面向长尾词义消歧的类量子表征方法

Quantum-like Representation Methods for Long-tail Word Sense Disambiguation

一级学科：计算机科学与技术
研究方向：自然语言处理
作者姓名：张俊伟
指导教师：贺瑞芳　教授

答辩日期	2023 年 5 月 24 日		
答辩委员会	姓名	职称	工作单位
主席	赵　鑫	教授	中国人民大学
委员	洪　宇	教授	苏州大学
	张　鹏	教授	天津大学
	熊德意	教授	天津大学
	王龙标	教授	天津大学

天津大学智能与计算学部
二〇二三 年 六 月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **天津大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名:　　签字日期:　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 **天津大学** 有关保留、使用学位论文的规定。特授权 **天津大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名:　　导师签名:

签字日期:　　年　　月　　日　签字日期:　　年　　月　　日

摘　要

词义消歧旨在依据给定的上下文信息为目标词确定词义，属于自然语言处理领域中基础性的研究课题，影响着绝大多数的下游任务。前人工作已经能够有效识别所有高频、常用词义，当前研究的重心转向更为困难的低频词义，即长尾词义消歧。长尾词义消歧的核心挑战在于其面临的低资源问题：匮乏的训练样本使得目标词表征难以学习到所属上下文的特征信息，有限且固定的词义注释文本导致难以学习到准确且易于区分的词义定义表征。其中获取词义定义表征最为困难，不仅要求准确表征词义注释文本，而且需要在语义空间中建立起词义之间的内在联系以及位置关系。在词义消歧过程中，目标词表征与词义定义表征共同决定着词义识别的准确性。本课题针对长尾词义消歧任务中表征学习模型的大数据量依赖问题、低资源下文本表征的不准确性问题、以及词义定义表征间的不可区分性问题，受量子理论中纠缠态、叠加态、量子测量和量子双缝干涉实验的启发，提出量子理论启发的系列表征学习方法和具备整合多种表征方法的多元词义消歧模型，尝试在类量子表征框架下解决低资源下的长尾词义消歧。研究内容与创新点归纳如下：

1. 针对表征学习模型的大数据量依赖问题，提出量子纠缠启发的解耦表征。考虑到基于变分自动编码器实现的解耦表征学习模型拥有着简单的结构、有限的参数数量，能够缓解表征学习模型的大数据量依赖，并且获得的特征仅与单个生成因子相关，能够实现针对性地挑选决定性的特征，本课题提出量子纠缠启发的解耦表征。该表征方法基于纠缠态的解纠缠原理构造强的解纠缠约束机制，用以约束变分自动编码器输出的重采样的特征表示，使采样特征之间彼此独立。该表征方法从表征学习模型优化的角度解决长尾词义消歧任务中低资源下的目标词表征和词义定义表征。

2. 针对低资源下文本表征的不准确性问题，提出量子叠加态启发的自适应调优表征。考虑到量子态与其反态构造的叠加态能够实现基于一个概率幅控制下的希尔伯特空间中的位置变换，本课题提出量子叠加态启发的自适应调优表征。该表征方法基于有限数据中学习到的一个特征描述概率幅，基于数据中学习到的表

征向量构建量子态与其反态，并借助概率幅、量子态、量子态的反态构造叠加态，以实现基于有限数据中学习到的一个有效特征改善表征向量的准确性。该表征方法从表征向量优化的角度解决长尾词义消歧任务中低资源下的目标词表征和词义定义表征。

3. 针对低资源下词义定义表征间的不可区分性问题，提出量子测量启发的多源融合表征。考虑到叠加态的构造过程能够实现量子态之间的融合，叠加态的测量过程所导出的干涉项能够实现量子态之间的空间几何形式下的约束，本课题提出量子测量启发的多源融合表征。该表征方法借助叠加态的构造过程实现融合多源数据获得的词义定义表征，借用叠加态的测量过程所导出的干涉项实现在语义空间中约束词义定义表征之间的位置关系，最终实现构建词义之间的内在联系以及增强词义之间的区分边界。该表征方法采用数据增强的方式解决长尾词义消歧任务中低资源下的词义定义表征。

4. 面向上述类量子表征方法，提出量子干涉实验启发的多元词义消歧模型，以实现整合多种表征方法应对低资源下的长尾词义消歧。考虑到双缝干涉实验揭示了微观粒子间复杂的相干现象，并且属于量子力学理论中最被熟知的实验模型，本课题提出量子干涉实验启发的多元词义消歧模型。该模型基于双缝干涉实验的数学模型所导出，用以整合自适应调优表征、多源融合表征获得的词义定义表征，以实现目标词的多元词义识别与校验，即整合多路词义表征的多元词义消歧模型。该词义消歧模型仿照双缝干涉实验的设置方式构造，以实现从量子认知的视角理解、审视词义识别过程。

最后，基于解耦表征、自适应调优表征、多源融合表征分别构造词义消歧模型并在标准词义消歧评估框架下实施验证，实验结果证实解耦表征有助于降低表征学习模型的大数据量依赖，自适应调优表征能够有效对抗低资源下的文本表征不准确性，多源融合表征有利于强化词义定义表征之间的可区分性。此外，在构造的长尾词义评估数据集上对整合多路词义表征的多元词义消歧模型实施验证，实验结果证实类量子表征框架有助于改善长尾词义识别的准确性。

关键词： 长尾词义消歧，类量子模型，解耦表征，自适应调优表征，多源融合表征，多元词义消歧模型

ABSTRACT

Word Sense Disambiguation (WSD) aims to determine the sense of the target word according to the given context, which is a basic research topic and affects most downstream tasks in the field of Natural Language Processing (NLP). Previous work has been able to effectively identify all high-frequency and commonly used senses, and the focus of current research has shifted to the more difficult low-frequency senses, that is, long-tail WSD. The core challenge of long-tail WSD lies in the low resource: the lack of training samples makes it difficult for the target word representation to learn the characteristic information of the context, and the limited and fixed word sense annotation text makes it difficult to learn the accurate and easily distinguishable word sense definition representation. Among them, obtaining word sense definition representations is the most difficult, not only requires the accuracy of representation vectors, but also needs to establish the internal connection and positional relationship between word senses in the semantic space. In the process of disambiguation, the target word representation and word sense definition representation jointly determine the accuracy of word sense recognition. Inspired by entangled states, superposition states, quantum measurements and quantum double-slit interference experiments in quantum theory, this paper aims at the large data volume dependence of representation learning models, the inaccuracy of text representations under low resources, and the indistinguishability between word sense definition representations in the long-tail WSD task, and proposes a series of representation learning methods and a multivariate WSD model that integrates multiple representation methods. They attempt to solve long-tail WSD under a quantum-like representation framework inspired by quantum theory. The research content and innovation points are summarized as follows:

1. A disentangled representation inspired by quantum entanglement is proposed for the large data volume dependence of representation learning models. Considering that

the disentangled representation learning model based on the variational autoencoder has a simple structure and a limited number of parameters, which can alleviate the large data volume dependence of the representation learning model, and the obtained features are only related to a single generation factor, which select decisive features in a targeted manner, this paper proposes a disentangled representation inspired by quantum entanglement. The representation method constructs a strong disentanglement constraint mechanism based on the disentanglement principle of entangled states, which is used to constrain the resampled feature representation output by the variational autoencoder, and make the sampling features independent of each other. The representation method solves the target word representation and word sense definition representation under low resources in the long-tail WSD task from the perspective of representation learning model optimization.

2. An adaptive tuning representation inspired by quantum superposition state is proposed for the inaccuracy of text representation under low resources. Considering that the superposition state constructed by the quantum state and its antistate can realize the position transformation in the Hilbert space under the control of a probability amplitude, this paper proposes an adaptive tuning representation inspired by the quantum superposition state. The representation method describes the probability amplitude based on a feature learned in the limited data, constructs the quantum state and its antistate based on the representation vector learned in the data, and uses the probability amplitude, quantum state, and the antistate of the quantum state to construct a superposition state, in order to improve the accuracy of the representation vector based on an effective feature learned from limited data. The representation method solves the target word representation and word sense definition representation under low resources in the long-tail WSD task from the perspective of representation vector optimization.

3. A multi-source fusion representation inspired by quantum measurement is proposed for the indistinguishability between word sense definition representations. Considering that the construction process of the superposition state can realize the fusion between quantum states, and the interference term derived from the measurement process of the superposition state can realize the constraint between the quantum states in the geometric form of space, this paper proposes a multi-source fusion representation inspired by quantum measurement. The representation method uses the superposition

state construction process to fuse the word sense definition representations obtained by multi-source data, borrows the interference item derived from the superposition state measurement process to constrain the positional relationship between the word sense definition representations in the semantic space, and finally establishes the inner connection and strengthens the distinction boundaries between word senses. This representation method uses data augmentation to solve the word sense definition representation under low resources in the long-tail WSD task.

4. Facing the above-mentioned quantum-inspired representation methods, a multivariate WSD model inspired by quantum interference experiments is proposed to integrate multiple representation methods to consistently deal with long-tail WSD under low resources. Considering that the double-slit interference experiment reveals the complex coherence phenomenon between microscopic particles, and belongs to the most well-known experimental model in quantum theory, this paper proposes a multivariate WSD model inspired by quantum interference experiments. This model is derived based on the mathematical model of the double-slit interference experiment, which is used to integrate the word sense definition representations obtained by the adaptive tuning representation and the multi-source fusion representation, and to realize the recognition and verification of multiple senses of the target word, that is, a multivariate WSD model under multi-way word sense representation. The model is constructed in the same way as the double-slit interference experiment, in order to understand and examine the word sense recognition process from the perspective of quantum cognition.

Finally, WSD models are constructed based on disentangled representation, adaptive tuning representation, and multi-source fusion representation, and verified under the standard WSD evaluation framework. The experimental results confirm that the disentangled representation helps reduce the large data volume dependence of the representation learning model, the adaptive tuning representation can effectively combat the inaccuracy of text representation under low resources, and the multi-source fusion representation is beneficial to enhance the distinguishability between word sense definition representations. In addition, the multivariate WSD model integrating multi-way word sense representations is verified on the constructed long-tail word sense evaluation dataset, and the experimental results confirm that the quantum-like representation framework can help improve the accuracy of long-tail word sense

recognition.

KEY WORDS: Long-tail Word Sense Disambiguation, Quantum-like Models, Disentangled Representation, Adaptive Tuning Representation, Multi-source Fusion Representation, Multivariate Word Sense Disambiguation Model

目　录

第1章 绪论

歧义是一种非常普遍的语言学现象，广泛存在于所有的自然语言之中。歧义现象会出现在自然语言的各个层级，如语音、词汇、语句、篇章等，影响着绝大多数的使用到自然语言的智能系统，可以说语言中的歧义现象是智能系统研究过程中面临着的重要挑战。本文围绕词义消歧任务中最为困难的长尾词义消歧，提出量子理论启发的系列表征方法，以实现在类量子表征框架下解决长尾词义消歧过程中面临的低资源问题。本章将逐一阐述论文的研究背景与意义、国内外研究现状、科学问题、研究内容、组织结构。

1.1 研究背景与意义

1.1.1 背景

人类在使用语言或文字进行交流的过程中往往力求准确地表达、正确地传达自己的意图或目的，但语言中存在着的歧义现象导致这一诉求并不容易达成。人类能够借助于长期学习过程中积累下来的经验知识实现自然地消除歧义，但对于智能系统而言是困难的，并且更为困难的是赋予智能系统像人类一样低资源下的学习能力，所以智能系统很难应对低资源下的消歧任务。

歧义的成因：歧义是指同一形式表达呈现出两种或两种以上意义表达的语言学现象[1]，在词汇层面，造成歧义的主要原因在于词汇的“一词多义”[2,3]。我们的社会在发展过程中会出现新的事物和概念，采用原有的符号对其进行定义或描述不仅能够降低记忆的成本，而且有利于理解和传播，因此在我们的语言中会出现大量的一词多义现象。这一现象不仅在以汉字为代表的表意文字中存在，而且在以英语为代表的表音文字中也普遍存在。歧义现象虽然在绝大多数情况下会造成表达上的混乱、理解上的错位，但也并不是一无是处的。在我们的生活中，人们常常会使用语言中的歧义现象来营造语言表达上的生动、活泼、有趣，这一方法常常会被用到相声、小品、脱口秀等艺术形式上，因此生活中并不缺乏歧义的

口语表达和文字内容[4]。

歧义的消减：在口语表达形式下，可以通过语气的停顿、语调的轻重来消除歧义；此外，也可以利于常识思维、语境信息（即语言表达过程中所处的环境，包括人物、时间、地点等具体信息）以及谈话内容的上下文信息实施歧义的消减。相比而言，文字表达形式下天然地缺乏语音层面上的消歧信息，可用的只有常识思维和上下文信息。但对于消歧系统而言，口语和文字表达形式的消歧方法是相似的，都是尽可能多地挖掘有利于消歧的信息以改善消歧系统的识别精度。除此之外，能够改善消歧系统性能的方法只有研发更为有效的表征学习模型，以实现高效、精准地抽取有效的特征信息。

词义消歧（Word Sense Disambiguation，WSD）旨在依据给定的上下文信息为目标词确定一个词义清单中最有可能的词义，其中目标词指将要消歧的词汇，词义清单指词典中目标词的所有词义的集合。词义消歧是一个标准的分类任务，属于自然语言处理（Natural Language Processing，NLP）领域中最为基础性的研究课题，影响着几乎所有的下游任务[5,6]，例如自然语言理解[7]、自然语言生成[8,9]、机器翻译[10,11]、信息检索[12,13]、文本分类[14,15]等。词义消歧早在 20 世纪 40 年代的机器翻译研究任务中被首次确立为一项独立的研究课题，并随之成为自然语言处理领域中最具挑战性的研究任务之一[16]。 20 世纪 70 年代，词义消歧被作为人工智能领域内语义解释系统的子任务，并首先尝试了基于规则和手工编码的方式实施词义消歧[17]。 20 世纪 90 年代，基于统计学方法的学习模型在自然语言处理领域中得到广泛应用，词义消歧任务成为统计学习模型攻克的重要目标[18]。进入 21 世纪，监督式机器学习技术得到迅猛发展，在各个领域中都取得了一些不错的成绩，针对词义消歧任务的研究工作也逐步转向更为细粒度的、基于语料库的研究方法上来[19]。

近年来，词义消歧系统对于高频词义的识别已经达到预期，能够有效应对绝大多数常用词义；当前研究的重心已经逐步转向最为困难的低频词义，也就是长尾词义消歧（Long-tail WSD）。长尾词义消歧的核心挑战在于其面临的低资源问题[20]：训练样本不足导致目标词表征难以学习到所属上下文的特征信息，有限且固定的词义注释文本导致难以获得准确且易于区分的词义定义表征。其中获取词义定义表征最为困难，不仅要求表征向量的准确性，而且需要在语义空间中建立起词义之间的内在联系以及位置关系。在词义消歧过程中，目标词表征与词义定义表征共同决定着词义识别的准确性。事实上，“词义”的概念本身就存在着争议，人类并不总是能够对所有的词义达成一致的共识，不同的专家可能会对同一

词义给出截然相反的定义描述。也正是因为长尾词义缺乏清晰一致的定义描述，在使用过程中会造成语言的歧义，所以人们在日常生活中会尽可能地去规避，进而导致其出现的频率下降。

传统的、以高频（常用或头部）词义为主要研究对象的词义消歧系统很难无障碍地迁移到长尾词义消歧任务上来，其核心原因在于长尾词义面临着更为严峻的低资源问题，传统的基于大数据驱动的词义消歧系统并不适用于长尾词义消歧任务。此外，对于并没有针对长尾词义给出应对方案的词义消歧系统来说，长尾词义相对于高频词义而言处于弱势地位，词义消歧系统会趋向于忽视长尾词义。然而，长尾词义的识别是重要的、有价值的，它不仅能够改善智能系统对复杂语言表达的理解能力，而且可以使智能系统输出更为准确、凝练、丰富、有趣的语言表达，对提升智能系统的综合能力和整体水平至关重要。因此，有必要面向长尾词义消歧任务开展更为针对性的研究工作，以改善词义消歧系统对于长尾词义的识别能力。

1.1.2 意义

对于自然语言处理领域中的下游任务而言，长尾词义消歧任务的研究将有助于改善下游任务对于文本内容的理解与处理能力，提升下游任务的整体水平。此外，长尾词义消歧任务面临的核心挑战是低资源问题，而低资源下的机器学习任务在很多领域中都存在，面向低资源下的表征学习技术的研究将有助于推动机器学习领域整体的发展，同时将促进工业技术领域的革新。

对于使用到自然语言的智能系统而言，词义消歧系统常被作为智能系统的基础性构件，负责智能系统输入内容的理解，输出内容的校准与优化，使智能系统的语言表达更加的自然，用词更加的准确、丰富。因此，针对于长尾词义消歧任务的研究将有助于提升智能系统的语言理解与表达能力，最终实现高效的人机交互与协作。

当前，人机交互的方式多为有限指令下的机械式输入，固定内容的机械式输出，即使能够使用自然语言进行交流也存在对话方式固定、语言表达单一的问题。导致智能系统与人类之间难以实现高效交流与协作的核心原因在于智能系统依然缺乏像人类一样的语言理解和表达能力，完善智能系统的语言理解与表达能力是迈向“类人智能”的重要一步。赋予词义消歧系统针对长尾词义的识别能力将有助于改善智能系统对于小众、低频语言表达的理解与使用能力，一个显著的优势

就是不再需要限定系统的输入必须是有限、固定的内容，系统的输出也可以根据具体场景的需要进行灵活地创造。

1.2 国内外研究现状

1.2.1 词义消歧

本节对词义消歧的研究进展与成果进行归纳总结，并对重要的模型和系统进行概述，以理清词义消歧任务下一步的研究思路和研究方向。依据过往分类方法，词义消歧系统可归纳为基于知识图谱的词义消歧和监督式的词义消歧，其中监督式的词义消歧又可归纳为纯训练数据驱动的方法、扩展词义注释信息的方法、扩展知识图谱信息的方法和扩展其它知识源的方法。

1.2.1.1 基于知识图谱的词义消歧

基于知识图谱的词义消歧系统利用了电子词典（如 WordNet [1] 和 BabelNet [2]）中提供的词汇知识图谱实施词义消歧，其中常被使用的是词汇关系知识图谱。基于词汇关系知识图谱的词义消歧系统采用图算法，如随机游走[21]、博弈论[22]等，挖掘知识图谱中词与词之间的同义词关系或上下位关系实施词义消歧，其性能直接取决于知识图谱的规模与质量。在该思路的基础之上，近年来先后提出扩展分布语义信息的词义消歧系统[23]、扩展上下文信息的词义消歧系统[24]、基于主题模型的词义消歧系统[25]、使用神经概念嵌入的词义消歧系统[26]、使用词义嵌入的词义消歧系统[27]等。其中性能表现最佳的词义消歧系统是两个设计理念截然不同的模型，SyntagRank 模型[28] 和 SREF 模型[29]。 SyntagRank 模型是一个完全基于知识图谱的词义消歧模型，其利用了个性化的 PageRank 算法挖掘电子词典 WordNet 中的词汇知识图谱，并借用语料库 WNG 和 SyntagNet 的词汇关系知识增强其使用的词汇知识图谱。SyntagRank 模型具有适应多种不同语言场景的能力，表现出优秀的泛化性能。 SREF 模型是一个基于词和文本嵌入技术的词义消歧模型，其利用预训练语言模型 BERT[30] 学习上下文化的目标词表征和词义定义表征，进而实施词义消歧。SREF 模型具有整合其它外部知识的能力，表现出优秀的鲁棒性。

综上所述，基于知识图谱的方法对于高频、常用词义而言是实用且有效的，

[1] http://wordnetweb.princeton.edu/perl/webwn

[2] https://babelnet.org

但对于低频、罕见词义而言并不适用。知识图谱的规模与质量直接决定着词义消歧系统的性能，绝大多数的知识图谱对于高频、常用词义都给出了准确的标注，但对于低频、罕见词义而言，多数的知识图谱并未给出标注。基于此，针对长尾词义消歧任务的研究工作难以采用基于知识图谱的词义消歧方法。

1.2.1.2 监督式的词义消歧

近几年取得突出成绩的词义消歧系统多数都使用预训练语言模型获取表征向量，本节主要关注神经监督模型，并依据使用的附加信息将其归纳为纯训练数据驱动的方法、扩展词义注释信息的方法、扩展知识图谱信息的方法、扩展其它知识源的方法。

1. 纯训练数据驱动的方法。 纯训练数据驱动的词义消歧系统在不利用任何超出训练语料信息的条件下实现对目标词的词义识别。最简单的词义消歧模型是使用训练语料库中相似词义的目标词表征获得词义定义表征，进而实现对未知目标词的词义识别，其中常被采用的获取词义定义表征的方法是取均值的方法[31]。该模型对常用、高频词义的识别是有效的[32]，但由于词义分布的极端不平衡现象，许多词义实际上并没有出现在训练语料库中，所以该模型只适用于高频词义、训练语料库中出现过的词义。

2. 扩展词义注释信息的方法。 电子词典 WordNet 中的一个重要信息源就是词义注释文本，其提供了一种简单的人类可读的用于澄清词义定义的文本描述。相关工作已经证实直接使用词义注释文本学习词义定义表征有助于改善词义消歧系统的表现[33]，如 SensEmBERT 模型[34]、ARES 模型[35]、SREF 模型[29]等。此外，Kumar 等人[36]通过词义注释来创建词义定义表征的训练编码器；Bevilacqua 等人[37]通过对预训练语言模型的微调获得上下文化的、信息增强的词义定义表征；Blevins 等人[38]则通过目标词编码器与词义定义编码器联合训练的方式实现借用训练样本中的语义知识增强词义定义表征的目的；Huang 等人[39]和 Yap 等人[40]将词义消歧任务重新定义为序列分类任务；Bevilacqua 等人[41]则将词义消歧任务作为自然语言生成任务来应对。

3. 扩展知识图谱信息的方法。 电子词典 WordNet 提供了除词义注释、词义例句之外更为丰富的边缘信息，如同义词、上下位词等，这些信息常被基于知识图谱的词义消歧系统所使用，如基于个性化 PageRank 算法的词义消歧系统[28]。尽管如此，许多监督式的词义消歧系统也从电子词典 WordNet 的知识图谱中获益，如 Loureiro 等人[33]通过对同义词表征取均值以实现对未知词义创建表征；Wang

等人[29]利用上下位词设计了一种校验机制以实现对系统的预测结果进行纠偏与校准；Vial 等人[42]通过上位词寻找根源词以减少输出类别的数量；Kumar 等人[36]利用电子词典 WordNet 的知识图谱离线训练词义定义表征；Bevilacqua 等人[37]则在 Kumar 等人[36]工作的基础之上通过电子词典 WordNet 的知识图谱将词义定义表征直接合并到其架构之中。总的来说，在监督式的词义消歧系统中使用知识图谱变得司空见惯，并逐渐与扩展外部知识的方法相融合。

4. 扩展其它知识源的方法。 词义消歧系统的性能已经被证实受益于所使用的知识库，以及额外知识源的数据[43]。 Calabrese 等人[44]利用 BabelPic 数据集中的图像信息来构建多模态的词义定义表征，并将其用于初始化分类矩阵的权重，其实验结果证实比纯文本下获得的表征更为有效。维基百科以及网络知识也常常被用作附加知识源来增强词义定义表征，实验结果显示它们同样有利于长尾词义消歧[29,34,35]。

综上所述，监督式的词义消歧系统多采用扩展外部知识的方式改善目标词表征和词义定义表征的准确性，或使用更为高效的表征学习方法获得优质的表征向量以改善词义识别的准确性，但对于低频、罕见词义而言，挖掘或构造训练样本并实施数据增强是困难的，并且构造的训练样本也很难再提供有效的未知信息。

1.2.2 长尾词义消歧

词义消歧早期的研究目标多关注于提升系统整体的性能表现，即解决对性能提升起决定性作用的常用、高频词义，并没有将关注点放在占绝对少数的长尾词义上来。虽然其中也存在一些解决方案对长尾词义消歧是有促进作用的，但并没有将长尾词义消歧作为研究的重点。可以理解，在词义消歧研究的早期将关注点放在高频、常用词义消歧上可以推进词义消歧系统上线的速度，并且解决了常用词义的识别问题就基本能够满足日常生活中绝大多数的需要。近年来，词义消歧系统已经能够高效应对几乎所有的常用、高频词义[37]，为了彻底解决词义消歧问题有必要攻克词义消歧领域中最为困难的长尾词义。本节将归纳长尾词义消歧的研究进展与研究成果，并对重要模型和系统进行概述，以理清长尾词义消歧下一步的研究思路和研究方向。依据近年来应对长尾词义消歧的技术方案，词义消歧系统可归纳为基于数据增强的方法、基于低资源学习技术的方法、基于语义空间约束机制的方法。

1. 基于数据增强的方法。 导致词义消歧系统难以有效应对长尾词义的根源在

于训练样本中长尾词义实例的占比较少、词义注释文本难以获得准确且易于区分的长尾词义定义表征[45]。对此，一个自然的想法就是寻找更多的训练数据和挖掘更多的词义注释信息。在获取词义定义表征方面，Ruas 等人[46]采用多词义定义表征的方式改善单表征方法下词义定义表征不准确的问题；Barba 等人[47]采用多语言知识来改善词义定义表征不准确的问题；Song 等人[48]则使用了更多的有助于改善词义定义表征的数据，如词义例句、同义词、上下位词等。在获取目标词表征方面，Yap 等人[40]通过词典中的例句对训练样本中的长尾词义实例数量实施增强；Hadiwinoto 等人[49]则采用上下文化的词表征改善目标词表征。并且上述方案都不约而同地使用了预训练语言模型去获取目标词表征和词义定义表征，以实现尽可能多地利用到预训练知识或获得上下文信息增强的表征向量。

2. 基于低资源学习技术的方法。 长尾词义消歧实质上就是一个低资源学习问题[50]，因此迁移或借鉴其它领域中的低资源学习方法将是最为直接的应对方案。 Kohli 等人[51]使用迁移学习和数据增强技术尝试在不同的自然语言处理任务和数据集上进行广泛的预训练以实现借助更多的外部知识。 Hu 等人[52]采用生成对抗技术尝试更大程度的利用、挖掘现有数据以实现改善词义消歧系统的性能。Holla 等人[53]和 Du 等人[54]采用元学习技术尝试通过在大量常用词义消歧任务上训练模型以实现在长尾词义消歧任务中快速学习。 Chen 等人[55]采用小样本学习技术尝试使用无参数的小样本学习方法应对长尾词义消歧。此外，Su 等人[56]、Kumar 等人[36]、Blevins 等人[57]针对性地研究了长尾词义消歧，并分别借助于 Z-Reweighting 方法、上下文信息增强方法和词典知识库增强方法应对长尾词义表征。

3. 基于语义空间约束机制的方法。 基于数据增强的和基于低资源学习技术的长尾词义消歧系统一定程度上有效改善了长尾词义的识别能力，但上述词义消歧系统并不是一种针对性的解决方案，并且不利于建立词义之间的内在联系以及增强词义之间的区分边界，与之对应的是基于语义空间约束机制的方法。 Kang 等人[58]将目标词表征与词义定义表征合并到同一个词义空间以使表征间更紧凑，寻找更高效。Kang 等人[59]在之前工作的基础之上提出新的目标词表征与词义定义表征结合方法以获得更加细粒度的词义定义表征。 Kumar 等人[36]对词义空间施加连续性约束以实现常用词义表征推断长尾词义表征的目的。 Barba 等人[60]采用目标词在文本中的相邻词汇辅助实施词义识别，其假设相邻词汇同样在一个词义空间之中。

综上所述，基于数据增强的和基于低资源学习技术的长尾词义消歧系统多尝

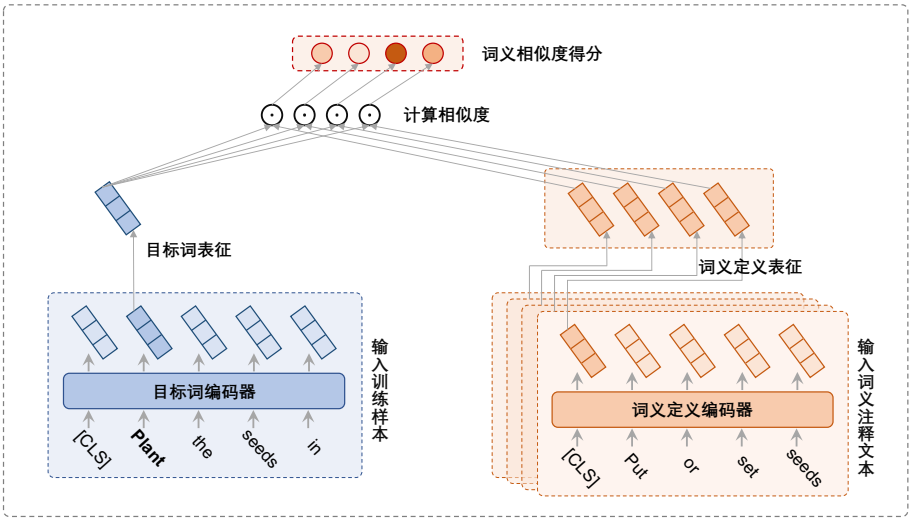

图 1-1　监督式词义消歧系统常用框架（即本文采用的基础框架）

试获取更为准确的表征向量以改善词义识别的准确性，但对于词义定义表征而言，上述方法获得的表征向量难以有效建立起词义之间的内在联系。词义定义表征不仅要求表征向量的准确度，而且需要在语义空间中建立起词义之间的内在联系以及位置关系，同时词义之间还需要具有一定的区分边界。相比而言，基于语义空间约束机制的长尾词义消歧系统在构建词义定义表征方面是有优势的，本文也将在此方向上进一步深入研究。此外，应对长尾词义消歧的主流方法是为包括长尾词义在内的所有词义学习词义定义表征，并分别与训练样本中获得的目标词表征计算相似度，相似度得分最高的词义定义表征所对应的词义被认定为该目标词所属的词义，如图 1-1 所示。在本文中，我们同样依照该思路构造词义消歧系统，同时需要强调的是本文构造的词义消歧模型将默认使用该词义消歧方法作为长尾词义消歧系统的基础框架。

1.2.3　量子理论的方法论

量子力学理论的数学基础，即量子概率理论[61,62]，近些年来一直被用作经典数学（指经典概率论、经典力学等）的替代方案应用于构建人类认知与决策行为的数学模型[63–66]。但对于使用量子力学理论建模传统任务的动机和依据问题一直难以达成一致的共识。学者们相继提出“还原主义[67]”、“整体主义[68]”、“量子认知论[69]”以及“量子信息论[70]”的观点阐释其正当性与合理性。

还原主义：还原主义主张人类认知与决策行为所呈现出来的宏观现象始于微

观粒子层次的量子行为，以量子力学为主要方法框架、经典量子物理实验确认的物理效应为主要实现机理，自底向上地解释和构建人类认知与决策行为的数学模型[71,72]。还原主义认为人类的认知与决策行为归根到底是由人脑神经元所操控的，模拟神经元的行为活动采用量子理论更为贴切。还原主义范式下的研究工作尝试从底层刻画人脑微观层次的意识活动，进而描述、解释人类宏观高层次的认知决策现象。

整体主义：整体主义对应于还原主义，主张人类认知与决策行为所呈现出来的宏观现象不应被还原（或归因）于微观层次上的量子效应或量子行为，而是在一定神经元规模上才可能发生的系统性的突现行为[73]。整体主义认为人脑神经元层次上产生的电信号是无意识的，并不存在量子行为，而产生人类认知与决策行为以及宏观类量子行为的根源在于成规模的群体现象。整体主义范式下的研究工作并不尝试从底层刻画、解释人类宏观层次的认知决策行为，更注重于描述系统性的群体现象。

量子认知论：量子认知论始于 20 年前，由一群极具创造力的物理学家和心理学家所提出，并在最近 10 年里得到迅猛发展，被正式冠名为量子认知科学[74]。量子认知论认为微观粒子的行为是不确定性的、概率性的，人类的认知决策行为同样存在着不确定性；微观粒子的行为受测量方式（和观测者）的影响，人类认知与决策行为同样受所处情境的影响。故此，采用专门为刻画微观系统不确定性的量子力学构建人类认知与决策行为的数学模型就显得如此的自然。基于量子认知论的研究工作多尝试采用量子现象，如叠加、干涉、纠缠等，来诠释经典理论难以解答的类量子、非经典难题，如著名的合取谬误和析取谬误等。

量子信息论：研究证实，量子力学的主要形式特征可以由几条简明的信息原则所导出，典型的工作包括 Spekkens 模型[75,76]、量子贝叶斯解释[77,78]、Khrennikov 解释[79]、极大化物理信息原则[80]和可信信息优先原则[81]等。虽然各

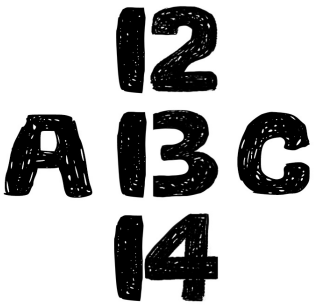

图 1-2　　词义消歧的类量子现象

种类量子特征的导出模型可能依赖于形式上不同的信息原则，但各种信息原则的共同内核可抽象为约束信息最大化原则。在 Spekkens 模型和 Khrennikov 解释中，约束信息最大化原则被显式地给出；在 Fuchs 的量子贝叶斯解释中，该原则可由特定约束下的对称信息完备性原则所体现；更一般地，Frieden 由极大化物理信息原则导出大量已知物理规律，包括若干主要的量子力学定律。基于量子信息论的研究工作将量子力学作为描述不确定现象的数学工具，并尝试使用该工具所特有的形式特征便捷地刻画人类的认知与决策行为。

本文采用量子信息论的研究视角，主张将量子力学提供的新颖、优雅而又完备的数学框架作为经典理论的替代品用于建模人类的认知与决策行为。在此之上，借用量子力学中特殊的形式特征刻画经典理论难以建模的类量子现象。如图 1-2 所示，图片来源于网络，对于图中间的符号（“B”或“13”）而言，在“A ～ C”情景下，该符号将被认为是“B”；在“12 ～ 14”的情景下，该符号将被认为是“13”。对于上述情况，经典理论难以描述同时处于两种状态下的叠加状态。此外，如果将情景“A ～ C”或“12 ～ 14”设定为观测者，观测结果设定为“B”和“13”，经典理论也难以建模观测结果受观测者影响的类量子现象。**对于词义消歧任务而言**，当词汇没有被用在具体的上下文之前，它可能属于任意一种词义（或多种词义下的叠加状态）；当词汇被用在特定的上下文之后，它的状态将被确定。对于上述现象，经典理论是难以应对的，但量子力学为描述、刻画上述现象提供了贴切的数学形式。在量子力学中，对于没有确定状态的事物可以采用量子叠加态进行描述；观测结果受观测者影响的行为可以采用量子测量操作进行刻画。综上所述，对于词义消歧任务而言，量子力学理论恰如其分的适合于描述消歧过程。本文尝试在量子力学的数学框架下给出一套新颖的词义消歧方案，并力求借助于量子力学攻克词义消歧任务中最为困难的长尾词义消歧。

1.3 科学问题

长尾词义消歧本质上是一个低资源下的表征学习问题，具体为低资源下的目标词表征和词义定义表征学习问题。**目标词表征**是基于匮乏的**训练样本**学习到的独立的词表征，该词表征学习问题可以借鉴小样本学习、元学习或迁移学习技术，它们面对的场景是相似的。**词义定义表征**是基于有限且固定的**词义注释文本**学习到的词义之间存在区分边界的文本表征，词义定义表征需要在语义空间中建立起

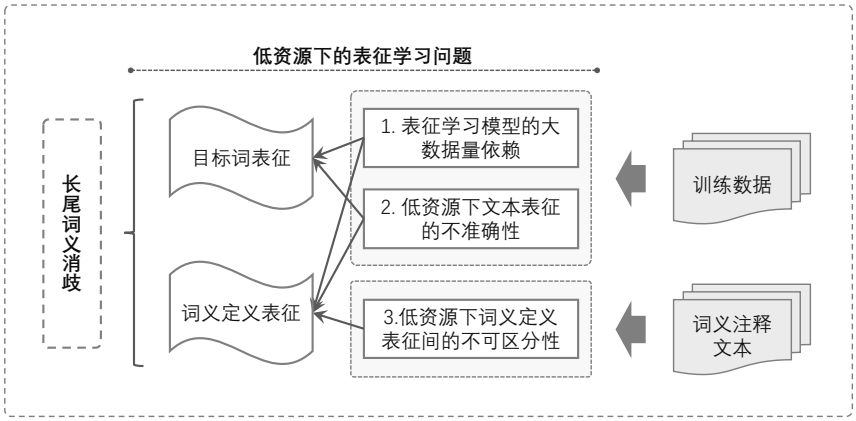

图 1-3 长尾词义消歧任务中的科学问题

词义之间的内在联系以及位置关系，也就是说该词义定义表征学习问题是独特的，很难单纯的使用或借鉴其它领域中低资源下的文本表征学习技术。长尾词义消歧任务同时面临上述两种低资源下的表征学习问题，因此长尾词义消歧是一项非常困难的研究课题。针对长尾词义消歧任务中的目标词表征与词义定义表征提出三个科学问题，其对应关系如图 1-3 所示：

1. 表征学习模型的大数据量依赖。 长尾词义消歧的核心困难在于其面临的低资源问题，即在匮乏的训练样本中学习目标词表征、在有限且固定的词义注释文本中学习词义定义表征。虽然目标词表征是基于其所属上下文信息获得的词表征，词义定义表征是基于词义注释文本获得的文本表征，但它们存在着相似之处，也就是从有限的文本数据中学习表征向量。如果尝试在表征学习模型的层次统一地应对低资源下的目标词表征与词义定义表征学习问题，可将其统筹为表征学习模型的大数据量依赖问题。长尾词义消歧任务下的表征学习模型的大数据量依赖问题旨在优化表征学习模型的学习策略，减少学习模型需要学习的参数数量，最终实现降低学习模型对大数据量的依赖。

2. 低资源下文本表征的不准确性。 长尾词义消歧的主要挑战是低资源场景下的表征学习问题，即在匮乏的训练样本中挖掘到准确且有效的目标词表征，在有限且固定的词义注释文本中挖掘到准确且易于区分的词义定义表征。虽然从训练样本中获得的是目标词的词表征，从词义注释文本中获得的是词义定义的文本表征，但它们存在着相似之处，也就是从已知文本数据中挖掘到尽可能多的有效信息进而校准、调优表征向量的准确性。如果尝试在表征向量的层次统一地应对低资源下的目标词表征与词义定义表征学习问题，可将其统筹为低资源下文本表征的不准确性问题。长尾词义消歧任务下的文本表征不准确性问题旨在基于已知

文本数据对已获得的表征向量实施进一步优化，最终实现依托低资源数据改善表征向量的准确性。

3. 低资源下词义定义表征间的不可区分性。 相比于目标词表征而言，词义定义表征更难获得。目标词表征是基于匮乏的训练样本获得的词表征，该词表征向量仅需要满足准确性的要求。词义定义表征是基于有限且固定的词义注释文本获得的文本表征，该文本表征向量不仅需要满足准确性的要求，还需要在语义空间中建立起词义之间的内在联系以及位置关系。从有限且固定的词义注释文本中挖掘词义之间的内在联系并建立词义之间的位置关系是困难的，一个可选的应对方案是借助于外部数据获得的词义信息增强词义定义表征之间的可区分性以及强化词义定义表征之间的区分边界。基于此，数据增强下的词义定义表征之间的融合与约束方法研究可归纳为低资源下词义定义表征间的不可区分性问题。

1.4 研究内容

本文围绕长尾词义消歧面临的低资源下的表征学习问题，提出量子理论启发的表征学习方法，以应对匮乏训练样本下的目标词表征、有限且固定词义注释文本下的词义定义表征。具体地，同时针对目标词表征与词义定义表征提出量子纠缠启发的解耦表征方法和量子叠加态启发的自适应调优表征方法，单独针对词义定义表征提出量子测量启发的多源融合表征方法；在此基础上，模拟量子双缝干涉实验的设置方式提出具备整合上述表征方法的多元词义消歧模型。本文整体研究框架如图 1-4 所示，研究内容与创新点概述如下：

1. 针对长尾词义消歧任务中表征学习模型的大数据量依赖问题，提出**量子纠缠启发的解耦表征**，在第三章中呈现。本章基于纠缠态的解纠缠原理构造强的解纠缠约束机制，用以约束重采样的特征表示，使采样特征之间彼此独立，即解耦表征。使用重采样技术学习特征表示的目的在于降低表征学习模型构造的复杂度和规模，以缓解模型学习过程中对大数据量的依赖；特征解耦的目的在于获得特征独立的表征向量，以实施进一步的特征选择。基于解耦表征获得目标词表征和词义定义表征并实施词义消歧，实验结果证实解耦表征有助于缓解表征学习模型对大数据量的依赖；此外，基于解耦表征构建的词义消歧模型有效改善了长尾词义的识别能力。该工作的贡献在于提供了一种量子理论启发的、具备强特征解耦的表征学习方法，其创新之处在于使用重采样结合特征选择的方式应对低资源下

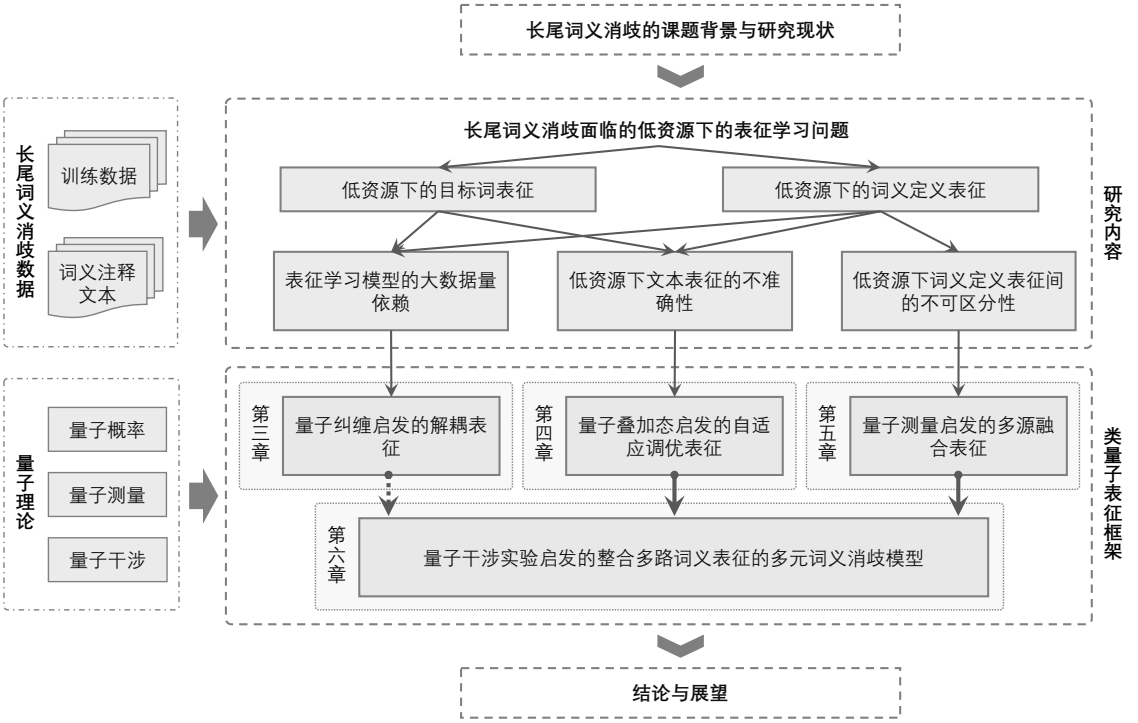

图 1-4　　论文整体研究框架

的表征学习问题。

2. 针对长尾词义消歧任务中低资源下文本表征的不准确性问题，提出**量子叠加态启发的自适应调优表征**，在第四章中呈现。本章基于已知表征向量构建叠加态，以实现在希尔伯特空间中自适应地调优表征向量的准确性，即自适应调优表征。需要强调的是该表征方法不需要增加额外的模型参数和附加更多的训练数据，仅仅是通过施加约束以及叠加态的构造过程完成对原表征向量的优化操作。基于自适应调优表征获得目标词表征和词义定义表征并实施词义消歧，实验结果证实自调优表征有助于改善目标词表征与词义定义表征的准确性；此外，基于自适应调优表征构建的词义消歧模型有效改善了长尾词义的识别能力。该工作的贡献在于提供了一种量子理论启发的、不需要增加训练数据的表征优化方法，其创新之处在于借用量子叠加态的构造原理实现由一个参数控制下的表征优化操作。

3. 针对长尾词义消歧任务中低资源下词义定义表征间的不可区分性问题，提出**量子测量启发的多源融合表征**，在第五章中呈现。本章借助叠加态的构造过程实现融合多源数据获得的词义定义表征，借用叠加态的测量过程所导出的干涉项实现在语义空间中约束词义定义表征之间的位置关系，即多源融合表征。词义注释文本的数据量有限，建立起词义之间的内在联系以及位置关系是困难的，对此，

一个自然的想法就是实施数据增强。多源融合表征试图整合外部数据中获得的表征向量增强词义定义表征之间的内在联系以及可区分性；此外，借助外部数据中获得的表征向量约束表征学习模型中参数搜索的范围，进而降低表征学习模型对大数据量的依赖。基于多源融合表征获得词义定义表征并实施词义消歧，实验结果证实多源融合表征有助于增强词义定义表征之间的可区分性；此外，基于多源融合表征构建的词义消歧模型有效改善了长尾词义的识别能力。该工作的贡献在于提供了一种整合外部数据的方式，其创新之处在于借助量子测量原理导出的干涉项实施表征向量之间的空间几何形式下的表征向量约束。

4. 面向上述量子理论启发的表征方法，提出**量子干涉实验启发的整合多路词义表征的多元词义消歧模型**，在第六章中呈现。本章基于双缝干涉实验的数学模型导出多元词义消歧模型，即双缝干涉模型的泛化形式，用以整合上述量子理论启发的表征方法一致性地应对低资源下的长尾词义消歧任务。最后，对整合多路词义定义表征的多元词义消歧模型实施实验，实验结果证实多元词义消歧模型具备整合多种表征方法的能力，并且整合多种表征方法的词义消歧模型有助于提升长尾词义识别的准确性。该工作的贡献在于提供了一套能够整合多种表征学习方法的多元词义消歧框架，同时该框架也可以作为新方法、新模型的基础框架迅速搭建词义消歧系统。

1.5 组织结构

本文面向长尾词义消歧任务中的量子理论启发的表征方法开展研究，涉及内容共分七个章节呈现，各章节的内容概述如下：

- 第一章，绪论。本章介绍长尾词义消歧的研究背景与意义，概述国内外的研究现状与进展，澄清长尾词义消歧面临的核心科学问题是低资源下的表征学习问题，阐述本文的研究内容和创新性，最后概括本文的组织结构。
- 第二章，量子理论基础。本章给出本文中使用到的量子力学基础知识，对量子力学中的数学符号、名词术语给出解释，所涉及内容被归纳为量子概率、量子测量、量子干涉三部分。
- 第三章，量子纠缠启发的解耦表征。本章介绍基于量子纠缠实现的解耦表征方法，该表征方法用于解决长尾词义消歧任务中表征学习模型的大数据量依赖问题。最后，解耦表征方法用于构造词义消歧模型并实施验证，实

验结果证实解耦表征方法在应对长尾词义消歧方面具有不错的表现。

- 第四章，量子叠加态启发的自适应调优表征。本章介绍基于叠加态实现的自适应调优表征方法，该表征方法用于解决长尾词义消歧任务中低资源下的文本表征不准确性问题。最后，自适应调优表征方法用于构造词义消歧模型并实施验证，实验结果证实自适应调优表征方法在应对长尾词义消歧方面具有出色的表现。
- 第五章，量子测量启发的多源融合表征。本章介绍基于量子叠加态的构造过程实现的多源表征融合方法、基于量子叠加态的测量过程所导出的干涉项实现的多源表征约束方法，统称多源融合表征。该表征方法用于解决长尾词义消歧任务中词义定义表征之间的不可区分性问题。最后，多源融合表征方法用于构造词义消歧模型并实施验证，实验结果证实多源融合表征方法在应对长尾词义消歧方面具有不俗的表现。
- 第六章，量子干涉实验启发的多元词义消歧模型。本章介绍基于双缝干涉实验启发的多元词义消歧模型，该多元词义消歧模型旨在整合上述量子理论启发的表征方法一致性地应对低资源下的长尾词义消歧任务。最后，对整合多路词义表征的多元词义消歧模型实施验证，实验结果证实整合多种表征方法的多元词义消歧模型在应对长尾词义消歧方面具有突出的表现。此外，在该章节的实验分析部分给出了各个章节中提出的词义消歧模型间的横向对比分析。
- 第七章，总结与展望。本章对本文研究内容进行回顾与总结，同时对将后来的研究工作进行规划与展望。

第2章 量子理论基础

本文采用量子力学理论的数学基础，即量子概率理论，构建词义消歧模型，所涉及到的量子概念、数学符号等在本章节呈现，所有内容被归纳为量子概率、量子测量与量子干涉三部分。

2.1 量子概率

量子概率理论由物理学家在 20 世纪初发现，并被主要应用于物理学领域。随后，冯 · 诺依曼将量子概率理论进行公理化，发现其蕴含着一种新的概率理论，即量子概率[61,62]，以及一种新的逻辑体系，即量子逻辑[82,83]。正如微分方程诞生于牛顿力学的纯物理领域，随之应用到整个社会的各个方面，量子概率同样将会在社会学、经济学、博弈论等领域发挥其价值。

2.1.1 量子态

量子态（也称量子系统）用以描述一个孤立的物理系统的状态，其包含系统能够提供的所有信息，如系统的初始条件、初始设置、所处状态等[84]。在描述物理现象的语境中常使用量子系统的说法，而在形式化定义的上下文语境中则常使用量子态的说法，也就是说，量子系统是随机物理现象的载体，而量子态是概率模型的载体。在本文中，上述概念可以通用，并不作细致的区分。

定义 2.1 （量子系统）：任一独立量子系统都有一个被称之为状态空间的希尔伯特空间（即复内积向量空间）与之对应，量子系统由状态空间中的一个单位向量描述，该向量被称之为状态向量。

最简单的量子系统是由一个量子比特构成的系统。量子比特是一个构建在二维状态空间中的单位向量，该空间由一组正交的、单位长度的基向量描述，即：

$$\left\{|0\rangle \rightarrow [1 \quad 0]^{\dagger},\ |1\rangle \rightarrow [0 \quad 1]^{\dagger}\right\} \tag{2-1}$$

其中记号“$|\cdot\rangle$”称为 Dirac 符号（符号说明见表 2-1 所示），该组基也被称之为计算基，则状态空间中的状态向量可被定义为：

$$|\psi\rangle = \varepsilon_0 \cdot |0\rangle + \varepsilon_1 \cdot |1\rangle \tag{2-2}$$

其中 ε_0 和 $\varepsilon_1 \in \mathbb{C}$ 被称为概率幅，$|\varepsilon_0|^2 + |\varepsilon_1|^2 = 1$。量子比特区别于经典比特的特征在于量子比特的状态向量可以是 $|0\rangle$ 和 $|1\rangle$ 的线性叠加形式，处于叠加形式的量子态也被称之为**叠加态**。

更为一般化的量子系统也可以被构建在 N 维希尔伯特空间，该空间同样可由一组正交的、单位长度的基向量描述，基向量可记为：$\{|e_i\rangle\}_{i=1}^{N}$，则状态空间中的状态向量可被定义为：

$$|\psi\rangle = \varepsilon_1 \cdot |e_1\rangle + \varepsilon_2 \cdot |e_2\rangle + ... + \varepsilon_i \cdot |e_i\rangle + ... \tag{2-3}$$

其中 $\varepsilon_i \in \mathbb{C}$ 表示概率幅，$\sum_{i=1}^{N} |\varepsilon_i|^2 = 1$。此外，量子系统的状态向量也可以由已知状态向量叠加而成，如

$$|\Psi\rangle = \varepsilon_1 \cdot |\psi_1\rangle + \varepsilon_2 \cdot |\psi_2\rangle + ... + \varepsilon_i \cdot |\psi_i\rangle + ... \tag{2-4}$$

其中 $\sum_{i=1}^{N} |\varepsilon_i|^2 = 1$。由 $\langle\Psi|\Psi\rangle = 1$ 可推知，$|\Psi\rangle$ 同样可以被分解为标准正交基下的展开形式。在构造量子系统的过程中，通过量子态的叠加原理可以还原真实情况下的量子系统，因此叠加态常被用于构造量子系统。

量子系统也可以通过多个量子比特复合而成，但多量子比特系统并没有在本文中涉及，因此多量子比特系统的构造方法将不在此处提及。

表 2-1　量子概率中涉及到的 Dirac 符号汇总

符号	含义
z^*	复数 z 的复共轭，$(1+i)^* = 1-i$
$\lvert\cdot\rvert$	标量的绝对值
$\lVert\cdot\rVert$	向量的范数，在量子力学中一般指 2 范数
$\lvert\psi\rangle$	向量，又称右态失
$\langle\psi\rvert$	向量 $\lvert\psi\rangle$ 的对偶向量，又称左态失
$\langle\varphi\rvert\psi\rangle$	向量 $\lvert\varphi\rangle$ 和向量 $\lvert\psi\rangle$ 的内积
$\lvert\varphi\rangle \otimes \lvert\psi\rangle$	向量 $\lvert\varphi\rangle$ 和向量 $\lvert\psi\rangle$ 的张量积
$\lvert\varphi\rangle\lvert\psi\rangle$	向量 $\lvert\varphi\rangle$ 和向量 $\lvert\psi\rangle$ 的张量积的缩写
A^*	矩阵 A 的复共轭
$A^\intercal$	矩阵 A 的转置
$A^\dagger$	矩阵 A 的共轭转置或 Hermite 共轭，$A^\dagger = (A^\intercal)^*$
$\langle\varphi\rvert A\rvert\psi\rangle$	向量 $\lvert\varphi\rangle$ 和向量 $A\lvert\psi\rangle$ 的内积，等价于向量 $A^\dagger\lvert\varphi\rangle$ 和向量 $\lvert\psi\rangle$ 的内积

2.1.2 事件

经典概率理论将概率分配给事件，并且为每个事件指派一个集合域中的集合。概率为 0 的事件对应于空集 $\varnothing$，表示不可能事件；概率为 1 的事件对应于全集 U，表示必然事件。此外，用集合的交、并、补运算可以构造出更多的新事件。对于给定的事件 A、B、C 而言，事件 A 与事件 B 都发生的事件可以由集合的交运算获得，被表示为：

$$AB = A \cap B; \tag{2-5}$$

事件 A 与事件 B 至少有一个发生的事件可由集合的并运算获得，被表示为：

$$A + B = A \cup B; \tag{2-6}$$

事件 A 不发生的事件可由集合的补运算获得，被表示为：

$$\overline{A} = U - A。 \tag{2-7}$$

并且经典概率的事件遵循布尔代数的运算逻辑，即：

- 交换律：$A \cup B = B \cup A$，$A \cap B = B \cap A$；
- 结合律：$(A \cup B) \cup C = A \cup (B \cup C)$，$(A \cap B) \cap C = A \cap (B \cap C)$；
- 互补律：$A \cup \overline{A}) = U$，$A \cap \overline{A} = \varnothing$；
- 吸收律：$A \cup (A \cap B) = A$，$A \cap (A \cup B) = A$；
- 分配律：$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$，$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$。

需要强调的是分配律对于区分经典概率与量子概率至关重要，并且量子概率中很多重要的特性都是由量子概率违背分配律所导出的。

量子概率理论同样将概率分配给事件，不同的是量子概率中的每个事件都对应于希尔伯特空间中的一个子空间，如直线、平面或超平面等。量子概率中，概率为 0 的事件，即不可能事件，对应于希尔伯特空间中的 $\mathbf{0}$ 向量，即 0 维的线性空间；概率为 1 的事件，即必然事件，对应于整个希尔伯特空间[84]。量子概率理论同样可以使用三种方式构造出更多的新事件：

定义 2.2 （否运算）：若事件 A 对应于子空间 L_A，则事件 A 不发生的事件 $\overline{A}$ 被定义为垂直于空间 L_A 的子空间，记为：

$$L_{\overline{A}} = L_A^{\perp}。 \tag{2-8}$$

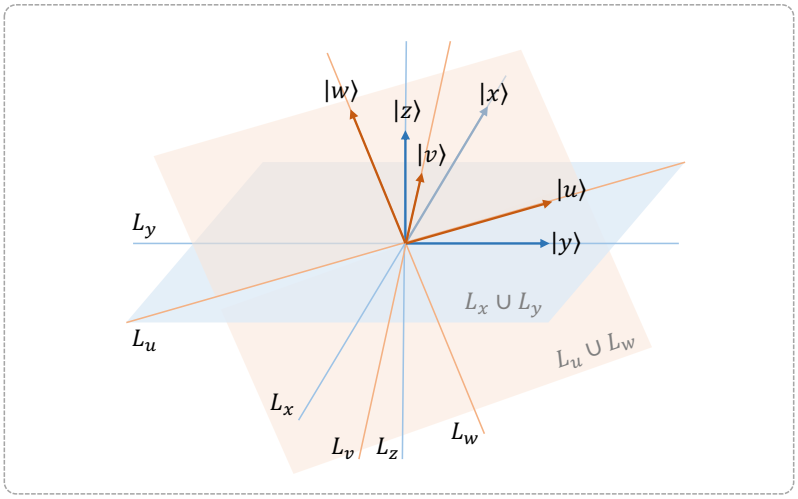

图 2-1　　量子概率违背布尔代数分配律的示意图

定义 2.3 （与运算）：若事件 A 对应于子空间 L_A，事件 B 对应于子空间 L_B，则事件 A 与事件 B 都发生的事件 $A \cap B$ 被定义为空间 L_A 与空间 L_B 的交集，记为：

$$L_{A\cap B} = L_A \cap L_B。 \tag{2-9}$$

定义 2.4 （或运算）：若事件 A 对应于子空间 L_A，事件 B 对应于子空间 L_B，则事件 A 与事件 B 至少有一个发生的事件 $A \cup B$ 被定义为空间 L_A 与空间 L_B 所张成的更大空间，记为：

$$L_{A\cup B} = Span(L_A, L_B)。 \tag{2-10}$$

经典概率中的并运算与量子概率中的张成运算 $Span(\cdot,\cdot)$ 完全不同，这也直接导致量子概率不遵循布尔代数的分配律。图 2-1 是一个量子概率不遵循分配律的示例，所涉及 Dirac 符号如表 2-1 所示。假设量子系统所属的希尔伯特空间是一个 3 维空间，该空间中的事件 $|x\rangle$、$|y\rangle$ 和 $|z\rangle$ 分别由标准坐标系 L_x、L_y 和 L_z 来描述。此外，该空间中的事件 $|u\rangle$、$|v\rangle$ 和 $|w\rangle$ 分别由另一个正交坐标系 L_u、L_v 和 L_w 来描述。对于事件

$$(L_u \cup L_w) \cap (L_x \cup L_y \cup L_z) \tag{2-11}$$

而言，由于 L_x、L_y 与 L_z 是该空间标准坐标系的基，则事件 $L_x \cup L_y \cup L_z$ 等于整个空间，所以

$$(L_u \cup L_w) \cap (L_x \cup L_y \cup L_z) = L_u \cup L_w。 \tag{2-12}$$

如果假设分配律在量子概率中成立，则

$$\begin{aligned}(L_u \cup L_w) \cap (L_x \cup L_y \cup L_z) &= (L_u \cup L_w) \cap \Big((L_x \cup L_y) \cup L_z\Big) \\ &= (L_u \cup L_w) \cap (L_x \cup L_y) \cup (L_u \cup L_w) \cap L_z。\end{aligned} \tag{2-13}$$

其中 $(L_u \cup L_w) \cap (L_x \cup L_y) = L_u$ 和 $(L_u \cup L_w) \cap L_z = \mathbf{0}$，则

$$(L_u \cup L_w) \cap (L_x \cup L_y) \cup (L_u \cup L_w) \cap L_z = L_u \cup \mathbf{0} = L_u。 \tag{2-14}$$

据此可知

$$(L_u \cup L_w) \cap (L_x \cup L_y \cup L_z) = L_u \cup L_w \neq L_u, \tag{2-15}$$

所以分配律在量子概率中不成立。

2.1.3 概率

经典概率理论中被广泛采用和认可的公理化定义是基于 Kolmogorov 公理体系[85]的概率公理，所涉及的公理包括：

- 标准概率公理：$0 \leq Pr(A) \leq 1$，$Pr(\varnothing) = 0$，$Pr(U) = 1$；
- 附加概率公理：若 $A \cap B = \varnothing$，则 $Pr(A \cup B) = Pr(A) + Pr(B)$。

当涉及多个观测量时，可使用条件概率进行描述，如在给定事件 A 的条件下，事件 B 的条件概率为：

$$Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)}。 \tag{2-16}$$

由条件概率可导出联合概率公式，即：

$$Pr(B \cap A) = Pr(A)Pr(B|A)。 \tag{2-17}$$

量子概率理论中的概率同样遵循 Kolmogorov 公理体系，但与之不同的是量子概率理论的概率是由投影规则计算而来的[84]。在计算概率的过程中，事件对应的子空间被构造为投影算子，该算子作用于量子系统对应的状态向量并获得概率，具体涉及如下三个步骤：

- 准备量子系统。在量子概率理论中，所有事件的概率都由量子系统确定。量子系统对应于希尔伯特空间中的一个单位向量，称为量子系统的状态向量，假设存在状态向量 $|z\rangle$。
- 构造投影算子。在量子概率理论中，每个事件所对应的子空间都可以被构造为相应的投影算子，投影算子可将状态向量投影到对应事件的子空间，

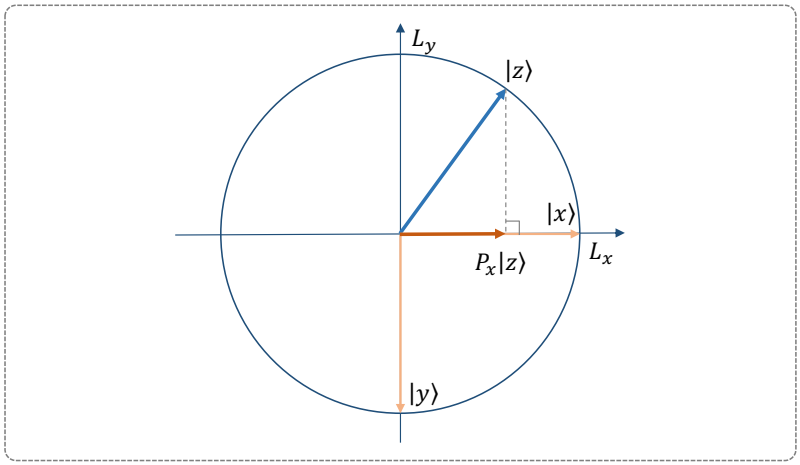

图 2-2　　量子概率投影规则的示意图

假设事件 L_x 的投影算子为 P_x。

- 计算投影概率。量子系统对应于指定事件的概率等于其投影的平方长度，则状态向量 $|z\rangle$ 对应于事件 L_x 的概率为：

$$Pr(x) = Pr(|z\rangle; P_x) = ||P_x|z\rangle||^2。 \tag{2-18}$$

图 2-2 描述了量子概率的投影过程，所涉及 Dirac 符号如表 2-1 所示。在图中，状态向量 $|z\rangle$ 在子空间 L_x 上投影的平方长度就是量子系统对应于事件 L_x 的概率。

2.1.4　与经典概率的区别

量子概率是一种更为泛化的概率理论，量子概率向下兼容经典概率。量子概率可以被视为包含不相容测度下的经典概率的推广。当所有观测量都相容时，量子概率降级为经典概率；但当观测量之间存在不相容情况时，量子概率与经典概率之间将存在着明显的差异。对于不相容测度下的量子概率将服从互易律、双重随机律等，此外量子概率不必像经典概率一样强制要求服从全概率定律[84]。量子概率与经典概率之间的区别与联系如表 2-2 所示。

与基于经典概率构建的数学模型相比，使用量子概率构建的数学模型具有以下几个优点：

- 首先，量子模型并不总是需要假设一个联合概率空间来从不同的测量中推导或关联边际概率。不同测度的边际概率都可以从公共状态向量中获得，而无需假设公共的联合分布。
- 其次，量子概率为测量的顺序效应提供了解释，这是社会学、行为学中普

遍存在着的建模问题。

- 再次，量子概率可以解释一种测量对另一种测量的干扰现象，这也是社会学、行为学中普遍需要的建模方法。
- 最后，量子概率允许同时存在确定性和概率性的现象，这比随机误差理论更适合于描述人类认知与决策行为。

表 2-2 量子概率与经典概率之间的区别和联系

性质	经典概率	量子概率（事件A与B不兼容）
联合概率	$Pr(A \cap B) = Pr(B \cap A)$	不成立
全概率公式	$Pr(A) = \sum_i Pr(A\|B_i)Pr(B_i)$	不成立
条件概率	$Pr(A\|B) = Pr(A \cap B)/Pr(B)$	$Pr(A\|B) = \|\|P_A P_B\|z\rangle\|\|^2 / \|\|P_B\|z\rangle\|\|^2$
条件概率	$Pr(B\|A) = Pr(A \cap B)/Pr(A)$	$Pr(B\|A) = \|\|P_B P_A\|z\rangle\|\|^2 / \|\|P_A\|z\rangle\|\|^2$
条件概率的互易性	不成立	$Pr(A\|B) = Pr(B\|A)$
双向随机性	不成立	$\sum_i Pr(A_i\|B_j) = \sum_j Pr(A_i\|B_j) = 1$

2.2 量子测量

量子测量存在多种假设，如一般测量假设、投影测量假设、POVM 测量假设[84]。一般测量是对量子测量的一般性约定，测量算子仅需满足完备性约束。POVM 测量是一般测量的特例，测量算子不仅需要满足一般测量的完备性约束外，还要求每个测量算子都是半正定的。投影测量同样是一般测量的特例，测量算子不仅需要满足完备性约束外，还需要满足正交性约束。此外，投影测量是一种特殊的 POVM 测量，因此，一般测量包含 POVM 测量，POVM 测量包含投影测量。

定义 2.5 （一般测量）：一般测量由一组满足完备性要求的测量算子 $\{M_m\}$ 定义，即：

$$\sum_m M_m^\dagger M_m = I \tag{2-19}$$

其中符号 I 表示单位矩阵，指标 m 表示实验过程中可能的测量结果，这组测量算子作用于被测量子系统的状态向量。若测量前，量子系统的状态向量为 $|\psi\rangle$，则测量结果 m 发生的概率为：

$$Pr(m) = Pr(|\psi\rangle; M_m) = \langle\psi|M_m^\dagger M_m|\psi\rangle = \langle\psi|M_m|\psi\rangle; \tag{2-20}$$

测量后，量子系统的状态向量为：

$$|\psi'\rangle = \frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^\dagger M_m|\psi\rangle}}。\tag{2-21}$$

测量算子的完备性保证了量子概率理论中概率的和为 1，进而确保了量子概率理论遵循 Kolmogorov 公理体系，即：

$$\sum_m Pr(m) = \sum_m \langle\psi|M_m^\dagger M_m|\psi\rangle = \sum_m \langle\psi|M_m|\psi\rangle = 1。\tag{2-22}$$

定义 2.6 （投影测量）：投影测量由状态空间中的一个 Hermite 算子 M 定义，该算子被称为可观测量，其谱分解为：

$$M = \sum_m mP_m \tag{2-23}$$

其中 P_m 是本征值 m 对应于本征空间的投影算子。投影测量的结果为投影算子对应的本征值 m。若测量前，量子系统的状态向量为 $|\psi\rangle$，则测量结果 m 发生的概率为：

$$Pr(m) = Pr(|\psi\rangle; P_m) = \langle\psi|P_m^\dagger P_m|\psi\rangle = \langle\psi|P_m|\psi\rangle;\tag{2-24}$$

测量后，量子系统的状态向量为：

$$|\psi'\rangle = \frac{P_m|\psi\rangle}{\sqrt{\langle\psi|P_m^\dagger P_m|\psi\rangle}}。\tag{2-25}$$

投影测量是一般测量的特例，所以投影测量的投影算子同样满足完备性约束，即：

$$\sum_m P_m^\dagger P_m = I;\tag{2-26}$$

此外，还满足 P_m 是正交投影算子的要求，即 P_m 是 Hermite 的。投影测量具有很多很好的性质，特别是容易计算测量结果的平均值，即：

$$E(M) = \sum_m mPr(m) = \sum_m m\langle\psi|P_m|\psi\rangle = \langle\psi|\left(\sum_m mP_m\right)|\psi\rangle = \langle\psi|M|\psi\rangle。\tag{2-27}$$

该公式可用于简化计算过程，在实践中具有重要的意义，此外可观测量 M 的平均值常被定义为：

$$\langle M\rangle \equiv \langle\psi|M|\psi\rangle。\tag{2-28}$$

定义 2.7 （POVM 测量，由一般测量导出）： 对于一组一般测量的测量算子 $\{M_m\}$ 而言，令

$$E_m \equiv M_m^\dagger M_m, \tag{2-29}$$

若 E_m 是满足

$$\sum_m E_m = I \tag{2-30}$$

和

$$Pr(m) = \langle\psi|E_m|\psi\rangle \tag{2-31}$$

的半正定算子，则算子 E_m 被称为 POVM 测量的元，集合 $\{E_m\}$ 被称为 POVM 测量的可观测量。

定义 2.8 （POVM 测量，由投影测量导出）： 对于一个投影测量的可观测量 M 而言，若算子 M 的本征值 m 对应的投影算子 P_m 满足

$$P_m P_{m'} = \delta_{mm'} P_m \tag{2-32}$$

和

$$\sum_m P_m = I, \tag{2-33}$$

则该投影测量就是一个 POVM 测量，其中

$$E_m \equiv P_m^\dagger P_m = P_m \tag{2-34}$$

被称为 POVM 测量的元，集合 $\{E_m\}$ 被称为 POVM 测量的可观测量。

2.3 量子干涉

托马斯 · 杨最早于 1801 年设计、实施了双缝实验，以验证光的波动性，该实验也被称为杨氏双缝干涉实验[86]。杨氏双缝干涉实验的贡献在于证实了光同时具有波动性与粒子性，即光的波粒二象性。 1927 年，戴维森和革默证实电子同样具有类似于光的波粒二象性，此后该实验被推广到原子和分子层次[87]。双缝实验与其变体因为成功地验证了量子力学领域中的核心难题而成为经典，随后该实验得到进一步发展，成为揭示量子现象的重要手段。

在双缝干涉实验中，光源照亮一个具有两条狭缝的幕布，即双缝幕布，光透

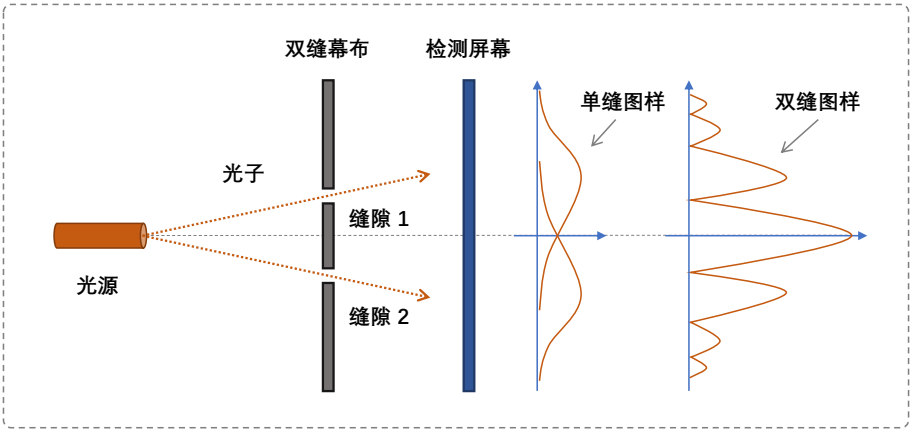

图 2-3　　量子双缝干涉实验的示意图

过双缝照射在幕布后面的检测屏幕上，如图 2-3 所示。光的波动性导致照射在检测屏幕上的光波呈现干涉条纹，这与光的经典粒子假设的结果相违背，因此该实验证实光具有波动性。此外，在狭缝处加装检测器时发现光是以粒子的形式一个一个地穿过狭缝，因此该实验也可以用于证实光的粒子性。然而，此类实验不仅能够验证光的波动性与粒子性，而且修改实验的设置可以发现更多与经典认知相违背的量子现象。例如光源以点射的方式一个一个地发射光子，最后在检测屏幕上依然能够发现干涉条纹，即光子的叠加效应。当在狭缝处加装监视器后，光呈现出粒子性，但当关掉监视器后，光又呈现出波动性，即观察者效应。此外，双缝干涉实验的延迟选择实验还揭示出“未来决定过去”的诡异量子现象。

双缝干涉实验揭示的量子干涉现象与经典物理中的波干涉存在显著的差异。假设量子系统处于叠加状态，如：

$$|\Psi\rangle = \varepsilon_1 \cdot |\psi_1\rangle + \varepsilon_2 \cdot |\psi_2\rangle \tag{2-35}$$

其中 $|\varepsilon_1|^2 + |\varepsilon_2|^2 = 1$，再假设量子系统的一个测量算子为 M_m，则测量结果 m 发生的概率为：

$$\begin{aligned} Pr(m) &= \langle\Psi|M_m^{\dagger}M_m|\Psi\rangle \\ &= \langle\Psi|M_m|\Psi\rangle \\ &= \||M_m|\Psi\rangle\|^2 \\ &= \|M_m\,(\varepsilon_1 \cdot |\psi_1\rangle + \varepsilon_2 \cdot |\psi_2\rangle)\|^2 \\ &= \|\varepsilon_1 \cdot M_m|\psi_1\rangle\|^2 + \|\varepsilon_2 \cdot M_m|\psi_2\rangle\|^2 + 2 \cdot \varepsilon_1 \cdot \varepsilon_2 \cdot \cos(\theta) \cdot |\langle\psi_1|M_m|\psi_2\rangle| \end{aligned} \tag{2-36}$$

其中 θ 为内积运算 $\langle\psi_1|M_m|\psi_2\rangle$ 的相位角。经典物理中的波干涉只具有两束波之间的相加干涉或相减干涉，即：

$$\|\varepsilon_1 \cdot M_m|\psi_1\rangle\|^2 + \|\varepsilon_2 \cdot M_m|\psi_2\rangle\|^2, \tag{2-37}$$

而量子干涉多了一个刻画系统内部干涉现象的干涉项，即：

$$2 \cdot \varepsilon_1 \cdot \varepsilon_2 \cdot \cos(\theta) \cdot |\langle\psi_1|M_m|\psi_2\rangle|, \tag{2-38}$$

因此量子干涉揭示的现象更为复杂。以下总结经典波干涉与量子干涉之间的一些差异：

- 经典波干涉是两束波之间的干涉，而量子干涉还存在波函数自身的干涉；
- 经典波干涉是两束波之间的位移相加或相减，而量子干涉发生在波函数相关的概率函数之上，因此波函数存在自身干涉现象；
- 经典波干涉由表示相对位置位移的实函数描述，而量子波函数是一个复杂的复数函数。

量子干涉虽然复杂，但量子干涉的形式化模型却为建模一些复杂的人类认知与决策行为提供了契机，如借助量子干涉模型可以刻画复杂的对话交互行为[88]、认知决策行为[89]等。

2.4 本章小结

本章归纳总结了论文中涉及到的量子力学理论方面的基础知识，分别介绍了量子概率，量子概率使用的测度方法，以及量子力学理论中重要的双缝干涉实验。其中，量子概率部分给出了量子系统的构造方法与形式化描述，量子概率中事件的定义与生成方法，量子概率中概率的定义与概率的投影规则，以及量子概率与经典概率之间的差异汇总；量子测量部分给出了一般测量、投影测量与 POVM 测量的形式化定义；量子干涉部分给出了双缝干涉实验的设置方式，以及经典波干涉与量子干涉的差异汇总；最后，对该章节的内容进行了小结。

第3章 量子纠缠启发的解耦表征

长尾词义消歧任务涉及到目标词表征与词义定义表征，二者协作共同实现词义识别。本章提出同时适用于目标词表征与词义定义表征的解耦表征方法，用以解决长尾词义消歧任务中表征学习模型的大数据量依赖问题。此外，基于该表征方法实现进一步的特征选择，用以降低输出表征向量的维度。最后，基于提出的量子纠缠启发的解耦表征方法构建词义消歧模型，并实施标准评估实验、消融分析实验和长尾分析实验。

3.1 研究动机

长尾分布（Long-tailed Distribution），也称重尾分布，是一种比正态分布更为普遍的随机变量分布[90]，其直观的特征是头部类别（高频、常见类别）占据绝大多数的样本数据，而尾部类别（低频、罕见类别）只拥有少量样本数据。来自于真实世界中的数据往往存在着更为严重的数据不平衡，常常呈现出明显的长尾分布现象[91]。

对于机器学习（Machine Learning，ML）领域中数据驱动的模型而言，模型的性能会受到数据分布的影响，模型往往在头部数据上表现优秀，而在尾部数据上表现不佳[92]。其原因在于尾部类别的训练样本数量不足，模型难以从有限的训练样本中抽取到有效的特征信息，进而完成具体的任务。更为核心的原因在于模型难以从有限的训练样本中充分更新、优化模型的参数。在自然语言处理（Natural Language Processing，NLP）领域中，同样面临着因训练样本数据分布不平衡所导致的模型失效问题，特别是当前主流的深度学习模型、预训练语言模型等更易受到数据长尾分布的影响[93]。对此，研究者们提出数据增强的方法用以改善尾部类别所面临的训练样本稀缺的问题，典型的数据增强方法有过采样[94]、数据加权[95]、样本加噪[96]等。此类方法虽然可以扩展训练样本的数量，但难以提供更为丰富的未知信息和未知知识，因此数据增强的方法适用领域有限。此外，研究者们还提出一些针对训练样本稀缺问题的学习策略和学习模型，例如对比学习[97]、

元学习[98]、小样本学习[99]、生成对抗网络[100]等。

当前，词义消歧系统已经能够有效地解决高频（头部、常用）词义的识别任务，但对于低频（长尾、罕见）词义，词义消歧系统还并不能够有效地应对[50]。长尾词义消歧面临的主要困难在于：（1）长尾词义大多有其固定的适用场景，并且适用场景多为罕见领域，所以长尾词义的训练样本严重缺乏；（2）长尾词义定义（或称词义注释）是由词典提供的一段有限且固定的文本描述，所以长尾词义难以获得清晰且易于区分的词义定义表征[37]。对此，研究者们的主要应对思路可归纳为数据增强的方法、知识迁移的方法和施加外部约束的方法。数据增强的方法尝试构造或引入外部数据增加长尾词义的训练样本数量。其中最具代表性的方法是整合词典中的例句以改善目标词表征的准确性[40]，或引入多语种词义注释文本以改善长尾词义定义表征的准确性[46]。知识迁移的方法尝试整合多场景下目标词的上下文信息以改善目标词表征的准确性，或融合其它领域中的语义知识以改善长尾词义定义表征的准确性。其中最受关注的方法是采用双编码器（即目标词编码器与词义定义编码器）联合训练的方式实现借助于训练样本中的语义知识强化长尾词义定义表征[38]。该方法的优势在于并没有引入外部知识，仅是迁移训练样本中的语义信息增强词义定义表征。施加外部约束的方法着眼于词义定义表征间内在的位置关系，尝试将词义定义表征映射到一个连续的语义空间，以实现表征之间的相互制约，最终实现由头部词义定义表征去校准长尾词义定义表征的目的。其中最有成效的方法是将原本的离散空间中的词义定义表征映射到连续空间，以实现对长尾词义定义表征或未知词义定义表征的有效推断[36]。

早在 2013 年，Bengio 等人[101]发表的关于表征学习方面的论文中就提出解耦表征的概念。解耦表征（Disentangled Representation，DR），也称解纠缠表征，旨在学习一个特征之间彼此分离的（即特征之间不存在相关性约束），并且每个特征只对单个生成因子敏感的特征表示。解耦表征的特征通常具有可解释的语义信息，并且能够针对性地操控以实现靶向性地调控目标对象的具体属性[102]。例如，对于照片抽取的解耦表征而言，可以通过修改具体的几个特征值实现改变照片中人像的肤色、脸型、发色等。对于长尾词义消歧任务而言，由于获取解耦表征不需要深度网络模型对特征进行深层次的抽取与融合，能够降低表征学习模型对大规模训练样本数量的依赖；同时基于解耦表征的长尾词义消歧系统能够有针对性地选择一些决定性的特征以实现表征向量降维的目的。

具体地，本章基于量子纠缠态的解纠缠原理构造强的解纠缠约束机制，用以约束重采样的特征表示，使采样特征之间彼此独立，即解耦表征。使用重采样技

术学习特征表示的目的在于降低表征学习模型构造的复杂度和规模，以缓解模型学习过程中对大数据量的依赖；特征解耦的目的在于获得特征独立的表征向量，以实施进一步的特征选择。本章的贡献在于：

- 针对长尾词义消歧任务中训练样本与词义注释文本的低资源问题，提出量子纠缠启发的解耦表征方法，用以解决长尾词义消歧任务中表征学习模型对大数据量的依赖问题；
- 基于量子纠缠启发的解耦表征方法构造词义消歧模型并实施验证，实验结果证实解耦表征有助于缓解表征学习模型对大数据量的依赖，此外，基于解耦表征构建的词义消歧模型有效改善了长尾词义的识别能力。

3.2 基于解耦表征的词义消歧模型

本节首先介绍量子纠缠启发的解耦表征方法，然后基于该解耦表征方法构造词义消歧模型，最后给出词义消歧模型的训练方法。

3.2.1 解耦表征

解耦表征的主要特点在于输出的表征向量的特征之间是相互独立的，因此为了获得独立的特征，需要对特征表示施加独立性约束，此外，为了使特征之间彼此无关，需要对特征表示施加解相关约束（或称解纠缠约束）。对于给定的随机向量，度量其分量之间相关性强度的常用工具是互信息[103]，其中互信息值越大，分量间相关性越强；互信息值越小，分量间相关性越弱；互信息值为零时，分量间彼此独立。对此，一个自然的想法就是约束采样生成的随机向量，使其分量间的相关性降低，即随机向量的互信息值下降，最终实现采样的随机向量的分量间彼此独立。对应于表征学习过程，表征学习模型的输入用于学习每个特征分布函数的参数，模型的输出为特征分布函数们采样生成的随机向量，度量随机向量的互信息将被添加到学习模型的损失函数之中。

假设解耦表征学习模型基于已知的输入数据获得输出的解耦表征，即一个采样生成的随机向量，其中输入数据为：

$$X = [x_1, x_2, ..., x_i, ...] \in \mathbb{R}^m, \tag{3-1}$$

输出的随机向量为：

$$Z = [z_1, z_2, ..., z_i, ...] \in \mathbb{R}^n。 \tag{3-2}$$

其中指标 m 和 n 分别表示输入、输出向量的维度，它们可以是相同的值，也可以根据实际情况调整为合适的取值。

随机向量 Z 中分量之间的互信息可借助于 Kullback-Leibler 散度来定义，

$$\mathbf{I}(Z) = \mathbf{KL}\left(Pr_1(Z) \| \prod_i^n Pr_2(z_i)\right) \tag{3-3}$$

其中下标 1 和 2 在下文中用于区分不同分布函数的参数，$Pr_1(Z)$ 表示联合概率分布，$\prod_i^n Pr_2(z_i)$ 表示边际概率分布。进一步假设随机向量 Z 的概率密度函数为多元高斯分布（也称多元正态分布或联合正态分布），则在多元高斯分布假设下，随机向量 Z 的联合概率分布可被表示为：

$$Pr_1(Z) \sim \mathcal{N}(\vec{\mu}_1, \Sigma) \tag{3-4}$$

其中 $\vec{\mu}_1 \in \mathbb{R}^n$ 是多元高斯分布的均值向量，

$$\vec{\mu}_1 = \mathbf{E}[Z] = \Big(\mathbf{E}[z_1], \mathbf{E}[z_2], ..., \mathbf{E}[z_i], ...\Big), \tag{3-5}$$

$\Sigma \in \mathbb{S}_{++}^{n \times n}$ 是多元高斯分布的协方差矩阵，

$$\Sigma_{i,j} = \mathbf{E}\left[(z_i - \vec{\mu}_i)(z_j - \vec{\mu}_j)\right] = Cov(z_i, z_j), \tag{3-6}$$

其中 $\mathbf{E}$ 表示用于求期望，$Cov(\cdot, \cdot)$ 表示用于求协方差。均值向量 $\vec{\mu}_1$ 可以借助于神经网络中的全连接层（或称为线性层）由已知输入数据 X 获得，

$$\vec{\mu}_1 = Linear^1(X)。 \tag{3-7}$$

协方差矩阵 Σ 同样可以由输入数据 X 获得，

$$\Sigma = \frac{1}{S} \sum_s^S X_s'^T X_s' \tag{3-8}$$

其中 S 表示组成 Σ 的向量个数，X_s' 被定义为：

$$X_s' = SSN\left(Linear_s^1(X)\right) = \frac{Linear_s^1(X)}{\|Linear_s^1(X)\|}, \tag{3-9}$$

其中函数 $SSN(\cdot)$ 表示平方和归一化函数。因为协方差矩阵是半正定矩阵，所以上述复杂运算的目的是为了得到一个合格的协方差矩阵 Σ。通过实验分析发现 S 的取值既不能太大也不能太小；当取值太大时，模型不易收敛；当取值太小时，获

得的解耦表征将退化为一般表征。这里之所以假设生成的随机向量的特征服从高斯分布的原因在于高斯分布有且仅有两个需要学习的参数，并且高斯分布是自然界存在着的最为广泛的分布。此外，当多元高斯分布的协方差矩阵为对角矩阵 Λ 时，随机向量的分量间相互独立。据此，在多元高斯分布假设下，公式 (3-3) 中随机向量 Z 的边际概率分布可被表示为：

$$\prod_{i}^{n} Pr_2(z_i) \sim \mathcal{N}(\vec{\mu}_2, \Lambda) \tag{3-10}$$

其中均值向量 $\vec{\mu}_2$ 可借助于神经网络中的全连接层（或称为线性层）由已知输入数据 X 获得，

$$\vec{\mu}_2 = Linear^2(X), \tag{3-11}$$

对角矩阵形式下的协方差矩阵 Λ 同样可以由输入数据 X 获得，

$$\Lambda = Diag\Big(Softmax\left(Linear^2(X)\right)\Big)。 \tag{3-12}$$

其中函数 $Softmax(\cdot)$ 表示归一化指数函数，函数 $Diag(\cdot)$ 的作用是将一个向量变换为对角矩阵，也就是用向量的元素作为新矩阵的对角线元素，矩阵的其它非对角线元素被设置为 0 值。

解耦表征学习模型在训练过程中，Kullback-Leibler 散度的两个参数 $Pr_1(Z)$ 和 $\prod_i^n Pr_2(z_i)$ 分别被多元高斯分布假设下的 $\mathcal{N}(\vec{\mu}_1, \Sigma)$ 和 $\mathcal{N}(\vec{\mu}_2, \Lambda)$ 取代，则公式 (3-3) 可被表示为：

$$\mathbf{KL}\left(Pr_1(Z) \| \prod_{i}^{n} Pr_2(z_i)\right) \approx \mathbf{KL}\Big(\mathcal{N}(\vec{\mu}_1, \Sigma) \| \mathcal{N}(\vec{\mu}_2, \Lambda)\Big), \tag{3-13}$$

并且该 Kullback-Leibler 散度将被作为约束项添加到损失函数中用以约束生成的随机向量。解耦表征学习模型在迭代、优化过程中，随机向量 Z 的全概率公式的多元高斯表示 $\mathcal{N}(\vec{\mu}_1, \Sigma)$ 将趋向于其边际概率公式的多元高斯表示 $\mathcal{N}(\vec{\mu}_2, \Lambda)$。至此，我们可以通过对各个特征的高斯分布进行采样用以获得输出的随机向量 Z。但是由于采样过程无法求导，而采样的结果是可以求导的，所以输出的随机向量 Z 可以近似地表示为：

$$Z \approx \frac{1}{2} \cdot \Big(\left(\vec{\mu}_1 + \vec{\mu}_2\right) + \vec{\varepsilon} * \Big(diag(\Sigma) + diag(\Lambda)\Big)\Big) \tag{3-14}$$

其中向量 $\vec{\varepsilon}$ 表示从多元高斯分布 $\mathcal{N}(0, diag(\Sigma))$ 采样得到的扰动（或称噪声），符号 $*$ 表示按位乘操作，函数 $diag(\cdot)$ 表示获取矩阵的对角线元素。函数 $Diag(\cdot)$ 的作用是将向量转化为矩阵，而函数 $diag(\cdot)$ 的作用是将矩阵转化为向量。

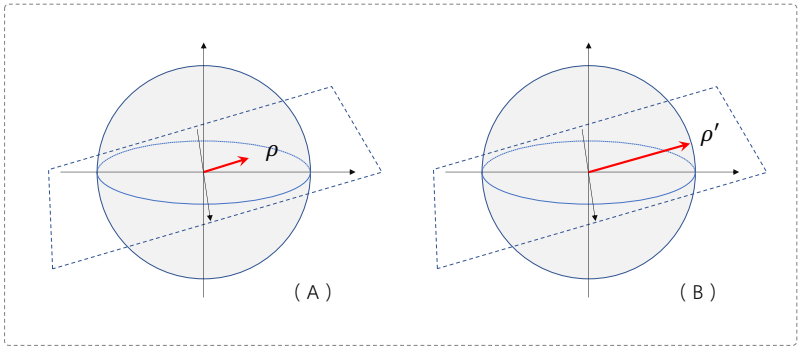

图 3-1　　量子纠缠启发的解耦表征原理图

另外，通过进一步推导上述多元高斯分布的 Kullback-Leibler 散度，可得约束项为：

$$\mathbf{KL}\Big(\mathcal{N}(\vec{\mu}_1,\Sigma)\|\mathcal{N}(\vec{\mu}_2,\Lambda)\Big)=\mathbf{E}\Big(\log(\mathcal{N}(\vec{\mu}_1,\Sigma))-\log(\mathcal{N}(\vec{\mu}_2,\Lambda))\Big) \tag{3-15}$$
$$=\frac{1}{2}\cdot\Big\{\log\frac{|\Lambda|}{|\Sigma|}+tr(\Lambda^{-1}\Sigma)+(\vec{\mu}_2-\vec{\mu}_1)\Lambda^{-1}(\vec{\mu}_2-\vec{\mu}_1)^T-n\Big\},$$

其中求逆矩阵需要消耗大量的算力，且 Λ 是对角矩阵，则 Λ^{-1} 可被表示为：

$$\Lambda^{-1}=\frac{\mathbf{1}}{\Lambda} \tag{3-16}$$

其中 $\Lambda_{i,i}\neq 0$，$\mathbf{1}$ 表示单位矩阵。最终，用于获得解耦表征的约束项被定义为：

$$\mathbf{KL}=\frac{1}{2}\cdot\Big\{\log\frac{|\Lambda|}{|\Sigma|}+tr(\frac{\Sigma}{\Lambda})+\frac{(\vec{\mu}_2-\vec{\mu}_1)(\vec{\mu}_2-\vec{\mu}_1)^T}{\Lambda}-n\Big\} \tag{3-17}$$

其中函数 $tr(\cdot)$ 表示求迹运算。

理论分析： 由于用于生成随机向量 Z 的联合概率分布的协方差矩阵由量子力学中的密度矩阵表示，如公式 (3-8) 所示，该形式下获得的密度矩阵在绝大多数情况下是一个纠缠态，即采样的随机向量的分量之间存在相关性。采用量子力学语言描述，纠缠态的状态向量长度小于 1，如图 3-1 (A) 所示。而用于生成随机向量 Z 的边际概率分布的协方差矩阵则由一个单位向量表示，如公式 (3-12) 所示，该形式下的量子态是一个非纠缠态，即采样的随机向量的分量之间不存在相关性。采用量子力学语言描述，非纠缠态的状态向量长度等于 1，如图 3-1 (B) 所示。在表征学习模型训练过程中，联合概率分布的协方差矩阵会趋向于边际概率分布的协方差矩阵，最终能够保证采样的随机向量的分量之间存在着较小或不存在相关性。

该设计的优势在于：初始状态下，用于学习联合概率分布的网络模型具备学习分量间存在相关性的协方差矩阵；随后经过迭代、优化，强制约束该网络模型

放弃掉学习分量间存在相关性的协方差矩阵，即强制约束学习一个对角矩阵形式下的协方差矩阵；最终使解耦表征学习模型采样生成的随机向量的分量是独立采样，同时分量之间彼此不存在相关性。该原理是获得解耦表征的重要保证。

3.2.2 模型结构

基于解耦表征的词义消歧模型框架结构如图 3-2 所示。该词义消歧模型仿照 Blevins 等人[104]提出的双编码器结构设计：一个编码器用于获取目标词嵌入（被称为传统表征方法下的目标词嵌入），在此之上借用解耦表征方法获取解耦表征方法下的目标词嵌入，该编码器被称为**目标词编码器**；另一个编码器用于获取词义定义嵌入（被称为传统表征方法下的词义定义嵌入），在此之上借用解耦表征方法获取解耦表征方法下的词义定义嵌入，该编码器被称为**词义定义编码器**；最后，分别计算传统表征方法下的词义相似度得分和解耦表征方法下的词义相似度得分，并将两者加权求和作为最终的词义相似度得分。该词义消歧模型同样采用双编码器结构的原因在于：包含目标词的待消歧文本与词义注释文本之间存在着差异，采用两个编码器分别学习各自的表征将有利于保留各自的独特性，更利于获得优质的解耦表征。在实施过程中，目标词编码器和词义定义编码器都采用预训练语言模型 BERT[30] 作为编码器。选用 BERT 模型作为编码器的原因在于 BERT 模型可以学习到上下文化的词嵌入，该优势有利于目标词嵌入学习到丰富的上下文信息以改善词义识别的准确性。各个编码器的实施细节如下所述：

目标词编码器将输入的待消歧文本中的每个单词编码为相应的词嵌入，待消歧文本为：

$$W^{text} = [w_1, w_2, ..., w_i, ...], \tag{3-18}$$

其中目标词包含于待消歧文本之中，即 $w_{target} \in W_{text}$，获得的词嵌入为：

$$V_{W^{text}} = [V_{w_1}, V_{w_2}, ..., V_{w_i}, ...], \tag{3-19}$$

同理 $V_{w_{target}} \in V_{W^{text}}$。根据预训练语言模型 BERT 的文本处理规则，输入模型的文本需要在开头与分割处分别添加相应的开始标记符 [*CLS*] 和分割标记符 [*SEP*]，所以输入的文本为：

$$W^{text} = \Big[[CLS], w_1, w_2, ..., w_i, ..., [SEP]\Big], \tag{3-20}$$

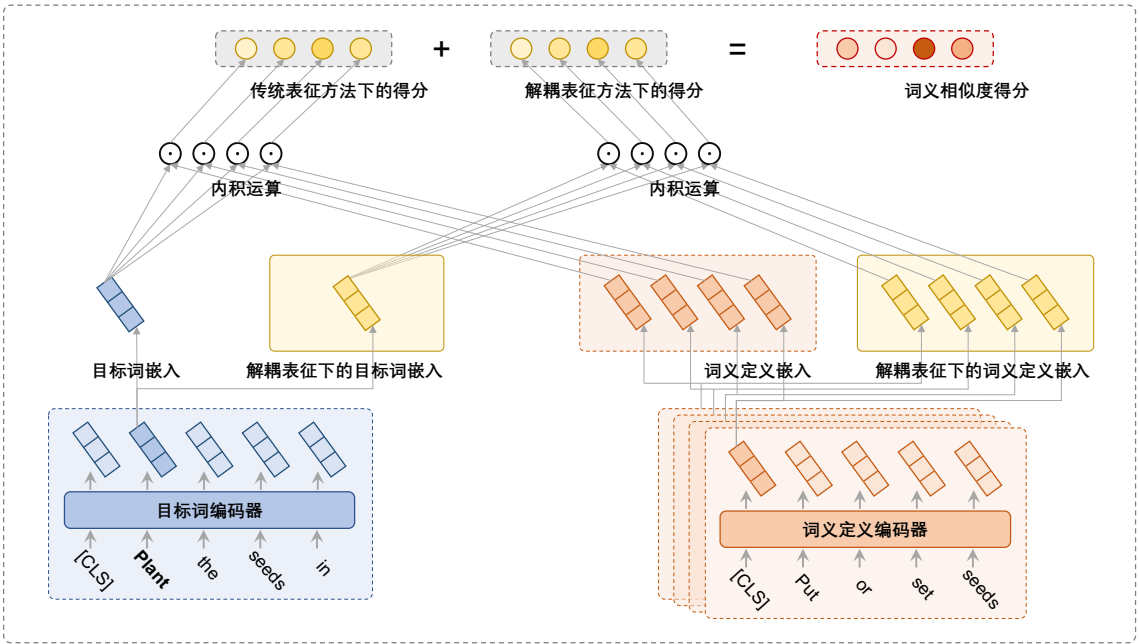

图 3-2　　基于解耦表征的词义消歧模型

输出的待消歧文本中的每个单词的词嵌入为：

$$V_{W^{text}} = \left[V_{[CLS]}, V_{w_1}, V_{w_2}, ..., V_{w_i}, ..., V_{[SEP]}\right]。\tag{3-21}$$

最后，对应于目标词的词嵌入，即目标词嵌入，将作为目标词编码器唯一的输出，

$$V_{target} \equiv V_{w_{target}} = BERT_{TargetEncoder}(W^{text}),\tag{3-22}$$

该目标词嵌入被称为**传统表征方法下的目标词嵌入**。

将目标词编码器的输出拷贝一份用于生成解耦表征方法下的目标词嵌入，其中 V_{target} 被作为解耦表征方法的输入，分别用于生成多元高斯分布假设下联合概率的均值向量 $\vec{\mu}_1$，

$$\vec{\mu}_1 = Linear^1(V_{target});\tag{3-23}$$

多元高斯分布假设下联合概率的协方差矩阵 Σ，

$$\Sigma = \frac{1}{S}\sum_{s}^{S} X_s'^T X_s', \quad X_s' = SSN\left(Linear_s^1(V_{target})\right);\tag{3-24}$$

多元高斯分布假设下边际概率的均值向量 $\vec{\mu}_2$，

$$\vec{\mu}_2 = Linear^2(V_{target});\tag{3-25}$$

多元高斯分布假设下边际概率的协方差矩阵 Λ，

$$\Lambda = Diag\left(Softmax\left(Linear^2(V_{target})\right)\right)。 \tag{3-26}$$

至此，基于公式 (3-14) 获得**解耦表征方法下的目标词嵌入**，

$$Q\text{-}V_{target} = \frac{1}{2}\cdot\left(\left(\vec{\mu}_1 + \vec{\mu}_2\right) + \vec{\varepsilon} * \left(diag(\Sigma) + diag(\Lambda)\right)\right)。 \tag{3-27}$$

词义定义编码器同样按照 BERT 模型的文本处理规则，将输入到词义定义编码器的词义注释文本中的单词编码为相应的词嵌入，添加开始标记符 [*CLS*] 和分割标记符 [*SEP*] 后的词义注释文本为：

$$W_k^{gloss} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., w_{k,i}, ..., [SEP]_k\right], \tag{3-28}$$

输出的词嵌入为：

$$V_{W_k^{gloss}} = \left[V_{[CLS]_k}, V_{w_{k,1}}, V_{w_{k,2}}, ..., V_{w_{k,i}}, ..., V_{[SEP]_k}\right], \tag{3-29}$$

其中索引 k 表示目标词的词义清单中第 k 个词义的词义注释文本，并且目标词不不含在词义注释文本之中，即 $w_{k,target} \notin W_k^{gloss}$，同理 $V_{w_{k,target}} \notin V_{W_k^{gloss}}$。由于词义定义编码器输出的要是词义注释文本的文本嵌入，按照业界通行作法，开始标记符 [*CLS*] 对应的词嵌入常被作为对应文本的文本嵌入，所以词义定义编码器的输出为：

$$V_k^{gloss} \equiv V_{[CLS]_k} = BERT_{GlossEncoder}(W_k^{gloss}), \tag{3-30}$$

该词义定义嵌入被称为**传统表征方法下的词义定义嵌入**。

将词义定义编码器的输出拷贝一份用于生成解耦表征方法下的词义定义嵌入，其中 V_k^{gloss} 被作为解耦表征方法的输入，分别用于生成多元高斯分布假设下联合概率的均值向量 $\vec{\mu}_1$，

$$\vec{\mu}_1^k = Linear^1(V_k^{gloss}); \tag{3-31}$$

多元高斯分布假设下联合概率的协方差矩阵 Σ，

$$\Sigma^k = \frac{1}{S}\sum_s^S X_s'^{kT} X_s'^k, \qquad X_s'^k = SSN\left(Linear_s^1(V_k^{gloss})\right); \tag{3-32}$$

多元高斯分布假设下边际概率的均值向量 $\vec{\mu}_2$，

$$\vec{\mu}_2^k = Linear^2(V_k^{gloss}); \tag{3-33}$$

多元高斯分布假设下边际概率的协方差矩阵 Λ，

$$\Lambda^k = Diag\Big(Softmax\Big(Linear^2(V_k^{gloss})\Big)\Big)。\tag{3-34}$$

至此，基于公式 (3-14) 获得**解耦表征方法下的词义定义嵌入**，

$$Q\text{-}V_k^{gloss} = \frac{1}{2}\cdot\Big(\big(\vec{\mu}_1^k + \vec{\mu}_2^k\big) + \vec{\varepsilon} * \Big(diag(\Sigma^k) + diag(\Lambda^k)\Big)\Big)\tag{3-35}$$

其中符号 $*$ 表示按位乘运算。

最后，分别计算传统表征方法下的词义相似度得分，

$$Score_k = V_{target} \odot V_k^{gloss},\tag{3-36}$$

与解耦表征方法下的词义相似度得分，

$$Q\text{-}Score_k = Q\text{-}V_{target} \odot Q\text{-}V_k^{gloss},\tag{3-37}$$

并将两者加权求和作为最终的**词义相似度得分**，

$$Score_k^{all} = \alpha \cdot Score_k + \beta \cdot Q\text{-}Score_k,\tag{3-38}$$

其中符号 $\odot$ 表示内积运算，索引 k 表示目标词的词义清单中第 k 个词义，α 和 $\beta \in \mathbb{R}$ 分别表示相应的权重。

3.2.3 模型训练

该词义消歧模型的损失函数由两部分组成，即最终结果的交叉熵损失和获取解缠表征的约束项，其中约束项进一步分为获得解耦表征方法下目标词表征的约束项和词义定义表征的约束项，约束项如公式 (3-17) 所示。词义消歧模型采用表现效果最好的 Adam 优化器[105] 来更新参数，优化器的具体设置将在实验部分给出。交叉熵损失函数为：

$$\begin{aligned} Loss(Score^{all}, index) &= -\log\left(\frac{\exp(Score_{[index]}^{all})}{\sum_{i=1}\exp(Score_{[i]}^{all})}\right) \\ &= -Score_{[index]}^{all} + \log\sum_{i=1}\exp(Score_{[i]}^{all}) \end{aligned}\tag{3-39}$$

其中 $Score^{all}$ 表示目标词的词义清单中各个词义的相似度得分，

$$Score^{all} = \Big[Score_1^{all}, Score_2^{all}, ..., Score_k^{all}, ...\Big],\tag{3-40}$$

变量 $index$ 表示目标词的词义清单列表索引。

3.3 实验准备

3.3.1 数据集

为了有效评估本章提出的基于解耦表征方法构造的词义消歧模型，我们使用了 Navigli 等人[106]提出的词义消歧评估框架作为标准评估实验设置，并在标准评估实验设置的基础之上构建了增强评估实验设置。提出增强评估实验设置的目的是为了满足更大规模、更多参数的深度学习模型。

标准评估实验设置的数据集被划分为训练集、开发集、测试集，其中训练集为 SemCor [1]，开发集为 SemEval-07[107]（*SE7*），测试集包括 Senseval-2[108]（*SE2*）、Senseval-3[109]（*SE3*）、SemEval-13[110]（*SE13*）、SemEval-15[111]（*SE15*）以及所有测试集的整合集（*ALL*）。此外，测试集的整合集 *ALL* 又按词性划分为名词集（*N.*）、动词集（*V.*）、形容词集（*Adj.*）、副词集（*Adv.*）。各个数据集的统计信息如表 3-1 所示。

增强评估实验设置的数据集同样包括训练集、开发集、测试集，其中训练集包括 SemCor 和 WNGT [2]（WordNet Gloss Tags），开发集和测试集与标准评估实验设置下的开发集和测试集保持一致。

此外，需要说明的是目标词的候选词义清单由电子词典 WordNet 3.0 [3] 提供，即电子词典 WordNet 针对目标词给出的所有词义被认定为目标词的候选词义清单。评估实验所使用的评估指标为百分比下的 F1 值（即 F1-score (%)）。其它未列出的有关数据集设置方面的信息可参见 Navigli 等人[106]提出的词义消歧评估框架的设置。

3.3.2 实验设置

模型运行平台为 Linux 操作系统下的 Ubuntu 18.04 LTS 版本，平台配置六块 NVIDIA GeForce RTX 3090 的显卡。模型的开发语言为 Python 3.9.0 [4]，模型的开发

[1] http://lcl.uniroma1.it/wsdeval/training-data

[2] https://wordnetcode.princeton.edu/glosstag.shtml

[3] http://wordnetweb.princeton.edu/perl/webwn

[4] https://www.python.org/downloads/release/python-390/

表 3-1　数据集的统计信息

数据集	文档数	句子数	单词数	标注单词数	词义类型	词汇类型	复杂度
Senseval-2	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3	3	352	5,541	1,850	1,167	977	6.8
SemEval-07	3	135	3,201	455	375	330	8.5
SemEval-13	13	306	8,391	1,644	827	751	4.9
SemEval-15	4	138	2,604	1,022	659	512	5.5
SemCor	352	37,176	802,443	226,036	33,362	22,436	6.8

框架为 Pytorch 1.8.2 [5]，WordNet 3.0 由 Python 库 NLTK 3.5 [6]提供，预训练语言模型 BERT 的 *bert-base-uncased* 和 *bert-large-uncased* 版本都由 Python 库 Transformers 4.5 [7]提供。

基于解耦表征方法的词义消歧模型被命名为 DR-WSD，并且基于预训练语言模型 BERT 的 *bert-base-uncased* 版本的词义消歧模型被命名为 DR-WSD$_{\text{Base}}$，基于 *bert-large-uncased* 版本的词义消歧模型被命名为 DR-WSD$_{\text{Large}}$。词义消歧模型 DR-WSD 的核心超参数包括迭代次数（*epoch*）、批次大小（*batch-size*）、学习率（*learning-rate*）、待消歧文本最大长度（*context-maximum-length*）、注释文本最大长度（*gloss-maximum-length*）、例句文本最大长度（*example-maximum-length*）等。模型 DR-WSD$_{\text{Base}}$ 和 DR-WSD$_{\text{Large}}$ 都使用了相似的超参数设置，迭代次数为 20、批次大小为 4、学习率为 $\{1E\text{-}5, 5E\text{-}5, 1E\text{-}6, 5E\text{-}6\}$、待消歧文本的最大长度为 128、注释文本的最大长度为 32、例句文本的最大长度为 32。其它未列出的超参数参见章节对应发表论文所公布的代码。

3.3.3　对比模型

DR-WSD 模型的对比模型分为前人工作和基线模型两部分。**前人工作**部分包括一些主流的、代表性的词义消歧模型，但由于模型设计方式上的差别，该部分的对比结果并不能真实反应模型的好坏，仅作参考使用。**基线模型**部分包括一些与我们的工作可比的词义消歧模型，该部分的实验结果将反应模型的真实水平。其中前人工作部分包括对比模型 GLU[112]、LMMS[113]、SREF[114]、ARES[115]、SyntagRank[116]、COF[117]、ESR[118]、SACE[119]。基线模型部分包括对

[5] https://pytorch.org/

[6] https://www.nltk.org/

[7] https://huggingface.co/transformers/v4.5.1/

比模型 GlossBERT[120]、BEM[104]、IMS[121]、EWISE[122]、EWISER[123]。此外，基于 IMS[121] 模型的框架，分别采用三种传统的表征学习方法（即 Word2Vec[124]、Context2Vec[125]、BERT[30]）获得的词表征或文本表征替代 IMS 模型的原表征（即 IMS+word2vec、IMS+context2vec、IMS+bert），该设置的实验结果用于佐证解耦表征方法的有效性和优越性。上述模型的实验结果均取自原论文公布的数据。

对比模型（包括接下来章节部分的对比模型）的研究内容简述如下：

GLU[112]：该模型验证了上下文化的词表征对于词义消歧系统的积极意义，并通过整合多种上下文化的词表征获得了超越前人工作的突出表现。

COF[117]：该模型分析了上下文信息对词义消歧的价值，并指出上下文化的目标词表征或词义定义表征有利于促进词义消歧系统的整体表现。

BEM[104]：该模型采用双编码器，即目标词编码器与词义定义编码器，联合训练的方式实现训练文本的知识增强词义定义表征。

IMS[121]：该模型将待消歧文本与词义注释文本作为一个整体输入到神经网络以实现词义的识别，其优势在于模型构造简单，易于整合到现有学习模型之中。

ESR[118]：该模型通过合并同义词、上下位词或示例短语实现相关词的语义信息增强词义定义表征，并通过实验分析证明了结合此类附加信息可以提高词义消歧系统的性能。

GAS[126]：该模型将目标词的上下文和词义注释整合到一个统一的记忆网络，借助于记忆网络对目标词的上下文和词义注释文本之间的语义关系进行建模。

SREF[114]：该模型利用同义词间的语义信息增强词义表征，实验结果表明同义词间存在着强的相关性，能够改善表征的准确性。

ARES[115]：该模型利用词汇知识库中的相似词汇增强词义表征，其优势在于知识库中的语义知识能够保证生成的向量处于相同的语义空间。

SACE[119]：该模型通过整合相似词义的表征以改善词义表征的准确性，其优势在于可避免单一词义注释学习到的表征会存在的片面性。

LMMS[113]：该模型表明上下文化的表征方式对词义消歧任务具有积极意义，并且表征的准确性直接决定了系统最终的表现。

QWSD[127]：该模型提出了一种基于量子概率理论的词义消歧新算法，尽管提出的方法非常简单，但在标准的词义消歧评估框架下表现出优越的性能。

HCAN[128]：该模型使用协同注意力机制获取目标词的上下文与词义注释文本之间相互依赖信息的表征，并以分层方式扩展注意力机制以实现捕获单词级和句子级的相关性。

DR-WSD[129]：该模型使用解耦表征方法替换传统的表征学习方法，以实现降低表征学习过程中对大数据量的依赖。

EWISE[122]：该模型将词义定义嵌入约束到一个连续空间，以实现改善词义表征的准确性，其优势在于能够对未知词义的表征进行预测。

EWISER[123]：该模型利用了知识图谱中的语义信息以增强词义消歧系统对同义词的预测能力。

GlossBERT[120]：该模型采用词典中的词义注释去学习词义定义表征，将词义消歧任务改造成为一个词义匹配模型，其贡献在于将词义消歧的研究重心转移到了词义定义表征的不可区分性上来。

SyntagRank[116]：该模型利用 SyntagNet 语料库中所包含的共现词的消歧知识增强目标词表征与词义定义表征，并证实词汇与语义的组合能够提供丰富的语义信息。

SparseLMMS[130]：该模型证明了通过利用稀疏词表征可以在细粒度的全词词义消歧任务上超越更复杂的特定任务模型。

SensEmBERT[131]：该模型将语言模型的表征能力和语义网络中包含的大量知识相结合，生成单词的潜在语义层次的具有多种语言含义的词表征。

Z-Reweighting[132]：该模型利用了词频排名与词义数分布之间的统计关系，提出了一种词级 Z 重加权方法来调整不平衡数据集的权重分配。

3.4 实验结果与分析

本节对基于解耦表征方法构造的词义消歧模型 DR-WSD 进行实验分析，分析实验涉及词义消歧模型的标准评估实验、消融分析实验和长尾分析实验三部分。标准评估实验侧重于分析模型的整体表现，并不区分长尾词义与高频词义；消融分析实验侧重于评估解耦表征方法的有效性以及对词义消歧模型整体表现的贡献；长尾分析实验则专注于分析解耦表征方法和词义消歧模型对长尾词义识别的效用。

3.4.1 标准评估实验

DR-WSD 模型的实验结果如表 3-2 所示，对比模型分为前人工作对比组和基线模型对比组，其中表现最佳的结果采用粗体显示。

与前人工作相比，DR-WSD 在多数测试集上的表征优于对于模型，其中在测试集的整合集 *ALL* 上的不俗表现佐证了 DR-WSD 的有效性与竞争力。在测试集 *SE13* 和 *Adj.* 上 DR-WSD 的表现略低于对比模型。从数据集的统计信息来看，如表 3-1 所示，测试集 *SE13* 的“复杂度”最低，说明该数据集中高频词义的占比较大，而长尾词义的占比较小；测试集 *Adj.* 为形容词集，相比于名词集与动词集而言，形容词集中的高频词义的占比较大，而长尾词义的占比较小。由于上述两个测试集都是高频词义占主导的测试集，并不适合于评估我们的解耦表征方法的优势。同时该实验结果也说明解耦表征方法会对高频词义的识别有一定的拖累。

表 3-2 DR-WSD 模型的标准评估实验结果

模型	开发集	测试集				测试集的整合集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*N.*	*V.*	*Adj.*	*Adv.*	*ALL*
前人工作：										
GLU	68.1	75.5	73.6	71.1	76.2	–	–	–	–	74.1
LMMS	68.1	76.3	75.6	75.1	77.0	–	–	–	–	75.4
SREF	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
ARES	71.0	78.0	77.1	77.3	**83.2**	80.6	68.3	80.5	83.5	77.9
SyntagRank	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR	**75.4**	80.6	78.2	79.8	82.8	82.5	69.5	82.5	87.3	79.8
SACE	74.7	80.9	79.1	**82.4**	84.6	83.2	71.1	**85.4**	87.9	80.9
基线模型：										
IMS+word2vec	62.6	72.2	70.4	65.9	71.5	71.9	56.6	75.9	84.7	70.1
IMS+context2vec	61.3	71.8	69.1	65.6	71.9	71.0	57.6	75.2	82.7	69.0
IMS+bert	68.6	75.9	74.4	70.6	75.2	75.7	63.7	78.0	85.8	73.7
GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	83.0	87.9	79.0
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
EWISER	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
DR-WSD$_{\text{Base}}$	74.7	80.8	78.0	80.0	82.7	82.7	69.5	82.9	86.6	80.4
DR-WSD$_{\text{Large}}$	**75.4**	**81.1**	**79.2**	81.1	**83.2**	**84.9**	**72.1**	84.4	**88.5**	**81.4**

与基线模型相比，DR-WSD 在大多数测试集上都表现出色，其证实了我们的解耦表征方法在应对长尾词义消歧方面的有效性。下面给出具体的对比分析：

- 与具有相似结构的模型 BEM 相比，DR-WSD 优于 BEM，说明添加解耦表征方法的词义消歧模型可以提升模型整体的表现水平，一定意义上可以说明 DR-WSD 比 BEM 在长尾词义识别上是优秀的。
- 与使用了相似外部资源的模型 GlossBERT 和 BEM 相比，DR-WSD 同样表

现出色，说明解耦表征方法确实是有效的。

- 与采用不同表征学习方法的模型 EWISE 和 EWISER，以及采用不同编码方法的模型 IMS+word2vec、IMS+context2vec、IMS+bert 相比，DR-WSD 同样有不错的表现，证明我们的词义消歧模型同时使用传统表征方法和解耦表征方法的设计方案是正确，优于使用单一表征方法的词义消歧模型。

3.4.2 消融分析实验

为检验解耦表征方法对词义消歧模型整体的贡献，我们采用删除构件的消融研究方法实施验证，删除解耦表征方法后的模型被称为**消融模型**。此外，为检验获得解耦表征的约束项是否发挥作用，我们采用冻结编码器参数更新的消融研究方法实施验证，冻结目标词编码器与词义定义编码器参数更新后的模型被称为**冰冻模型**。对比模型为：

- 原模型 DR-WSD$_{\text{Base}}$，原模型采用基于预训练语言模型 BERT 的 *bert-base-uncased* 版本实现的模型。
- 消融模型 DR-WSD$^{ablation}_{\text{Base}}$，消融模型是在原模型的基础之上删除解耦表征方法后的模型。
- 冰冻模型 DR-WSD$^{frozen}_{\text{Base}}$，冰冻模型是在原模型的基础之上阻止因获取解耦表征对目标词编码器与词义定义编码器的参数更新操作后的模型。
- 模型 BEM，由于去除解耦表征方法后的模型结构与模型 BEM 的几乎相似，所以模型 BEM 被纳入对比实验。

模型 BEM 的配置和实验结果取自原论文，消融模型和冰冻模型的参数设置以及运行平台与标准评估实验部分保持一致。

表 3-3 DR-WSD 模型的消融分析实验结果

模型	开发集	测试集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：SemCor						
BEM	74.5	79.4	77.4	79.7	81.7	79.0
DR-WSD$_{\text{Base}}$	**74.7**	**80.8**	**78.0**	**80.0**	**82.7**	**80.4**
DR-WSD$^{ablation}_{\text{Base}}$	73.5	76.1	75.2	77.7	81.4	77.0
DR-WSD$^{frozen}_{\text{Base}}$	73.6	76.7	76.4	76.9	80.3	76.6

实验结果与分析：实验结果如表 3-3 所示。从整体表现来看，原模型 DR-

WSD_{Base} 优于三个对比模型，说明基于解耦表征方法的词义消歧模型在设计上是合理的。详细对比如下：

- 对比消融模型 $\text{DR-WSD}^{ablation}_{\text{Base}}$ 和原模型 $\text{DR-WSD}_{\text{Base}}$，$\text{DR-WSD}_{\text{Base}}$ 在测试集 *SE2* 和 *SE3* 上的表现要优于在测试集 *SE13* 和 *SE15* 上的表现。测试集 *SE2* 和 *SE3* 的“复杂度”较低，即数据集中高频词义占比多于长尾词义，在测试集 *SE2* 和 *SE3* 上的出色表现说明解耦表征方法有助于改善词义消歧模型对于长尾词义的识别能力，同时也说明解耦表征方法对词义消歧模型的整体表现是有贡献的。
- 对比冰冻模型 $\text{DR-WSD}^{frozen}_{\text{Base}}$ 和原模型 $\text{DR-WSD}_{\text{Base}}$，$\text{DR-WSD}_{\text{Base}}$ 的出色表现说明获取解耦表征方法的约束项确实有利于编码器参数更新，即获得更有利于词义消歧的表征，同时也说明使用解耦表征方法的词义消歧模型要比单纯使用 BERT 模型的词义消歧模型更有优势。
- 对比消融模型 $\text{DR-WSD}^{ablation}_{\text{Base}}$ 和冰冻模型 $\text{DR-WSD}^{frozen}_{\text{Base}}$，实验结果表明 $\text{DR-WSD}^{frozen}_{\text{Base}}$ 在测试集 *SE2* 和 *SE3* 上的表现更好，而 $\text{DR-WSD}^{ablation}_{\text{Base}}$ 在测试集 *SE13* 和 *SE15* 上的表现更好，说明解耦表征方法在简单任务上会有一定的作用，但在复杂任务上却有可能拖累词义消歧模型整体的表现。

3.4.3 长尾分析实验

为检验解耦表征方法在长尾词义消歧任务中的有效性，我们分别在构造的长尾词义数据集（Long-tail Sense Datasets，LTS）和最新的交叉语种数据集（Cross-lingual Datasets，CL）上实施实验。长尾词义数据集是基于标准估计实验设置的数据集构建的，通过挑选出标准估计实验设置的训练集、开发集和测试集中词义出现次数少于等于 3 次的词义构造相应的长尾词义训练集、开发集和测试集。交叉语种数据集来自于 Pasini 等人[133]在 2021 年最新提出的跨语言评估框架 [8]，该框架包含了除英文之外一些小语种数据集，可以被认为是一个低资源下的评估数据集。

对比模型： 长尾词义数据集上的对比模型采用消融分析实验部分构造的实验模型，即原模型 $\text{DR-WSD}_{\text{Base}}$、消融模型 $\text{DR-WSD}^{ablation}_{\text{Base}}$、冰冻模型 $\text{DR-WSD}^{frozen}_{\text{Base}}$。交叉语种数据集上的对比模型采用消融分析实验部分的原模型 $\text{DR-WSD}_{\text{Base}}$ 和消融模型 $\text{DR-WSD}^{ablation}_{\text{Base}}$ 以及原论文中使用的模型 XLMR-Base[134] 作为实验模型。

[8] https://sapienzanlp.github.io/xl-wsd/

由于跨语言评估框架是基于电子词典 BabelNet [9] 构建的，因此实验中的词义注释同样取自电子词典 BabelNet。此外，由于电子词典 BabelNet 中的大多数小语种都使用英语的词义注释，因此我们直接使用英语中的词义注释来提供其目标词的候选词义清单。

模型设置：原模型的实验设置与标准评估实验部分的保持一致，消融模型和冰冻模型的实验设置与消融分析实验部分的保持一致；此外三个对比模型的硬件运行平台和软件开发框架与标准评估实验部分的设置保持一致。模型 XLMR-Base 的实验结果来自于跨语言评估框架原文中公布的实验数据。

表 3-4　DR-WSD 模型在长尾词义数据集上的实验结果

模型	开发集	测试集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：LTS						
$\text{DR-WSD}_{\text{Base}}$	**43.9**	**48.5**	**47.1**	**48.8**	**50.4**	**49.2**
$\text{DR-WSD}^{ablation}_{\text{Base}}$	42.4	43.3	42.8	43.8	45.0	44.4
$\text{DR-WSD}^{frozen}_{\text{Base}}$	41.9	42.0	41.9	42.2	43.9	43.0

实验结果与分析：长尾词义数据集上的实验结果如表 3-4 所示。从整体表现来看，原模型优于消融模型和冰冻模型，说明当前基于解耦表征方法的词义消歧模型的设计是合理的。对比 $\text{DR-WSD}_{\text{Base}}$ 和 $\text{DR-WSD}^{ablation}_{\text{Base}}$，原模型优于消融模型，说明解耦表征方法有助于改善词义消歧模型的识别能力，同时也说明解耦表征方法有助于长尾词义的识别。对比 $\text{DR-WSD}_{\text{Base}}$ 和 $\text{DR-WSD}^{frozen}_{\text{Base}}$，原模型优于冰冻模型，同样佐证了解耦表征方法的有效性。对比 $\text{DR-WSD}^{ablation}_{\text{Base}}$ 和 $\text{DR-WSD}^{frozen}_{\text{Base}}$，消融模型优于冰冻模型，说明获取解耦表征过程中要是不更新编码器的参数反而会干扰词义消歧模型整体的效果，也说明获取解耦表征过程中对编码器参数的更新是有意义的。

交叉语种数据集上的实验结果如图 3-3 所示。对比 $\text{DR-WSD}_{\text{Base}}$ 和 XLMR-Base，原模型在多个数据集上优于模型 XLMR-Base，这说明我们的模型具有一定的鲁棒性，也说明解耦表征方法具有广泛的适用性。对比 $\text{DR-WSD}_{\text{Base}}$ 和 $\text{DR-WSD}^{ablation}_{\text{Base}}$，原模型优于消融模型，这说明解耦表征方法确实发挥了作用，间接指出解耦表征方法有助于低资源任务中的表征学习和词义识别问题。在 Spanish、Japanese、Estonian、Dutch、Croatian、Catalsn 等语种上，我们的模型表现不佳，

[9] https://babelnet.org/

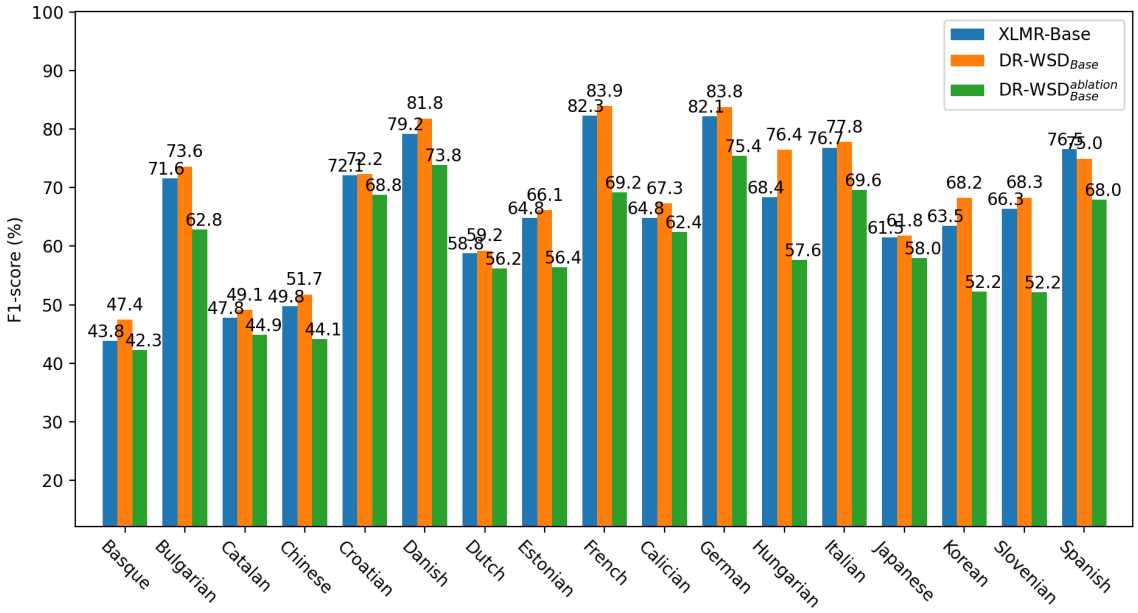

图 3-3 DR-WSD 模型在交叉语种数据集上的实验结果

核心原因在于上述语种的数据集体量比较小，而且大部分都是高频词义。

3.5 本章小结

本章针对训练样本与词义注释文本的低资源问题提出量子纠缠启发的解耦表征，用以应对低资源下的目标词表征与词义定义表征。该表征方法尝试降低表征学习模型对大数据量的依赖，同时实现针对性地挑选决定性的特征。基于该表征方法构造词义消歧模型并实施验证，实验结果证实解耦表征方法能够有效缓解表征学习模型的大数据量依赖问题。此外，为了更进一步明确解耦表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施消融分析实验和长尾分析实验，实验结果证实解耦表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果发表在 CIKM 会议论文中。

对于长尾词义消歧任务整体而言，本章提出的解耦表征方法从表征学习模型优化的角度而言有利于缓解表征学习模型对大数据量的依赖，但该方法对于改善表征向量的准确性方面依然存在不足。下一章将从表征向量优化的角度提出量子叠加态启发的自适应调优表征，用以解决低资源下的文本表征不准确性问题。

第4章　量子叠加态启发的自适应调优表征

在前一章中从表征学习模型优化的角度提出了同时适用于目标词表征与词义定义表征的解耦表征，该表征方法能够有效降低表征学习模型对大数据量的依赖，同时实现针对性地挑选决定性的特征。本章将从表征向量优化的角度提出同时适用于目标词表征与词义定义表征的自适应调优表征，该表征方法用于解决低资源下文本表征的不准确性问题。最后，基于量子叠加态启发的自适应调优表征方法构建词义消歧模型，并实施标准评估实验、消融分析实验和长尾分析实验。

4.1　研究动机

表征学习（Representation Learning，RL），也称特征学习，属于机器学习领域中基础性的研究课题，其目标是将原始数据映射为一种能够被机器学习算法直接使用的数学对象[135,136]。在表征学习提出之前，研究人员往往采用人工的方式手动地从原始数据中抽取特征以构建输入到算法的特征表示。人工的方式虽然一定程度上能够实现特征抽取的功能，但昂贵的人力成本和大量的时间开销使其难以应对数据规模指数级的增长速度和智能系统快速的更新迭代，同时也难以处理像音频、视频、图片之类的复杂数据类型。表征学习有效地弥补了人工方式的不足，使特征抽取自动化的进行，大幅降低了人力成本和时间开销，也为处理高维复杂数据类型提供了可能。最为重要的是表征学习的输出可以直接作为机器学习算法的输入，实现无缝衔接，并且此类设计能够显著改善系统整体的表现。

表征学习方法可以归纳为有监督学习和无监督学习。有监督的表征学习方法是从被标注的数据中抽取特征用以生成算法所需的特征表示，包括神经网络[137]（Neural Networks）、多层感知机[138]（Multi-Layer Perception）等。无监督的表征学习方法是从未标注的数据中抽取特征进而生成算法所需的特征表示，包括独立成分分析[139]（Independent Component Analysis）、主成分分析[140]（Principal Component Analysis）、矩阵分解[141]（Matrix Factorization）、自编码器[142]（Auto Encoders）等。当前，表征学习方法已经能够有效地从大量的训练样本中挖掘出

数据的特征表示，实现对复杂、高维数据的精准描述，但对于样本量稀疏的小样本任务而言，依赖于大样本驱动的表征学习方法将素手无策。导致表征学习方法失效的主要原因是：（1）有限的训练样本不足以使表征学习模型的全部参数得到充分的学习；（2）有限的训练样本只能支撑挖掘少量的有效特征。研究者们提出数据增强、知识迁移等方法应对小样本任务下的表征学习。数据增强的方法直接从源头入手，借助于外部资源改善小样本任务训练样本稀疏的问题，其核心思路是增加训练样本的数量。知识迁移的方法借助于外部知识强化表征学习方法抽取特征的能力与水平，其核心思想是将其它领域或任务中学习到的知识迁移到当前表征学习方法，以协助完成特征的抽取。数据增强的方法通过增加训练样本的数量来改善表征学习模型参数的学习效果，知识迁移的方法则通过引入外部知识来协助表征学习方法挖掘更多的有效特征，但两者都不同程度地依赖于外部资源。

长尾词义消歧任务本质上是一个低资源下的表征学习任务，其面临的主要困难在于：（1）目标词缺乏足够的训练样本，导致目标词表征难以获得准确的词表征；（2）词义定义缺乏足够的词义注释文本，导致词义定义表征难以获得准确的文本表征。对此，研究者们的主要应对思路可归纳为数据增强的方法[40]、知识迁移的方法[46]和施加外部约束的方法[38]。数据增强的方法尝试构造或引入外部数据增加长尾词义的训练样本数量。知识迁移的方法尝试整合多场景下目标词的上下文信息以改善目标词表征的准确性，或融合其它领域中的语义知识以改善长尾词义定义表征的准确性。施加外部约束的方法着眼于词义定义表征间内在的位置关系，尝试将词义定义表征映射到一个连续的语义空间，以实现表征之间相互的制约，最终实现由头部词义定义表征去校准长尾词义定义表征的目的。

在量子力学理论中，叠加态的构造过程能够实现在受约束的希尔伯特空间中对原表征向量实施进一步地调优与校准，可用于改善低资源下的表征向量的不准确性。在叠加态的形式化体系下，基于传统表征学习模型从训练数据中学习一个量子态，也就是原表征向量，并由该量子态获得其反态；此外，从训练数据中学习一个系数用作构建叠加态的概率幅；最后借助于量子态、量子态的反态以及概率幅共同构成叠加态，可实现由小样本中学习到的一个系数调整、优化原表征向量。对于长尾词义消歧任务而言，基于稀疏的训练数据学习到足够数量的有效特征是困难的，但学习到一个或几个有效的特征是可行的。量子力学中叠加态的构造过程具备基于原表征向量并借助于一个有效特征发掘更优表征的能力。

具体地，本章基于量子叠加态的构造过程提出量子叠加态启发的自适应调优表征，用以解决低资源下的目标词表征与词义定义表征的不准确性问题。该表征

方法尝试基于传统表征学习模型获得原表征向量，并基于叠加态的构造过程实现对原表征向量实施进一步的调优与校准。本章的贡献在于：

- 针对长尾词义消歧任务中训练样本与词义注释文本的低资源问题，提出量子叠加态启发的自适应调优表征方法，用以解决低资源下目标词表征与词义定义表征的不准确性问题；
- 基于量子叠加态启发的自适应调优表征构造词义消歧模型并实施验证，实验结果证实自适应调优表征有助于改善低资源下表征向量的不准确性问题，此外，基于自适应调优表征构建的词义消歧模型有效改善了长尾词义的识别能力。

4.2 基于自适应调优表征的词义消歧模型

本节首先介绍量子叠加态启发的自适应调优表征方法，然后基于该自适应调优表征方法构造词义消歧模型，最后给出词义消歧模型的训练方法。

4.2.1 自适应调优表征

在机器学习领域中，表征学习，或称特征学习，是一个重要的研究方向，对自然语言处理、图形图像处理领域中的众多下游任务都有着深远的影响。小样本、低资源情况下如何学习到有效的表征，或如何改善、优化已获得的表征（也称原表征）是一个具有挑战性的研究课题。量子叠加态启发的自适应调优表征方法旨在基于数据中抽取到的少数几个有效特征改善原表征的准确性，具体包括如下步骤：

- 由传统学习模型从训练样本中学习数据的一般形式下的特征表示，即原表征，再将其构造为量子态形式下的特征表示；
- 由量子非门计算量子态的反态；
- 由传统学习模型从训练样本中学习一个用于构造叠加态的概率幅；
- 基于量子态、量子态的反态以及概率幅获得叠加态形式下的表征，以实现在叠加态构成的空间中找到一个优于原表征的新表征。

具体实施细节如下所述：

1. 假设量子叠加态启发的自适应调优表征方法的输入为一个基于传统表征学

习方法获得的词表征或文本表征，

$$V = [v_1, v_2..., v_i, ...] \in \mathbb{R}^m, \tag{4-1}$$

其中标量 m 表示输入向量的维度；借助于神经网络中的全连接层（或称为线性层）将其调整为构造量子态所需的维度 n，

$$V' = Linear^1(V) \in \mathbb{R}^n, \tag{4-2}$$

其中上标 1 用于区分不同的线性层；并采用平方和归一化函数 $SSN(\cdot)$ 将其约束为**量子态**，

$$|\psi\rangle = SSN(V') = \frac{V'}{\|V'\|}, \tag{4-3}$$

则 $\||\psi\rangle\| = 1$。非纠缠的量子态是希尔伯特空间中的单位向量，上述方法可以作为构造量子态的通用方法。

2. 基于量子态 $|\psi\rangle$，并借助量子力学理论中的非门获得该**量子态的反态**，

$$|\overline{\psi}\rangle = \mathbf{X}|\psi\rangle \tag{4-4}$$

其中非门 $\mathbf{X}$ 的高维形式为：

$$\mathbf{X} = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}。 \tag{4-5}$$

该运算的目的在于获得量子态 $|\psi\rangle$ 在希尔伯特空间中相反方向的量子态，即反态。

3. 基于输入的表征 V 获得构造叠加态所需的**概率幅**，此处同样借助于神经网络中的全连接层（或称为线性层）获得一个用于构造概率幅的标量 θ，

$$\theta = Linear^2(V) \in \mathbb{R} \tag{4-6}$$

其中上标 2 用于区分不同的线性层。该标量 θ 也可借助于传统的学习模型直接从训练样本中获取。相比而言，直接取自训练数据的方法会更有利于获取准确的表征，但出于形式描述上的可读性，这里选择从输入表征中抽取。

4. 至此，基于上述获得的量子态 $|\psi\rangle$、量子态的反态 $|\overline{\psi}\rangle$ 和构成概率幅的标量 θ 可以构造**叠加态**，

$$|\Psi\rangle = \cos(\theta) \cdot |\psi\rangle + \sin(\theta) \cdot |\overline{\psi}\rangle \tag{4-7}$$

其中 $|\cos(\theta)|^2 + |\sin(\theta)|^2 = 1$，$\||\psi\rangle\| = 1$，$\||\overline{\psi}\rangle\| = 1$，所以叠加态同样是希尔伯特空

间中的单位向量，即 $|||\Psi\rangle|| = 1$。从输入表征到叠加态形式下的表征之间的映射可以被形式化地描述为：

$$\begin{aligned}\mathcal{F}(V) &= \cos\left(Linear^2(V)\right)\cdot SSN\left(Linear^1(V)\right) \\ &\quad + \sin\left(Linear^2(V)\right)\cdot \mathbf{X}SSN\left(Linear^1(V)\right) \\ &= |\Psi\rangle。\end{aligned} \tag{4-8}$$

该自适应调优表征方法涉及到的需要学习参数的地方只有两个线性层，即 $Linear^1(\cdot)$ 与 $Linear^2(\cdot)$。其中 $Linear^1(\cdot)$ 用于调整输入表征的维度，但并不能够实质性地改善表征的准确性；$Linear^2(\cdot)$ 虽然只需要学习一个标量，该标量能够实质性地改善表征的准确性。由 $Linear^2(\cdot)$ 学习到的标量 θ 能够控制生成的叠加态在其构成的空间中找到一个优于原表征（即输入表征）的新表征。其可行性在理论分析部分给出。

理论分析： 当训练样本数量不足时，直接获得一个准确的表征是困难的，往往学习到的表征是一个偏离正确表征的次优表征，借助于量子叠加态启发的自适应调优表征方法可以在次优表征的基础之上获得一个优于次优表征的新表征。由于获取优于次优表征的代价只是学习一个有效的参数，并不需要依赖大量的训练样本和学习大量的模型参数，所以对于小样本、低资源任务而言，自适应调优表征方法是一个可行的表征优化方案。该结论的正式定义和证明如下所示：

推论 4.1 假设 $|\phi\rangle$ 是正确表征的量子态形式，$|\psi\rangle$ 是在训练样本量不足时获得的次优表征的量子态形式，则基于量子态 $|\psi\rangle$ 可获得其叠加态 $|\Psi\rangle$，

$$|\Psi\rangle = \cos(\theta)\cdot|\psi\rangle + \sin(\theta)\cdot|\overline{\psi}\rangle。 \tag{4-9}$$

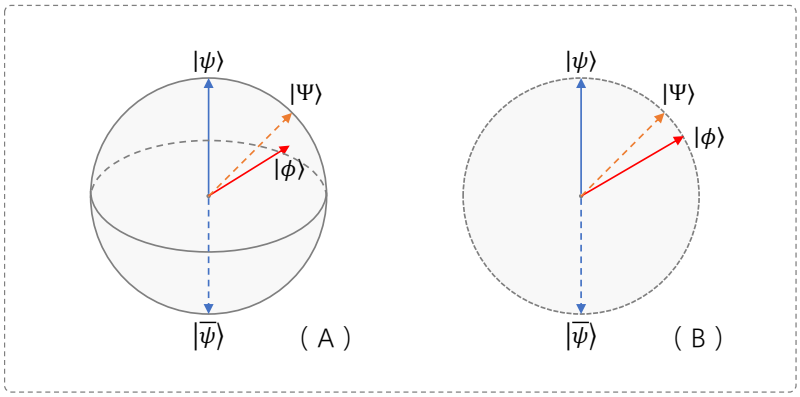

图 4-1 量子叠加态启发的自适应调优表征原理图

对于叠加态而言，只需要学习到一个有效的参数 θ 就可以获得优于 $|\psi\rangle$ 的量子态形式下的表征，即存在下述关系：

$$Pr(|\psi\rangle; |\phi\rangle\langle\phi|) \leq Pr(|\Psi\rangle; |\phi\rangle\langle\phi|) \leq 1 \tag{4-10}$$

其中 $|\phi\rangle\langle\phi|$ 表示测量算子。可令 $M_\phi = |\phi\rangle\langle\phi|$，则概率的计算公式为：

$$Pr(|\cdot\rangle; M_\phi) = \langle\cdot|M_\phi^\dagger M_\phi|\cdot\rangle = \langle\cdot|\phi\rangle\langle\phi|\cdot\rangle = |\langle\phi|\cdot\rangle|^2。 \tag{4-11}$$

证明： 在高维希尔伯特空间中，量子态 $|\psi\rangle$ 与其反态 $|\overline{\psi}\rangle$ 构成的叠加态可以形成一个平面，即希尔伯特空间球面中的大圆围成的平面，如图 4-1 所示。

- 当正确表征的量子态形式 $|\phi\rangle$ 不在叠加态形成的平面上时，如图 4-1 (A) 所示，平面上存在一些叠加态比 $|\psi\rangle$ 更接近于 $|\phi\rangle$，则存在 $Pr(|\psi\rangle; M_\phi) \leq Pr(|\Psi\rangle; M_\phi)$。
- 当正确表征的量子态形式 $|\phi\rangle$ 在叠加态形成的平面上时，如图 4-1 (B) 所示，叠加态能够准确地表示 $|\phi\rangle$，则存在 $Pr(|\Psi\rangle; M_\phi) = 1$；在其它情况下，则存在 $Pr(|\Psi\rangle; M_\phi) < 1$。
- 整合上式可导出存在以下关系：

$$Pr(|\psi\rangle; |\phi\rangle\langle\phi|) \leq Pr(|\Psi\rangle; |\phi\rangle\langle\phi|) \leq 1。 \tag{4-12}$$

□

说明： 该证明是存在性证明，即证明存在最优解，表明基于自适应调优表征方法具备找到比原表征更准确的新表征的条件，但也存在导致无法获得更优表征的可能性，如学习模型的学习能力不足，或并未从训练样本中挖掘到有效的原表征或概率幅。

4.2.2 模型结构

基于自适应调优表征的词义消歧模型框架结构如图 4-2 所示。词义消歧模型的整体结构可分为基于传统表征方法的**传统识别方法**和基于自适应调优表征方法的**量子识别方法**。此外，为增强词义消歧的准确性，两种识别方法又可以被分为词义定义识别方式和词义场景识别方式。其中词义场景识别方式是通过挖掘词典中词义的例句获得词义适用的上下文，在词义消歧过程中判断目标词是否适用于该上下文进而实现词义消歧。词义定义识别方式适合于能够得到准确词义定义表征的高频（头部、常用）词义消歧，而长尾词义由于常被应用于固定的场景，所

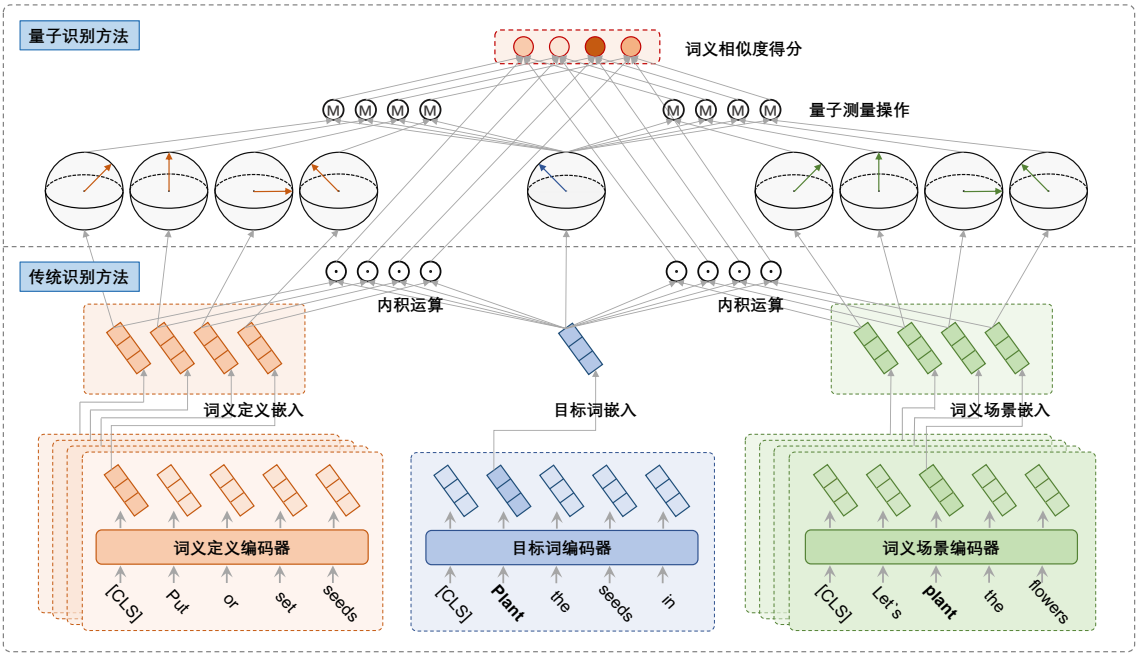

图 4-2 基于自适应调优表征的词义消歧模型

以词义场景识别方式更适合于长尾（低频、罕见）词义消歧。获取目标词嵌入的编码器被称为**目标词编码器**，获取词义定义嵌入的编码器被称为**词义定义编码器**，获取词义场景嵌入的编码器被称为**词义场景编码器**，它们都由预训练语言模型 BERT[30] 实现。

目标词编码器输入待消歧文本，输出目标词的词嵌入，即**目标词嵌入**。依照 BERT 模型的文本处理规则，会在文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入的待消歧文本为：

$$W^{text} = [w_1, w_2, ..., w_i, ...], \tag{4-13}$$

添加标记符后的文本为：

$$W^{text} = \left[[CLS], w_1, w_2, ..., w_i, ..., [SEP]\right], \tag{4-14}$$

其中待消歧文本包含目标词，即 $w_{target} \in W^{text}$。 BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W^{text}} = \left[V_{[CLS]}, V_{w_1}, V_{w_2}, ..., V_{w_i}, ..., V_{[SEP]}\right], \tag{4-15}$$

其中对应于目标词的词嵌入，即目标词嵌入，将被作为目标词编码器的输出，

$$V_{target} \equiv V_{w_{target}} = BERT_{TargetEncoder}(W^{text})。 \tag{4-16}$$

词义定义编码器输入词典中词义注释文本，输出词义定义的文本嵌入，即**词义定义嵌入**。BERT 模型同样会在输入文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入的词义注释文本为：

$$W_k^{gloss} = [w_{k,1}, w_{k,2}, ..., w_{k,i}, ...], \tag{4-17}$$

添加标记符后的文本为：

$$W_k^{gloss} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., w_{k,i}, ..., [SEP]_k\right], \tag{4-18}$$

其中索引 k 表示目标词的词义清单中第 k 个词义的词义注释文本。 BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W_k^{gloss}} = \left[V_{[CLS]_k}, V_{w_{k,1}}, V_{w_{k,2}}, ..., V_{w_{k,i}}, ..., V_{[SEP]_k}\right]。 \tag{4-19}$$

BERT 模型通常将开始标记符对应的词嵌入作为该文本的文本嵌入，所以开始标记符对应的词嵌入将被作为词义定义编码器的输出，

$$V_k^{gloss} \equiv V_{[CLS]_k} = BERT_{GlossEncoder}(W_k^{gloss})。 \tag{4-20}$$

词义场景编码器输入词典中词义例句文本，输出**词义场景嵌入**。由于词义场景编码器是要获取词义适用的上下文信息，所以在 BERT 模型编码过程中会将例句中包含的目标词遮盖住，以获得目标词适用的场景信息。输入到 BERT 模型中的文本首先会在开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，然后将文本中的目标词使用遮盖标记符 [*MASK*] 覆盖，如输入的词义例句文本为：

$$W_k^{example} = [w_{k,1}, w_{k,2}, ..., w_{k,i}, ...], \tag{4-21}$$

添加开始与分割标记符后的文本为：

$$W_k^{example} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., w_{k,i}, ..., [SEP]_k\right]。 \tag{4-22}$$

假设单词 $w_{k,i}$ 为目标词，则添加遮盖标记符后的文本为：

$$W_k^{example} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., [MASK]_k, ..., [SEP]_k\right]。 \tag{4-23}$$

其中索引 k 表示目标词的词义清单中第 k 个词义对应的词义例句。通常情况下，词典中的词义例句个数不止一个，当多于一个时默认选择第一个，当不存在时使用待消歧文本替代。 BERT 模型对输入的文本进行编码，并生成每个单词对应的

词嵌入（包括开始标记符、分割标记符和遮盖标记符），

$$V_{W_k^{example}} = \left[V_{[CLS]_k}, V_{w_{k,1}}, V_{w_{k,2}}, ..., V_{[MASK]_k}, ..., V_{[SEP]_k} \right]。 \tag{4-24}$$

由于词义场景编码器输出的要是词义适用的上下文，所以这里选择遮盖标记符对应的词嵌入作为词义场景编码器的输出，

$$V_k^{context} \equiv V_{[MASK]_k} = BERT_{ContextEncoder}(W_k^{example})。 \tag{4-25}$$

传统识别方法下的词义相似度得分： 至此，可以计算传统表征方法下词义定义识别方式的词义相似度得分，

$$Score_k^{gloss} = V_{target} \odot V_k^{gloss}, \tag{4-26}$$

和词义场景识别方式的词义相似度得分，

$$Score_k^{context} = V_{target} \odot V_k^{context}, \tag{4-27}$$

其中符号 $\odot$ 表示内积运算。

量子识别方法下的词义相似度得分： 将目标词编码器生成的目标词嵌入 V_{target} 复制一份，用自适应调优表征方法 $\mathcal{F}(\cdot)$，即公式 (4-8)，将其构造为量子态形式下的目标词嵌入，

$$|\Psi_{target}\rangle = \mathcal{F}(V_{target}), \tag{4-28}$$

再将其构造为测量算子，

$$M_{target} = |\Psi_{target}\rangle\langle\Psi_{target}|。 \tag{4-29}$$

其次，将词义定义编码器生成的词义定义嵌入 V_k^{gloss}、词义场景编码器生成的词义场景嵌入 $V_k^{context}$ 各复制一份，分别用自适应调优表征方法将其构造为量子态形式下的词义定义嵌入，

$$|\Psi_k^{gloss}\rangle = \mathcal{F}(V_k^{gloss}), \tag{4-30}$$

和量子态形式下的词义场景嵌入，

$$|\Psi_k^{context}\rangle = \mathcal{F}(V_k^{context})。 \tag{4-31}$$

至此，可以计算自适应调优表征方法下词义定义识别方式的词义相似度得分，

$$Q\text{-}Score_k^{gloss} = Pr(|\Psi_k^{gloss}\rangle; M_{target}), \tag{4-32}$$

和词义场景识别方式的词义相似度得分，

$$Q\text{-}Score_k^{context} = Pr(|\Psi_k^{context}\rangle; M_{target}), \tag{4-33}$$

其中概率计算方法如公式 (4-11) 所示。

4.2.3 模型训练

该词义消歧模型采用表现效果最好的 Adam 优化器[105]更新模型的参数，优化器的具体设置在实验部分给出。词义消歧模型使用交叉熵损失函数评估模型，损失函数为：

$$\begin{aligned} Loss(Score^{all}, index) &= -\log\left(\frac{\exp(Score_{[index]}^{all})}{\sum_{i=1}\exp(Score_{[i]}^{all})}\right) \\ &= -Score_{[index]}^{all} + \log\sum_{i=1}\exp(Score_{[i]}^{all}) \end{aligned} \tag{4-34}$$

其中 $Score^{all}$ 表示目标词的词义清单中各个词义的相似度得分，

$$Score^{all} = \left[Score_1^{all}, Score_2^{all}, ..., Score_k^{all}, ...\right], \tag{4-35}$$

其包括传统表征方法下的词义定义识别方式和词义场景识别方式、自适应调优表征方法下的词义定义识别方式和词义场景识别方式的词义相似度得分，

$$Score_k^{all} = Score_k^{gloss} + Score_k^{context} + Q\text{-}Score_k^{gloss} + Q\text{-}Score_k^{context}, \tag{4-36}$$

变量 *index* 表示目标词的词义清单列表索引。

4.3 实验准备

数据集：本章提出的基于自适应调优表征的词义消歧模型同样使用第三章的实验设置方法，即标准评估实验设置和增强评估实验设置。**标准评估实验设置**的数据集被划分为训练集、开发集、测试集，其中训练集为 SemCor，开发集为 SemEval-07[107]（*SE7*），测试集包括 Senseval-2[108]（*SE2*）、Senseval-3[109]（*SE3*）、SemEval-13[110]（*SE13*）、SemEval-15[111]（*SE15*）以及所有测试集的整合集（*ALL*）。此外，测试集的整合集 *ALL* 又按词性划分为名词集（*N.*）、动词集（*V.*）、形容词集（*Adj.*）、副词集（*Adv.*）。**增强评估实验设置**的数据集同样包括训练集、开发集、测试集，其中训练集包括 SemCor 和 WNGT（WordNet Gloss

Tags），开发集和测试集与标准评估实验设置下的开发集和测试集保持一致。此外，需要注意的是目标词的候选词义清单为电子词典 WordNet 3.0 中给出的目标词的所有词义。评估实验所使用的评估指标为百分比下的 F1 值，即 F1-score (%)。

实验设置：本章提出的词义消歧模型同样运行在第三章描述的硬件和软件平台上。基于量子叠加态启发的自适应调优表征的词义消歧模型被命名为 SR-WSD，并且基于预训练语言模型 BERT 的 *bert-base-uncased* 版本的词义消歧模型被命名为 SR-WSD_{Base}，基于 *bert-large-uncased* 版本的词义消歧模型被命名为 SR-$\text{WSD}_{\text{Large}}$。词义消歧模型的超参数分别为：迭代次数（*epoch*）为 20、批次大小（*batch-size*）为 4、学习率（*learning-rate*）为 $\{1E\text{-}5, 5E\text{-}5, 1E\text{-}6, 5E\text{-}6\}$、待消歧文本的最大长度（*context-maximum-length*）为 128、注释文本的最大长度（*gloss-maximum-length*）为 32、例句文本的最大长度（*example-maximum-length*）为 32。其它未列出的超参数可参见章节对应发表论文所公布的代码。

对比模型：SR-WSD 模型的对比模型分为标准实验对比模型和增强实验对比模型两部分。**标准实验对比模型**都采用预训练语言模型 BERT 的 *bert-base-uncased* 版本构建编码器，标准实验对比模型部分包括对比模型 EWISE[122]、LMMS[113]、SREF[114]、ARES[115]、BEM[104]、EWISER[123]、SyntagRank[116]、COF[117]、ESR[118]、Z-Reweighting[132]、QWSD[127]、DR-WSD[129]。**增强实验对比模型**则采用预训练语言模型 BERT 的 *bert-large-uncased* 版本构建编码器，增强实验对比模型部分包括对比模型 SparseLMMS[130]、EWISER[123]、ESR[118]。对比模型研究内容的简介参见第三章的对比模型部分，此外上述模型的实验结果均取自原论文公布的数据。

4.4 实验结果与分析

本节对基于自适应调优表征构造的词义消歧模型 SR-WSD 进行实验分析，分析实验涉及词义消歧模型的标准评估实验、消融分析实验和长尾分析实验三部分。标准评估实验侧重于分析模型的整体表现，并不区分长尾词义与高频词义；消融分析实验侧重于评估自适应调优表征方法的有效性以及对模型整体表现的贡献；长尾分析实验则专注于分析词义消歧模型和自适应调优表征方法对长尾词义识别的效用。

4.4.1 标准评估实验

SR-WSD 模型的实验结果如表 4-1 所示，对比模型分为标准实验对比模型组和增强实验对比模型组，标准实验对比模型组包含量子启发的对比模型，其中与 $\text{SR-WSD}_{\text{Base}}$ 相比表现最佳的结果采用粗体显示，与 $\text{SR-WSD}_{\text{Large}}$ 相比表现最佳的结果添加下划线。

表 4-1 SR-WSD 模型的标准评估实验结果

模型	开发集	测试集				测试集的整合集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*N.*	*V.*	*Adj.*	*Adv.*	*ALL*
标准实验对比模型：										
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
LMMS	68.1	76.3	75.6	75.1	77.0	–	–	–	–	75.4
SREF	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
ARES	71.0	78.0	77.1	77.3	83.2	80.6	68.3	80.5	83.5	77.9
BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	**87.9**	79.0
EWISER	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
SyntagRank	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR	**75.4**	80.6	78.2	79.8	82.8	82.5	69.5	82.5	87.3	79.8
Z-Reweighting	71.9	79.6	76.5	78.9	82.5	–	–	–	–	78.6
量子启发的模型：										
QWSD	-	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
DR-WSD	74.7	80.8	78.0	80.0	82.7	82.7	69.5	82.9	86.6	80.4
$\text{SR-WSD}_{\text{Base}}$	74.5	**80.6**	**79.1**	**80.0**	**84.7**	**83.7**	**71.4**	82.8	86.7	**80.5**
$\text{SR-WSD}_{\text{Large}}$	75.4	81.8	80.2	81.1	85.1	84.9	72.1	84.4	88.0	81.7
增强实验对比模型：										
SparseLMMS	73.0	79.6	77.3	79.4	81.3	–	–	–	–	78.8
EWISER	75.2	80.8	**79.0**	80.7	81.8	81.7	66.3	81.2	85.8	80.1
ESR	77.4	81.4	78.0	81.5	83.9	**83.1**	71.1	**83.6**	**87.5**	80.7
$\text{SR-WSD}_{\text{Base}}$	**78.1**	**81.5**	**79.0**	**82.0**	**84.4**	**83.1**	**71.2**	80.9	84.1	**81.0**
$\text{SR-WSD}_{\text{Large}}$	78.8	82.4	81.0	83.4	85.6	84.0	73.0	84.7	87.0	82.1

与标准实验对比模型相比，$\text{SR-WSD}_{\text{Base}}$ 与 $\text{SR-WSD}_{\text{Large}}$ 在多个测试集上都优于对比模型，其佐证了自适应调优表征方法的有效性与竞争力。

- 与量子启发的模型 QWSD、 DR-WSD 相比，SR-WSD 远优于 QWSD，其中一个值得注意的原因是 QWSD 没有使用可训练的表征学习模型，而 SR-WSD 选择了当前主流的预训练语言模型；SR-WSD 在多个指标上小幅优于 DR-WSD，其原因在于上述模型都是针对于长尾词义消歧，长尾词义数据

占比较少，难以明显的体现出优势。

- 与空间约束机制下的模型 EWISE 相比，SR-WSD 也要优于 EWISE，其原因是 SR-WSD 不仅使用连续空间来约束表征，而且使用的自适应调优表征方法能在原表征的基础之上发掘更优表征。
- 在开发集 *SE7* 上，SR-WSD 不及对比模型 ESR，其原因在于模型 ESR 具有远超平均水平的拟合能力，但需要说明的是学习模型过度拟合训练集与开发集将影响模型的泛化水平。
- 在测试集 *Adj.* 与 *Adv.* 上，$\text{SR-WSD}_{\text{Base}}$ 不及对比模型 BEM。测试集 *Adj.* 与 *Adv.* 分别为形容词集与副词集，相比于名词集和动词集而言，形容词集和副词集中的高频词义的占比较大，而长尾词义的占比较小。由于上述测试集是高频词义占主导的测试，并不适合于评估我们的词义消歧模型的优势，因为该词义消歧模型中采用了适合于长尾词义消歧的自适应调优表征方法。此外，该实验结果也说明自适应调优表征方法对于高频词义占主导的数据而言并不具有优势。

与增强实验对比模型相比，SR-WSD 的整体表征优于绝大多数对比模型，但在测试集 *Adj.* 和 *Adv.* 上表现不及对比模型 ESR。

- 测试集 *Adj.* 与 *Adv.* 分别为形容词集与副词集，相比于名词集和动词集而言，形容词集和副词集中的高频词义的占比较大，而长尾词义的占比较小。SR-WSD 的创新之处在于改善了对长尾词义的识别能力，而测试集 *Adj.* 与 *Adv.* 中的长尾词义占比较小，所以上述测试集并不能展现 SR-WSD 的优势。
- 在重要的测试集 *ALL* 上的表现要优于对比模型 ESR，说明自适应调优表征方法对于高频词义的识别同样具有一定的促进作用。

4.4.2 消融分析实验

为检验自适应调优表征方法对词义消歧模型整体表现的贡献，我们采用删除构件的消融研究方法实施验证，删除自适应调优表征方法后的模型被称为**消融模型**。对比模型为：

- 原模型 $\text{SR-WSD}_{\text{Base}}$，原模型采用基于预训练语言模型 BERT 的 *bert-base-uncased* 版本实现的模型。
- 消融模型 $\text{SR-WSD}^{ablation}_{\text{Base}}$，消融模型是在原模型的基础之上删除量子识别方

法部分后的模型，即只保留传统识别方法部分。

消融模型的参数设置以及运行平台与标准评估实验部分的保持一致。

表 4-2　SR-WSD 模型的消融分析实验结果

模型	测试集				
	SE2	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：SemCor					
SR-WSD$_{\text{Base}}$	**80.6**	**79.1**	**80.0**	**84.7**	**80.5**
SR-WSD$^{ablation}_{\text{Base}}$	77.7	75.3	76.1	82.0	77.1

实验结果与分析：实验结果如表 4-2 所示。从整体表现来看，原模型 SR-WSD$_{\text{Base}}$ 优于对比模型，说明基于自适应调优表征方法的量子识别方法部分对词义消歧模型的整体表征有着重要的贡献。对比消融模型 SR-WSD$^{ablation}_{\text{Base}}$ 和原模型 SR-WSD$_{\text{Base}}$，原模型 SR-WSD$_{\text{Base}}$ 在测试集 *SE2*、*SE3*、*SE15* 和 *ALL* 上的表现相对于消融模型 SR-WSD$^{ablation}_{\text{Base}}$ 都有 2-3 个百分点的提高；在测试集 *SE13* 上的表现相对于消融模型 SR-WSD$^{ablation}_{\text{Base}}$ 有 4 个百分点的提高。其中测试集 *SE13* 和 *SE15* 是高频词义占主导的测试集，而测试集 *SE2* 和 *SE3* 是长尾词义占主导的测试集。该结果说明自适应调优表征方法确实在长尾词义消歧方面具有重要贡献，同时也说明自适应调优表征方法对于高频词义的识别具有促进作用。

4.4.3　长尾分析实验

为检验自适应调优表征方法在长尾词义消歧任务中的有效性，我们分别在构造的长尾词义数据集（Long-tail Sense Datasets，LTS）和最新的交叉语种数据集（Cross-lingual Datasets，CL）上实施实验。长尾词义数据集的构造方法参见章节 3.4.3，交叉语种数据集的简介同样参见章节 3.4.3。

表 4-3　SR-WSD 模型在长尾词义数据集下的实验结果

模型	测试集				
	SE2	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：LTS					
SR-WSD$_{\text{Base}}$	**48.3**	**48.6**	**49.6**	**50.5**	**49.7**
SR-WSD$^{ablation}_{\text{Base}}$	43.1	43.4	44.2	45.7	44.8

对比模型：长尾词义数据集上的对比模型采用消融分析实验部分构造的实验

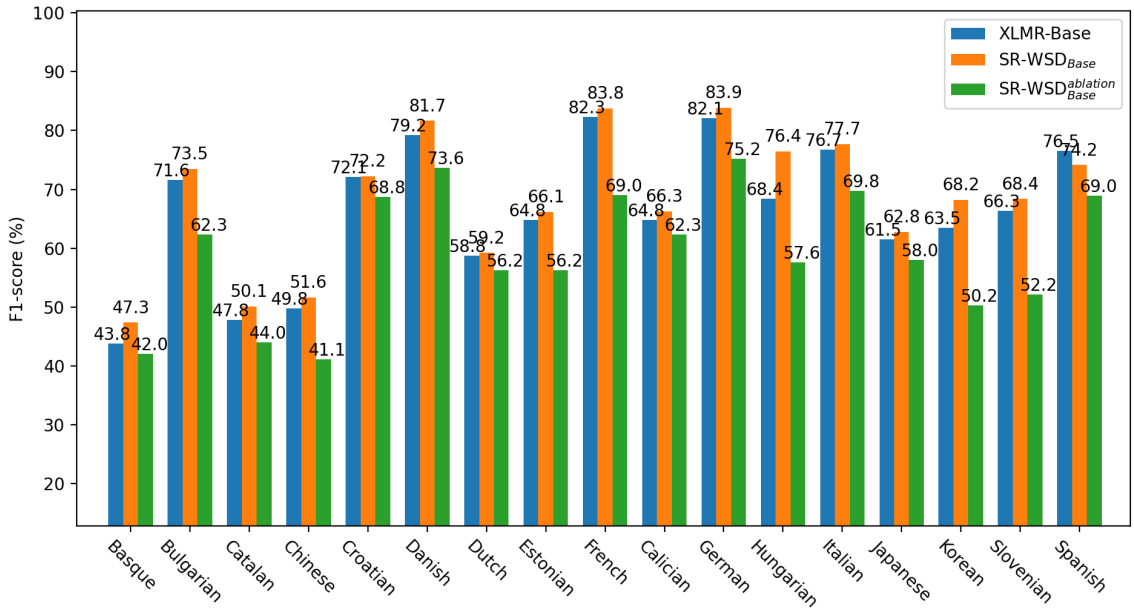

图 4-3 SR-WSD 模型在交叉语种数据集下的实验结果

模型，即原模型 SR-WSD$_{\text{Base}}$、消融模型 SR-WSD$^{ablation}_{\text{Base}}$。交叉语种数据集上的对比模型采用消融分析实验部分的原模型 SR-WSD$_{\text{Base}}$ 和消融模型 SR-WSD$^{ablation}_{\text{Base}}$ 以及原论文中使用的模型 XLMR-Base[134] 作为实验模型。由于跨语言评估框架是基于电子词典 BabelNet 构建的，因此实验中的词义注释同样取自电子词典 BabelNet。此外，由于电子词典 BabelNet 中的大多数小语种都使用英语的词义定义，因此我们直接使用英语中的词义注释来提供其目标词的候选词义清单。

模型设置：原模型的实验设置与标准评估实验部分的保持一致，消融模型的实验设置与消融分析实验部分的保持一致；此外两个对比模型的硬件运行平台和软件开发框架与标准评估实验部分的设置保持一致。模型 XLMR-Base 的实验结果来自于跨语言评估框架原文中公布的实验数据。

实验结果与分析：长尾词义数据集上的实验结果如表 4-3 所示。从整体表现来看，原模型优于消融模型，说明当前基于自适应调优表征方法的词义消歧模型的设计是合理的。对比 SR-WSD$_{\text{Base}}$ 和 SR-WSD$^{ablation}_{\text{Base}}$，原模型优于消融模型，说明自适应调优表征方法有助于改善词义消歧模型的整体表现，同时也说明自适应调优表征方法有助于改善长尾词义的识别能力。

交叉语种数据集上的实验结果如图 4-3 所示。对比 SR-WSD$_{\text{Base}}$ 和 XLMR-Base，原模型在多个数据集上优于模型 XLMR-Base，说明我们的模型具有一定的鲁棒性，也表明自适应调优表征方法具有广泛的适用性。对比 SR-WSD$_{\text{Base}}$ 和

SR-WSD$_{\text{Base}}^{ablation}$，原模型优于消融模型，说明自适应调优表征方法确实发挥了作用，间接指出自适应调优表征方法有助于低资源任务中的表征学习和词义识别问题。在 Basque、Croatian、Dutch、Calician、Japanese 等语种上，我们的模型表现不佳。通过分析表现不佳的数据集，我们发现它们大多是小众或非字母类型的语言。结果不佳的原因在于我们使用英语的词义注释而不是其它语言对应的词义注释，因此字符类型与英语相似的语言会表现更好。

4.5 本章小结

本章针对训练样本与词义注释文本的低资源问题提出量子叠加态启发的自适应调优表征，用以应对低资源下的目标词表征与词义定义表征。该表征方法尝试基于有限数据中学习到的一个有效特征改善表征向量的准确性，用以应对长尾词义消歧任务中低资源下的文本表征不准确性问题。基于该表征方法构造词义消歧模型并实施验证，实验结果证实自适应调优表征方法能够有效改善低资源下的文本表征不准确性问题。此外，为了更进一步明确自适应调优表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施消融分析实验和长尾分析实验，实验结果证实自适应调优表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果发表在 AAAI 会议论文中。

对于长尾词义消歧任务整体而言，本章提出的自适应调优表征方法从表征向量优化的角度而言有助于改善低资源下的文本表征不准确性问题，但该方法对于构建词义表征之间的内在联系以及位置关系方面依然存在不足。下一章将针对于词义定义表征之间的不可区分性问题提出量子测量启发的多源融合表征，用以强化词义之间的可区分性以及增强词义之间的区分边界。

第5章　量子测量启发的多源融合表征

在前二章中分别从表征学习模型优化的角度提出了解耦表征、从表征向量优化的角度提出了自适应调优表征，上述表征方法分别解决了表征学习模型的大数据量依赖问题和低资源下的文本表征不准确性问题。本章将针对词义定义表征之间的不可区分性问题提出基于数据增强的多源融合表征，用以强化词义之间的可区分性以及增强词义之间的区分边界。基于提出的量子测量启发的多源融合表征方法构建词义消歧模型，并实施标准评估实验、消融分析实验和长尾分析实验。

5.1　研究动机

长尾词义消歧面临的核心挑战在于：（1）匮乏的训练文本导致目标词表征难以学习到所属上下文的特征信息；（2）有限且固定的词义注释文本导致词义定义表征难以学习到词义定义的特征信息。总得来说，长尾词义消歧任务面临的核心挑战是低资源下的表征学习问题[20]。其中针对训练样本匮乏问题的学习策略或学习模型可以借鉴小样本学习[99]、元学习[98]、迁移学习[97]等领域的研究成果，而词义注释文本匮乏所导致的表征学习问题是长尾词义消歧所独有的。词义消歧模型不仅要从有限且固定的词义注释文本中学习到有效的词义定义表征，而且要求表征间具有明确的、可区分的边界。更为困难的是词义注释文本并不总是能够提供准确的、易于区分的词义定义描述。事实上，“词义”的概念本身就存在着争议，人类并不总是能够对所有的词义定义达成一致的共识。在人类的认知体系中，一个概念往往是由于区别于已有概念而出现的，在定义或描述时也常常是通过建立与已知概念之间的联系或区别以实现对新概念的澄清。但在此过程中，由于人们的知识储备、认知能力存在差异，对不同概念建立起来的认知也存在着差异，所以对同一个概念，不同的人会有不同的理解。对于长尾词义注释文本而言，有限且片面的描述信息难以建立起词义之间存在着的内在联系与区别，所以难以支撑表征学习模型学习到清晰且易于区分的词义定义表征。

对此，研究者们提出如下三种应对思路：（1）借用词典中提供的词汇知识图

谱建立不同词义之间的内在联系与区别；（2）借助于词义注释文本以外的知识增强词义定义表征，进而强化不同词义之间的区分边界；（3）借助于连续语义空间建立相似词义之间的内在联系。基于词汇知识图谱的工作中最具代表性的是Scozzafava 等人[28]提出的 SyntagRank 模型，其借助于电子词典 WordNet 中的词汇知识图谱，利用个性化的 PageRank 算法挖掘知识图谱中词义之间的联系进而实施词义消歧。SyntagRank 模型还借用了 WNG 和 SyntagNet 语料库中的词汇关系知识完善电子词典 WordNet 中的词汇知识图谱。此外，Vial 等人[42]通过上位词寻找根源词汇以减少输出类别的数量，其实质上是挖掘了词汇知识图谱中词义的上位词关系；Wang 等人[29]则利用上下位词设计了一套校验机制以实现对词义消歧系统的预测结果进行纠偏与校准。基于知识增强的工作中最具新颖性的是 Calabrese 等人[44]提出的利用 BabelPic 数据集中的图像信息增强词义定义表征，实现借助于多模态信息完善词义定义表征。此外，Loureiro 等人[33]通过对同义词表征取均值以实现对未知词义定义创建表征； Ruas 等人[46]采用多词义定义表征的方式对抗单表征方式下词义定义表征不准确的问题； Barba 等人[47]采用多语言知识来改善词义定义表征的准确性； Song 等人[48]则利用了更多的有助于改善词义定义表征的数据，如词义例句、同义词、上下位词等。基于连续语义空间的工作中最具影响力的是 Kumar 等人[36]提出的采用连续语义空间施加约束以实现借助于高频词义定义表征校准长尾词义定义表征，其优势在于能够推断未知词义定义表征。此外，Kang 等人[58]将目标词表征与词义定义表征合并到同一个语义空间以使表征间更紧凑，寻找更高效； Kang 等人[59]在之前工作的基础上提出新的目标词表征与词义定义表征结合的方法以获得更加细粒度的词义定义表征。

在量子力学理论中，叠加态的构造过程能够实现融合多源数据下获得的表征向量，并依据各个表征向量的重要程度赋予相应的权重值；此外，叠加态的测量过程所导出的干涉项能够约束表征向量在希尔伯特空间中的位置关系，进而建立起表征向量之间的内在联系。对于长尾词义消歧任务而言，基于有限且固定的词义注释文本学习到准确的且词义之间存在区分边界的词义定义表征是困难的，对此，一个自然的想法是实施数据增强。量子力学中的叠加态的构造过程与叠加态的测量过程为整合（多源）外部数据增强词义定义表征之间的内在联系以及位置关系提供了完备的数学框架。

具体地，本章借助叠加态的构造过程实现融合多源数据中获得的词义定义表征，借用叠加态的测量过程所导出的干涉项实现在语义空间中约束词义之间的位置关系以及构建词义之间的内在联系，即多源融合表征。该表征方法试图整合外

部数据中获得的表征向量增强词义定义表征之间的可区分性；此外，借助外部数据中获得的表征向量约束表征学习模型中参数搜索的范围，进而降低表征学习过程中对大数据量的依赖。本章的贡献在于：

- 针对长尾词义消歧任务中有限且固定的词义注释文本难以建立起词义之间内在联系以及位置关系的挑战，提出量子测量启发的多源融合表征，用以实现通过整合外部数据的方式解决词义定义表征之间的不可区分性问题；
- 基于量子测量启发的多源融合表征构造词义消歧模型并实施验证，实验结果证实多源融合表征有助于改善词义定义表征之间的不可区分性问题，此外，基于多源融合表征构建的词义消歧模型有效改善了长尾词义的识别能力。

5.2 基于多源融合表征的词义消歧模型

本节首先介绍量子测量启发的多源融合表征方法，然后基于该多源融合表征方法构造词义消歧模型，最后给出词义消歧模型的训练方法。

5.2.1 多源融合表征

对于小样本、低资源任务而言，面临的核心挑战在于训练样本缺乏，不足以支撑表征学习模型学习到有效的表征向量。面对训练样本稀缺问题最直接的应对方案是构造或挖掘更多的训练数据，进而实施数据增强。本节基于叠加态的构造过程实现对多个不同来源的外部数据（即多源数据）下获得表征向量实施融合，以建立起表征向量之间的内在联系，基于叠加态的测量过程实现对多源数据下获得表征向量实施空间几何形式上的约束，以实现在语义空间中建立起表征向量之间的位置关系，最终实现强化表征向量之间的可区分性。

量子叠加态的构造过程实施多源表征融合： 假设借助于多个经典的表征学习模型能够从多个数据源中获得相应目标（或对象）的词表征或文本表征，如表征 V_i，

$$V_i = [v_1, v_2, ..., v_j, ...] \in \mathbb{R}^n \tag{5-1}$$

其中索引 i 指代各个数据源。可利用平方和归一化函数 $SSN(\cdot)$ 将获得的表征构造

为量子态，

$$|\psi_i\rangle = SSN(V_i) = \frac{V_i}{\|V_i\|} \in \mathbb{R}^n \tag{5-2}$$

其中 $\||\psi_i\rangle\| = 1$。借助于量子叠加态的构造过程可以将多个数据源下获得的表征整合为一个叠加态，即叠加态形式下的特征表示，

$$|\Psi\rangle = \varepsilon_1 \cdot |\psi_1\rangle + \varepsilon_2 \cdot |\psi_2\rangle + ... + \varepsilon_i \cdot |\psi_i\rangle + ... \tag{5-3}$$

其中 $\varepsilon_i \in \mathbb{C}$ 表示构成叠加态的概率幅，$\sum_i |\varepsilon_i|^2 = 1$。基于输入的表征 V_i 可获得构造叠加态所需的概率幅，此处可借助于神经网络中的全连接层（或称为线性层）获得各个量子态 $|\psi_i\rangle$ 对应的概率幅 ε_i，

$$[\varepsilon_1, \varepsilon_2, ..., \varepsilon_i, ...] = SSN\Big(Linear([V_1, V_2, ..., V_i, ...])\Big)。 \tag{5-4}$$

量子叠加态的构造过程可以统一各个表征向量中分量的属性，同时能够根据各个表征向量的准确程度、重要性赋予不同的权重值，即概率幅。此外，更为重要的是量子态是构建在希尔伯特空间下的表征形式，能够抵消表征向量长度对表征准确性的干扰。

量子叠加态的测量过程实施多源表征约束：对多源数据下获得的叠加态实施测量操作将导出量子干涉项，

$$Pr(|\Psi\rangle; M_m) = \|M_m|\Psi\rangle\|^2 = \sum_i \|\varepsilon_i \cdot M_m|\psi_i\rangle\|^2 + Int, \tag{5-5}$$

其中 M_m 表示具体任务中需要识别的向量所构造的测量算子，如向量 $V \in \mathbb{R}^n$，量子态为：

$$|\psi\rangle = SSN(V) \in \mathbb{R}^n, \tag{5-6}$$

测量算子为：

$$M_m = |\psi\rangle\langle\psi|, \tag{5-7}$$

量子干涉项 Int 为：

$$Int = \sum_i \sum_j \varepsilon_i \cdot \varepsilon_j \cdot \cos(\theta_{ij}) \cdot |\langle\psi_i|M_m|\psi_j\rangle| \tag{5-8}$$

其中 $i \neq j$，参数 θ_{ij} 表示量子态之间内积运算的相位角。该相位角揭示了多源数据获得的表征之间的空间几何关系。

由于基于该多源融合表征方法的学习模型在每次训练过程中都存在这样一个

量子干涉项，其能够确保多源数据下获得的表征之间的空间位置关系，同时能够约束目标表征学习、搜索的空间，最终实现降低表征学习过程中对大数据量的依赖。具体原理在接下来的理论分析部分给出。

理论分析： 假设基于两个数据源 A 和 B 获得的相应目标（或对象）的量子态分别为 $|\psi_A\rangle$ 和 $|\psi_B\rangle$，构造的叠加态为：

$$|\Psi_{AB}\rangle = \varepsilon_A \cdot |\psi_A\rangle + \varepsilon_B \cdot |\psi_B\rangle \tag{5-9}$$

其中 ε_A，$\varepsilon_B \in \mathbb{R}$ 表示构造叠加态的概率幅，$|\varepsilon_A|^2 + |\varepsilon_B|^2 = 1$。假设测量算子 M_m 由已知的一个量子态 $|\phi\rangle$ 构建，

$$M_m = |\phi\rangle\langle\phi|, \tag{5-10}$$

则测量结果 m 的概率为：

$$\begin{aligned} Pr(m) &= \| M_m |\Psi_{AB}\rangle \|^2 \\ &= \| M_m(\varepsilon_A \cdot |\psi_A\rangle + \varepsilon_B \cdot |\psi_B\rangle) \|^2 \\ &= \| \varepsilon_A \cdot M_m|\psi_A\rangle + \varepsilon_B \cdot M_m|\psi_B\rangle \|^2 \\ &= \| \varepsilon_A \cdot M_m|\psi_A\rangle \|^2 + \| \varepsilon_B \cdot M_m|\psi_B\rangle \|^2 + Int \end{aligned} \tag{5-11}$$

其中量子干涉项 Int 为：

$$\begin{aligned} Int &= 2 \cdot \varepsilon_A \cdot \varepsilon_B \cdot \cos(\theta) \cdot |\langle\psi_A|M_m|\psi_B\rangle| \\ &= 2 \cdot \varepsilon_A \cdot \varepsilon_B \cdot \cos(\theta) \cdot |\langle\psi_A|\phi\rangle\langle\phi|\psi_B\rangle| \end{aligned} \tag{5-12}$$

其中参数 θ 是量子态之间内积运算的相位角。

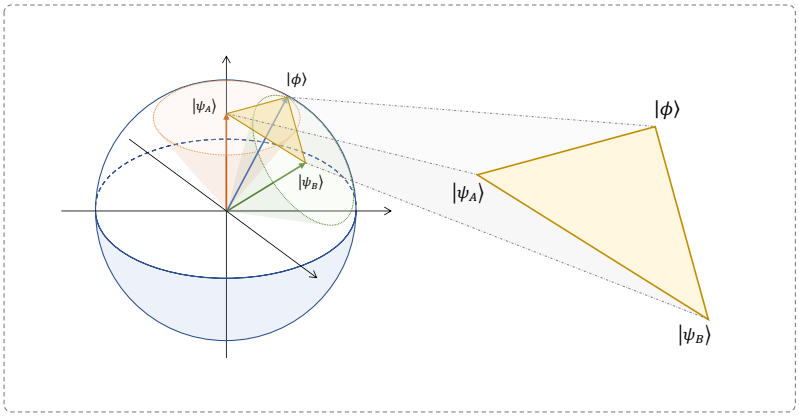

图 5-1　量子测量启发的多源融合表征原理图

量子干涉项是源自于量子力学理论的独有特征，它揭示了希尔伯特空间中量子态之间的几何关系，如图 5-1 所示。干涉项中的 ε_A 和 ε_B 是构建叠加态的概率幅，$|\langle\psi_A|\phi\rangle|$ 和 $|\langle\phi|\psi_B\rangle|$ 描述了图中三角形两条边的边长，$\cos(\theta)$ 描述了两边的夹角，它们共同描述了三角形的面积。因此干涉项本身可以被看作是对希尔伯特空间中量子态之间几何关系的刻画。

在基于上述量子力学形式语言构建的学习模型中，干涉项在模型的训练过程中会被作为损失函数中的一个约束项，约束量子态之间空间几何关系。此外，与传统的没有量子干涉项的学习模型相比，具有空间几何约束的损失函数可以限制表征在表征空间中位置，进而减小模型参数的取值范围、缓解表征学习过程中对大数据量的依赖。可以理解，描述几何关系的干涉项限制了表征在空间中的位置关系，相对于无约束的表征而言会改善表征学习的难度。

5.2.2 模型结构

基于多源融合表征方法的词义消歧模型框架结构如图 5-2 所示。词义消歧模型包括基于经典表征方法的**传统识别方法**和基于多源融合表征方法的**量子识别方法**。传统识别方法适用于高频词义消歧，而量子识别方法则更适用于长尾词义消歧。该词义消歧模型整合两种识别方法的目的在于使词义消歧模型同时兼顾到高频与长尾词义，并发挥各自的优势改善词义消歧模型的性能。传统识别方法的输入为**目标词编码器**生成的目标词嵌入和**词义定义编码器**生成的词义定义嵌入；量子识别方法的输入不仅包括上述目标词嵌入和词义定义嵌入，还包括基于词典中词义例句学习到的词义场景嵌入，其中词义场景嵌入由**词义场景编码器**生成。上述编码器都由预训练语言模型 BERT[30] 实现。

目标词编码器输入待消歧文本，输出目标词的词嵌入，即**目标词嵌入**。依照 BERT 模型的文本处理规则，会在文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入的待消歧文本为：

$$W^{text} = [w_1, w_2, ..., w_i, ...], \tag{5-13}$$

添加标记符后的文本为：

$$W^{text} = \Big[[CLS], w_1, w_2, ..., w_i, ..., [SEP]\Big], \tag{5-14}$$

其中待消歧文本包含目标词，即 $w_{target} \in W^{text}$。 BERT 模型对输入的文本进行编

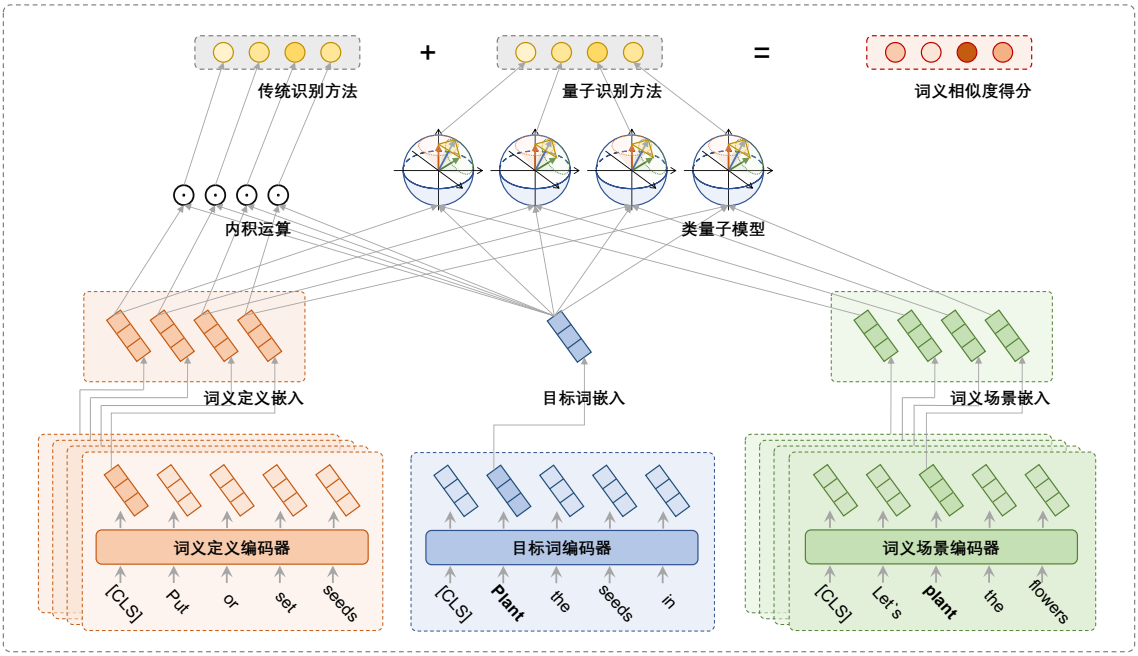

图 5-2　基于多源融合表征的词义消歧模型

码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W^{text}} = \left[V_{[CLS]}, V_{w_1}, V_{w_2}, ..., V_{w_i}, ..., V_{[SEP]}\right], \tag{5-15}$$

其中对应于目标词的词嵌入，即目标词嵌入，将被作为目标词编码器的输出，

$$V_{target} \equiv V_{w_{target}} = BERT_{TargetEncoder}(W^{text})。 \tag{5-16}$$

词义定义编码器输入词典中词义注释文本，输出词义定义的文本嵌入，即**词义定义嵌入**。BERT 模型同样会在输入中文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入的词义注释文本为：

$$W_k^{gloss} = [w_{k,1}, w_{k,2}, ..., w_{k,i}, ...], \tag{5-17}$$

添加标记符后的文本为：

$$W_k^{gloss} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., w_{k,i}, ..., [SEP]_k\right], \tag{5-18}$$

其中索引 k 表示目标词的词义清单中第 k 个词义的词义注释文本。 BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W_k^{gloss}} = \left[V_{[CLS]_k}, V_{w_{k,1}}, V_{w_{k,2}}, ..., V_{w_{k,i}}, ..., V_{[SEP]_k}\right]。 \tag{5-19}$$

BERT 模型通常将开始标记符对应的词嵌入作为该文本的文本嵌入，所以开始标记符对应的词嵌入将被作为词义定义编码器的输出，

$$V_k^{gloss} \equiv V_{[CLS]_k} = BERT_{GlossEncoder}(W_k^{gloss})。 \tag{5-20}$$

词义场景编码器输入词典中词义例句文本，输出**词义场景嵌入**。由于词义场景编码器是要获取词义适用的上下文信息，所以在 BERT 模型编码过程中会将例句中包含的目标词遮盖住，以获得其所属的场景信息。输入到 BERT 模型中的文本首先会在开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，然后将文本中的目标词使用遮盖标记符 [*MASK*] 覆盖，如输入的词义例句文本为：

$$W_k^{example} = [w_{k,1}, w_{k,2}, ..., w_{k,i}, ...], \tag{5-21}$$

添加开始与分割标记符后的文本为：

$$W_k^{example} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., w_{k,i}, ..., [SEP]_k\right]。 \tag{5-22}$$

假设单词 $w_{k,i}$ 为目标词，则添加遮盖标记符后的文本为：

$$W_k^{example} = \left[[CLS]_k, w_{k,1}, w_{k,2}, ..., [MASK]_k, ..., [SEP]_k\right]。 \tag{5-23}$$

其中索引 k 表示目标词的词义清单中第 k 个词义对应的词义例句。通常情况下，词典中的词义例句个数不止一个，当多于一个时默认选择第一个词义例句，当不存在时使用待消歧文本替代。 BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符、分割标记符和遮盖标记符），

$$V_{W_k^{example}} = \left[V_{[CLS]_k}, V_{w_{k,1}}, V_{w_{k,2}}, ..., V_{[MASK]_k}, ..., V_{[SEP]_k}\right]。 \tag{5-24}$$

由于词义场景编码器输出的要是词义适用的上下文，所以这里选择遮盖标记符对应的词嵌入作为词义场景编码器的输出，

$$V_k^{context} \equiv V_{[MASK]_k} = BERT_{ContextEncoder}(W_k^{example})。 \tag{5-25}$$

传统识别方法下的词义相似度得分： 至此，可以计算经典表征方法下词义相似度得分，

$$Score_k = V_{target} \odot V_k^{gloss} \tag{5-26}$$

其中符号 $\odot$ 表示内积运算。

量子识别方法下的词义相似度得分： 量子识别方法首先需要将输入的目标词嵌入、词义定义嵌入和词义场景嵌入分别构造为量子态；其次需要将目标词嵌

入对应的量子态构造为测量算子，词义定义嵌入和词义场景嵌入分别对应的量子态构造为叠加态；最后输出目标词的测量结果的概率值作为词义相似度得分。这里需要明确的是词义注释与词义例句可被认为是两个数据源，即多源数据，因为词义注释文本是不包含目标词的词义描述信息，而词义例句文本是包含目标词的词义应用场景信息。词义场景嵌入更像是从词义适用场景的角度给出的词义定义表征。量子识别方法下词义相似度得分的计算过程如下：

首先，利用平方和归一化函数将目标词嵌入、词义定义嵌入和词义场景嵌入分别构造为量子态，目标词嵌入 V_{target} 的量子态为：

$$|\psi_{target}\rangle = SSN(V_{target}), \tag{5-27}$$

词义定义嵌入V_k^{gloss} 的量子态为：

$$|\psi_k^{gloss}\rangle = SSN(V_k^{gloss}), \tag{5-28}$$

词义场景嵌入$V_k^{context}$ 的量子态为：

$$|\psi_k^{context}\rangle = SSN(V_k^{context})。 \tag{5-29}$$

其次，利用目标词嵌入对应的量子态 $|\psi_{target}\rangle$ 构造测量算子，

$$M_{target} = |\psi_{target}\rangle\langle\psi_{target}|, \tag{5-30}$$

利用词义定义嵌入对应的量子态 $|\psi_k^{gloss}\rangle$ 和词义场景嵌入对应的量子态 $|\psi_k^{context}\rangle$ 构造叠加态，

$$|\Psi_k^{gloss+context}\rangle = \cos(\sigma)\cdot|\psi_k^{gloss}\rangle + \sin(\sigma)\cdot|\psi_k^{context}\rangle \tag{5-31}$$

其中 $\cos(\sigma)$ 和 $\sin(\sigma)$ 表示构造叠加态的概率幅，$|\cos(\sigma)|^2 + |\sin(\sigma)|^2 = 1$。参数 σ 可借助于神经网络中的全连接层（或称为线性层）从输入的词义定义嵌入 V_k^{gloss} 和词义场景嵌入 $V_k^{context}$ 中获得，

$$\sigma = Linear\left([V_k^{gloss}, V_k^{context}]\right) \in \mathbb{R}。 \tag{5-32}$$

最后，计算目标词的测量结果的概率值，并将其作为**量子识别方法下的词义相似度得分**，

$$Q\text{-}Score_k = Pr(|\Psi_k^{gloss+context}\rangle; M_{target}) = \||M_{target}|\Psi_k^{gloss+context}\rangle\|^2 \tag{5-33}$$

其中量子干涉项 Int 为：

$$Int_k = 2\cdot\cos(\sigma)\cdot\sin(\sigma)\cdot\cos(\theta)\cdot|\langle\psi_k^{gloss}|\psi_{target}\rangle\langle\psi_{target}|\psi_k^{context}\rangle|。 \tag{5-34}$$

5.2.3 模型训练

该词义消歧模型采用表现效果最好的 Adam 优化器[105]更新模型的参数，优化器的具体设置在实验部分给出。词义消歧模型使用交叉熵损失函数评估模型，损失函数为：

$$\begin{aligned} Loss(Score^{all}, index) &= -\log\left(\frac{\exp(Score^{all}_{[index]})}{\sum_{i=1}\exp(Score^{all}_{[i]})}\right) \\ &= -Score^{all}_{[index]} + \log\sum_{i=1}\exp(Score^{all}_{[i]}) \end{aligned} \tag{5-35}$$

其中变量 *index* 表示目标词的候选词义清单列表的索引，参数 $Score^{all}$ 为：

$$Score^{all} = \left[Score^{all}_1, Score^{all}_2, ..., Score^{all}_k, ...\right]。 \tag{5-36}$$

$Score^{all}_k$ 表示目标词的候选词义清单中第 k 个词义的最终词义定义相似度得分，由传统识别方法下的词义定义相似度得分和量子识别方法下的词义定义相似度得分组成，

$$Score^{all}_k = \alpha \cdot Score_k + \beta \cdot Q\text{-}Score_k \tag{5-37}$$

其中 α 和 β 表示传统识别方法与量子识别方法的权重，具体取值在实验设置部分给出。

5.3 实验准备

数据集：本章提出的基于多源融合表征的词义消歧模型同样使用第三章的实验设置方法，即标准评估实验设置和增强评估实验设置。 **标准评估实验设置**的数据集被划分为训练集、开发集、测试集，其中训练集为 SemCor，开发集为 SemEval-07[107]（*SE7*），测试集包括 Senseval-2[108]（*SE2*）、Senseval-3[109]（*SE3*）、SemEval-13[110]（*SE13*）、SemEval-15[111]（*SE15*）以及所有测试集的整合集（*ALL*）。此外，测试集的整合集 *ALL* 又按词性划分为名词集（*N.*）、动词集（*V.*）、形容词集（*Adj.*）、副词集（*Adv.*）。 **增强评估实验设置**的数据集同样包括训练集、开发集、测试集，其中训练集包括 SemCor 和 WNGT（WordNet Gloss Tags），开发集和测试集与标准评估实验设置下的开发集和测试集保持一致。此外，需要注意的是目标词的候选词义清单为电子词典 WordNet 3.0 中给出的目标词的所有词义。评估实验所使用的评估指标为百分比下的 F1 值，即 F1-score (%)。

实验设置：本章提出的词义消歧模型同样运行在第三章描述的硬件和软件平台上。基于多源融合表征的词义消歧模型被命名为 MF-WSD，并且基于预训练语言模型 BERT 的 *bert-base-uncased* 版本的词义消歧模型被命名为 $\text{MF-WSD}_{\text{Base}}$，基于 *bert-large-uncased* 版本的词义消歧模型被命名为 $\text{MF-WSD}_{\text{Large}}$。词义消歧模型的超参数分别为：迭代次数（*epoch*）为 20、批次大小（*batch-size*）为 4、学习率（*learning-rate*）为 $\{1E\text{-}5, 5E\text{-}5, 1E\text{-}6, 5E\text{-}6\}$、待消歧文本的最大长度（*context-maximum-length*）为 128、注释文本的最大长度（*gloss-maximum-length*）为 32、例句文本的最大长度（*example-maximum-length*）为 32。其它未列出的超参数可参见章节对应发表论文所公布的代码。

对比模型：MF-WSD 模型的对比模型分为标准实验对比模型和增强实验对比模型两部分。**标准实验对比模型**都采用预训练语言模型 BERT 的 *bert-base-uncased* 版本构建编码器，标准实验对比模型部分包括对比模型 EWISE[122]、LMMS[113]、SREF[114]、ARES[115]、BEM[104]、EWISER[123]、SyntagRank[116]、COF[117]、ESR[118]、Z-Reweighting[132]、QWSD[127]、DR-WSD[129]。**增强实验对比模型**则采用预训练语言模型 BERT 的 *bert-large-uncased* 版本构建编码器，增强实验对比模型部分包括对比模型 SparseLMMS[130]、EWISER[123]、ESR[118]。对比模型研究内容的简介参见第三章的对比模型部分，此外上述模型的实验结果均取自原论文公布的数据。

5.4 实验结果与分析

本节对基于多源融合表征构造的词义消歧模型 MF-WSD 进行实验分析，分析实验涉及模型的标准评估实验、消融分析实验和长尾分析实验三部分。标准评估实验侧重于分析模型的整体表现，并不区分长尾词义与高频词义；消融分析实验侧重于评估多源融合表征方法的有效性以及对模型整体表现的贡献；长尾分析实验则专注于分析词义消歧模型和多源融合表征方法对长尾词义识别的效用。

5.4.1 标准评估实验

MF-WSD 模型的实验结果如表 5-1 所示，对比模型分为标准实验对比模型组和增强实验对比模型组，标准实验对比模型组包含了量子启发的对比模型，其中与 $\text{MF-WSD}_{\text{Base}}$ 相比表现最佳的结果添加下划线，与 $\text{MF-WSD}_{\text{Large}}$ 相比表现最佳

的结果采用粗体显示。

表 5-1　MF-WSD 模型的标准评估实验结果

模型	开发集	测试集				测试集的整合集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*N.*	*V.*	*Adj.*	*Adv.*	*ALL*
标准实验对比模型：										
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
LMMS	68.1	76.3	75.6	75.1	77.0	–	–	–	–	75.4
SREF	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
ARES	71.0	78.0	77.1	77.3	**83.2**	80.6	68.3	80.5	83.5	77.9
BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	**87.9**	79.0
EWISER	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
SyntagRank	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR	**75.4**	80.6	78.2	79.8	82.8	82.5	69.5	82.5	87.3	79.8
Z-Reweighting	71.9	79.6	76.5	78.9	82.5	–	–	–	–	78.6
量子启发的模型：										
QWSD	-	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
DR-WSD	74.7	80.8	78.0	80.0	82.7	82.7	69.5	82.9	86.6	80.4
MF-WSD$_{\text{Base}}$	74.8	81.0	79.3	80.8	82.7	83.7	71.5	82.8	87.6	80.8
MF-WSD$_{\text{Large}}$	75.2	**82.5**	**80.5**	**81.2**	**83.2**	**84.2**	**71.7**	**83.0**	87.7	**81.8**
增强实验对比模型：										
SparseLMMS	73.0	79.6	77.3	79.4	81.3	–	–	–	–	78.8
EWISER	75.2	80.8	79.0	80.7	81.8	81.7	66.3	81.2	85.8	80.1
ESR	**77.4**	81.4	78.0	**81.5**	**83.9**	83.1	71.1	**83.6**	87.5	80.7
MF-WSD$_{\text{Base}}$	75.0	81.8	79.5	81.0	83.0	84.0	72.2	82.8	87.6	81.1
MF-WSD$_{\text{Large}}$	75.4	**83.3**	**80.8**	**81.5**	**83.9**	**85.3**	**73.0**	83.0	**87.7**	**82.1**

与标准实验对比模型相比，MF-WSD$_{\text{Base}}$ 与 MF-WSD$_{\text{Large}}$ 在多个测试集上都优于对比模型，其佐证了多源融合表征的有效性与竞争力。详细对比如下：

- 与量子启发的模型 QWSD 和 DR-WSD 相比，MF-WSD 远优于 QWSD、优于 DR-WSD，其原因在于 QWSD 没有使用可学习的表征学习模型，而 MF-WSD 与 DR-WSD 都使用了当前主流的预训练语言模型去学习表征，而且 MF-WSD 要比 DR-WSD 多使用了词典中提供的词义例句资源。
- 在开发集 *SE7* 上，MF-WSD 不及对比模型 ESR，其原因在于模型 ESR 具有远超平均水平的拟合能力。同时需要说明的是学习模型过度拟合训练集与开发集将影响模型的泛化水平。
- 在测试集 *Adj.* 与 *Adv.* 上，MF-WSD$_{\text{Base}}$ 不及对比模型 BEM，在测试集 *SE15* 上，MF-WSD$_{\text{Base}}$ 不及对比模型 ARES。测试集 *Adj.* 与 *Adv.* 分别为形容词

集与副词集，相比于名词集和动词集而言，形容词集和副词集中的高频词义的占比较大，而长尾词义的占比较小。由于上述测试集是高频词义占主导的测试，并不适合于评估我们的词义消歧模型的优势。同理，在测试集 *SE15* 也是一个高频词义占主导的数据集，所以我们的模型难以发挥其优势。此外，该实验结果也说明多源融合表征方法对于高频词义占主导的数据而言并不具有优势。

与增强实验对比模型相比，MF-WSD 在开发集 *SE7* 上不及对比模型 ESR，在测试集 *SE13* 与 *SE15* 上，MF-WSD$_{\text{Base}}$ 不及模型 ESR，在测试集 *Adj.* 上，MF-WSD 不及模型 ESR，该结果说明模型 ESR 要远超 MF-WSD。但在重要的测试集 *ALL* 上，MF-WSD$_{\text{Base}}$ 和 MF-WSD$_{\text{Large}}$ 都优于对比模型，包括模型 ESR，说明多源融合表征方法对于高频、长尾词义识别都有促进作用。

5.4.2 消融分析实验

为检验多源融合表征方法对词义消歧模型整体表现的贡献，我们采用删除构件的消融研究方法实施验证，删除多源融合表征方法后的模型被称为**消融模型**。对比模型为：

- 原模型 MF-WSD$_{\text{Base}}$，原模型采用基于预训练语言模型 BERT 的 *bert-base-uncased* 版本实现的模型。
- 消融模型 MF-WSD$^{ablation}_{\text{Base}}$，消融模型是在原模型的基础之上删除多源融合表征方法部分后的模型，即删除掉量子识别方法部分只保留传统识别方法部分。

消融模型的参数设置以及运行平台与标准评估实验部分的保持一致。

表 5-2 MF-WSD 模型的消融分析实验结果

模型	开发集	测试集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：SemCor						
MF-WSD$_{\text{Base}}$	**74.8**	**81.0**	**79.3**	**80.8**	**82.7**	**80.8**
MF-WSD$^{ablation}_{\text{Base}}$	71.1	77.7	75.3	76.3	78.0	77.7

实验结果与分析：实验结果如表 5-2 所示。从整体表现来看，原模型 MF-WSD$_{\text{Base}}$ 优于对比模型，说明多源融合表征方法对词义消歧模型的整体表征有着重要的贡献，同时也说明添加量子识别方法部分是有价值的。对比消融模型 MF-

$\text{WSD}_{\text{Base}}^{ablation}$ 和原模型 MF-WSD_{Base}，原模型 MF-WSD_{Base} 在测试集 *SE2*、*SE3*、*SE13* 和 *SE15* 上的实验结果相对于消融模型 MF-$\text{WSD}_{\text{Base}}^{ablation}$ 都有 4 个百分点左右的提高，提升幅度相似，说明多源融合表征方法对于高频词义识别和长尾词义识别都是有价值的。由于长尾词义在四个测试集中的占比不同，词义消歧模型在四个测试集上的实验结果有相似的提升幅度说明多源融合表征方法对长尾词义的识别能力同样出色。

5.4.3 长尾分析实验

为检验多源融合表征方法在长尾词义消歧任务中的有效性，我们分别在构造的长尾词义数据集（Long-tail Sense Datasets，LTS）和最新的交叉语种数据集（Cross-lingual Datasets，CL）上实施实验。长尾词义数据集的构造方法参见章节 3.4.3，交叉语种数据集的简介同样参见章节 3.4.3。

对比模型： 长尾词义数据集上的对比模型采用消融分析实验部分构造的实验模型，即原模型 MF-WSD_{Base}、消融模型 MF-$\text{WSD}_{\text{Base}}^{ablation}$。交叉语种数据集上的对比模型采用消融分析实验部分的原模型 MF-WSD_{Base} 和消融模型 MF-$\text{WSD}_{\text{Base}}^{ablation}$ 以及原论文中使用的模型 XLMR-Base[134] 作为实验模型。此外，由于电子词典于 BabelNet 中的大多数小语种都使用英语的词义定义，因此我们直接使用英语中的词义注释来提供目标词的候选词义列表。

模型设置：原模型的实验设置与标准评估实验部分的保持一致，消融模型的实验设置与消融分析实验部分的保持一致；此外，两个对比模型的硬件运行平台和软件开发框架与标准评估实验部分的保持一致。模型 XLMR-Base 的实验结果来自于跨语言评估框架中公布的实验数据。

表 5-3 MF-WSD 模型在长尾词义数据集下的实验结果

模型	开发集	测试集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：LTS						
MF-WSD_{Base}	**51.0**	**52.3**	**48.6**	**50.1**	**51.5**	**49.3**
MF-$\text{WSD}_{\text{Base}}^{ablation}$	33.3	37.7	35.0	35.9	37.6	34.9

实验结果与分析： 长尾词义数据集上的实验结果如表 5-3 所示。从整体表现来看，原模型优于消融模型，说明当前基于多源融合表征方法的词义消歧模型的设计是合理的。对比 MF-WSD_{Base} 和 MF-$\text{WSD}_{\text{Base}}^{ablation}$，原模型优于消融模型，说明

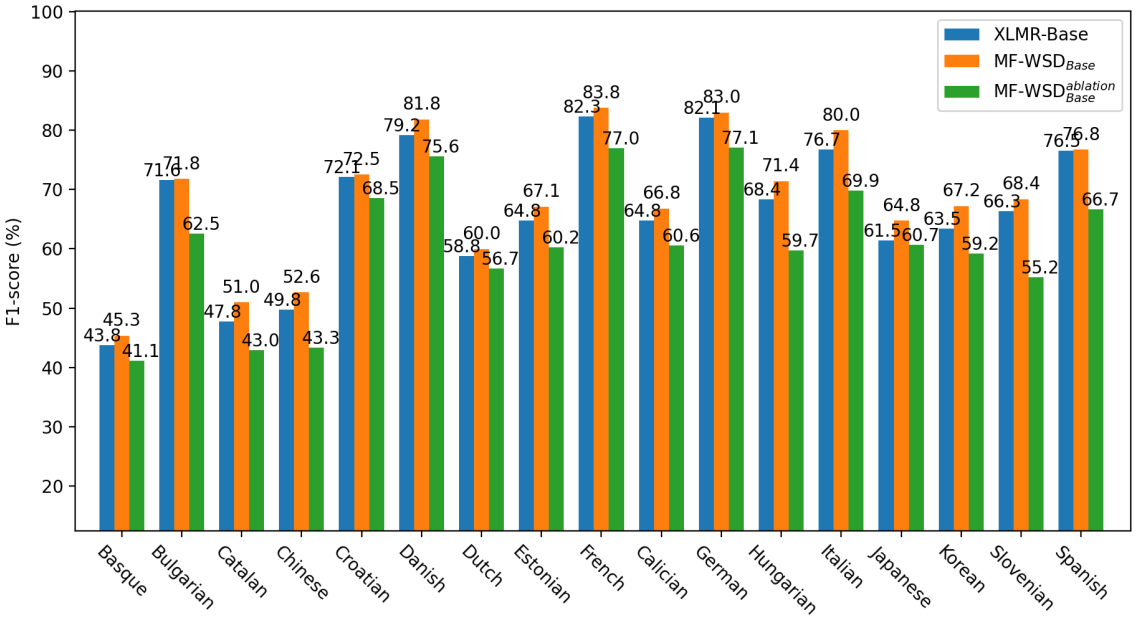

图 5-3 MF-WSD 模型在交叉语种数据集下的实验结果

多源融合表征方法有助于改善词义消歧模型的识别能力，同时也说明多源融合表征方法有助于长尾词义的识别。

交叉语种数据集上的实验结果如图 5-3 所示。对比 MF-WSD$_{\text{Base}}$ 和 XLMR-Base，原模型在多个数据集上优于模型 XLMR-Base，这说明我们的模型具有一定的鲁棒性，也说明多源融合表征方法具有广泛的适用性。对比 MF-WSD$_{\text{Base}}$ 和 MF-WSD$_{\text{Base}}^{ablation}$，原模型优于消融模型，这说明多源融合表征方法确实发挥了作用，间接指出多源融合表征方法有助于改善低资源任务中的表征学习问题。在 Basque、Croatian、Dutch、Japanese 等语种上，我们的模型表现不佳，核心原因是上述语种的数据集体量比较小，而且大部分都是高频词义。

5.5 本章小结

本章针对词义注释文本的低资源问题，提出量子测量启发的多源融合表征，用以实现通过整合（多源）外部数据的方式解决词义定义表征之间的不可区分性问题。该表征方法借助叠加态的构造过程融合多源数据下获得的词义表征向量，借用叠加态的测量过程所导出的干涉项实现在语义空间中约束词义之间的位置关系以及构建词义之间的内在联系，最终实现增强词义定义表征之间的区分边界。基于该表征方法构造词义消歧模型并实施验证，实验结果证实多源融合表征能够

有效改善词义定义表征之间的不可区分性问题。此外，为了更进一步明确多源融合表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实多源融合表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果投稿在IJCAI 会议论文中。

对于长尾词义消歧任务整体而言，前二章节提出的两个表征方法有利于改善目标词表征与词义定义表征面临的表征学习模型的大数据量依赖问题和低资源下文本表征的不准确性问题，本章提出的表征方法有利于改善低资源下词义定义表征之间的不可区分性问题，下一章将面向上述表征方法提出量子干涉实验启发的多元词义消歧模型，用以整合各个表征方法的优势一致性地应对低资源下的长尾词义消歧。

第6章　量子干涉实验启发的多元词义消歧模型

在第三、四章中针对目标词表征与词义定义表征面临的表征学习模型的大数据量依赖问题和文本表征的不准确性问题分别提出解耦表征和自适应调优表征，在第五章中针对低资源下词义定义表征之间的不可区分性问题提出多源融合表征。本章将面向上述表征方法提出多元词义消歧模型，用以整合上述表征方法的优势一致性地应对低资源下的长尾词义消歧任务，最后对词义消歧模型实施标准评估实验、消融分析实验、长尾分析实验、以及各个章节中提出的词义消歧模型间的横向对比分析。

6.1　研究动机

对于词义消歧系统而言，很难将高频（头部、常见）词义与长尾（低频、罕见）词义实施针对性的词义消歧，其原因在于目标词在没有被识别之前难以推测出可能的类型，所以词义消歧系统常常将高频词义与长尾词义进行混合识别。对于并没有针对长尾词义给出应对方案的词义消歧系统来说，长尾词义相对于高频词义而言在数量上处于弱势地位，词义消歧系统会趋向于忽视长尾词义。

对此，研究者们的应对思路大致可以归纳为如下三种方案：（1）数据增强的方法，该方法尝试挖掘更多的词义注释文本或有利于改善词义定义表征的数据，进而对长尾词义定义表征实施增强；（2）知识迁移的方法，该方法尝试借助于其它词义场景中的语义知识改善长尾词义定义表征的准确性；（3）词汇知识图谱的方法，该方法放弃掉构造词义定义表征，转向使用同样具备词义识别能力的词汇知识图谱，尝试公平的对待高频与长尾词义。数据增强方法中最具新颖性的工作是 Calabrese 等人[44]提出的利用 BabelPic 数据集中的图像信息增强词义定义表征，其尝试借助于多模态知识改善表征向量的准确性。此外，Ruas 等人[46]采用多词义定义表征的方式改善单表征方法下词义定义表征不准确的问题；Song 等人[48]则使用了更多的有助于改善词义定义表征的数据，如词义例句、同义词、上下位词等。并且上述方案都不约而同地使用了预训练语言模型去获取词义定义表征，以实现

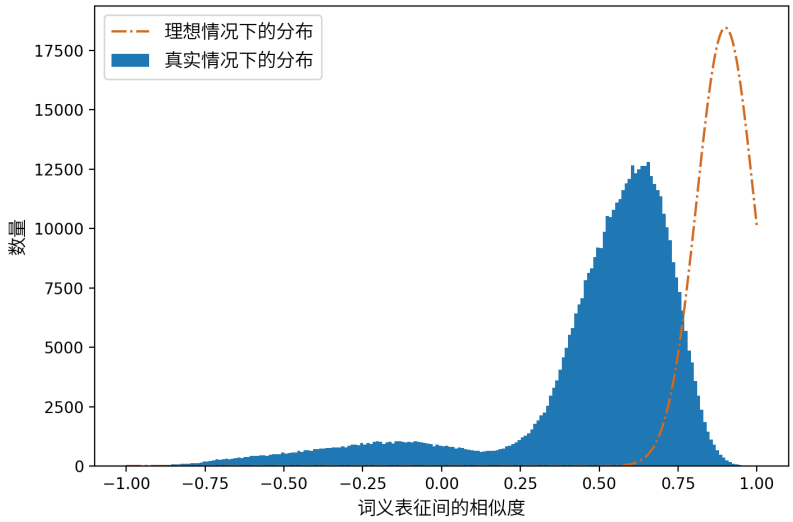

图 6-1　同一词义不同注释文本间词义表征的相似度统计

尽可能多地借助到外部知识获得上下文化的表征。知识迁移方法中最具影响力的工作多借助于小样本学习、元学习、迁移学习领域中的研究成果应对长尾词义定义表征。 Kohli 等人[51]使用迁移学习和数据增强技术尝试在不同的自然语言处理任务和数据集上进行广泛的预训练以实现借助到更多的外部知识。 Hu 等人[52]采用生成对抗技术尝试更大程度地利用、挖掘现有数据以改善词义消歧系统的性能。Holla 等人[53]和 Du 等人[54]采用元学习技术尝试通过在大量常用词义消歧任务上训练模型以实现在长尾词义消歧任务中改善词义识别的准确性。 Chen 等人[55]采用小样本学习技术尝试采用无参数的小样本学习方法应对长尾词义消歧。知识图谱方法中最具代表性的工作是 Scozzafava 等人[28]提出的 SyntagRank 模型，其借助个性化的 PageRank 算法挖掘电子词典 WordNet 中的词汇知识图谱实施词义消歧，并借用 WNG 和 SyntagNet 语料库中的词汇关系扩展电子词典 WordNet 的词汇知识图谱。此外，Loureiro 等人[33]通过对同义词表征取均值以实现对未知词义创建表征；Wang 等人[29]利用上下位词设计了一种校验机制以实现对系统的预测结果进行校准；Vial 等人[42]通过上位词寻找根源词以减少输出类别的数量。

本章尝试从多个电子词典中挖掘更多的词义注释文本，用以增强包括长尾词义在内的词义训练数据。对获得的词义注释文本进行统计分析发现，不同词典对同一词义给出的注释文本间存在着较大的语义偏差，如图 6-1 所示。在该图示中，我们使用预训练语言模型 BERT[30] 对同一词义的多个注释文本进行向量化，并计算多个词义定义表征之间相似度。统计结果揭示，同一词义的多个注释文本间的相似度分布与理想情况下的相似度分布存在着较大的偏差。基于此，将所有的同

一词义的注释文本都放到同一个词义定义表征学习模型中进行无差别的训练，并不能够有效改善词义定义表征的准确性。

具体地，本章基于双缝干涉实验的数学模型导出多元词义消歧模型，即双缝干涉模型的泛化形式，用以实现目标词的多元词义识别与校验，最终实现词义消歧系统能够尽可能公平的对待高频与长尾词义。此外，提出多元词义消歧模型的目的在于借助于自适应调优表征、多源融合表征从同一词义的多个词义注释文本中获得同一词义的多个词义定义表征，进而实现对目标词的同一词义实施多元词义识别与校验，即整合多路词义表征的多元词义消歧模型。本章的贡献在于：

- 针对词义消歧系统难以公平对待长尾词义的问题，提出量子双缝干涉实验启发的多元词义消歧模型，以实现整合多种低资源下的表征方法增强长尾词义定义表征，最终实现借助于多元词义识别与校验的方式改善词义消歧过程中长尾词义的弱势地位；
- 对整合多路词义表征的多元词义消歧模型实施验证，实验结果证实多元词义消歧模型具备整合多种表征学习方法的能力，此外，整合多路词义表征的多元词义消歧模型有效改善了长尾词义的识别能力。

6.2 整合多路词义表征的多元词义消歧模型

本节首先介绍量子干涉实验启发的多元词义消歧模型，然后整合前几章中提出的表征方法获得词义定义表征，进而构造整合多路词义表征的多元词义消歧模型，最后给出词义消歧模型的训练方法。

6.2.1 多元词义消歧模型

量子干涉效应是最有趣的量子现象之一，在物理学领域中有着重要的研究价值。双缝干涉实验（简称干涉实验）是揭示量子干涉现象的重要手段，同时也被用于揭示微观粒子的波粒二相性等重要物理现象，实验设置如图 6-2 所示。 电子发射抢一个一个地发射电子，电子穿过双缝幕布的两个狭缝，如狭缝 A 和 B，并射向检测屏幕。电子落在检测屏幕上的具体位置被标记为 x。当关闭两个狭缝之一时，如狭缝 B，可以计算电子通过狭缝 A 并落在位置 x 的概率，

$$Pr_A(x) = ||M_x|\psi_A\rangle||^2 \tag{6-1}$$

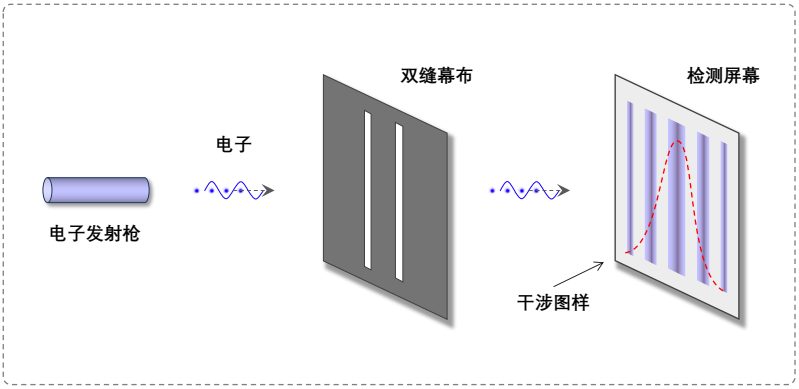

图 6-2　　量子双缝干涉实验的示意图

其中 M_x 表示对应于位置 x 的测量算子，$|\psi_A\rangle$ 表示通过狭缝 A 的电子。相应地，当缝隙 A 关闭时，电子通过缝隙 B 并落在位置 x 的概率为:

$$Pr_B(x) = \||M_x|\psi_B\rangle\|^2 \tag{6-2}$$

$|\psi_B\rangle$ 表示通过狭缝 B 的电子。基于经典概率理论，两个狭缝都打开并且电子通过狭缝 A 或 B 并落在位置 x 的概率为：

$$Pr_{AB}(x) = Pr_A(x) + Pr_B(x) = \||M_x|\psi_A\rangle\|^2 + \||M_x|\psi_B\rangle\|^2。 \tag{6-3}$$

但实际测得的实验结果表明，干涉实验的结果并不能用经典概率理论来描述，即公式 (6-3) 不成立。下面基于量子概率理论导出符合干涉实验现象的量子干涉模型，并给出其广义形式，即**量子干涉实验启发的多元词义消歧模型**。

干涉实验表明微观粒子（包括电子）可以以叠加的形式存在，即在双缝干涉实验中，一个电子同时通过狭缝 A 和 B 并发生彼此干涉现象。在量子力学中，以叠加形式存在的系统可以由叠加态描述，因此被测系统可以被定义为：

$$|\Psi_{AB}\rangle = \varepsilon_A \cdot |\psi_A\rangle + \varepsilon_B \cdot |\psi_B\rangle \tag{6-4}$$

其中 ε_A，$\varepsilon_B \in \mathbb{C}$ 表示概率幅，$|\varepsilon_A|^2 + |\varepsilon_B|^2 = 1$。检测屏幕落点 x 的测量算子记为 M_x 时，则符合双缝干涉实验现象的**量子干涉模型**可以被定义为：

$$\begin{aligned} Pr_{AB}(x) =& \||M_x|\Psi_{AB}\rangle\|^2 \\ =& \||M_x(\varepsilon_A \cdot |\psi_A\rangle + \varepsilon_B \cdot |\psi_B\rangle)\|^2 \\ =& \||\varepsilon_A \cdot M_x|\psi_A\rangle + \varepsilon_B \cdot M_x|\psi_B\rangle\|^2 \\ =& \||\varepsilon_A \cdot P_x|\psi_A\rangle\|^2 + \||\varepsilon_B \cdot M_x|\psi_B\rangle\|^2 + 2 \cdot \varepsilon_A \cdot \varepsilon_B \cdot \cos(\theta) \cdot |\langle\psi_A|M_x|\psi_B\rangle| \end{aligned} \tag{6-5}$$

其中参数 θ 表示内积运算 $\langle\psi_A|M_x|\psi_B\rangle$ 的相位角，$2\cdot\varepsilon_A\cdot\varepsilon_B\cdot\cos(\theta)\cdot|\langle\psi_A|M_x|\psi_B\rangle|$ 描述量子叠加系统自身的干涉，即量子干涉项。

量子干涉模型的广义形式，即**量子干涉实验启发的多元词义消歧模型**，可以被定义为：

$$Pr(x)=\sum_{i}^{n}\|\varepsilon_i\cdot M_x|\psi_i\rangle\|^2+\sum_{i,j;i\neq j}^{i=j=n}Int_{i,j}(x) \tag{6-6}$$

其中 n 表示构造叠加态的量子态个数，$|\psi_i\rangle$ 表示构造叠加态的量子态，ε_i 表示构造叠加态的概率幅，

$$\sum_{i,j;i\neq j}^{i=j=n}Int_{i,j}(x)=\varepsilon_i\cdot\varepsilon_j\cdot\cos(\theta)\cdot|\langle\psi_i|M_x|\psi_j\rangle| \tag{6-7}$$

表示叠加系统自身的干涉项。在词义消歧任务中，$|\psi_i\rangle$ 可由目标词的某一词义的一种词义定义表征构建，表示量子态形式下的词义定义表征，并且 $|\psi_i\rangle$ 可以由多种表征学习方法获得；M_x 可由目标词表征构建，表示目标词的词义测量算子。

6.2.2 模型结构

整合多路词义表征的多元词义消歧模型框架结构如图 6-3 所示。双缝干涉实验本身就是一个循环系统，其通过反复发射电子来统计电子落点的概率分布。机器学习模型同样通过反复迭代训练数据并学习输出结果的概率分布。基于此，词义消歧模型可以仿照量子干涉实验的设置方式进行构造，其优势在于更容易理解，而且符合量子干涉实验的构件组织过程。

电子：电子发射枪发射出来的电子可能处于任意状态下的叠加态，穿过带缝隙的幕布后，电子被调整为与缝隙状态相关的叠加态。在词义消歧任务中，由同一词义的多个词义定义表征（即多元词义表征）构成的叠加态就可以对应于穿过幕布后处于叠加状态下的电子。假设目标词有 K 个词义，每个词义有 M 个来自于不同词典的词义注释文本，并且分别输入到各自的**词义定义编码器**中获取词义注释的文本表征。词义定义编码器由预训练语言模型 BERT[30] 实现，编码器的输入为词义注释文本，输出为词义定义的文本嵌入，即**词义定义嵌入**。按照 BERT 模型的文本处理规则会在输入中文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入编码器的目标词的第 k 个词义的第 m 个词义注释文本为：

$$W_{k,m}^{gloss}=[w_{k,m,1},w_{k,m,2},...,w_{k,m,i},...], \tag{6-8}$$

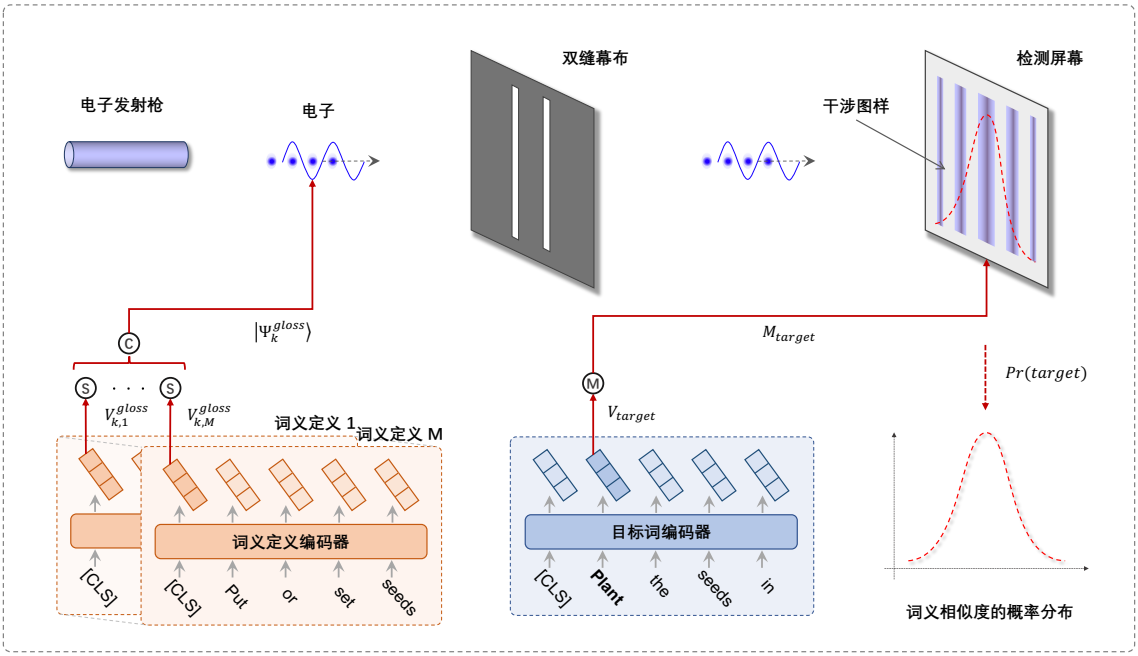

图 6-3　　整合多路词义表征的多元词义消歧模型

添加标记符后的文本为：

$$W_{k,m}^{gloss} = \left[[CLS]_{k,m}, w_{k,m,1}, w_{k,m,2}, ..., w_{k,m,i}, ..., [SEP]_{k,m}\right], \tag{6-9}$$

其中索引 k 表示目标词的词义清单中第 k 个词义，索引 m 表示第 k 个词义的第 m 个词义注释文本，索引 i 表示该词义注释文本的第 i 个单词。BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W_{k,m}^{gloss}} = \left[V_{[CLS]_{k,m}}, V_{w_{k,m,1}}, V_{w_{k,m,2}}, ..., V_{w_{k,m,i}}, ..., V_{[SEP]_{k,m}}\right]。 \tag{6-10}$$

BERT 模型通常将开始标记符对应的词嵌入作为该文本的文本嵌入，所以开始标记符对应的词嵌入将被作为词义定义编码器的输出，

$$V_{k,m}^{gloss} \equiv V_{[CLS]_{k,m}} = BERT_{GlossEncoder}^{m}(W_{k,m}^{gloss})。 \tag{6-11}$$

接着，采用第五章节提出的**多源融合表征方法**将目标词嵌入们构造为叠加态：首先，利用平方和归一化函数将词义定义嵌入构造为量子态，词义定义嵌入 $V_{k,m}^{gloss}$ 的量子态形式为：

$$|\psi_{k,m}^{gloss}\rangle = SSN(V_{k,m}^{gloss}); \tag{6-12}$$

此外，部分的词义定义嵌入 $V_{k,m}^{gloss}$ 的量子态形式为可由第四章节提出的**自适应调**

优表征方法获得，

$$|\psi_{k,m}^{gloss}\rangle = \mathcal{F}(V_{k,m}^{gloss}) \tag{6-13}$$

其中 $\mathcal{F}(\cdot)$ 表示自适应调优表征方法，如公式 (4-8) 所示；其次，将词义 k 的 M 个词义定义嵌入对应的量子态 $|\psi_{k,m}^{gloss}\rangle$ 构造为叠加态，

$$|\Psi_{k}^{gloss}\rangle = \sum_{m=1}^{M} \varepsilon_{k,m} \cdot |\psi_{k,m}^{gloss}\rangle \tag{6-14}$$

其中 $\varepsilon_{k,m}$ 表示构造叠加态的概率幅，$\sum_{m=1}^{M} |\varepsilon_{k,m}|^2 = 1$。 $\varepsilon_{k,m}$ 可借助于神经网络中的全连接层 $Linear(\cdot)$ 从输入的词义定义嵌入 $V_{k,m}^{gloss}$ 中获得，

$$[\varepsilon_{k,1}, \varepsilon_{k,2}, ..., \varepsilon_{k,m}, ...] = SSN\left(Linear\left([V_{k,1}^{gloss}, V_{k,2}^{gloss}, ..., V_{k,m}^{gloss}, ...]\right)\right) \in \mathbb{R}^{M}。 \tag{6-15}$$

落点的测量算子：检测屏幕落点 x 的测量算子记为 M_x，可由待消歧文本中的目标词嵌入获得，其中待消歧文本的编码器称为**目标词编码器**，由预训练语言模型 BERT[30] 实现。目标词编码器输入待消歧文本，输出目标词的词嵌入，即**目标词嵌入**。依照 BERT 模型的文本处理规则，会在文本的开头与分割处分别添加开始标记符 [*CLS*] 和分割标记符 [*SEP*]，如输入的待消歧文本为：

$$W^{text} = [w_1, w_2, ..., w_i, ...], \tag{6-16}$$

添加标记符后的文本为：

$$W^{text} = \left[[CLS], w_1, w_2, ..., w_i, ..., [SEP]\right], \tag{6-17}$$

其中待消歧文本包含目标词，即 $w_{target} \in W^{text}$。 BERT 模型对输入的文本进行编码，并生成每个单词对应的词嵌入（包括开始标记符和分割标记符），

$$V_{W^{text}} = \left[V_{[CLS]}, V_{w_1}, V_{w_2}, ..., V_{w_i}, ..., V_{[SEP]}\right], \tag{6-18}$$

其中对应于目标词的词嵌入，即目标词嵌入，将被作为目标词编码器的输出，

$$V_{target} \equiv V_{w_{target}} = BERT_{TargetEncoder}(W^{text})。 \tag{6-19}$$

接着，采用第四章节提出的**自适应调优表征方法**将目标词嵌入构造为量子态，

$$|\Psi_{target}\rangle = \mathcal{F}(V_{target}) \tag{6-20}$$

其中 $\mathcal{F}(\cdot)$ 表示自适应调优表征方法，如公式 (4-8) 所示；再在该量子态的基础上将

其构造为测量算子

$$M_{target} = |\Psi_{target}\rangle\langle\Psi_{target}|。 \tag{6-21}$$

最后，落点 x 的概率也就是目标词的第 k 个词义的**词义相似度得分**，计算公式为：

$$Score_k = Pr(target) = Pr(|\Psi_k^{gloss}\rangle; M_{target})。 \tag{6-22}$$

6.2.3 模型训练

该词义消歧模型采用表现效果最好的 Adam 优化器[105]更新模型的参数，优化器的具体设置在实验部分给出。词义消歧模型使用交叉熵损失函数评估模型，损失函数为：

$$\begin{aligned} Loss(Score, index) &= -\log\left(\frac{\exp(Score_{[index]})}{\sum_{i=1}\exp(Score_{[i]})}\right) \\ &= -Score_{[index]} + \log\sum_{i=1}\exp(Score_{[i]}) \end{aligned} \tag{6-23}$$

其中参数 $Score$ 为：

$$Score = [Score_1, Score_2, ..., Score_k, ...]\,, \tag{6-24}$$

变量 $index$ 表示目标词的候选词义清单列表的索引。

6.3 实验准备

数据集：本章提出的整合多路词义表征的多元词义消歧模型同样使用第三章节的实验设置方法，即标准评估实验设置和增强评估实验设置。**标准评估实验设置**的数据集被划分为训练集、开发集、测试集，其中训练集为 SemCor，开发集为 SemEval-07[107]（*SE7*），测试集包括 Senseval-2[108]（*SE2*）、Senseval-3[109]（*SE3*）、SemEval-13[110]（*SE13*）、SemEval-15[111]（*SE15*）以及所有测试集的整合集（*ALL*）。此外，测试集的整合集 *ALL* 又按词性划分为名词集（*N.*）、动词集（*V.*）、形容词集（*Adj.*）、副词集（*Adv.*）。**增强评估实验设置**的数据集同样包括训练集、开发集、测试集，其中训练集包括 SemCor 和 WNGT（WordNet Gloss Tags），开发集和测试集与标准评估实验设置下的开发集和测试集保持一致。此外，需要注意的是目标词的候选词义清单为电子词典 WordNet 3.0 中给出的目标词

的所有词义，词义注解则均来自于电子词典 BabelNet [1]。评估实验所使用的评估指标为百分比下的 F1 值，即 F1-score (%)。

实验设置：本章提出的词义消歧模型同样运行在第三章描述的硬件和软件平台上。基于多词义识别框架的词义消歧模型被命名为 MS-WSD，并且基于预训练语言模型 BERT 的 *bert-base-uncased* 版本的词义消歧模型被命名为 $\text{MS-WSD}_{\text{Base}}$，基于 *bert-large-uncased* 版本的词义消歧模型被命名为 $\text{MS-WSD}_{\text{Large}}$。词义消歧模型的超参数分别为：迭代次数（*epoch*）为 20、批次大小（*batch-size*）为 4、学习率（*learning-rate*）为 $\{1E\text{-}5, 5E\text{-}5, 1E\text{-}6, 5E\text{-}6\}$、待消歧文本的最大长度（*context-maximum-length*）为 128、注释文本的最大长度（*gloss-maximum-length*）为 32。其它未列出的超参数可参见章节对应发表论文所公布的代码。

对比模型：MS-WSD 模型的对比模型分为**前人工作**和**基线模型**两部分。前人工作部分包括一些主流的、有代表性的词义消歧模型，但由于模型设计方式上的差别，对比结果并不能真实反应模型的好坏，仅作参考使用。基线模型部分包括一些与我们的工作相似的词义消歧模型，该部分的实验结果将反应模型的真实水平。其中前人工作部分包括对比模型 HCAN[128]、GAS[126]、GLU[112]、LMMS[113]、SREF[114]、SyntagRank[116]、COF[117]、ESR[118]、Z-Reweighting[132]。基线模型部分包括对比模型 GlossBERT[120]、BEM[104]、EWISE[122]、SensEmBERT[131]、ARES[115]、QWSD[127]。对比模型研究内容的简介参见第三章的对比模型部分，此外上述模型的实验结果均取自原论文公布的数据。

6.4 实验结果与分析

本节对整合多路词义表征的多元词义消歧模型 MS-WSD 进行实验分析，分析实验涉及词义消歧模型的标准评估实验、消融分析实验和长尾分析实验三部分。标准评估实验侧重于分析模型的整体表现，并不区分长尾词义与高频词义；消融分析实验侧重于评估多元词义消歧模型的有效性以及整合其它表征方法的能力；长尾分析实验则专注于分析该词义消歧模型以及整合的表征方法对长尾词义消歧的效用。

[1] https://babelnet.org/

6.4.1 标准评估实验

MS-WSD 模型的实验结果如表 6-1 所示，对比模型分为前人工作对比组和基线模型对比组，基线模型对比组包含了量子启发的对比模型，其中与 MS-WSD$_{\text{Base}}$ 相比表现最佳的结果采用粗体显示，与 MS-WSD$_{\text{Large}}$ 相比表现最佳的结果添加下划线。

表 6-1　MS-WSD 模型的标准评估实验结果

模型	开发集	测试集				测试集的整合集				
	SE7	*SE2*	*SE3*	*SE13*	*SE15*	*N.*	*V.*	*Adj.*	*Adv.*	*ALL*
前人工作：										
HCAN	–	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
GAS	–	72.2	70.5	67.2	72.6	72.2	57.7	76.6	85.0	70.6
GLU	68.1	75.5	73.6	71.1	76.2	–	–	–	–	74.1
LMMS	68.1	76.3	75.6	75.1	77.0	–	–	–	–	75.4
SREF	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
SyntagRank	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR	**75.4**	80.6	78.2	79.8	82.8	82.5	**69.5**	82.5	87.3	79.8
Z-Reweighting	71.9	79.6	76.5	78.9	82.5	–	–	–	–	78.6
基线模型：										
GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	**87.9**	79.0
SensEmBERT	73.6	80.6	70.3	74.8	80.2	–	–	–	–	75.9
ARES	71.0	78.0	77.1	77.3	**83.2**	80.6	68.3	80.5	83.5	77.9
量子启发的模型：										
QWSD	–	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
MS-WSD$_{\text{Base}}$	74.4	**80.8**	**78.5**	**80.0**	82.9	**82.7**	**69.5**	**83.0**	86.5	**80.1**
MS-WSD$_{\text{Large}}$	75.0	**81.5**	**79.2**	**81.2**	**83.2**	**84.4**	**71.7**	**84.4**	87.5	**81.6**

与前人工作的模型相比，MS-WSD 在多数测试集上的表现优于对比模型，其中在测试集的整合集 *ALL* 上的不俗表现佐证了 MS-WSD 的有效性与竞争力。在开发集 *SE7* 上的表现低于对比模型 ESR，在测试集上的表现优于对比模型 ESR，说明 MS-WSD 相比于 ESR 而言并没有过度拟合训练集与开发集，反映了 MS-WSD 具有不错的泛化能力。在测试集 *Adv.* 上，MS-WSD 的表现略低于对比模型。测试集 *Adv.* 为副词集，相比于名词集、动词集和形容词集而言，副词集中的高频词义的占比最大，而长尾词义的占比最少。由于该测试集是高频词义占主导的测试，

并不适合于评估我们的词义消歧模型的优势，因为该词义消歧模型中采用了适合于长尾词义消歧的表征学习方法，即自适应调优表征和多源融合表征。

与基线模型相比，MS-WSD 在大多数测试集上都表现出色，其证实了集成自适应调优表征和多源融合表征的词义消歧模型的有效性。下面给出具体的对比分析：

- 与具有相似结构的模型 BEM 相比，MS-WSD 在多数测试集上优于 BEM，说明集成多种表征方法的词义消歧模型的设计是合理的，一定程度上反应了 MS-WSD 比 BEM 在长尾词义识别上是有优势的。
- 与使用了相似的外部数据源的模型 GlossBERT 和 BEM 相比，MS-WSD 同样在多数测试集上优于对比模型，说明集成多种表征方法的词义消歧模型确实是有效的。
- 与同样是量子启发的模型 QWSD 相比，MS-WSD 远优于 QWSD，其原因在于 QWSD 并没有采用当前主流的预训练语言模型去学习表征。
- 在测试集 *Adj.* 和 *Adv.* 上，MS-WSD 不及模型 BEM，其原因在于上述测试集中高频词义占主导，整合了自适应调优表征和多源融合表征的 MS-WSD 并不具有优势。

6.4.2 消融分析实验

为检验多元词义消歧模型整合不同表征学习方法对词义消歧模型最终结果的影响，我们采用删除构件的消融研究方法实施验证，删除掉对应构件的模型被称为**消融模型**。当前多元词义消歧模型采用自适应调优表征方法获取目标词嵌入、采用自适应调优表征方法和多源融合表征方法获取词义定义嵌入，我们将分别删除掉自适应调优表征方法和多源融合表征方法后剩余的模型作为对比模型。对比模型为：

- 原模型 $\text{MS-WSD}_{\text{Base}}$，原模型是采用基于预训练语言模型 BERT 的 *bert-base-uncased* 版本实现的模型，其中目标词嵌入采用自适应调优表征方法获得，词义定义嵌入采用多源融合表征方法获得。
- 目标词嵌入消融模型 $\text{MS-WSD}^{target}_{\text{Base}}$，目标词嵌入消融模型是在原模型的基础之上，将基于**自适应调优表征方法**获取的目标词嵌入修改为基于传统的预训练语言模型直接输出的嵌入。此外，词义定义嵌入的获取方法保持不变。

- 词义定义嵌入消融模型 MS-WSD$^{gloss}_{\text{Base}}$，词义定义嵌入消融模型是在原模型的基础之上，将基于**自适应调优表征方法**与**多源融合表征方法**获得的词义定义嵌入修改为基于传统的预训练语言模型直接输出的嵌入，同时目标词的同一个词义的多个词义定义嵌入采用线性叠加的方式组合。此外，目标词嵌入的获取方法保持不变。

消融模型的参数设置以及运行平台与标准评估实验部分保持一致。

表 6-2　MS-WSD 模型的消融分析实验结果

模型	测试集				
	SE2	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：SemCor					
MS-WSD$_{\text{Base}}$	**80.8**	**78.5**	**80.0**	**82.9**	**80.1**
MS-WSD$^{target}_{\text{Base}}$	79.4	77.4	79.7	81.7	79.0
MS-WSD$^{gloss}_{\text{Base}}$	77.8	75.5	76.5	80.0	77.1

实验结果与分析：实验结果如表 6-2 所示。从整体表现来看，原模型 MS-WSD$_{\text{Base}}$ 优于对比模型，说明基于自适应调优表征方法获取目标词嵌入、基于自适应调优表征方法和多源融合表征方法获取词义定义嵌入对词义消歧模型的整体表现有着重要的价值，同时也说明多元词义消歧模型具备整合多种表征学习方法的能力。详细对比如下：

- 对比消融模型 MS-WSD$^{target}_{\text{Base}}$ 和原模型 MS-WSD$_{\text{Base}}$，原模型在所有测试集上的实验结果都有一个百分点的提升，说明采用自适应调优表征方法获取目标词嵌入是有价值的，对词义消歧模型最终表现是有贡献的，同时也说明多元词义消歧模型具备整合自适应调优表征方法的能力。
- 对比消融模型 MS-WSD$^{gloss}_{\text{Base}}$ 和原模型 MS-WSD$_{\text{Base}}$，原模型在所有测试集上的实验结果都有 2-4个百分点的提升，说明采用自适应调优表征方法和多源融合表征方法获取词义定义嵌入是有作用的，对词义消歧模型最终表现是有重要贡献的，同时也说明多元词义消歧模型具备整合自适应调优表征方法和多源融合表征方法的能力。
- 对比消融模型 MS-WSD$^{target}_{\text{Base}}$ 和消融模型 MS-WSD$^{gloss}_{\text{Base}}$，在所有测试集上目标词嵌入消融模型的实验结果都优于词义定义嵌入消融模型的实验结果，说明多源融合表征方法比自适应调优表征方法对词义消歧模型整体表现的贡献更大，同时也说明对于词义消歧任务而言，正确的词义定义表征是起决定性作用的。

6.4.3 长尾分析实验

为检验整合多路词义表征的多元词义消歧模型在长尾词义消歧任务中的有效性，我们分别在构造的长尾词义数据集（Long-tail Sense Datasets，LTS）和最新的交叉语种数据集（Cross-lingual Datasets，CL）上实施实验。长尾词义数据集的构造方法参见章节 3.4.3 ，交叉语种数据集的简介同样参见章节 3.4.3。

对比模型： 长尾词义数据集上的对比模型采用消融分析实验部分构造的实验模型，即原模型 MS-WSD$_{\text{Base}}$、目标词嵌入消融模型 MS-WSD$^{target}_{\text{Base}}$、词义定义嵌入消融模型 MS-WSD$^{gloss}_{\text{Base}}$。交叉语种数据集上的对比模型采用跨语言评估框架原论文中使用的模型 XLMR-Base[134] 作为基线模型。由于跨语言评估框架是基于电子词典 BabelNet [2] 构建的，因此实验中的词义注释同样选自电子词典 BabelNet。此外，由于电子词典 BabelNet 中的大多数小语种都使用英语的词义定义，因此我们直接使用英语中的词义注释来提供目标词的候选词义列表。

模型设置：原模型的实验设置与标准评估实验部分的保持一致，消融模型的实验设置与消融分析实验部分的保持一致；此外三个对比模型的硬件运行平台和软件开发框架与标准评估实验部分的保持一致。模型 XLMR-Base 的实验结果来自于跨语言评估框架中公布的实验数据。

表 6-3 MS-WSD 模型在长尾词义数据集下的实验结果

模型	测试集				
	SE2	*SE3*	*SE13*	*SE15*	*ALL*
训练数据：LTS					
MS-WSD$_{\text{Base}}$	**47.7**	**46.9**	**48.9**	**51.1**	**49.2**
MS-WSD$^{target}_{\text{Base}}$	44.4	43.1	44.5	46.8	45.0
MS-WSD$^{gloss}_{\text{Base}}$	43.7	41.9	42.8	44.0	43.7

实验结果与分析： 长尾词义数据集上的实验结果如表 6-3 所示。从整体表现来看，原模型优于两个消融模型，说明整合了自适应调优表征方法和多源融合表征方法的多元词义消歧模型的设计是合理的。对比 MS-WSD$_{\text{Base}}$ 和 MS-WSD$^{target}_{\text{Base}}$，原模型优于消融模型，说明自适应调优表征方法有助于改善词义消歧模型的识别能力，同时也说明自适应调优表征方法有助于长尾词义的识别。对比 MS-WSD$_{\text{Base}}$ 和 MS-WSD$^{gloss}_{\text{Base}}$，原模型优于消融模型，说明自适应调优表征方法和多源融合表征

[2] https://babelnet.org/

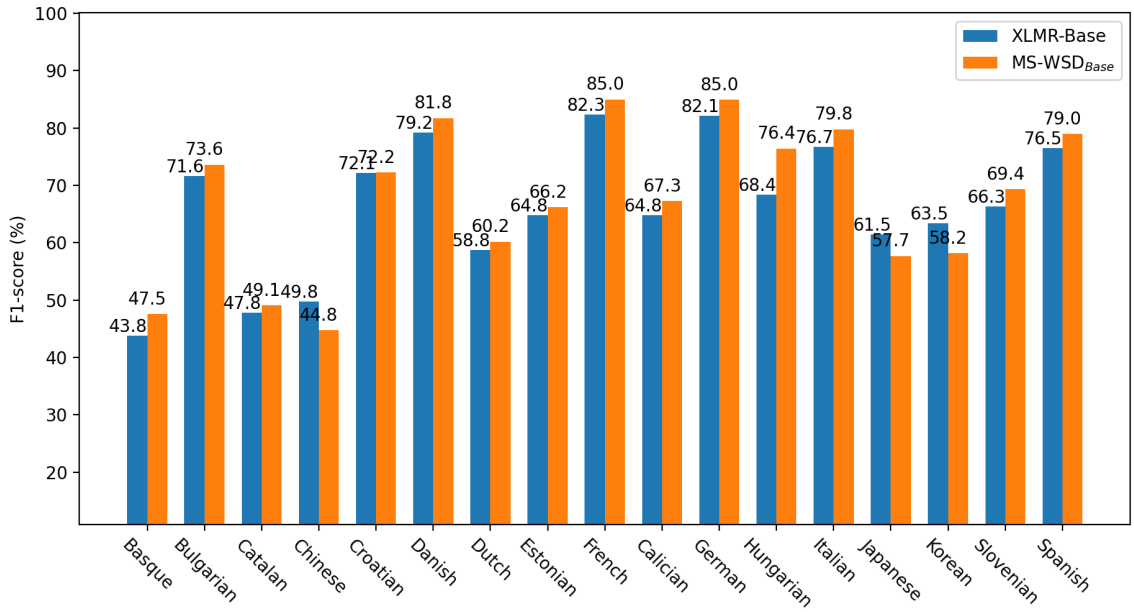

图 6-4　　MS-WSD 模型在交叉语种数据集下的实验结果

方法有助于改善词义消歧模型的识别能力，同时也说明自适应调优表征方法和多源融合表征方法有助于长尾词义的识别。对比 MS-WSD$^{target}_{\text{Base}}$ 和 MS-WSD$^{gloss}_{\text{Base}}$，目标词嵌入消融模型优于词义定义嵌入消融模型，说明正确的、有区分边界的词义定义表征更利于改善词义消歧模型的识别能力。

交叉语种数据集上的实验结果如图 6-4 所示。对比 MS-WSD$_{\text{Base}}$ 和 XLMR-Base，原模型在多个数据集上优于模型 XLMR-Base，这说明我们的模型具有一定的鲁棒性，也说明整合了自适应调优表征方法和多源融合表征方法的多元词义消歧模型具有广泛的适用性。在 Chinese、Croatian、Japanese、Korean 等语种上，我们的模型表现不佳。通过分析表现不佳的数据集，我们发现它们大多是小众或非字母类型的语言。结果不佳的原因是我们使用英语的词义注释而不是其他语言对应的词义注释，因此字符类型与英语相似的语言会表现更好。

6.4.4　模型间对比分析

由于各个章节中提出的词义消歧模型都采用了相同的实验设置，其中模型的主要贡献在于提出的各种表征方法，所以模型间存在可比性。但也需要强调的是，模型在设计机理上存在差异，所以模型间的横向对比也只能从一个侧面反映模型的优劣。

对比模型：模型间的横向对比分析采用各个章节中标准评估实验下的词义消

歧模型，其中第四、五章节采用标准评估实验设置下的模型，对比模型分别为第三章节提出的基于解耦表征的词义消歧模型 DR-WSD ，第四章节提出的基于自适应调优表征的词义消歧模型 SR-WSD，第五章节提出的基于多源融合表征的词义消歧模型 MF-WSD，以及本章节提出的整合多路词义表征的多元词义消歧模型 MS-WSD，并且模型都采用了基于预训练语言模型 BERT 的 *bert-base-uncased* 版本和 *bert-large-uncased* 版本实现的词义消歧模型。

各个模型的**模型设置**与**实验设置**都采用对应章节中的设置方式，具体细节参见相应章节中的描述内容。

表 6-4 标准评估实验下各个模型的实验结果

模型	模型章节	开发集	测试集				
		SE7	*SE2*	*SE3*	*SE13*	*SE15*	*ALL*
$\text{DR-WSD}_{\text{Base}}$	第三章	74.7	80.8	78.0	80.0	82.7	80.4
$\text{SR-WSD}_{\text{Base}}$	第四章	74.5	80.6	79.1	80.0	**84.7**	80.5
$\text{MF-WSD}_{\text{Base}}$	第五章	**74.8**	**81.0**	**79.3**	**80.8**	82.7	**80.8**
$\text{MS-WSD}_{\text{Base}}$	第六章	74.4	80.8	78.5	80.0	82.9	80.1
$\text{DR-WSD}_{\text{Large}}$	第三章	**75.4**	81.1	79.2	81.1	83.2	81.4
$\text{SR-WSD}_{\text{Large}}$	第四章	**75.4**	81.8	80.2	81.1	**85.1**	81.7
$\text{MF-WSD}_{\text{Large}}$	第五章	75.2	**82.5**	**80.5**	**81.2**	83.2	**81.8**
$\text{MS-WSD}_{\text{Large}}$	第六章	75.0	81.5	79.2	**81.2**	83.2	81.6

实验结果与分析：标准评估实验设置下各个模型的实验结果如表 6-4 所示，实验结果分别取自对应章节中的实验数据。从整体表现来看，第五章节提出的基于多源融合表征的词义消歧模型 MF-WSD 优于其它章节提出的词义消歧模型，说明基于量子测量启发的多源融合表征能够在语义空间中建立起词义之间的内在联系以及加强词义之间的区分边界，并且对于词义消歧任务而言，强化词义间的区分边界是有利于改善包括长尾词义在内的词义识别能力。详细对比如下：

- 模型 DR-WSD 与模型 MF-WSD 对比，$\text{MF-WSD}_{\text{Base}}$ 在各个测试集上都优于 $\text{DR-WSD}_{\text{Base}}$，$\text{MF-WSD}_{\text{Large}}$ 同样在各个测试集上都优于 $\text{DR-WSD}_{\text{Large}}$，但 $\text{DR-WSD}_{\text{Large}}$ 在开发集上优于 $\text{MF-WSD}_{\text{Large}}$。依据模型的设计机理分析，量子纠缠启发的解耦表征从模型整体来看是降低了所需参数数量，但由于长尾词义训练数据稀缺，并不能够提供足够数量的有效信息，依然不能得到准备的表征向量，而量子测量启发的多源融合表征实质上是实施了数据增强，所以对于长尾词义消歧任务而言，扩充更多的（长尾词义）训练数据是必要的。

- 模型 SR-WSD 与模型 MF-WSD 对比，$\text{MF-WSD}_{\text{Base}}$ 在多个测试集上优于 $\text{SR-WSD}_{\text{Base}}$，$\text{MF-WSD}_{\text{Large}}$ 同样在多个测试集上优于 $\text{SR-WSD}_{\text{Large}}$，但在测量集 *SE15* 上 SR-WSD 优于 MF-WSD。测量集 *SE15* 是一个高频词义占主导的数据集，上述结果说明量子叠加态启发的自适应调优表征虽然只需要一个参数就能够调整表征向量的准确性，但该原理对于高频词义更有效。此外，在重要的测试集 *ALL* 上，SR-WSD 的实验结果相较于其它模型而言要更接近 MF-WSD 的实验结果，说明基于一个参数控制下的自适应调优表征是具有优势的。
- 模型 MS-WSD 与模型 MF-WSD 对比，$\text{MF-WSD}_{\text{Base}}$ 在各个测试集上优于 $\text{MS-WSD}_{\text{Base}}$，$\text{MF-WSD}_{\text{Large}}$ 在多个测试集上优于 $\text{MS-WSD}_{\text{Large}}$，但在测试集 *SE13* 和 *SE15* 上，$\text{MF-WSD}_{\text{Large}}$ 与 $\text{MS-WSD}_{\text{Large}}$ 的表现是相似的。测试集 *SE13* 和 *SE15* 是高频词义占主导的数据集，而模型 MS-WSD 是一个整合了自适应调优表征的词义消歧模型，说明自适应调优表征发挥了效用，也说明基于一个参数控制下的自适应调优表征在高频词义上的表现更明显。此外，MS-WSD 是一个整合多路词义表征的多元词义消歧模型，MS-WSD 在长尾词义占主导的测试集 *SE2* 和 *SE3* 上的表现不及 MF-WSD 说明整合过多的表征学习方法是会拖累模型整体的表现，并且也不利于模型充分的学习与参数优化。

综上所述，对于词义消歧任务而言，量子测量启发的多源融合表征更适用于改善长尾词义的识别能力，并且采用单独的、针对性的表征学习方法更利于提高词义消歧模型整体的表现。但需要说明的是，由于上述实验分析依然存在局限性，并不能全面、客观的反映表征学习方法以及基于表征学习方法的词义消歧模型的真实水平。

6.5 本章小结

本章针对词义消歧系统难以公平对待长尾词义的问题，提出量子双缝干涉实验启发的多元词义消歧模型，以实现整合多种表征学习方法一致性地应对低资源下的长尾词义消歧任务。多元词义消歧模型基于双缝干涉实验的数学模型所导出，即双缝干涉模型的泛化形式，用以实现目标词的多元词义识别与校验，同时实现改善词义消歧过程中长尾词义的弱势地位。最后，基于自适应调优表征获取目标

词表征、基于自适应调优表征和多源融合表征获取词义定义表征，以构造整合多路词义表征的多元词义消歧模型并实施验证，实验结果证实多元词义消歧模型具备整合多种表征方法的能力。为了更进一步明确整合的多种表征方法对模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实整合的多种表征方法有助于改善模型整体的表现，也有利于提高长尾词义的识别能力。此外，在该章节的实验分析部分给出了各个章节中提出的词义消歧模型间的横向对比分析。相关成果投稿在 ACL 会议论文中。

对于长尾词义消歧任务整体而言，本章提出的多元词义消歧模型有助于整合前面章节的表征方法，以实现一致性地应对低资源下的长尾词义消歧任务，但整合了上述表征方法的词义消歧模型需要消耗大量的时间训练、优化模型。在将后来的工作中，我们将优化量子元素的构造方法，以改善词义消歧模型学习、训练的效率。

第7章 总结与展望

7.1 总结

本文针对长尾词义消歧面临的低资源问题，提出量子理论启发的面向低资源场景的系列表征学习方法，尝试在类量子表征框架下应对长尾词义消歧任务。长尾词义消歧的核心挑战在于其面临的低资源问题，其中匮乏的训练样本使得目标词表征难以学习到所属上下文的特征信息，有限且固定的词义注释文本导致难以获得准确且易于区分的词义定义表征。其中获取词义定义表征更为困难，不仅要求表征向量的准确性，而且需要在语义空间中建立起词义之间的内在联系以及位置关系。目标词表征与词义定义表征在词义消歧过程中共同决定着词义识别的准确性。本课题针对长尾词义消歧任务中目标词表征与词义定义表征面临的表征学习模型大数据量依赖问题、低资源下文本表征不准确性问题分别提出量子理论启发的解耦表征、自适应调优表征，针对词义定义表征之间的不可区分性问题提出量子理论启发的多源融合表征，以及具备整合上述表征方法的多元词义消歧模型，尝试逐级深入的优化表征向量和改善长尾词义识别的准确性。研究内容与贡献概述如下：

1. 针对表征学习模型的大数据量依赖问题，从表征学习模型优化的角度提出同时适用于目标词表征与词义定义表征的量子纠缠启发的解耦表征，以实现降低表征学习过程中对大数据量的依赖，同时实现针对性地挑选决定性的特征。本课题基于纠缠态的解纠缠原理构造强的解纠缠约束机制，用以约束变分自动编码器输出的重采样的特征表示，使采样特征之间彼此独立，即量子纠缠启发的解耦表征。基于该表征方法构造词义消歧模型并实施验证，实验结果证实解耦表征方法能够有效缓解表征学习模型对大数据量的依赖。此外，为了更进一步明确解耦表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实解耦表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果发表在 CIKM 会议论文中。

2. 针对低资源下的文本表征不准确性问题，从表征向量优化的角度提出同时适用于目标词表征与词义定义表征的量子叠加态启发的自适应调优表征，以实现基于有限数据中学习到的一个有效特征优化表征向量，并改善低资源下的目标词表征与词义定义表征的不准确性。本课题基于已知表征向量构建叠加态，以实现在希尔伯特空间中自适应地调优表征向量的准确性，即量子叠加态启发的自适应调优表征。基于该表征方法构造词义消歧模型并实施验证，实验结果证实自适应调优表征方法能够有效改善低资源下的文本表征不准确性问题。此外，为了更进一步明确自适应调优表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实自适应调优表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果发表在 AAAI 会议论文中。

3. 针对低资源下词义定义表征之间的不可区分性问题，提出量子测量启发的多源融合表征，以实现借助多源数据增强词义定义表征，并在语义空间中建立起词义之间的内在联系以及加强词义之间的区分边界。本课题借助叠加态的构造过程实现融合多源数据获得的词义定义表征，借用叠加态的测量过程所导出的干涉项实现在语义空间中约束词义定义表征之间的位置关系，即量子测量启发的多源融合表征。基于该表征方法构造词义消歧模型并实施验证，实验结果证实多源融合表征能够有效改善低资源下词义定义表征之间的不可区分性问题。此外，为了更进一步明确多源融合表征方法对词义消歧模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实多源融合表征方法有助于改善词义消歧模型整体的表现，也有利于提高长尾词义的识别能力。相关成果投稿在 IJCAI 会议论文中。

4. 面向上述量子理论启发的表征方法，提出量子干涉实验启发的多元词义消歧模型，以实现整合多种表征方法一致性地应对低资源下的长尾词义消歧，最终实现词义消歧模型能够公平的对待高频与长尾词义。本课题基于双缝干涉实验的数学模型导出双缝干涉模型的泛化形式，将其构造为词义消歧模型以实现目标词的多词义识别与校验，即量子干涉实验启发的多元词义消歧模型。其中基于自适应调优表征获取目标词表征、基于自适应调优表征和多源融合表征获取词义定义表征，以构建整合多路词义表征的多元词义消歧模型并实施验证，实验结果证实多元词义消歧模型具备整合多种量子启发的表征方法的能力。此外，为了更进一步明确整合的多种表征方法对模型整体表现的贡献和对长尾词义识别的效用，分别实施了消融分析实验和长尾分析实验，实验结果证实整合的多种表征方法有

助于改善模型整体的表现，也有利于提高长尾词义的识别能力。相关成果投稿在ACL 会议论文中。

综上所述，本课题提出的量子启发的表征学习方法和多元词义消歧模型有效改善了长尾词义识别的准确性，但表征学习模型普遍存在学习效率偏低、损失函数难以收敛的问题。在将后来的工作中有必要进一步研究量子元素的高效构造与快速优化方法。此外，需要说明的是本课题提出的应对长尾词义消歧的系列表征学习方法并不局限于词义消歧任务，我们鼓励在其它领域的低资源任务中使用、验证上述表征学习方法。

7.2 展望

本文针对长尾词义消歧任务面临的低资源下的表征学习问题提出量子理论启发的解耦表征方法、自适应调优表征方法、多源融合表征方法，以及具备整合多路词义表征的多元词义消歧模型。上述表征学习方法从原理上讲是一些通用的低资源下的表征学习方法，在将后来的工作中有必要将上述表征学习方法应用到更广泛的领域，以检验其适用领域和使用范围。同时，上述表征学习方法是基于量子力学理论的数学框架构建的机器学习模型，在将后来的工作中有必要进一步研究更为高效地获取量子叠加态、量子测量算子的方法，以实现高效地构建量子机器学习模型。

课题研究内容虽然着眼于长尾词义消歧任务，但更多的侧重于（类）量子机器学习方法的研究。量子机器学习属于量子力学理论与人工智能相结合的交叉学科，是一个新兴的研究领域，在诞生之初就受到广泛的关注。量子机器学习有望赋予人工智能系统像人类一样小样本下的学习、推理（或判断）能力，最终使人工智能系统具备“类人智能”。课题研究内容虽然是量子机器学习领域中一个小小的探索，但量子力学理论在应对长尾词义消歧任务时表现出来的力量是震憾的。量子力学理论不仅提供了建模长尾词义消歧模型的全新数学工具，而且该数学工具是有效的、具备可解释性的。此外，从微观量子世界的角度，即量子认知论，重新理解词义消歧任务是自然的，更符合词义消歧任务解决问题的思路。在将后来的工作中有必要进一步开展量子建模方法论的研究，同时也有必要研究如何从量子认知论的视角构建机器学习模型。

参考文献

[1] Kennedy C. Ambiguity and vagueness: An overview [J]. Semantics: Lexical structures and adjectives. Berlin/Boston: De Gruyter, 2019: 236–271.

[2] Edmonds P. Lexical disambiguation [J]. Encyclopedia of Language and Linguistics. Oxford: Elsevier Science, 2005: 607–623.

[3] Laporte É. Reduction of lexical ambiguity [J]. Lingvisticae Investigationes, 2001, 24 (1): 67–103.

[4] Piantadosi S T, Tily H, Gibson E. The communicative function of ambiguity in language [J]. Cognition, 2012, 122 (3): 280–291.

[5] Navigli R. Word Sense Disambiguation: A Survey [J]. ACM Comput. Surv., 2009, 41 (2): 10:1–10:69.

[6] Bhala V, Abirami S. Trends in Word Sense Disambiguation [J]. Artif. Intell. Rev., 2014, 42 (2): 159–171.

[7] Dewadkar A, Haribhakta V, Kulkarni P, et al. Unsupervised Word Sense Disambiguation in Natural Language Understanding [C]. In ICAI, 2010: 888–893.

[8] Plaza L, Díaz A. Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization [J]. Proces. del Leng. Natural, 2011, 47: 97–105.

[9] Rahman N, Borah B. Improvement of Query-based Text Summarization using Word Sense Disambiguation [J]. Complex & Intelligent Systems, 2020, 6 (1): 75–85.

[10] Gonzales A, Mascarell L, Sennrich R. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings [C]. In WMT, 2017: 11–19.

[11] Pu X, Pappas N, Henderson J, et al. Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation [J]. Trans. Assoc. Comput. Linguistics, 2018, 6: 635–649.

[12] Abderrahim M, Abderrahim M. Arabic Word Sense Disambiguation for Information Retrieval [J]. ACM Trans. Asian Low Resour. Lang. Inf. Process., 2022, 21 (4): 69:1–69:19.

[13] Zhong Z, Ng H. Word Sense Disambiguation Improves Information Retrieval [C]. In ACL, 2012: 273–282.

[14] Hadni M, Ouatik S, Lachkar A. Word Sense Disambiguation for Arabic Text Categorization [J]. Int. Arab J. Inf. Technol., 2016, 13 (1A): 215–222.

[15] Hung C, Chen S. Word Sense Disambiguation based Sentiment Lexicons for Sentiment Classification [J]. Knowl. Based Syst., 2016, 110: 224–232.

[16] Weaver W, Locke W N, Booth A D. Machine translation of languages [M]. Massachusetts Institute of Technology Press, 1949.

[17] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone [C]. In Proceedings of the 5th annual international conference on Systems documentation, 1986: 24–26.

[18] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods [C]. In ACL, 1995: 189–196.

[19] Agirre E, Edmonds P. Word Sense Disambiguation: Algorithms and Applications [M]. Springer Science & Business Media, 2007.

[20] Bevilacqua M, Pasini T, Raganato A, et al. Recent Trends in Word Sense Disambiguation: A Survey [C]. In IJCAI, 2021: 4330–4338.

[21] Agirre E, de Lacalle O L, Soroa A. Random Walks for Knowledge-Based Word Sense Disambiguation [J]. Comput. Linguistics, 2014, 40 (1): 57–84.

[22] Moro A, Raganato A, Navigli R. Entity Linking meets Word Sense Disambiguation: a Unified Approach [J]. Trans. Assoc. Comput. Linguistics, 2014, 2: 231–244.

[23] Rouhizadeh H, Shamsfard M, Rouhizadeh M. Knowledge-Based Word Sense Disambiguation with Distributional Semantic Expansion [C]. In ACL, 2019: 329–335.

[24] Simov K I, Osenova P, Popov A. Using Context Information for Knowledge-Based Word Sense Disambiguation [C]. In AIMSA, 2016: 130–139.

[25] Chaplot D S, Salakhutdinov R. Knowledge-based Word Sense Disambiguation using Topic Models [C]. In AAAI, 2018: 5062–5069.

[26] Sabbir A, Jimeno-Yepes A, Kavuluru R. Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings [C]. In BIBE, 2017: 163–170.

[27] Vial L, Lecouteux B, Schwab D. Sense Embeddings in Knowledge-Based Word Sense Disambiguation [C]. In IWCS, 2017: 1–9.

[28] Scozzafava F, Maru M, Brignone F, et al. Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation [C]. In ACL, 2020: 37–46.

[29] Wang M, Wang Y. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation [C]. In EMNLP, 2020: 6229–6240.

[30] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]. In NAACL, 2019: 4171–4186.

[31] Peters M E, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations [C]. In NAACL, 2018: 2227–2237.

[32] Raganato A, Bovi C D, Navigli R. Neural Sequence Learning Models for Word Sense Disambiguation [C]. In EMNLP, 2017: 1156–1167.

[33] Loureiro D, Jorge A. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation [C]. In ACL, 2019: 5682–5691.

[34] Scarlini B, Pasini T, Navigli R. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation [C]. In AAAI, 2020: 8758–8765.

[35] Scarlini B, Pasini T, Navigli R. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation [C]. In EMNLP, 2020: 3528–3539.

[36] Kumar S, Jat S, Saxena K, et al. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings [C]. In ACL, 2019: 5670–5681.

[37] Bevilacqua M, Navigli R. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information [C]. In ACL, 2020: 2854–2864.

[38] Blevins T, Zettlemoyer L. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders [C]. In ACL, 2020: 1006–1017.

[39] Huang L, Sun C, Qiu X, et al. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge [C]. In EMNLP, 2019: 3507–3512.

[40] Yap B P, Koh A, Chng E S. Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences [C]. In EMNLP, 2020: 41–46.

[41] Bevilacqua M, Maru M, Navigli R. Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss" [C]. In EMNLP, 2020: 7207–7221.

[42] Vial L, Lecouteux B, Schwab D. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation [C]. In GWC, 2019: 108–117.

[43] Luan Y, Hauer B, Mou L, et al. Improving Word Sense Disambiguation with Translations [C]. In EMNLP, 2020: 4055–4065.

[44] Calabrese A, Bevilacqua M, Navigli R. EViLBERT: Learning Task-Agnostic Multimodal Sense Embeddings [C]. In IJCAI, 2020: 481–487.

[45] Pasini T. The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation [C]. In IJCAI, 2020: 4936–4942.

[46] Ruas T, Grosky W I, Aizawa A. Multi-sense embeddings through a word sense disambiguation process [J]. Expert Syst. Appl., 2019, 136: 288–303.

[47] Barba E, Procopio L, Campolungo N, et al. MuLaN: Multilingual Label Propagation for Word Sense Disambiguation [C]. In IJCAI, 2020: 3837–3844.

[48] Song Y, Ong X C, Ng H T, et al. Improved Word Sense Disambiguation with Enhanced Sense Representations [C]. In EMNLP, 2021: 4311–4320.

[49] Hadiwinoto C, Ng H T, Gan W C. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations [C]. In EMNLP, 2019: 5296–5305.

[50] Chen Y, Zhang J, He Q. PosWSD: Low-Resource Word Sense Disambiguation Model using Part Of Speech Information [C]. In IALP, 2022: 26–31.

[51] Kohli H. Transfer Learning and Augmentation for Word Sense Disambiguation [C]. In ECIR, 2021: 303–311.

[52] Hu Z, Luo F, Tan Y, et al. WSD-GAN: Word Sense Disambiguation Using Generative Adversarial Networks [C]. In AAAI, 2019: 9943–9944.

[53] Holla N, Mishra P, Yannakoudakis H, et al. Learning to Learn to Disambiguate: Meta-Learning for Few-Shot Word Sense Disambiguation [C]. In EMNLP, 2020: 4517–4533.

[54] Du Y, Holla N, Zhen X, et al. Meta-Learning with Variational Semantic Memory for Word Sense Disambiguation [C]. In ACL, 2021: 5254–5268.

[55] Chen H, Xia M, Chen D. Non-Parametric Few-Shot Learning for Word Sense Disambiguation [C]. In NAACL, 2021: 1774–1781.

[56] Su Y, Zhang H, Song Y, et al. Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting [C]. In ACL, 2022: 4713–4723.

[57] Blevins T, Joshi M, Zettlemoyer L. FEWS: Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary [C]. In EACL, 2021: 455–465.

[58] Kang M Y, Kim B, Lee J S. Word Sense Disambiguation Using Embedded Word Space [J]. Journal of computing science and engineering, 2017, 11 (1): 32–38.

[59] Kang M Y, Min T, Lee J S. Sense Space for Word Sense Disambiguation [C]. In International Conference on Big Data and Smart Computing, 2018: 669–672.

[60] Barba E, Procopio L, Navigli R. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension [C]. In EMNLP, 2021: 1492–1503.

[61] Rédei M, Summers S J. Quantum probability theory [J]. Studies in History and Philosophy of Modern Physics, 2006, 38: 390–417.

[62] Janotta P, Hinrichsen H. Generalized probability theories: what determines the structure of quantum theory? [J]. Journal of Physics A: Mathematical and Theoretical, 2014, 47 (32): 323001.

[63] Pothos E M, Busemeyer J R. A quantum probability explanation for violations of ‘rational’ decision theory [J]. Proceedings of the Royal Society B: Biological Sciences, 2009, 276: 2171–2178.

[64] Pothos E M, Busemeyer J R. Can quantum probability provide a new direction for cognitive modeling? [J]. The Behavioral and brain sciences, 2013, 36: 255–740.

[65] Barnum H, Caves C M, Finkelstein J, et al. Quantum probability from decision theory? [J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1999, 456: 1175–1182.

[66] Bruza P D, Wang Z, Busemeyer J R. Quantum cognition: a new theoretical approach to psychology [J]. Trends in cognitive sciences, 2015, 19 (7): 383–393.

[67] Primas H. Chemistry, quantum mechanics and reductionism: perspectives in theoretical chemistry [M]. Springer Science & Business Media, 2013.

[68] Esfeld M. Quantum holism and the philosophy of mind [J]. Journal of Consciousness Studies, 1999, 6 (1): 23–38.

[69] Pothos E M, Busemeyer J R. Quantum cognition [J]. Annual review of psychology, 2022, 73: 749–778.

[70] Keyl M. Fundamentals of quantum information theory [J]. Physics reports, 2002, 369 (5): 431–548.

[71] Bokulich A. Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism [C]. 2008: 29–48.

[72] Khrennikov A Y. Information Dynamics in Cognitive, Psychological, Social, and Anomalous Phenomena [M]. Springer Science & Business Media, 2013.

[73] Ismael J, Schaffer J. Quantum holism: nonseparability as common ground [J]. Synthese, 2020: 1–30.

[74] Yearsley J M, Busemeyer J R. Quantum cognition and decision theories: A tutorial [J]. Journal of Mathematical Psychology, 2016, 74: 99–116.

[75] Spekkens R W. Evidence for the epistemic view of quantum states: A toy theory [J]. Physical Review A, 2007, 75 (3): 032110.

[76] Van Enk S. A toy model for quantum mechanics [J]. Foundations of Physics, 2007, 37 (10): 1447–1460.

[77] Schack R, Brun T A, Caves C M. Quantum bayes rule [J]. Physical Review A, 2001, 64 (1): 014305.

[78] Moreira C, Wichert A. Quantum-like bayesian networks for modeling decision making [J]. Frontiers in psychology, 2016, 7: 11.

[79] Khrennikov A. Quantum-like modeling of cognition [J]. Frontiers in Physics, 2015, 3: 77.

[80] Fuchs C A. Nonorthogonal quantum states maximize classical information capacity [J]. Physical Review Letters, 1997, 79 (6): 1162.

[81] D'Ariano G M, Chiribella G, Perinotti P. Quantum theory from first principles: an informational approach [M]. Cambridge University Press, 2017.

[82] Svozil K. Quantum logic [M]. Springer Science & Business Media, 1998.

[83] Pitowsky I. Quantum probability-quantum logic [M]. Springer, 1989.

[84] Nielsen M A, Chuang I. Quantum computation and quantum information [M]. American Association of Physics Teachers, 2002.

[85] Kolmogorov A N, Bharucha-Reid A T. Foundations of the theory of probability: Second English Edition [M]. Courier Dover Publications, 2018.

[86] Young T. The Bakerian Lecture. Experiments and calculations relative to physical optics [J]. Philosophical Transactions of the Royal Society of London, 1801, 94: 1–16.

[87] Eibenberger S, Gerlich S, Arndt M, et al. Matter–wave interference of particles selected from a molecular library with masses exceeding 10000 amu [J]. Physical Chemistry Chemical Physics, 2013, 15 (35): 14696–14700.

[88] Moreira C, Wichert A M. Quantum-Like Bayesian Networks for Modeling Decision Making [J]. Frontiers in Psychology, 2016, 7.

[89] Zhang P, Niu J, Su Z, et al. End-to-End Quantum-like Language Models with Application to Question Answering [C]. In AAAI, 2018: 5666–5673.

[90] Reed W J. The Pareto, Zipf and other power laws [J]. Economics Letters, 2001, 74 (1): 15–19.

[91] Wang Y, Ramanan D, Hebert M. Learning to Model the Tail [C]. In NIPS, 2017: 7029–7039.

[92] Buda M, Maki A, Mazurowski M A. A systematic study of the class imbalance problem in convolutional neural networks [J]. Neural Networks, 2018, 106: 249–259.

[93] Tahvili S, Hatvani L, Ramentol E, et al. A novel methodology to classify test cases using natural language processing and imbalanced learning [J]. Engineering applications of artificial intelligence, 2020, 95: 103878.

[94] Bej S, Davtyan N, Wolfien M, et al. LoRAS: An oversampling approach for imbalanced datasets [J]. Mach. Learn., 2021, 110 (2): 279–301.

[95] Hu Z, Tan B, Salakhutdinov R, et al. Learning Data Manipulation for Augmentation and Weighting [C]. In NeurIPS, 2019: 15738–15749.

[96] Li B, Hou Y, Che W. Data Augmentation Approaches in Natural Language Processing: A Survey [J]. AI Open, 2022, 3: 71–90.

[97] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning [J]. Advances in Neural Information Processing Systems, 2020, 33: 18661–18673.

[98] Hospedales T M, Antoniou A, Micaelli P, et al. Meta-Learning in Neural Networks: A Survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44: 5149–5169.

[99] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning [J]. ACM computing surveys (csur), 2020, 53 (3): 1–34.

[100] Creswell A, White T, Dumoulin V, et al. Generative Adversarial Networks: An Overview [J]. IEEE Signal Processing Magazine, 2018, 35 (1): 53–65.

[101] Bengio Y, Courville A C, Vincent P. Representation Learning: A Review and New Perspectives [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35 (8): 1798–1828.

[102] Sanchez E H, Serrurier M, Ortner M. Learning Disentangled Representations via Mutual Information Estimation [C]. In ECCV, 2020: 205–221.

[103] Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization [C]. In ICLR, 2019: 897–907.

[104] Blevins T, Zettlemoyer L. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders [C]. In ACL, 2020: 1006–1017.

[105] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [C]. In ICLR, 2015: 99–110.

[106] Navigli R, Camacho-Collados J, Raganato A. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison [C]. In EACL, 2017: 99–110.

[107] Pradhan S, Loper E, Dligach D, et al. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words [C]. In Proceedings of the Fourth International Workshop on Semantic Evaluations, 2007: 87–92.

[108] Edmonds P, Cotton S. SENSEVAL-2: Overview [C]. In Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, 2001: 1–5.

[109] Snyder B, Palmer M. The English all-words task [C]. In The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004: 41–43.

[110] Navigli R, Jurgens D, Vannella D. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation [C]. In Second Joint Conference on Lexical and Computational Semantics, 2013: 222–231.

[111] Moro A, Navigli R. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking [C]. In Proceedings of the 9th International Workshop on Semantic Evaluation, 2015: 288–297.

[112] Hadiwinoto C, Ng H T, Gan W C. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations [C]. In EMNLP, 2019: 5296–5305.

[113] Loureiro D, Jorge A M. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation [C]. In ACL, 2019: 5682–5691.

[114] Wang M, Wang Y. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation [C]. In EMNLP, 2020: 6229–6240.

[115] Scarlini B, Pasini T, Navigli R. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation [C]. In EMNLP, 2020: 3528–3539.

[116] Scozzafava F, Maru M, Brignone F, et al. Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation [C]. In ACL, 2020: 37–46.

[117] Wang M, Zhang J, Wang Y. Enhancing the Context Representation in Similarity-based Word Sense Disambiguation [C]. In EMNLP, 2021: 8965–8973.

[118] Song Y, Ong X C, Ng H T, et al. Improved Word Sense Disambiguation with Enhanced Sense Representations [C]. In EMNLP, 2021: 4311–4320.

[119] Wang M, Wang Y. Word Sense Disambiguation: Towards Interactive Context Exploitation from Both Word and Sense Perspectives [C]. In ACL, 2021: 5218–5229.

[120] Huang L, Sun C, Qiu X, et al. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge [C]. In EMNLP, 2019: 3507–3512.

[121] Iacobacci I, Pilehvar M T, Navigli R. Embeddings for Word Sense Disambiguation: An Evaluation Study [C]. In ACL, 2016: 897–907.

[122] Kumar S, Jat S, Saxena K, et al. Zero-Shot Word Sense Disambiguation using Sense Definition Embeddings [C]. In ACL, 2019: 5670–5681.

[123] Bevilacqua M, Navigli R. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information [C]. In ACL, 2020: 2854–2864.

[124] Church K W. Word2Vec [J]. Natural Language Engineering, 2017, 23 (1): 155–162.

[125] Melamud O, Goldberger J, Dagan I. context2vec: Learning generic context embedding with bidirectional lstm [C]. In CoNLL, 2016: 51–61.

[126] Luo F, Liu T, Xia Q, et al. Incorporating Glosses into Neural Word Sense Disambiguation [C]. In ACL, 2018: 2473–2482.

[127] Tamburini F. A Quantum-Like Approach to Word Sense Disambiguation [C]. In RANLP, 2019: 1176–1185.

[128] Luo F, Liu T, He Z, et al. Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention [C]. In EMNLP, 2018: 1402–1411.

[129] Zhang J, He R, Guo F, et al. Disentangled Representation for Long-tail Senses of Word Sense Disambiguation [C]. In CIKM, 2022: 2569–2579.

[130] Berend G. Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations [C]. In EMNLP, 2020: 8498–8508.

[131] Scarlini B, Pasini T, Navigli R. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation [C]. In AAAI, 2020: 8758–8765.

[132] Su Y, Zhang H, Song Y, et al. Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting [C]. In ACL, 2022: 4713–4723.

[133] Pasini T, Raganato A, Navigli R. XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation [C]. In AAAI, 2021: 13648–13656.

[134] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised Cross-lingual Representation Learning at Scale [C]. In ACL, 2020: 8440–8451.

[135] Zhang D, Yin J, Zhu X, et al. Network representation learning: A survey [J]. IEEE transactions on Big Data, 2018, 6 (1): 3–28.

[136] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35 (8): 1798–1828.

[137] Bishop C M. Neural networks and their applications [J]. Review of scientific instruments, 1994, 65 (6): 1803–1832.

[138] Singh J, Banerjee R. A study on single and multi-layer perceptron neural network [C]. In ICCMC, 2019: 35–40.

[139] Stone J V. Independent component analysis: an introduction [J]. Trends in cognitive sciences, 2002, 6 (2): 59–64.

[140] Abdi H, Williams L J. Principal component analysis [J]. Wiley interdisciplinary reviews: computational statistics, 2010, 2 (4): 433–459.

[141] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42 (8): 30–37.

[142] Kingma D P, Welling M, et al. An introduction to variational autoencoders [J]. Foundations and Trends in Machine Learning, 2019, 12 (4): 307–392.

发表论文和参加科研情况说明

（一）已发表的学术论文

[1] **Junwei Zhang**, Ruifang He, Fengyu Guo. Quantum-inspired Representation for Long-tail Senses of Word Sense Disambiguation [C]. In AAAI, 2023. (CCF-A类会议长文)

[2] **Junwei Zhang**, Ruifang He, Fengyu Guo, et al. Disentangled Representation for Long-tail Senses of Word Sense Disambiguation [C]. In CIKM, 2022: 2569-2579. (CCF-B类会议长文)

[3] **Junwei Zhang**, Ruifang He, Fengyu Guo. Bi-matching Mechanism to Combat Long-tail Senses of Word Sense Disambiguation [C]. In ECML-PKDD (2), 2022: 621-637. (CCF-B类会议长文)

[4] **Junwei Zhang**, Yuexian Hou, Zhao Li, et al. Strong statistical correlation revealed by quantum entanglement for supervised learning [C]. In ECAI, 2020: 1650-1657. (CCF-B类会议长文)

[5] **Junwei Zhang**, Zhao Li. Quantum contextuality for training neural networks [J]. Chinese Journal of Electronics (CJE), 2020, 29(6): 1178-1184. (SCI 四区，CCF-T1国内顶级期刊)

[6] **Junwei Zhang**, Zhao Li, Hao Peng, et al. Feedforward Neural Network Reconstructed from High-order Quantum Systems [C]. In IJCNN, 2022: 1-8. (CCF-C类会议长文)

[7] **Junwei Zhang**, Zhao Li, Jianmao Xiao, et al. Neural Network Model Reconstructed from Entangled Quantum States [C]. In IJCNN, 2022: 1-8. (CCF-C类会议长文)

[8] **Junwei Zhang**, Zhao Li, Juan Wang, et al. Quantum Entanglement Inspired Correlation Learning for Classification [C]. In PAKDD, 2022: 58-70. (CCF-C类会议长文)

[9] **Junwei Zhang**, Zhao Li, Ruifang He, et al. Interactive quantum classifier inspired by quantum open system theory [C]. In IJCNN, 2021: 1-7. (CCF-C类会议长文)

[10] **Junwei Zhang**, Ruifang He, Zhao Li, et al. Quantum correlation revealed by bell state for classification tasks [C]. In IJCNN, 2021: 1-8. (CCF-C类会议长文)

（二）在投稿的学术论文

[1] **Junwei Zhang**, Ruifang He, Fengyu Guo. Quantum Interference Model for Semantic Deviations of Glosses in Word Sense Disambiguation [C]. In ACL, 2023. (CCF-A类会议长文)

[2] **Junwei Zhang**, Ruifang He, Fengyu Guo. Quantum-inspired Non-homologous Representation Constraint Mechanism for Long-tail Senses of Word Sense Disambiguation [C]. In IJCAI, 2023. (CCF-A类会议长文)

（三）申请与授权的专利

[1] 贺瑞芳, **张俊伟**. 一种融合定义与场景匹配机制的长尾词义消歧方法. 申请人: 天津大学, 申请号: CN202211131383.7, 申请日: 2022.09.16.

[2] 贺瑞芳, **张俊伟**. 一种融合解耦表征的长尾词义消歧方法. 申请人: 天津大学, 申请号: CN202211265279.7, 申请日: 2022.10.16.

[3] 贺瑞芳, **张俊伟**. 一种适用于小样本的类量子表征学习方法. 申请人: 天津大学, 申请号: CN2023100540795, 申请日: 2023.02.03.

（四）参与的科研项目

[1] 基于用户交互行为深度表示的社会媒体微博摘要, 国家自然科学基金面上项目, 项目编号: 61976154, 项目期限: 2020.01–2023.12.

致　谢

二十六载求学路，路途漫长，磕磕绊绊。感恩父母的鞭策、鼓励以及源源不断的经济支持；同时感谢姐姐、姐夫以及小外甥帮我侍父母、敬孝道。

六载读博路，步步艰辛，处处是坎。感谢恩师贺瑞芳教授危难之际施以援手、谆谆教诲、拨云见日；感谢盼盼师姐、响哥、亚洲师兄、凤羽师姐以及所有的师弟师妹们的陪伴与帮助；感谢我的同学们，此路能与你们同行我并不孤独；同时，感谢侯老师、张老师、李老师、安老师以及人生中遇到的每一位，你们的出现让我的生活充满了色彩！

2023年 5月　于北洋园