在重构过程中，适当引入噪音，减少翻译错误对分类性能的影响，提高分类模型的鲁棒性和抗噪音能力。中英文双视图融合，能够修正部分单一视图的分类错误，发挥两种语言的互补优势，减少语言鸿沟对跨语言信息抽取性能的影响。

3.1　基于降噪自动编码机的双视图跨语言信息抽取

基于降噪自动编码机的双视图跨语言信息抽取流程如图 3.1 所示，可以分为训练和测试两个阶段。

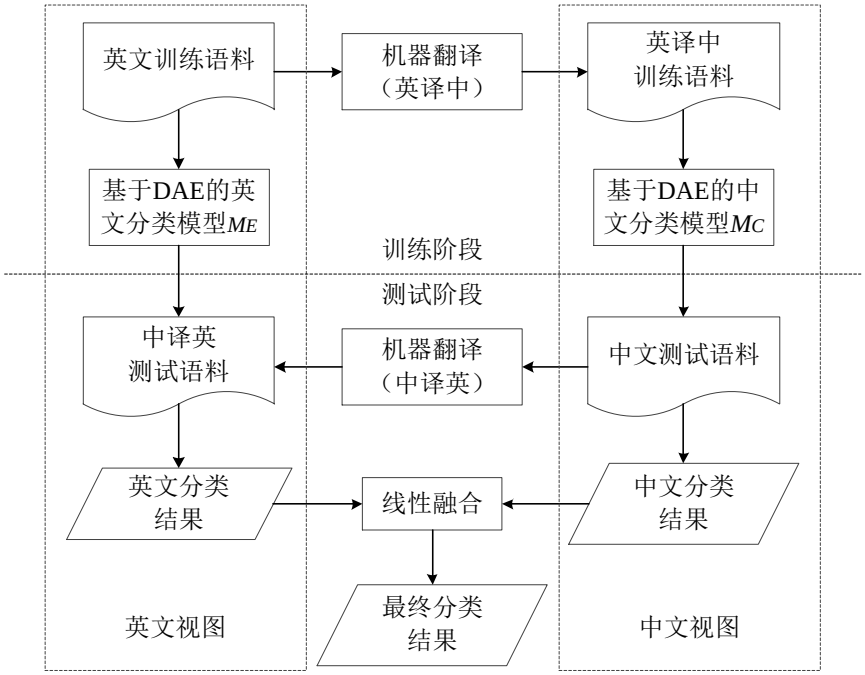


图 3.1　基于降噪自动编码机的双视图跨语言信息抽取流程

Fig. 3.1　The process of Two-view DAE approach

在训练阶段，首先基于机器翻译系统将英文训练语料翻译成中文。然后在英文视图下，利用降噪自动编码机，基于英文训练语料训练获得英文分类模型。同时在中文视图下，基于英译中训练语料训练获得中文分类模型。

在测试阶段，首先基于机器翻译系统将中文测试语料翻译成英文。然后在英文视图下，利用英文分类模型对中译英测试语料进行分类，得到英文分类结果。同时在中文视图下，利用中文分类模型对中文测试语料进行分类，获得中文分类结果。最后线性融合两个视图的分类结果，得到最终的分类结果。

具体过程如 Two-view DAE 算法所示：

Two-view DAE 算法

输入：英文训练语料 D_{LE}，对应的英译中训练语料 D_{LC}

中文测试语料 D_{UC}，对应的中译英测试语料 D_{UE}

算法步骤：

1、基于 D_{LE} 训练英文分类模型 M_E，基于 D_{LC} 训练中文分类模型 M_C；

2、**LOOP** $s_E(i) \in D_{UE}$， $s_C(i) \in D_{UC}$：

（1） 利用 M_E 对 $s_E(i)$ 分类，获得 $P(s_E(i))$ 和 $N(s_E(i))$；

（2） 利用 M_c 对 $s_C(i)$ 分类，获得 $P(s_C(i))$ 和 $N(s_C(i))$；

（3） 计算 $P(s(i)) = (P(s_E(i)) + P(s_C(i)))/2$；

（4） 计算 $N(s(i)) = (N(s_E(i)) + N(s_C(i)))/2$；

（5） **IF** $P(s(i)) > N(s(i))$：判断 $s(i)$ 为正例；**ELSE**：判断 $s(i)$ 为负例；

（$P(s(i))$ 和 $N(s(i))$ 分别为样例 $s(i)$ 被判断为正例和负例的概率）

END LOOP

3.2 特征抽取

采用“词袋”（Bag of Words，BOW）方法将样例表示为英文向量和中文向量，用于训练 DAE 分类模型。首先抽取英文特征词典，包含 M 个特征词；同时抽取中文特征词典，包含 N 个特征词。再将英文训练样例和中译英测试样例表示为“one-hot”形式的英文向量，如公式（3.1）所示；将英译中训练样例和中文测试样例表示为“one-hot”形式的中文向量，如公式（3.2）所示。

$$x_E = \{f_E(1), f_E(2), ..., f_E(i), ..., f_E(M)\} \quad (3.1)$$

$$x_C = \{f_C(1), f_C(2), ..., f_C(i), ..., f_C(N)\} \quad (3.2)$$

式中 $f_E(i)$ 和 $f_C(i)$ 分别是英文和中文向量的特征。本文采用了以下 3 种方法计算特征值：

（1） BOOL：采用布尔值作为特征值，0 表示样例中不含该特征，1 表示样例含有该特征。

（2） TF：采用特征的词频作为特征值，衡量特征的重要程度。

（3） TF-IDF：在多数样例中出现的特征词叫做“常用词”。采用 TF 计算方法可能会将缺乏区分度的“常用词”赋予很高的权重。在 TF 的基础上引入逆文档数（Inversion Document Frequency，IDF）作为特征的重要性权重系数[39]，可以赋予“常用词”较小的权重，更加合理地衡量特征的重要程度。计算公式如公式（3.3）所示。

$$weight_{f(i)} = TF_{f(i)} \cdot \log \frac{N}{n_{f(i)}} \quad (3.3)$$

式中，$weight_{f(i)}$ 是特征 $f(i)$ 的特征值，$TF_{f(i)}$ 为 $f(i)$ 的词频，N 为样例总数，$n_{f(i)}$ 为包含特征 $f(i)$ 的样例数。当 $TF_{f(i)}$ 值很大时，包含该特征的样例数 $n_{f(i)}$ 会很大，这样 IDF 值会较小，减少了特征 $f(i)$ 的权重，这样解决了高词频特征值造成的缺乏区分度问题。

下面将分别介绍跨语言情感分类和跨语言模糊限制语识别两个任务的特征抽取方法。

3.2.1 跨语言情感分类

在情感分类任务中，情感词作为表达情感的重要因素，直接影响情感分类的性能。本文的重点是探索深度学习在跨语言情感分类任务中的应用，因此仅采用中英文情感词特征及其对应的否定特征作为跨语言情感分类模型的基本特征。

（1） 英文情感词特征

MPQA（Multi-perspective Question Answering）[40]情感词典涵盖 8221 个英文情感词，包括情感词的情感强弱（Type）、长度（Len）、情感词（Word）、词性（POS）、词干（Stemmed）以及词典情感倾向（Prior Polarity）。为了减小向量的维度、降低计算复杂度，本文基于英文训练语料和中译英测试语料，利用 χ^2 统计的方法[41]筛选英文情感词，用于表示英文训练样例和中译英测试样例。χ^2 统计法可以抽取与情感类别相关度较高的特征，筛选方法如公式（3.4）所示：

$$\chi^2(t_i, C_j) = \frac{N \times (A \times D - B \times C)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (3.4)$$

式中，$\chi^2(t_i,C_j)$表示情感词t_i用于情感类别C_j的相关度。A表示含有情感词t_i属于情感类别C_j的文档数，B表示含有情感词t_i不属于情感类别C_j的文档数，C表示不含有情感词t_i属于情感类别C_j的文档数，D表示不含有情感词t_i不属于情感类别C_j文档数。N表示文档的总数。

本文计算词典中所有情感词的χ^2值，进行排序，筛选出前 2000 个英文情感词作为英文情感词特征。

（2） 中文情感词特征

由于评测任务数据集没有提供中文情感词典，本文仅利用机器翻译系统将抽取的 2000 个英文情感词翻译成中文，作为中文情感词特征。

（3） 否定特征

否定词在情感分类任务中起着至关重要的作用。如例句 3.1 中，说话人持消极情感倾向，但如果只考虑句中的情感词“好”和“满意”，而忽略了前面的否定修饰词“不”，则该句的情感倾向很有可能会被判断为积极的。

（例句 3.1） 看着像盗版，纸张质量不好，快递包装不满意。

鉴于此，本文引入否定特征提高跨语言情感分类的准确性。对于每个情感词特征，本文设置窗口$[-2,2]$用于检测该情感词是否被否定词修饰。如果在窗口内含有否定词，则认为该情感词被否定词修饰，记为该情感词的否定特征，而非该情感词特征本身。将否定特征分别加在情感词特征之后，表示该情感词特征的否定特征。在识别否定词时，本文收集了常用的 14 个英文否定词作为英文否定词典，如表 3.1 所示。对于中文否定词，仅考虑了“不”、“不会”以及“没有”对情感词的修饰情况。

基于英（中）文情感词特征和对应的否定特征可以将一个文档表示为英文向量和中文向量。

表 3.1 英文否定词词典

Tab. 3.1 Lexicon of English negations

序号	否定词	序号	否定词	序号	否定词
1	not	6	little	11	neither
2	n't	7	few	12	seldom
3	no	8	nobody	13	nowhere
4	nor	9	nothing	14	without
5	never	10	none		

3.2.2 跨语言模糊限制语识别

包含模糊限制语的句子称为模糊限制性句子。模糊限制语识别要求准确找到表达模糊限制信息的词语或短语，而模糊限制性句子识别只要求判断句子是否具有模糊性，无需识别具体的模糊限制语。因此，模糊限制语识别比模糊限制性句子识别难度更大。

本章将模糊限制语识别转换为模糊限制性句子识别。为了降低向量的维度，减小计算复杂度，我们将模糊限制语候选词作为特征表示句子。首先基于英文模糊限制语训练语料抽取英文模糊限制语词典，基于中文模糊限制语测试语料抽取中文模糊限制语词典。再基于抽取后的英文和中文模糊限制语候选词将句子分别表示为英文向量和中文向量。摘要中，抽取英文模糊限制语候选词 138 个，中文模糊限制语候选词 183 个；全文中，抽取英文模糊限制语候选词 118 个，中文模糊限制语候选词 86 个。

获得训练和测试句子向量，基于二值分类识别模糊限制性句子：若为正例，句子具有模糊性，反之，句子不具有模糊性。如果句子具有模糊性，句中的所有候选词均被判断为模糊限制语；如果句子不具有模糊性，句中的所有候选词均不是模糊限制语。

3.3 实验结果及分析

实验采用 ICTCLAS 工具包[42]对中文语料进行分词。Theano[43]是 python 的库函数，专用于运算和优化数学表达式，适用于多维数组运算。本文采用 Theano 构建基于降噪自动编码机的双视图跨语言信息抽取系统。降噪自动编码机的损失率ν设置为 0.1，即重构过程中将输入向量的 10%随机变为 0。分别在跨语言情感分类和跨语言模糊限制语识别两个任务上，验证本方法的有效性。

3.3.1 特征值计算方法对分类性能的影响

在中英文单视图下分别比较特征值计算方法对跨语言情感分类和跨语言模糊限制语识别性能的影响。

表 3.2 分别比较了不同特征值计算方法在中英文视图下的跨语言情感分类性能。

从表 3.2 可以看出，在 3 种特征值计算方法中，TF-IDF 在跨语言情感分类中取得了最好的平均准确率，说明 TF-IDF 可以反映情感词特征在文档中的情感强弱，有助于提高情感分类的性能。

表 3.2　特征值计算方法对跨语言情感分类的影响

Tab. 3.2　The effect of feature weight calculation method on CLSC performance

视图	特征值计算法	书籍	DVD	音乐	*Average*
中文视图	BOOL	72.88	72.20	70.63	71.90
	TF	73.10	**78.23**	64.63	71.98
	TF-IDF	**78.15**	75.05	**74.30**	**75.83**
英文视图	BOOL	69.98	75.83	73.25	73.02
	TF	72.35	74.58	74.13	73.68
	TF-IDF	**73.03**	**76.93**	**75.15**	**75.04**

表 3.3 分别比较了不同特征值计算方法在中英文视图下的跨语言模糊限制语识别性能。

表 3.3　特征值计算方法对跨语言模糊限制语识别的影响

Tab. 3.3　The effect of feature weight calculation method on CLHCD performance

识别任务	视图	特征值	摘要			全文		
			P	*R*	*F*	*P*	*R*	*F*
句子识别	中文视图	BOOL	68.30	36.69	47.74	68.66	73.25	70.88
		TF	79.22	35.63	49.15	72.50	79.33	**75.76**
		TF-IDF	78.18	38.73	**51.80**	73.49	77.51	75.45
	英文视图	BOOL	48.68	51.87	**50.22**	63.85	57.45	60.48
		TF	49.06	48.71	48.88	63.76	57.75	**60.61**
		TF-IDF	38.73	56.31	45.89	63.76	57.75	**60.61**
词语识别	中文视图	BOOL	33.31	41.11	36.80	50.58	77.15	61.10
		TF	39.16	38.26	38.70	50.29	84.77	**63.13**
		TF-IDF	38.32	41.59	**39.89**	50.42	82.81	62.68
	英文视图	BOOL	26.87	54.80	**36.06**	50.45	65.43	56.97
		TF	27.14	51.57	35.56	50.37	65.63	**57.00**
		TF-IDF	23.81	58.85	33.90	50.37	65.63	**57.00**

从表 3.3 可以看出，TF-IDF 特征值计算方法在模糊限制语识别任务中并没有表现出明显的优势。原因在于一个句子是否具有模糊性，取决于使用的模糊限制语本身，并非取决于句子中同一个模糊限制语重复使用的次数。所以，对于模糊限制语识别任务，特征的选取起到了关键性的作用，而特征值计算方法的不同并没有起到明显的作用。

3.3.2 单视图与双视图方法比较

实验比较了单视图和双视图的跨语言信息抽取性能。表 3.4 比较了跨语言情感分类中双视图与中英文单视图模型的分类性能。

表 3.4 基于双视图和单视图模型的跨语言情感分类性能比较

Tab. 3.4 Two-view model vs. Single-view model in CLSC

视图	书籍	DVD	音乐	*Average*
中文视图	78.15	75.05	74.30	75.83
英文视图	73.03	76.93	75.15	75.04
双视图	**79.68**	**78.33**	**78.08**	**78.70**

从表 3.4 可以看出，在书籍、DVD 和音乐领域，双视图模型分别比它们各自最好的单视图模型提高了 1.53%、1.40%和 2.93%。实验结果说明，在跨语言情感分类任务中，双视图模型有助于发挥中英文语言的互补优势，修正由于语言鸿沟造成的分类错误，提高最终的分类性能。

表 3.5 比较了跨语言模糊限制语识别中双视图与中英文单视图模型的性能。

表 3.5 基于双视图和单视图模型的跨语言模糊限制语识别性能比较

Tab. 3.5 Two-view model vs. Single-view model in CLHCD

识别任务	视图	摘要			全文		
		P	*R*	*F*	*P*	*R*	*F*
句子识别	中文视图	78.18	38.73	**51.80**	73.49	77.51	**75.45**
	英文视图	38.73	56.31	45.89	63.76	57.75	60.61
	双视图	69.04	39.60	50.33	65.00	63.22	64.10
词语识别	中文视图	38.32	41.59	**39.89**	50.42	82.81	**62.68**
	英文视图	23.81	58.85	33.90	50.37	65.63	57.00
	双视图	34.13	42.62	37.91	50.49	70.51	58.84

从表 3.5 可以看出，在跨语言模糊限制语识别任务中双视图模型并没有提高分类性能，而是介于两个单视图性能之间。通过分析实验数据我们发现，中英文在表达模糊限制信息的方式上差异较大，翻译过程中，两种语言的模糊限制语不是一一对应的，如例句 3.2 为一个中文模糊限制性句子，例句 3.3 为对应的英文翻译。例句 3.2 中的“一定”为中文模糊限制语，表示数量变动，在对应的英文翻译中被翻译成了“some”（一些），

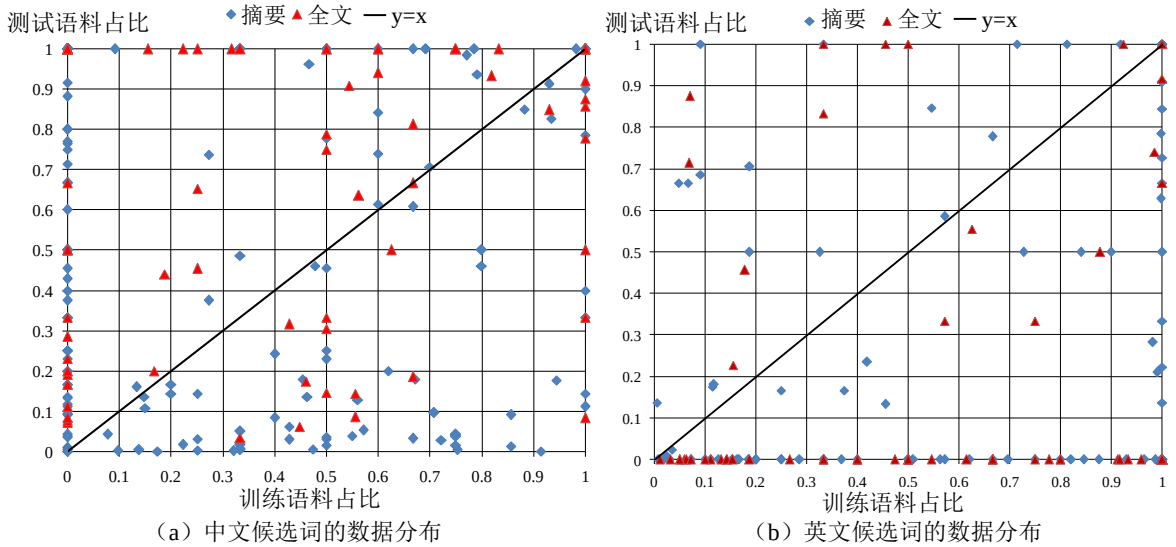
而“some”不是英文模糊限制语候选词。在分类过程中，该样例在中文视图中被识别为正例的概率值为 88%；在英文视图中，却被识别为负例的概率为 99%。因此，在双视图融合过程中，由于在英文视图中分类结果的概率较大，该样例最终被确立为负例，即不具有模糊性的句子。通过上面的分析，可以看出，在双视图融合过程中，英文视图的分类结果阻碍了模糊限制语的识别，最终造成上述现象。

（例句 3.2） 国内外学者通过传统的组织工程手段进行了大量牙再生的相关研究，已取得<*ccue*>一定</*ccue*>成绩。

（例句 3.3） Scholars through traditional means of tissue engineering research related to a large number of tooth regeneration, we have achieved **some** results.

为进一步解释上述现象，分别统计了跨语言情感分类和跨语言模糊限制语识别任务中特征词的数据分布。图 3.2 统计了跨语言模糊限制语识别中、英文候选词在训练和测试语料的分布情况，图中横坐标为模糊限制语在训练语料中的占比，纵坐标为候选词在测试语料中的占比，计算方法如公式（3.5）所示。$Ratio(cand_i)$ 为第 i 个候选词的占比，$Freq(cand_i)$ 和 $Freq(cue_i)$ 分别为候选词在语料中的频数和该词被标记为模糊限制语的频数。

$$Ratio(cand_i)=\frac{Freq(cue_i)}{Freq(cand_i)}, cand_i \in \{corpus_{Train}, corpus_{Test}\} \qquad (3.5)$$



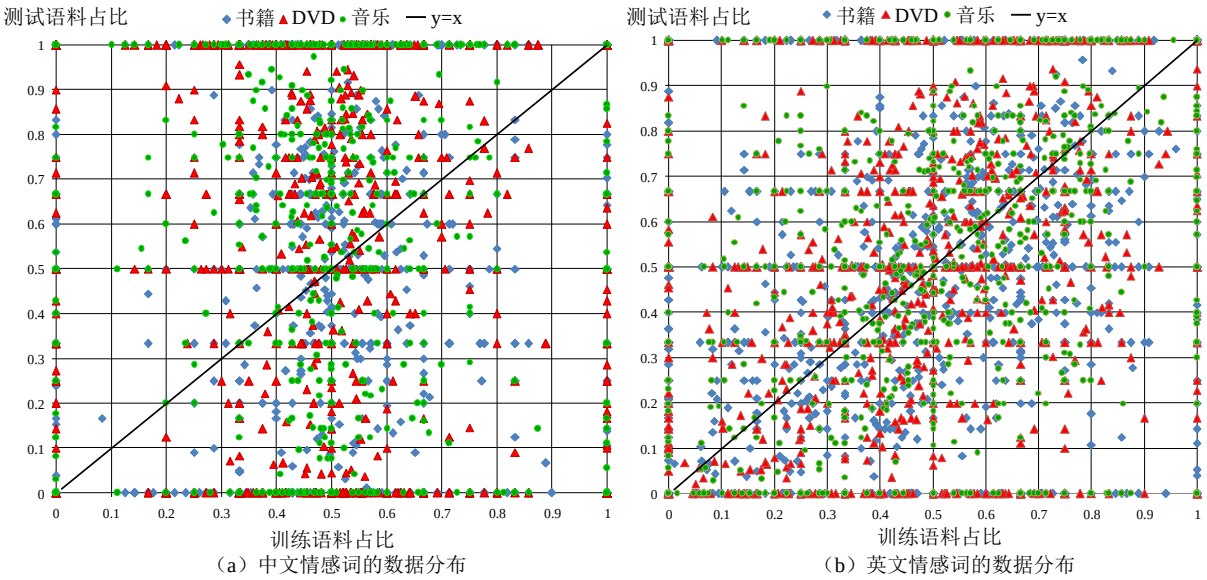
（a）中文候选词的数据分布　　（b）英文候选词的数据分布

图 3.2　中英文模糊限制语候选词在训练和测试语料中的数据分布

Fig. 3.2　The distributions of English and Chinese candidates in training and test corpora

图 3.3 统计了跨语言情感分类中，英文情感词在训练和测试语料的分布情况。如公式（3.6）所示，$Ratio(senti_i)$ 为第 i 个情感词的占比，$Freq(senti_i)$ 和 $Freq(senti_i \in doc_{pos})$ 分别是该情感词在语料中的频数和在积极情感倾向文档的频数。

$$Ratio(senti_i) = \frac{Freq(senti_i \in doc_{pos})}{Freq(senti_i)}, senti_i \in \{corpus_{Train}, corpus_{Test}\} \quad (3.6)$$



（a）中文情感词的数据分布　　（b）英文情感词的数据分布

图 3.3　中英文情感词在训练和测试语料中的数据分布

Fig. 3.3　The distributions of English and Chinese sentiment words in training and test corpora

基于机器学习的信息抽取方法在训练语料和测试语料数据分布相同的情况下，会取得较好的分类性能。在理想情况下，当候选词（情感词）分布在直线 $y = x$ 附近时，说明候选词（情感词）在训练语料和测试语料的分布相同。但从图 3.2 和图 3.3 可以看出，在跨语言信息抽取任务，候选词和情感词在训练语料和测试语料中正例的占比并不相同。与图 3.3 中情感词的分布相比，图 3.2 中的候选词数量较少，且大多数距离 $y = x$ 直线较远，仅有少数候选词在训练和测试语料中分布相似。这会导致基于英文训练语料训练的分类模型不适用于中文测试语料，降低了分类性能。进一步证明在跨语言模糊限制语识别任务中，英文视图的识别结果并没有对双视图融合的最终结果起到辅助作用。造成这种现象的原因一方面是翻译过程中带来的翻译错误，另一方面是中英文在表达模糊限制信息时，用词习惯差异很大。

3.3.3 损失率对跨语言信息抽取的影响

通过调整 DAE 的损失率ν，可以控制重构过程中向训练数据中加入噪音的程度。图 3.4 给出了跨语言情感分类准确率与 DAE 损失率的关系曲线，损失率$\nu \in [0,0.9]$。

从图 3.4 可以看出，当损失率$\nu = 0$时，在重构过程中不向训练数据加入噪音，降噪自动编码机（DAE）变为传统的自动编码机（AE），此时书籍、DVD 和音乐三个领域的平均准确率为 79.08%。随着损失率ν的逐渐增加，三个领域的分类性能也发生变化。当$\nu \in [0.2,0.5]$时，平均准确率超过传统自动编码机的平均准确率，说明适当地加入噪音，可以提高模型的鲁棒性和抗噪音能力，提高分类性能。当$\nu = 0.2$时，平均准确率达到最优，即 80.02%。

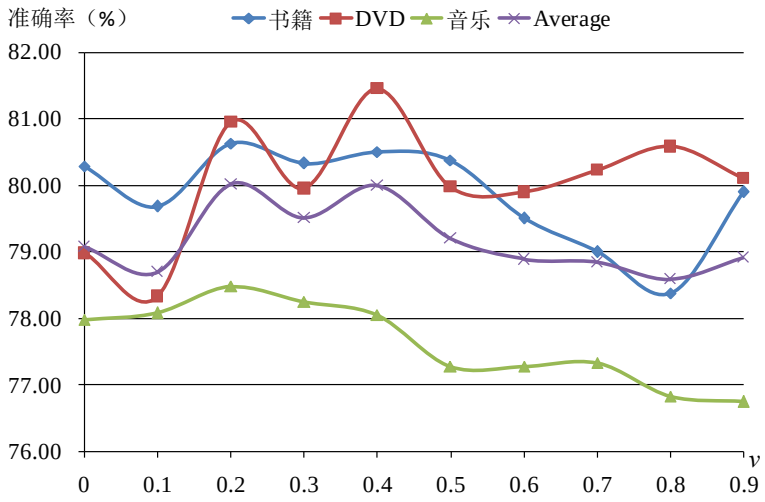


图 3.4　损失率对跨语言情感分类性能的影响

Fig. 3.4　Accuracy vs. Destruction fraction in CLSC

图 3.5 和图 3.6 分别给出了跨语言模糊限制语识别任务中摘要和全文的 *F* 值与 DAE 损失率的关系曲线，损失率$\nu \in [0,0.9]$。

图 3.5 中“英-句”、“中-句”和“双-句”分别表示模糊限制性句子识别在英文视图、中文视图以及双视图下的性能。从图 3.5 可以看出，在摘要中，当损失率$\nu = 0$时，基于英文视图、中文视图以及双视图的方法的跨语言模糊限制性句子识别 *F* 值分别为 50.04%、50.85%和 49.19%。随着损失率的增长，跨语言模糊限制性句子识别的性能发生变化，说明加入的噪音对识别性能产生了影响。对于中文视图和双视图的方法来说，在模型训练过程中加入噪音，可以提高模型抵抗噪音的能力，提高识别性能。当损失率$\nu = 0.8$时，中文视图的方法性能达到最高，当$\nu = 0.7$时，双视图的方法性能达到最高，

F 值分别为 53.56%和 50.49%。而对于英文视图的方法，加入的噪音降低了识别性能，没有起到提高分类模型的鲁棒性的作用。原因在于中英文常用的模糊限制语及其使用频率不同，中文测试语料在翻译成英文后，中译英句子的模糊性发生了改变。同时，由于翻译质量的影响，中文模糊限制语在翻译过程中可能丢失，导致分类模型无法识别原有的中文模糊限制语。虽然在 DAE 重构过程中向训练语料中加入噪音，可以增强模型的鲁棒性，但无法弥补由翻译过程导致的测试语料候选词丢失造成的影响，降低了识别性能。

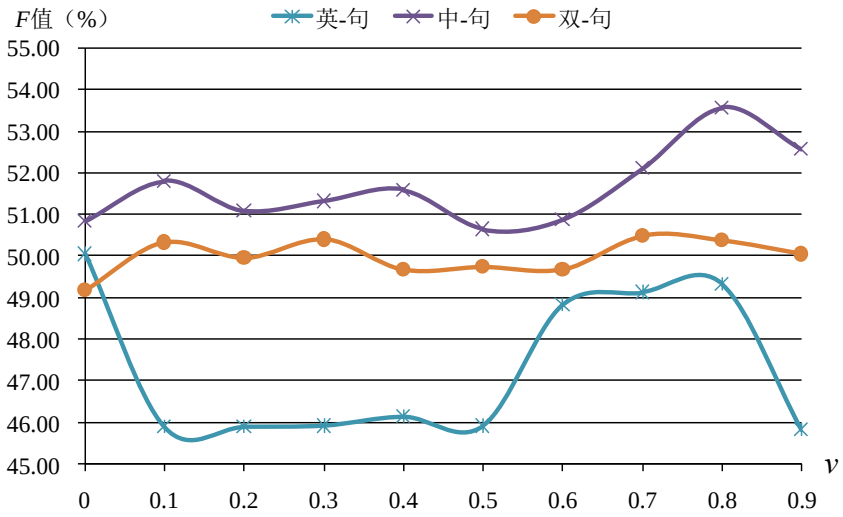


图 3. 5　摘要中损失率对跨语言模糊限制语识别的影响

Fig. 3.5　F-score vs. Destruction fraction in CLHCD (abstract)

与图 3.5 类似，从图 3.6 可以看出，在全文中，随着损失率ν的增长，基于中文视图和双视图方法的跨语言模糊限制性句子识别 *F* 值得到了提高，分别从 75.59%提高到 77.08%，从 62.36%提高到 67.27%。但对于英文视图方法来说，噪音仅对模型的分类概率产生了很小的影响，但对模糊限制性句子识别结果没有起到作用，一直保持在 60.61%。

从上述实验可以看出，在跨语言模糊限制语识别任务中，适当地在 DAE 重构过程中加入噪音，有助于提高中文视图和双视图模型的鲁棒性和抗噪音能力，而对于英文视图模型，加入的噪音阻碍了模型的正常训练，降低了识别性能。

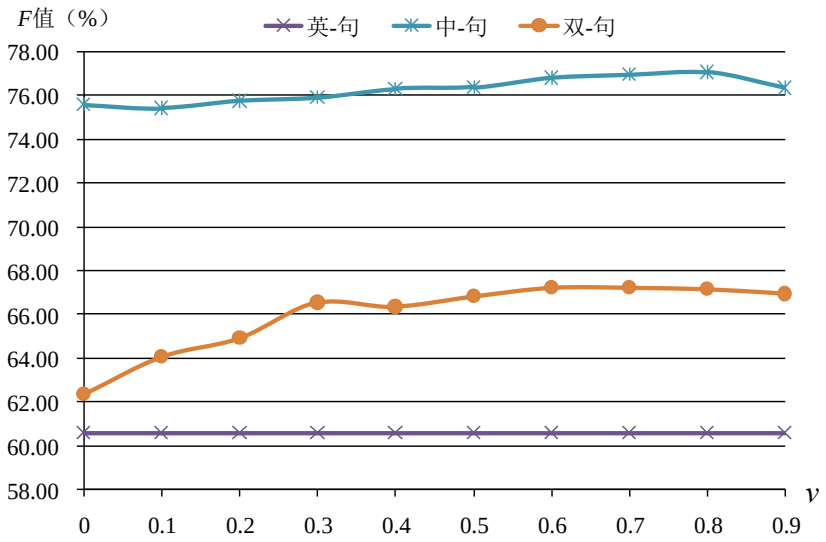


图 3.6　全文中损失率对跨语言模糊限制语识别的影响

Fig. 3.6　F-score vs. Destruction fraction in CLHCD (fullpaper)

3.4　本章小结

本章提出基于降噪自动编码机的双视图跨语言信息抽取方法。在源语言和目标语言向量重构过程中，降噪自动编码机适当引入噪音，提高信息抽取系统的鲁棒性和抗噪音能力。同时，在中英文两个视图，分别构建分类模型，最后融合两个分类模型的结果，获得最终分类结果。

在跨语言情感分类和跨语言模糊限制语识别两个任务上进行了实验，证明在降噪自动编码机重构过程中，适当引入噪音，可以提高模型的鲁棒性和抗噪音能力，减少翻译错误对分类性能的影响。同时，双视图的方法可以充分发挥中英文语言的互补优势，减少单视图分类错误，提高分类性能。但在跨语言模糊限制语识别任务中，由于中英文在表达模糊限制信息时有很大的差异，双视图的方法并没有起到提高识别性能的作用。

4 基于双语词表示的跨语言信息抽取

为了深入融合中英文两种语言的语义信息，提出一种基于中英文双语词表示的跨语言信息抽取方法（Cross-lingual Information Extraction Based on Bilingual Word Representations，BWR）。学习过程分为两个阶段：无监督学习阶段和有监督学习阶段。无监督学习阶段，利用降噪自动编码机同时进行中英文双语词表示学习，捕获中英文双语语义信息。有监督学习阶段，将训练语料的标注信息嵌入双语词表示，提高信息抽取性能。双语词表示能够有效地捕获双语语义信息和训练语料的标注信息，克服双视图方法分别训练中、英文分类模型存在的难以深入融合两种语言的不足。

4.1 无监督学习阶段

无监督学习阶段利用英文训练语料及其中文翻译，学习双语词表示。学习过程中，没有利用训练语料的标注信息。

对于训练样例 x 的英文向量 $x_E=\{f_E(1), f_E(2),..., f_E(M)\}$ 及其对应的中文向量 $x_C=\{f_C(1),f_C(2),..,f_C, N($，可以得到样例 x 的中英文对齐的样例对 (x_E,x_C)，将它用于学习双语词表示。采用 DAE，首先基于样例的英文向量，重构该样例的英文向量和中文向量；基于样例的中文向量，重构该样例的英文向量和中文向量。同时，基于中文向量和英文向量的连接（中文-英文向量对），重构该样例的中文-英文向量对。图 4.1 为无监督学习阶段的框架。

如 2.3.1 节所述，为了增强模型的鲁棒性，在 DAE 进行重构之前，需要以一定概率向输入数据加入噪音。假定 $(\tilde{x}_E,\tilde{x}_C)$ 是给定样例对 (x_E,x_C) 加入噪音后的向量，DAE 分别基于英文和中文两个编码器，利用 sigmoid 激活函数编码得到英文和中文向量的隐藏表示 h_E 和 h_C，过程如公式（4.1）和公式（4.2）所示：

$$h_E = f_\theta(\tilde{x}_E) = s(W_E \cdot \tilde{x}_E + b) \quad (4.1)$$

$$h_C = f_\theta(\tilde{x}_C) = s(W_C \cdot \tilde{x}_C + b) \quad (4.2)$$

式中，f_θ 为编码函数，s 为 sigmoid 激活函数，W_E 和 W_C 分别为英文和中文的转换矩阵。由于中英文隐藏表示 h_E 和 h_C 的维度相同，因此中英文编码器共享一个偏移量 b。

在获得两种语言的隐藏表示 h_E 和 h_C 后，基于 DAE 分别对两种语言的隐藏表示进行解码。如图 4.1（a）所示，对于英文隐藏表示 h_E，本文采用英文和中文两个解码器进行

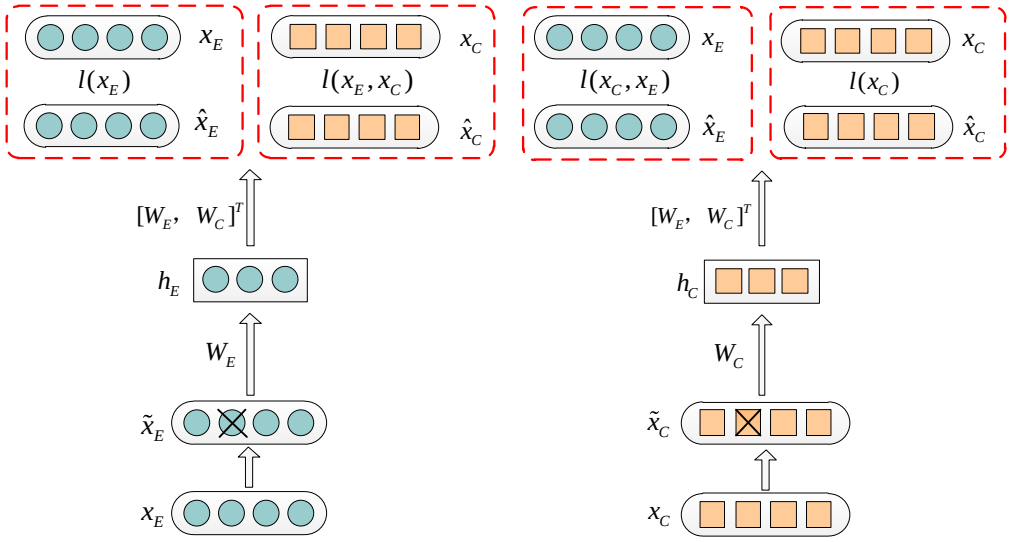
解码：将英文隐藏表示 h_E 分别解码为英文重构向量 $\hat{x}_E$ 和中文重构向量 $\hat{x}_C$，如公式（4.3）和公式（4.4）所示。

$$\hat{x}_E = f_{\theta'}(h_E) = s(W_E^T \cdot h_E + c_E) \tag{4.3}$$

$$\hat{x}_C = f_{\theta'}(h_E) = s(W_C^T \cdot h_E + c_C) \tag{4.4}$$

式中，$f_{\theta'}$ 为解码函数，c_E 和 c_C 分别为英文解码器和中文解码器的偏移量。

中文隐藏表示 h_C 的解码过程与英文隐藏表示 h_E 的解码过程类似，获得由 h_C 解码生成的英文重构向量 $\hat{x}_E$ 和中文重构向量 $\hat{x}_C$，如图 4.1（b）所示。同时，基于样例的英文向量 x_E 和中文向量 x_C 的连接 $[x_E, x_C]$（英文-中文向量对），重构该样例的英文-中文向量对 $[\hat{x}_E, \hat{x}_C]$。



（a）基于英文向量x_E的重构　　（b）基于中文向量x_C的重构

图 4.1　无监督学习框架

Fig. 4.1　The framework of unsupervised learning

这种“编码-解码”的结构可以使一种语言的输入向量重构回自身在本语言的向量，也可以将其重构为另一种语言的向量。由于中英文语言之间存在数据差异，在重构过程中存在重构误差。对于给定的输入样例对 (x_E, x_C)，本文将重构误差定义为交叉熵的形式，并将下面 5 种重构误差之和定义为无监督学习阶段的损失函数：

（1） 从英文向量 x_E 重构回自身的误差 $l(x_E)$；

（2） 从中文向量 x_C 重构回自身的误差 $l(x_C)$；

（3） 从英文向量 x_E 重构中文向量 x_C 的误差 $l(x_E, x_C)$；

（4） 从中文向量 x_C 重构英文向量 x_E 的误差 $l(x_C, x_E)$；

（5） 从英文-中文向量对 $[x_E, x_C]$ 重构回自身的误差 $l([x_E, x_C],[\hat{x}_E, \hat{x}_C])$。

损失函数的定义如公式（4.5）所示：

$$L_{semantic} = l(x_E) + l(x_C) + l(x_E, x_C) + l(x_C, x_E) + l([x_E, x_C],[\hat{x}_E, \hat{x}_C]) \quad (4.5)$$

在无监督学习阶段，定义模型参数 $\theta = \{W_E, W_C, b, c_E, c_C\}$，采用梯度下降算法对参数 θ 进行更新，使损失函数 $L_{semantic}$ 最小化，训练获得 W_E 和 W_C。

英文转换矩阵 W_E 的每一列对应输入向量 x_E 的每一个特征，中文转换矩阵 W_C 的每一列对应输入向量 x_C 的每一个特征，将 W_E 的每一列看作与 x_E 每个特征对应的英文词表示；将 W_C 的每一列看作与 x_C 每个特征对应的中文词表示。在无监督学习阶段，DAE 可以由样例的一种语言向量重构另一种语言向量，建立了中英文两种语言之间的联系，捕获两种语言的语义信息。因此，本文将在无监督学习阶段获得的英文词表示和中文词表示统称为无监督双语词表示。

4.2 有监督学习阶段

在有监督学习阶段，将标注信息嵌入到无监督双语词表示中，训练有监督双语词表示，提高双语词表示性能。图 4.2 为有监督学习阶段的框架。

在无监督学习阶段，转换矩阵 W_E 和 W_C 以最小的重构误差，将样例从一种语言的向量转换成另一种语言的向量，捕获两种语言的语义信息。此外，英文-中文向量对 $[x_E, x_C]$ 包含了 x 在中英文两种语言的所有信息。因此，本文将英文-中文向量对 $[x_E, x_C]$ 通过连接转换矩阵 $[W_E, W_C]$ 进行编码，得到隐藏表示 h_b，可以包含该样例的中英文信息。鉴于此，将 h_b 称为样例 x 的双语隐藏表示，如公式（4.6）所示。

$$h_b = f_\theta([x_E, x_C]) = s([W_E, W_C] \cdot [x_E, x_C] + b) \quad (4.6)$$

式中，b 为编码器的偏移量。

在获得双语隐藏表示 h_b 后，我们将其用于获得样例的分类概率。h_b 被输入到逻辑回归层，分别计算样例的正例概率 $P(y=1|x;\xi)$ 和负例概率 $P(y=0|x;\xi)$，如公式（4.7）所示。

$$P(y=1|x;\xi)=s(x)=s(\varphi^T h_b+b_l)$$
$$P(y=0|x;\xi)=1-P(y=1|x;\xi) \quad (4.7)$$

式中，φ 为逻辑回归的权重向量，b_l 为逻辑回归层的偏移量，$\xi=\{W_E,W_C,b,\varphi,b_l\}$ 表示有监督学习阶段的模型参数。

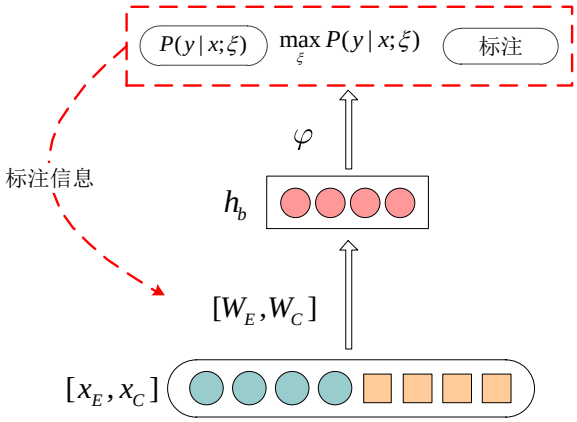


图 4.2　有监督学习框架

Fig. 4.2　The framework of supervised learning

监督学习阶段的目标是最大化分类概率，本文将有监督学习阶段的目标函数 $L(\xi)$ 定义为对数似然函数的形式，如公式（4.8）所示。最后采用梯度下降算法学习模型参数。

$$\begin{aligned} L(\xi) &= y\log(s(x))+(1-y)\log(1-s(x)) \\ &= y\log(s([x_E,x_C];\xi))+(1-y)\log(1-s([x_E,x_C];\xi)) \end{aligned} \quad (4.8)$$

经过有监督的学习阶段后，得到嵌入标注信息的转换矩阵 W_E 和 W_C，它们的每一列被看作对应特征的有监督双语词表示。

将训练样例和测试样例表示为有监督双语词表示的 TF-IDF 加权和形式，用于分类模型的训练和测试。将英文训练样例和中译英测试样例表示为 W_E 的 TF-IDF 加权和形式，将英译中训练样例和中文测试样例表示为 W_C 的 TF-IDF 加权和形式。关于 TF-IDF 的计算方法详见 3.3.1 节。对于样例 x，其英文向量和中文向量分别如公式（4.9）所示。本文提出一种中英文连接表示的方法（Concatenation Representation，CR），将样例 x 的英文向量 ϕ_{x_E} 和中文向量 ϕ_{x_C} 连接起来，即 $\phi_x=[\phi_{x_E},\phi_{x_C}]$，作为线性 SVM 分类模型的输入。中英文连接表示的方法包含了中英文两种语言丰富的的语义和标注信息，有助于提高分类性能。

$$\phi_{x_E} = \sum_{i=1}^{m} \text{TF-IDF}(f_E(i)) \cdot W_{E.,i}$$
$$\phi_{x_C} = \sum_{j=1}^{n} \text{TF-IDF}(f_C(j)) \cdot W_{C.,j} \quad (4.9)$$

4.3 实验结果及分析

在跨语言模糊限制语识别任务中，本章采用第 3 章的方法，将模糊限制语识别任务转换成模糊限制性句子识别：基于机器学习的方法识别模糊限制性句子，然后计算模糊限制语识别结果。实验基于 SVM-light 工具包[44]进行分类。

4.3.1 无监督和有监督的双语词表示的性能比较

我们训练 400 维双语词表示进行跨语言信息抽取实验。在训练无监督双语词表示过程中，将降噪自动编码机的损失率ν设置为 0.1，即将输入向量的 10%随机变为 0。

在情感分类任务中，仅依靠词的语义信息很难获得一个文档的情感倾向。例如“好”和“坏”，它们的词性相同，语义相近[45]，但情感倾向相反。在跨语言情感分类任务中，标注信息即为情感信息，引入有监督学习阶段，情感信息被嵌入到无监督双语词表示中。图 4.3 比较了无监督双语词表示和有监督双语词表示在跨语言情感分类任务中的性能。

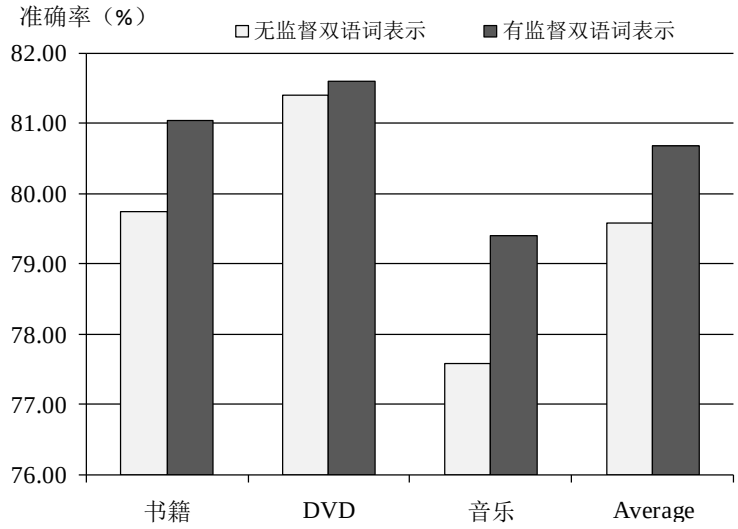


图 4.3 跨语言情感分类中无监督双语词表示与有监督双语词表示性能比较

Fig. 4.3 Unsupervised bilingual word representations vs. Supervised bilingual word representations in CLSC

从图 4.3 可以看出，无监督双语词表示可以捕获到双语语义信息，达到 79.58%的性能，高于 3.3.2 节双视图方法的平均准确率 78.70%。但无监督的方法无法充分表达文档的情感倾向。加入有监督学习阶段后，分类性能得到了提高：在书籍、DVD 和音乐三个领域，分类性能分别提高到 81.05%，81.60%和 79.40%，平均准确率达到了 80.68%，比无监督双语词表示提高了 1.10%。说明有监督的方法有助于增强双语词表示的情感表达能力，提高跨语言情感分类性能。

表 4.1 比较了跨语言模糊限制语识别中无监督双语词表示与有监督双语词表示的性能。

表 4.1　跨语言模糊限制语识别中无监督双语词表示与有监督双语词表示性能比较

Tab. 4.1　Unsupervised bilingual word representations vs. Supervised bilingual word representations in CLHCD

双语词表示	识别任务	摘要			全文		
		P	*R*	*F*	*P*	*R*	*F*
无监督双语词表示	句子识别	38.48	85.86	53.14	65.73	78.12	71.39
	词语识别	21.55	89.07	34.70	48.70	84.18	61.70
有监督双语词表示	句子识别	38.48	85.86	53.14	65.73	78.12	71.39
	词语识别	21.55	89.07	34.70	48.70	84.18	61.70

在跨语言模糊限制语识别任务中，与上一章的双视图方法相比，双语词表示的方法提高了模糊限制性句子识别的 *F* 值。但从表 4.1 可以看出，无监督双语词表示与有监督双语词表示取得了相同的分类性能。根据 2.1 节介绍的 SVM 原理可知，样例的分类边界距离越大，分类可信度越高。通过分析结果我们发现，在摘要中，基于无监督方法得到的模糊限制性句子识别结果，有 1112 个测试样例的边界距离大于 1，而基于有监督方法得到的识别结果，有 6140 个测试样例的边界距离大于 1；在全文中，基于无监督方法得到的模糊限制性句子识别结果，有 95 个测试样例的边界距离大于 1，而基于有监督方法得到的识别结果，有 154 个测试样例的边界距离大于 1。说明虽然无监督双语词表示与有监督双语词表示取得了相同的分类性能，但是经过有监督学习阶段调整双语词表示，增加了识别结果的可信度。

基于训练获得的有监督双语词表示，表 4.2 和表 4.3 分别给出以下 4 种样例表示方法在跨语言情感分类和跨语言模糊限制语识别任务中的性能：

（1） EN-EN：基于W_E表示英文训练样例和中译英测试样例；

（2） CN-CN：基于W_C表示中译英训练样例和中文测试样例；

（3） EN-CN：基于W_E表示英文训练样例，基于W_C表示中文测试样例；

（4） CR：本文提出的中英文连接表示方法。

表 4.2 跨语言情感分类中不同表示方法对分类性能的影响

Tab. 4.2 The effects of different representations on CLSC performance

表示方法	书籍	DVD	音乐	*Average*
EN-EN	76.05	77.90	76.78	76.91
CN-CN	78.95	79.03	74.85	77.61
EN-CN	77.85	79.30	73.28	76.81
CR	**81.05**	**81.60**	**79.40**	**80.68**

从表 4.2 可以看出，本文提出的 CR 样例表示方法的性能明显高于其他 3 种表示方法。说明 CR 表示方法可以充分包含中英文的语义和情感信息，有助于提高跨语言情感分类性能。

表 4.3 跨语言模糊限制语识别中不同表示方法对分类性能的影响

Tab. 4.3 The effects of different representations on CLHCD performance

识别任务	表示方法	摘要			全文		
		P	*R*	*F*	*P*	*R*	*F*
句子识别	EN-EN	38.59	56.43	45.84	63.76	57.75	60.61
	CN-CN	61.85	9.75	16.84	79.59	11.85	20.63
	EN-CN	26.55	100.00	41.96	59.07	100.00	**74.27**
	CR	38.48	85.86	**53.14**	65.73	78.12	71.39
词语识别	EN-EN	23.76	58.98	33.87	50.37	65.63	57.00
	CN-CN	36.12	8.18	13.34	64.65	8.98	15.77
	EN-CN	17.27	100.00	29.45	45.92	100.00	**62.94**
	CR	21.55	89.07	**34.70**	48.70	84.18	61.70

从表 4.3 可以看出，在摘要中，CR 取得了最好的识别性能，而在全文中，识别性能略低于 EN-CN 表示方法。在表 4.3 中，CN-CN 方法的召回率非常低，仅有 10%左右，说明测试样例中很多的模糊限制性句子并没有被正确地识别。原因在于 CN-CN 方法基于W_C表示训练样例和测试样例，采用英译中训练样例学习分类模型，对中文测试样例进行分类。然而中英文模糊限制语差异较大，许多含有中文候选词的英译中训练样例被

标注为不含有模糊性的句子。这样，分类模型会将包含这些中文候选词的中文测试样例分为负例，即不具有模糊性的句子，造成低召回率的现象。

4.3.2 双语词表示的跨语言语义表达能力

基于双语词表示的学习方法可以将中英文词表示映射到同一个空间。在这个空间内，通过欧氏距离可以找到语义最近和最远的两个词。同时，本文在训练双语词表示过程中加入有监督学习阶段，可以有效地将标注信息嵌入词表示中。以跨语言情感分类任务为例，加入监督学习，可以将情感信息嵌入双语词表示。这样，词表示的距离不仅可以衡量情感词的语义关系，还可以衡量情感词的情感倾向：距离越近，情感倾向相近；距离越远，情感倾向相反。表 4.4 给出了在跨语言情感分类中，与“disappointed”和“good”距离最近和最远的 5 个英文情感词和中文情感词。

表 4.4 与“disappointed”和“good”距离最近和最远的 5 个中文和英文情感词

Tab. 4.4 Top 5 similar and opposite English and Chinese sentiment words to “disappointed” and “good”

查询词	距离最近的英文情感词	距离最远的英文情感词	距离最近的中文情感词	距离最远的中文情感词
disappointed	broke	great	遗憾	爱
	views	comedy	失望	最差
	annoyed	love	观念	笑
	calming	fun	偏离	伟大
	dismayed	blah	含蓄	乐趣
good	beliefs	worst	最好	可怕
	annoyed	waste	好	垃圾
	educated	fun	更好	浪费
	comprehend	awful	不错	乐趣
	ranting	beware	很好	糟糕

从表 4.4 可以看出，与查询词距离最近的词，情感倾向相近，与查询词距离最远的词，情感倾向相反。以情感词“disappointed”为例，与它距离最近的英文情感词 “broke”、“views”、“annoyed”、“calming”和“dismayed”，几乎都有着相同的消极情感倾向。而与它距离最远的“great”、“comedy”、“love”、“fun”和“blah”，几乎都有着与其相反的积极情感倾向。同时，与“disappointed”距离最近和最远的中文情感词也存在着这样的现象：距离越近，情感倾向越相近；距离越远，情感倾向越相反。

但同时也可以看出，有一些与查询词距离较近的词，它们的情感倾向与查询词并不相近，甚至相反。原因在于有些文档包含与文档情感倾向相反的情感词，而在有监督学习阶段，分类模型基于文档的情感标注计算分类损失，然后对双语词表示进行“优化”。这样使消极（积极）情感倾向的双语词表示向积极（消极）情感方向调整，出现上述现象。图 4.4 是一则关于书籍的评价，标注的情感倾向是消极的。然而在评论中涵盖多个积极倾向的情感词，影响该文档中双语词表示的优化过程。

<item>
<summary>**Excellent** Writer.....Poor Plot</summary>
<polarity>N</polarity>
<text>West is without a doubt a superb writer, however I found The Shoes of The Fisherman to be a **great** disappointment. In spite of the fact that most of the reviews I have read highly **praise** the book, it never really got to me.
…
I can **honestly** say it never got my full attention.
</text>
</item>

图 4.4　跨语言情感分类书籍评价示例

Fig. 4.4　An example of book reviews in CLSC task

表 4.5 给出了跨语言模糊限制语识别中，与英文（中文）查询词距离最近的中文（英文）候选词。

表 4.5　与中文（英文）模糊限制语候选词距离最近的英文（中文）模糊限制语候选词

Tab. 4.5　Top 5 similar English (Chinese) candidates of the Chinese (English) candidate

语料	查询词	距离最近的英文候选词	查询词	距离最近的中文候选词
摘要	认为	thought	possible	可能
		believed		有可能
		considered		和/或
		feel		或许
		think		尚不明确
全文	表明	suggests	might	可能
		indicated that		有些
		suggesting		一旦
		indicate that		也许
		suggest		众多

从表 4.5 可以看出，距离最近的候选词虽然与查询词所属不同的语言，但表达的语义信息基本一致，进一步说明本章提出的双语词表示的有效性。

4.4 本章小结

提出基于双语词表示的跨语言信息抽取方法，双语词表示学习分为无监督学习阶段和有监督学习阶段。实验证明本章提出的方法在学习无监督双语词表示时，可以捕获中英文双语语义信息，克服了双视图方法难以深入融合两种语言的不足。此外，通过计算词表示的距离，还能够找到跨语言的近义词。有监督学习阶段将训练语料的标注信息嵌入双语词表示，可以有效提高跨语言信息抽取性能。

5　基于联合表示学习的跨语言信息抽取

双语词表示方法缺乏词的上下文语境信息，同时，双语语义信息学习的无监督学习阶段与标注信息学习的有监督学习阶段是完全独立的。

提出一种联合表示学习的跨语言信息抽取方法（Cross-language Information Extraction based on Joint Representation Learning，JRL）。采用长短时记忆递归网络（LSTM），学习中英文双语表示。在词语义表示（Word Semantic Representations，WSR）基础上，引入上下文情感（模糊）信息表示，联合训练情感词（模糊限制语）在特定语境下的语义和情感（模糊）信息。同时，将双语语义学习和标注信息学习合为一个学习阶段，进一步优化跨语言信息抽取性能。

5.1　上下文情感（模糊）信息表示学习

引入上下文情感（模糊）信息表示（context sentiment/hedge information representations，CSIR/CHIR），用于联合表示学习。

5.1.1　上下文情感信息表示（CSIR）

在情感分类任务中，情感词的上下文信息对判断文档的情感倾向起着重要作用。本文基于长短时记忆递归网络（LSTM）训练获得 CSIR。

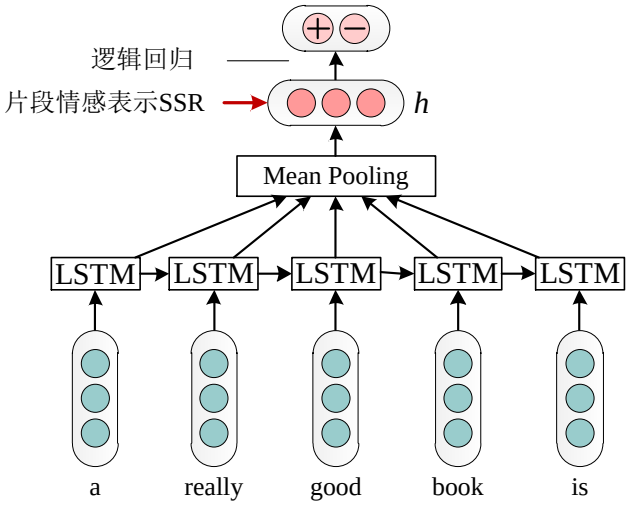


图 5.1　基于 LSTM 训练获得 SSR 的示意图

Fig. 5.1　An example of SSR trained based on LSTM

首先，以情感词 $senti_i$ 为中心，窗口为 $[-2,2]$，抽取中英文情感词的上下文序列，获得情感片段 $x=\{word_{i-2}, word_{i-1}, senti_i, word_{i+1}, word_{i+2}\}$。

然后，将获得序列的 50 维词表示作为 LSTM 每一时刻 t 的输入，计算获得 t 时刻的隐藏表示 h_t，计算过程详见 2.3.2 节。

再将所有时刻的隐藏表示进行平均池化（Mean Pooling）操作，获得情感片段 50 维的隐藏表示 h，我们将 h 称为片段情感表示（Segment Sentiment Representations，SSR）。将 h 输入逻辑回归层，基于情感词的情感极性，计算分类损失。

最后，基于 Adadelta 算法[46]更新模型参数，训练获得 SSR。图 5.1 为基于 LSTM 训练 SSR 的示意图。

在获得 SSR 后，将同一个句子的片段情感表示相加，获得句子级别的上下文情感信息表示（CSIR）。

5.1.2 上下文模糊信息表示（CHIR）

在模糊限制语识别任务中，候选词的共现特征有助于提高识别性能，即如果一个句子出现两个或两个以上候选词，那么这些候选词很有可能是模糊限制语，该句子极有可能是模糊限制性句子[26]。基于此，本文将跨语言模糊限制语识别任务的上下文模糊信息表示定义为：如果句子中出现两个或两个以上候选词，上下文模糊信息表示即为该句中候选词的语义表示相加。如果句子中只有一个候选词，上下文模糊信息表示为与词表示维度相同的零向量 $\vec{0}$。

5.2 语义信息和情感（模糊）信息的联合表示学习

在获得 CSIR（CHIR）后，将句子中的所有词的语义表示与该句的 CSIR（CHIR）连接，用于语义信息和情感（模糊）信息的联合表示学习。

图 5.2 为联合表示学习框架，我们将英文训练样例看作英文词序列，输入到 LSTM，训练获得英文长短时记忆递归网络 LSTM_E；将英译中训练样例看作中文词序列，输入到 LSTM，训练中文长短时记忆递归网络 LSTM_C。再分别将英文序列和英译中序列中每一时刻生成的隐藏表示进行平均池化，获得英文隐藏表示 h^E 和中文隐藏表示 h^C。关于 LSTM 中每一时刻隐藏表示的生成过程，详见 2.3.2 节。最终将训练获得的英文和中文隐藏表示 h^E 和 h^C 输入到逻辑回归层，计算分类损失。

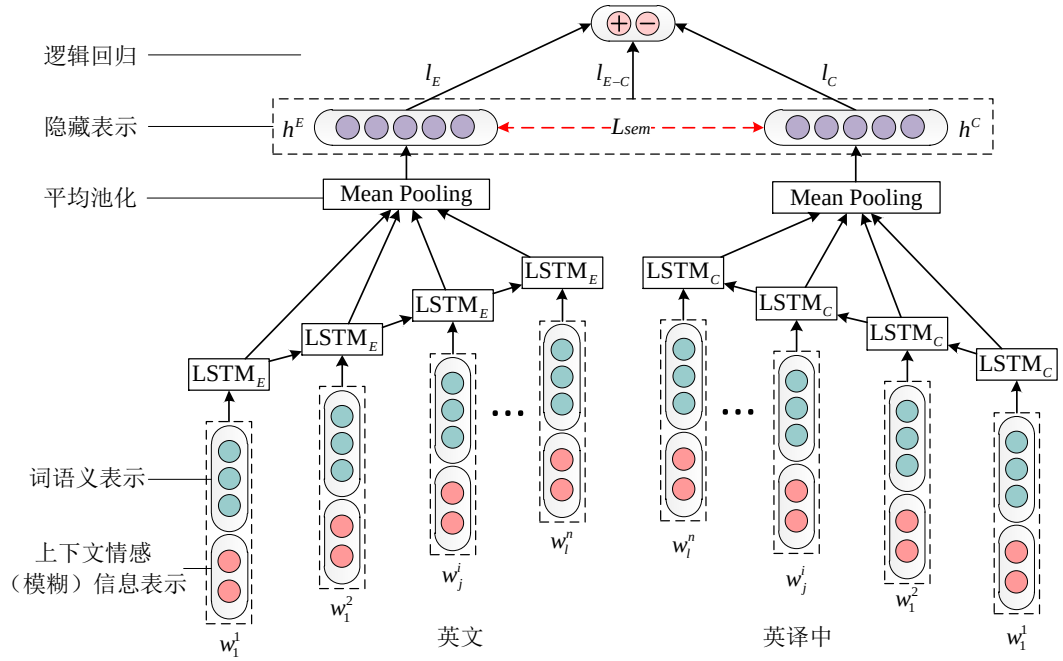


图 5.2　联合表示学习框架

Fig. 5.2　The framework of joint representation learning

本文基于对数似然函数定义三种分类损失：l_E、l_C和l_{E-C}。l_E为英文隐藏表示h^E的分类损失，l_C为中文隐藏表示h^C的分类损失，l_{E-C}为中英文隐藏表示连接$[h^E, h^C]$的分类损失。将上述三种分类损失相加作为模型的分类损失L_{pred}，如公式（5.1）所示。

$$L_{pred} = l_E + l_C + l_{E-C} \tag{5.1}$$

此外，根据神经网络机器翻译的相关研究，LSTM 可以将源语言的向量编码成隐藏表示，同时可以解码成目标语言的向量。受文献[47]的启发，我们认为隐藏表示h^E和h^C的差异应尽可能地小，这样可以减小样例的中英文向量差异。因此定义双语语义损失L_{sem}减少两种语言之间的鸿沟，如公式（5.2）所示。

$$L_{sem} = \left\| h^E - h^C \right\|_2^2 \tag{5.2}$$

最终，本文计算分类损失与双语语义损失的加权和，作为整个模型的目标函数，如公式（5.3）所示，其中α为分类损失的权重），并基于 Adadelta 算法学习模型参数。

$$L = \alpha \cdot L_{pred} + (1-\alpha) \cdot L_{sem} \tag{5.3}$$

5.2.1 跨语言情感分类

在跨语言情感分类中，将文档中所有词的语义表示与上下文情感信息表示进行连接，作为 LSTM 的输入，用于联合学习跨语言语义信息和情感信息。

5.2.2 跨语言模糊限制语识别

在跨语言模糊限制语识别中，与第 3、4 章不同的是，本章基于 LSTM 将模糊限制语识别看作机器学习的二值分类问题，将模糊限制性句子识别转换为模糊限制语识别任务。在进行模糊限制语识别之前，本文采用词典匹配的方法进行候选样例筛选：将匹配出来的英文候选词作为训练样例，模糊限制语作为正例，非模糊限制语作为负例；将匹配出来的中文候选词作为测试样例，用于进一步识别。通过候选样例筛选，可以过滤掉许多非模糊限制语的词，提高识别性能。

...	...	...	...	...	...	...
program	on	an	**unknown**	trna	family	which
trna	family	which	**may**	have	atypical	sequence
and	structure	is	**unclear**	thereby	rendering	their
NULL	NULL	NULL	**assuming**	that	the	23rd
codon	we	systematically	**predicted**	proteins	that	contain
...	...	...	...	...	...	...

（a）英文训练样例示例图

...	...	...	...	...	...	...
造成	损伤	,	**如果**	不能	使	细胞
损伤	降至	最低	**或**	零损伤	,	打印
随着	细胞	的	**大量**	死亡	而	失去
,	与	文献所	**报道**	的	其他	同类
,	细胞	存活率	**较**	高	。	NULL
...	...	...	...	...	...	...

（b）中文测试样例示例图

图 5.3 预处理跨语言模糊限制语识别语料示例

Fig. 5.3 An example of preprocessed corpora in CLHCD

在获得测试样例的分类结果后，计算模糊限制性句子识别结果：

（1） 如果句子中不包含候选词，则该句被判断为非模糊限制性句子；

（2） 如果句子中包含一个候选词，且该候选词被分为正例，则判断该句子为模糊限制性句子；如果候选词被分为负例，则判断该句子为非模糊限制性句子；

（3） 如果句子中包含多个候选词，且只要有一个候选词被分为正例，则判断该句子为模糊限制性句子，否则为非模糊限制性句子。

为了提高模糊限制语识别性能，本文设置窗口[−3,3]抽取候选词的上下文信息，并将训练样例和测试样例表示为词序列的形式，作为 LSTM 的输入。图 5.3（a）为英文训练样例示例，图 5.3（b）为中文测试样例示例。

5.3 实验结果及分析

基于 Theano 库函数构建长短时记忆递归网络（LSTM）。

5.3.1 预训练词表示

在进行联合表示学习之前，采用预训练词表示初始化词的语义表示。下面介绍中英文预训练的词表示。

（1） 英文词表示

在跨语言情感分类任务中，采用 Pennington 等[48]公开发表的 GloVe 词表示初始化英文词语义表示。词表示基于 Twitter 语料训练获得的，包括 1.2M 个词汇，本文采用 50、100 和 200 维的词表示。

在跨语言模糊限制语识别任务中，下载了约 9G 的英文生物医学领域的语料，并采用 Word2Vec 工具[49]基于这些语料训练 25、50、100 和 150 维的词表示。

（2） 中文词表示

在跨语言情感分类任务中，下载了约 60M 的中文产品评论相关语料，利用 Word2Vec 工具训练 50、100 和 200 维的词表示。

在跨语言模糊限制语识别任务中，下载了约 2G 的中文语料，涉及新闻、微博等领域，利用 Word2Vec 工具训练 25、50、100 和 150 维词表示。

5.3.2 上下文情感（模糊）信息表示的有效性

比较了上下文情感（模糊）信息表示对跨语言信息抽取性能的影响。在联合表示学习时，目标函数为分类损失和双语语义损失之和，且二者之比为 1:1，即 $\alpha = 0.5$。

表 5.1 和表 5.2 分别给出了不同词表示方法对应的跨语言情感分类准确率和跨语言模糊限制语识别的 P、R 和 F 值。表中“d-WSR”表示 d 维词的语义表示，“d-WSR+50-CSIR”表示 d 维词的语义表示与 50 维上下文情感信息表示的连接，“d-WSR+50-CHIR”表示 d 维词的语义表示与 50 维上下文模糊信息表示的连接。

从表 5.1 可以看出，在跨语言情感分类任务中，随着词语义表示的维度增加，在书籍、DVD 和音乐三个领域，准确率逐步提高，平均准确率呈上升趋势，当语义表示维度达到 100 维时，平均准确率达到最高 80.20%。从上述实验结果可知高维度词的语义

表示包含更多的语义信息，有助于提高分类性能。同时，在词的语义表示基础上加入上下文情感信息表示，可以更加准确地反映情感词特定语境下的情感信息，提高跨语言情感分类性能。

表 5.1　不同维度词表示的跨语言情感分类准确率比较

Tab. 5.1　The comparison of CLSC performances with different word representations

词表示	书籍	DVD	音乐	*Average*
50-WSR	76.35	76.05	74.90	75.77
100-WSR	77.43	78.73	77.33	77.83
200-WSR	77.90	78.80	77.78	78.16
50-WSR+50-CSIR	79.40	80.60	77.85	79.28
100-WSR+50-CSIR	**81.18**	80.73	**78.70**	**80.20**
200-WSR+50-CSIR	79.88	**81.50**	77.58	79.65

表 5.2　不同维度词表示的跨语言模糊限制语识别性能比较

Tab. 5.2　The comparison of CLHCD performances with different word representations

识别任务	词表示	摘要			全文		
		P	*R*	*F*	*P*	*R*	*F*
句子识别	25-WSR	71.71	51.52	59.96	67.10	79.33	72.70
	50-WSR	63.68	57.36	60.36	66.99	83.89	74.49
	100-WSR	51.49	80.72	**62.88**	65.00	90.88	75.79
	150-WSR	68.34	56.37	61.78	62.66	91.79	74.48
	100-WSR+50-CHIR	69.34	56.13	62.04	79.61	73.56	**76.46**
	150-WSR+50-CHIR	63.35	56.13	59.52	68.81	81.16	74.48
词语识别	25-WSR	63.13	42.01	50.45	50.60	65.43	57.07
	50-WSR	56.50	47.35	51.52	52.10	70.31	59.85
	100-WSR	43.53	70.04	53.69	48.89	77.73	60.03
	150-WSR	59.47	46.23	52.02	47.04	80.66	59.42
	100-WSR+50-CHIR	53.95	53.85	**53.90**	57.77	74.80	**65.19**
	150-WSR+50-CHIR	49.84	53.12	51.43	51.82	83.20	63.87

从表 5.2 可以看出，在跨语言模糊限制语识别中，与基于 DAE 训练的双语词表示性能相比，基于 LSTM 训练的词语义表示获得更高的识别性能。原因在于基于 DAE 训练的双语词表示虽然包含中英文的双语语义信息，但缺乏词的上下文语境信息。而基于 LSTM 训练中英文双语表示可以充分利用候选词的上下文，捕获词序列深层的语义信

息。同时，由于共现特征能够提高模糊限制语识别性能，将句子中共现的候选词语义表示相加作为上下文模糊信息表示，有助于提高模糊限制语识别性能。

5.3.3 双语语义信息与标注信息对跨语言信息抽取性能的影响

通过调整目标函数中分类损失和双语语义损失的权重系数，调整标注信息和双语语义信息在模型训练中的重要程度。图 5.4 给出了跨语言情感分类任务中，在书籍、DVD 和音乐领域$\alpha \in [0.1,1]$对应的分类准确率。其中，书籍、DVD 和音乐领域采用 100 维的词语义表示和 50 维上下文情感信息表示进行实验。

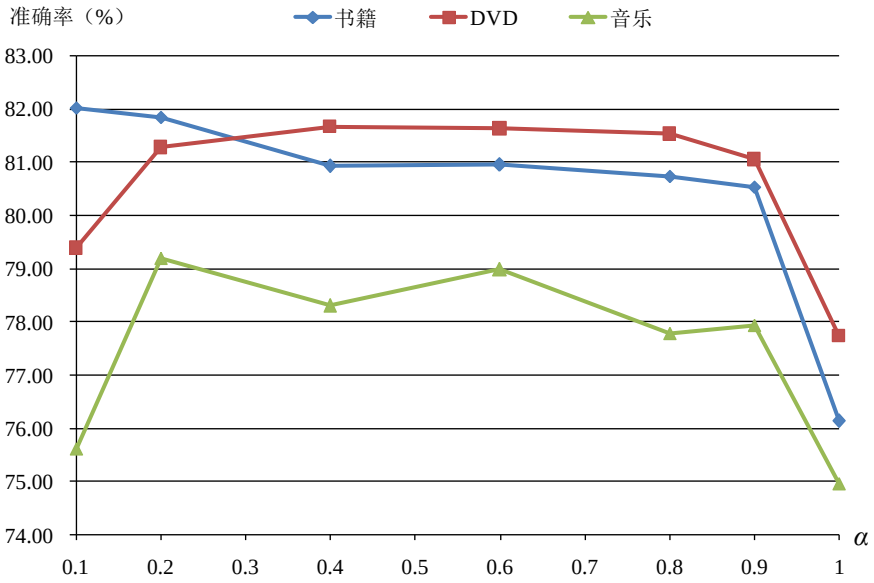


图 5.4 跨语言情感分类中双语语义信息的影响

Fig. 5.4 The effects of bilingual semantic information in CLSC

当$\alpha = 0$时，目标函数即为双语语义损失，此时未使用标注信息进行学习。我们发现三个领域的分类准确率仅为 50%左右。当$\alpha = 1$时，目标函数即为分类损失，书籍、DVD 和音乐三个领域的分类准确率分别为 76.13%、77.73%和 74.95%。从图 5.4 可以看出，三个领域的分类曲线分别从$\alpha = 0$时的 50%左右上升到各自分类准确率的最高值，然后下降到$\alpha = 1$时各自的分类准确率。上述实验结果可以说明，双语语义信息和标注信息对跨语言情感分类均有效，而且二者组合的分类性能优于单纯的双语语义或标注信息的分类性能。同时，当α分别为 0.1、0.4 和 0.2 时，书籍、DVD 和音乐三个领域的分

类准确率分别达到最高，即 82.03%、81.68%和 79.20%，说明在三个领域中，双语语义信息对提高分类性能更为重要。

图 5.5 给出了跨语言模糊限制语识别任务，在摘要和全文中$\alpha \in [0.1,1]$对应的识别 F 值。摘要和全文中均采用 100 维词的语义表示和 50 维上下文模糊信息表示进行实验。

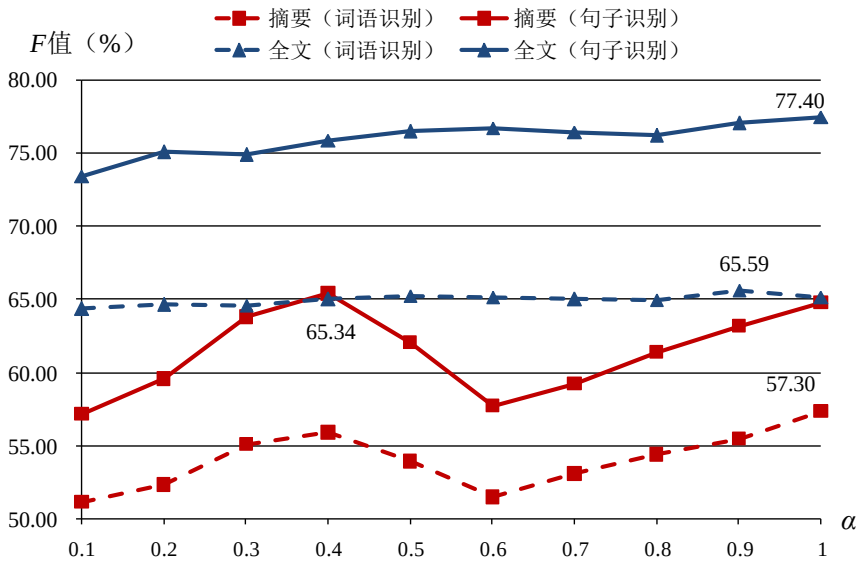


图 5. 5　跨语言模糊限制语识别中双语语义信息的影响

Fig. 5.5　The effects of bilingual semantic information in CLHCD

从图 5.5 可以看出，在全文中，双语语义信息的改变并没有对分类性能起到明显的作用，模糊限制语识别的 F 值几乎保持不变，最高达到 65.59%。对应的模糊限制性句子识别 F 值随着分类损失权重系数α 的增加而缓慢升高，当$\alpha=1$时，即目标函数为分类损失时，F 值达到最高 77.40%。在摘要中，双语语义信息对分类性能影响较大，识别 F 值随权重系数的增加上下波动。当$\alpha=0.4$时，模糊限制性句子识别的 F 值达到最高，为 65.34%，当$\alpha=1$时，模糊限制语识别的 F 值最高，为 57.30%。从实验结果可以看出，当α 接近于 1 时，两个领域均获得较高的 F 值，即目标函数为分类损失时识别效果较好。这说明中英文模糊限制语之间存在较大差异，缩小候选词中英文隐藏表示的差异会影响模糊限制语的识别性能。

5. 3. 4　与相关研究的比较

基于 NLPCC 2013 会议的“跨语言情感分类”评测任务提供的语料进行实验，该评测任务的评价指标即为本文所给出的准确率。表 5.3 比较了近年来基于该语料的跨语言情感分类相关研究与本文提出的方法。

表 5.3　与相关研究比较

Tab. 5.3　Comparison between related works

文献	书籍	DVD	音乐	*Average*
文献[16]	78.70	79.65	78.30	78.89
文献[18]	80.10	81.60	78.60	80.10
文献[19]	84.65	85.18	80.93	83.59
本文方法	82.03	81.68	79.20	80.97

文献[16]提出了一种将自训练（Self-training）和协同训练（Co-training）相结合的跨语言情感分类方法，取得了 NLPCC 2013 评测任务的第一名。文献[18]和文献[19]均基于迁移学习的方法进行跨语言情感分类。文献[18]在迁移学习过程中引入样例迁移监测机制，避免发生负迁移，减少噪音样例对分类性能的影响。文献[19]基于双视图的方法进行迁移学习。他们将英文视图看作“导师”，通过训练英文分类模型，将高可信度英文样例的中文翻译推荐给中文视图；同时，将中文视图看作“学生”，通过引入一种知识验证函数判断“导师”提供的中文翻译样例的可信度，并将高可信度的中文样例加入中文训练集中，最终取得 83.59%的平均准确率。当未标注数据与测试数据分布相同时，加入的未标注数据会大幅度提高分类性能，反之，则会发生负迁移，降低信息抽取性能。因此，迁移学习的方法非常依赖未标注的数据分布情况。

本文方法仅采用英文训练语料及其中文翻译联合学习词的语义表示和上下文情感信息表示，提高跨语言情感分类性能，具有更好的通用性和可扩展性。

5.4　本章小结

提出一种基于联合表示学习的跨语言信息抽取方法。采用 LSTM 学习词语义表示，同时，引入上下文情感（模糊）信息表示，联合训练捕获情感词（模糊限制语）在特定语境下的语义信息和情感（模糊）信息，提高跨语言信息抽取性能。实验表明，通过 LSTM 学习获得中英文双语语义表示，能够有效挖掘特征深层次的语义信息。同时，词语义表示与上下文情感（模糊）信息进行联合表示学习，能够进一步提高跨语言信息抽取的性能。

结　论

本文研究了深度学习在跨语言信息抽取任务上的应用，分为以下三部分：

（1） 基于降噪自动编码机的双视图跨语言信息抽取

在跨语言信息抽取中，源语言和目标语言之间的语言鸿沟会影响信息抽取性能。同时，机器翻译产生的翻译错误也会进一步降低系统性能。提出一种基于降噪自动编码机的双视图跨语言信息抽取方法。在降噪自动编码机重构过程中，适当引入噪音，提高信息抽取系统的鲁棒性和抗噪音能力。同时分别在中英文两个视图下构建分类模型，最后线性融合两个视图各自的分类结果，减少语言鸿沟对分类性能的影响。实验表明，在重构过程中，适当引入噪音，可以减小翻译错误的影响，提高分类性能。同时，在跨语言情感分类任务中，双视图的方法可以充分发挥中英两种语言的互补优势，修正单语视图下的分类错误，进一步提高分类性能。

（2） 基于双语词表示的跨语言信息抽取

为了克服双视图方法难以深入融合中英文两种语言的不足，提出基于双语词表示的跨语言信息抽取方法，学习过程分为无监督学习阶段和有监督学习阶段。在无监督学习阶段，利用降噪自动编码机进行中英文双语重构，捕获双语语义信息。在有监督学习阶段，将训练语料的标注信息嵌入双语词表示。在跨语言情感分类和跨语言模糊限制语识别任务上的实验表明，双语词表示能够有效捕获双语语义信息，通过计算词表示的距离，还能够找到跨语言的近义词。同时，通过引入有监督学习阶段，将标注信息嵌入双语词表示中，可以提高跨语言信息抽取性能。

（3） 基于联合表示学习的跨语言信息抽取

为了解决基于降噪自动编码机方法产生的数据稀疏问题，提出一种基于联合表示学习的跨语言信息抽取方法。采用长短时记忆递归网络（LSTM），学习中英文的词语义表示。在此基础上，引入上下文情感（模糊）信息表示，联合训练获得情感词（模糊限制语）在特定语境下的语义信息和情感（模糊）信息。在跨语言情感分类和跨语言模糊限制与识别任务上的实验表明，基于 LSTM 学习获得中英文双语语义表示，能够挖掘特征深层次的语义信息，获得比基于降噪自动编码机方法更好的分类性能。同时，词语义表示与上下文情感（模糊）信息的联合表示学习，进一步了提高跨语言信息抽取的性能。

本文研究了深度学习模型在跨语言信息抽取任务上的应用，提出了基于降噪自动编码机的双视图跨语言信息抽取方法，减小了翻译错误和语言鸿沟对跨语言信息抽取性能的影响。提出了基于双语词表示的跨语言信息抽取方法，克服了双视图方法难以深入融合中英文两种语言的不足。提出了基于联合表示学习的跨语言信息抽取方法，充分挖掘

特征深层次语义信息。这些研究有效地提高跨语言情感分类和跨语言模糊限制语识别系统的性能，为后续的跨语言信息抽取研究提供有益借鉴。

参 考 文 献

[1] 李宝利,陈玉忠,俞士汶.信息抽取研究综述[J].计算机工程与研究,2003(10):1-5.

[2] Wan X. Co-training for cross-lingual sentiment classification[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009:235-243.

[3] Tang X, Wan X. Learning Bilingual Embedding Model for Cross-Language Sentiment Classification[C]//Web Intelligence (WI) and Intelligent Agent Technologies (IAT), International Joint Conferences on IEEE/WIC/ACM, 2014:134-141.

[4] Zhou G, He T, Zhao J. Bridging the language gap: Learning distributed semantics for cross-lingual sentiment classification[C]//Proceedings of Natural Language Processing and Chinese Computing, 2014:138-149.

[5] Li S, Wang R, Liu H, et al. Active Learning for Cross-Lingual Sentiment Classification[C]//Proceedings of Natural Language Processing and Chinese Computing, 2013:236-246.

[6] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504-507.

[7] Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(8):1915-1929.

[8] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1):30-42.

[9] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011(12):2493-2537.

[10] 刘建伟,刘媛,罗雄麟.深度学习研究进展[J].计算机应用研究, 2014(7):1921-1930.

[11] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning, 2009, 2(1):1-127.

[12] Chandar S, Khapra M, Ravindran B, et al. Multilingual Deep Learning[J]. Deep Learning Workshop (NIPS), 2013.

[13] Lauly S, Larochelle H, Khapra M, et al. An autoencoder approach to learning bilingual word representations[C]//Advances in Neural Information Processing Systems. 2014:1853-1861.

[14] Rajendran J, Khapra M, Chandar S, et al. Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning[J]. arXiv preprint arXiv: 1510.03519, 2015.

[15] Huang J T, Li J, Yu D, et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers[C]//Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, 2013:7304-7308.

[16] Gui L, Xu R, Xu J, et al. A mixed model for cross lingual opinion analysis[C]//Proceedings of Natural Language Processing and Chinese Computing, 2013:93-104.

[17] Xu R, Xu J, Wang X. Instance level transfer learning for cross lingual opinion analysis[C]//Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics, 2011: 182-188.

[18] Gui L, Xu R, Lu Q, et al. Cross-lingual Opinion Analysis via Negative Transfer Detection[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (short paper), 2014:860-865.

[19] Chen Q, Li W, Lei Y, et al. Learning to Adapt Credible Knowledge in Cross-lingual Sentiment Analysis[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015:419-429.

[20] Zeng D, Liu K, Lai S, et al. Relation Classification via Convolutional Deep Neural Network[C]//Proceedings of COLING, 2014:2335-2344.

[21] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015:1785-1794.

[22] LeCun Y, Jackel L D, Bottou L, et al. Comparison of learning algorithms for handwritten digit recognition[C]//International conference on artificial neural networks, 1995:53-60.

[23] Zou W Y, Socher R, Cer D M, et al. Bilingual Word Embeddings for Phrase-Based Machine Translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013:1393-1398.

[24] Yang Z l, Salakhutdinov R, Cohen W. Multi-Task Cross-Lingual Sequence Tagging from Scratch[J]. arXiv preprint arXiv:1603.06270, 2016.

[25] Lakoff G. Hedges: A study in mearning criteria and the logic of fuzzy concepts[J]. Journal of philosophical logic, 1973, 2(4):458-508.

[26] 周惠巍,杨欢,张静,等. 中文模糊限制语语料库的研究与构建[J]. 中文信息学报, 2015(6):83-89.

[27] Farkas R, Vincze V, Móra G, et al. The CoNLL 2010 shared task: learning to detect hedges and their scope in natural language text[C]//Proceedings of the Fourteenth

Conference on Computational Natural Language Learning (CoNLL 2010): Shared Task, 2010:1-12.

[28] Cortes C, Vapnik V. Support vector networks[J]. Machine learning, 1995, 20(3):273-297.

[29] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002:79-86.

[30] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications, 2015, 42(4):1857-1863.

[31] Liu S, Li F, Li F, et al. Adaptive co-training SVM for sentiment classification on tweets[C]//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, 2013:2079-2088.

[32] 李航.统计学习方法[M].北京:清华大学出版社,2012.

[33] 胡候立,魏维,谢青松.深层自动编码机的文本分类算法改进[J].计算机应用研究,2015(4):992-995.

[34] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning, 2008:1096-1103.

[35] Mesnil G, He X, Deng L, et al. Investigation of recurrent neural network architectures and learning methods for spoken language understanding[C]// INTERSPEECH, 2013:3771-3775.

[36] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML-15),2015:2342-2350.

[37] Hochreiter,S. and Schmidhuber,J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[38] Gers F A, Schmidhuber J. Recurrent nets that time and count[C]//Neural Networks. Proceedings of the IEEE-INNS-ENNS International Joint Conference on, 2000(3):189-194.

[39] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5):513-523.

[40] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C]//Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005:347-354.

[41] Galavotti L, Sebastiani F, Simi M. Feature selection and negative evidence in automated text categorization[C]//Proceedings of KDD, 2000.

[42] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]//Proceedings of the second SIGHAN workshop on Chinese language processing, 2003:184-187.

[43] Bergstra J, Breuleux O, Bastien F, et al. Theano: a CPU and GPU math expression compiler[C]//Proceedings of the Python for scientific computing conference (SciPy), 2010(4):3.

[44] Schölkopf B, Burges C J C. Advances in kernel methods: support vector learning[M]. MIT press, 1999.

[45] Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 1555-1565.

[46] Zeiler M D. ADADELTA: an adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.

[47] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1412-1421.

[48] Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014:1532-1543.

[49] 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.

攻读硕士学位期间发表学术论文情况

1 Cross-lingual Sentiment Classification based on Denoising Autoencoder. Huiwei Zhou, **Long Chen**, and Degen Huang. Proceedings of Natural Language Processing and Chinese Computing (NLPCC), 2014: 181-192. 主办单位：China Computer Federation. EI 检索会议，本文 EI 检索号：20145000322473。（本硕士学位论文第 3 章）

2 Learning Bilingual Sentiment Word Embeddings for Cross-language Sentiment Classification. Huiwei Zhou, **Long Chen**, Fulin Shi, and Degen Huang. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL), 2015: 430–440. 主办单位：Association for Computational Linguistics. A 类会议. EI 检索会议，本文 EI 检索号：20154201387017。（本硕士学位论文第 4 章）

3 Jointly Learning Bilingual Sentiment and Semantic Representations for Cross-language Sentiment Classification（已投稿）. Huiwei Zhou, **Long Chen**, Yunlong Yang, Chen Jia, Huijie Deng, and Degen Huang. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016。（本硕士学位论文第 5 章）

4 Combining Feature-based and Instance-based Transfer Learning Approaches for Cross-domain Hedge Detection with Multiple Sources. Huiwei Zhou, Huan Yang, **Long Chen**, Zhenwei Liu, Jianjun Ma, and Degen Huang. Proceedings of 4th National Conference of Social Media Processing (SMP), 2015: 225-232. EI 检索会议，本文 EI 检索号：20161002063566。

5 Exploiting Syntactic and Semantics Information for Chemical–disease Relation Extraction. Huiwei Zhou, Huijie Deng, **Long Chen**, Yunlong Yang, Chen Jia, and Degen Huang. Database, 2016. 文献号：baw048. SCI 检索期刊，本文 SCI 检索号：DJ3IF。

致　　谢

三年的研究生旅途马上就要画上句号了。充实而又难忘的学习和科研生活让我收获很多：拼搏、钻研、乐观，这些宝贵的品质都将成为我不断前行的强大力量，值得我一生珍藏。至此论文即将完成之际，我要向陪伴我度过研究生三年的老师、同学、家人和朋友们表示由衷的感谢！

感谢我的导师周惠巍副教授！三年走来，从研究方向的制定，到实验方法的选择，再到学术论文的发表，最后到毕业论文的撰写，每一步的科研成果都凝结着周老师的心血。每一次认真地与我讨论方法，耐心地听我汇报实验结果，细致地帮我修改论文，都令我难以忘怀。我也从对科研的一无所知，到可以针对某一问题独立思考，再到可以在学术会议的讲台上分享团队的研究成果，这些点滴的成长都离不开周老师的精心栽培！在此，我向周老师表达最真挚的谢意！

感谢黄德根教授。通过聆听黄老师的课程，感受到黄老师为人正直的品格、严谨的治学态度、开阔的视野。黄老师经常在科研之余，组织爬山、徒步等系列活动，增强了实验室成员的凝聚力。相信自然语言处理实验室在黄老师的带领下，会越来越好。

感谢实验室所有的兄弟姐妹们。正是因为大家共同创造浓厚的学术氛围和融洽的生活环境，才让我在科研的道路上有了进步，在生活上成长了许多。感谢你们让我的研究生生活有了美好而难忘的回忆，三年来建立的友情也将是我一辈子珍藏的。

感谢一直支持和陪伴着我的家人和朋友们。你们为我建立了可以依靠的港湾，感谢你们在无数个日日夜夜的陪伴，陪我渡过生活的低谷，分享成功的喜悦，为我加油，给我力量。

最后，再一次感谢在我求学道路上支持和帮助过我的所有人，祝你们健康、快乐！

本论文的研究得到国家自然科学基金项目（基于翻译学习和核方法的中文模糊限制信息检测研究，No.61272375）的资助。

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：____年____月____日

导 师 签 名：_____ 日期：____年____月____日