重庆师范大学硕士学位论文

基于深度学习的本体匹配方法及其优化算法研究

硕士研究生：　段旭祥
指 导 教 师：　吴至友　教授
学 科 专 业：　计算数学
所 在 学 院：　数学科学学院

重庆师范大学

2020 年5 月

A Thesis Submitted to Chongqing Normal University in Partial Fulfillment of the Requirements for the Degree of Master

Research on ontology alignment method and optimization algorithm based on deep learning

Candidate：Duan Xuxiang
Supervisor：Professor Wu Zhiyou
Major：Computational Mathematics
College：School of Mathematical Sciences

Chongqing Normal University

May, 2020

基于深度学习的本体匹配方法及其优化算法研究

摘 要

自1989年Tim Berners Lee发明万维网后，人类便真正进入了信息爆炸式增长的时代。1.0时期的万维网由网页互相链接而成，但万维网中的网页内容都是通过文档呈现的。在万维网1.0时期，计算机将网页信息呈现给用户，但信息本身所包含的语义无法转化为计算机可理解的计算机语言，方便计算机进行理解，处理。为解决万维网1.0中的存在的以上问题，万维网之父Tim Berners Lee在2001年提出语义网，语义网是一个能够让计算机理解互联网中数据的语义信息，以及数据之间的逻辑关系，使得网页中的数据实现互相关联的数据智能网络。为了构建语义网，国际万维网组织(W3C)在2007年发起了开放链接数据项目，该项目旨在将推动万维网由网页互联走向成知识互联，也有助于推动语义网的发展。项目发起后，研究者们可以对语义数据进行关联和独立发布，从而形成了众多的知识库。本体可以看作知识库的结构框架，定义了概念与概念之间层次关系，形式化地定义了同一领域内共同认可的知识。由于知识库可以分布式地发布到万维网中，多个人对同一或相关联的领域知识构建的不同知识库，在发布数据时，个人的差异会导致对同一个概念命名不相同，取值范围不同等问题，即是本体的异构问题。本体异构问题，在两个本体合并过程中，会导致语义网中许多知识互相关联的缺失并且产生冗余信息。

为解决本体异构的问题，自20世纪90年代末开始，国内外学者便开始了对本体匹配的研究，历经二十多年的积累，已有大量针对本体匹配的方法。这些方法大致可分为基于统计学习的方法和基于深度学习的方法。

基于统计的方法需要借助人工构建特征，而该工作通常比较耗时且难以找到合适的特征；现有的基于深度学习的本体匹配方法克服了基于统计学习类方法中人工构建特征的困难，而忽略了本体本身的结构信息，本体本身的结构信息也包含了大量的语义，会直接影响本体匹配方法的性能。另一方面，基于深度学习的本体匹配方法采用的随机向量或者通过文本预训练模型得到的概念向量表示通常不符合概念在知识图谱中的表示。在知识图谱表示模型中，概念或者关系的表示非常依赖模型的参数。如何在知识图谱表示学习中找到更好的参数从而改善待匹配概念的表示并提高本体匹配的效果也是需要克服的问题。

针对现有基于深度学习的方法存在的两个问题，本文提出了一个新的本体匹配模型MOMWH，MOMWH利用多视角的方法来提取概念信息，从概念的文本、概念在多

个本体间的映射和概念所处的本体本身三个视角来计算概念的表示，从多视角不同的维度对待匹配的概念对进行计算，进而提升了本体匹配的性能，也有效减少人工对特征进行筛选的成本。其次为解决如何找到更好参数得到概念向量表示的问题。本体匹配的性能，为解决该问题，本文首次将黑箱优化和知识图谱表示学习进行结合。通过HORD算法自动调节多个表示学习模型的参数，进一步使得多视角表示学习模型可以学习更好的特征。最后，通过在MA-NCI-FMA与MA-FMA-SNOMED两个医学本体匹配数据集上进行的多个实验，本文设计的方法超过了现有的研究，也在一定程度上解决了现有方法的问题。

关键词： 本体匹配，多视角学习，表示学习，黑箱优化，HORD算法

Research on ontology alignment method and optimization algorithm based on deep learning

ABSTRACT

Since the invention of the World Wide Web(WWW) by Tim Berners Lee in 1989, human beings have really entered into the era of explosive growth of information. During the 1.0 of WWW, web pages are linked to each other, forming WWW. However, the content of web pages in the WWW are presented through documents. In the period of WWW 1.0, the computer presents the web information to the users, but the information contains some semantics, the semantics cannot be converted into computer language, therefore it can be comprehended by the computer for the convenience of its comprehending and processing. In order to solve the above problems in the WWW 1.0, Tim Berners Lee, the father of the WWW, put forward the Semantic Web in 2001, which is a data intelligent network that enables computers to mutually understand the semantic information of the data in the Internet and the logical relationship between the data, thus the data in the web pages can be interconnected. For the sake of constructing the semantic web, the international WWW Organization(W3C) launches the linked open data project in 2007, which aims to promote WWW from web page linking to knowledge linking, conducive to the development of the Semantic Web. After the project was launched, the researchers could associate and publish semantic data independently, thus forming a large number of knowledge bases. Ontology can be regarded as the structure framework of knowledge base, which defines the hierarchical relationship between concepts, and formalize the knowledge jointly recognized within the same domain. Since knowledge base can be distributed to the WWW, many people build different knowledge base for the same or related domain knowledge. When publishing data, the difference of individuals will lead to different naming of the same concept, different value range, etc. , which is the heterogeneous problem of ontology. In the process of ontology merging, the problem of ontology heterogeneity will lead to the lack of correlation and the generation of redundant information in the Semantic Web.

For the sake of resolving the problem of ontology heterogeneity, since the end of 1990s, scholars at home and abroad began to study ontology matching. After more than 20 years of

accumulation, there are a lot of methods in allusion to ontology matching. These methods can be roughly divided into the methods on the basis of statistical learning method and depth learning.

The methods on the basis of statistical learning need to build features manually, which is usually time-consuming and difficult to find suitable features; The existing ontology matching method based on deep learning overcomes the difficulty of Artificial feature Engineering in the method based on statistical learning, but ignores the structure information in ontology. the structure information in ontology contains a lot of semantics, so will directly affect the performance of ontology matching method. On the other hand, the random vectors used in ontology matching method based on deep learning or the concept vectors obtained from text pre-training model usually do not conform to the representation of concepts in knowledge graph. In the knowledge graph representation model, the representation of concepts or relationships depends on the parameters of the model. How to find better parameters in representation learning of knowledge graph to improve the representation of matching concepts and improve the effect of ontology matching is also a problem to be dealt with.

In this paper, a new ontology matching model, MOMWH, is proposed to solve the two problems of existing methods based on deep learning, which uses the multi-view method to extract concept information, calculates the representation of concept from three perspectives: the text of concept, the mapping of concept among multiple bodies and the body itself where the concept is located, and counts the matching concept pairs from different dimensions of multiple perspectives, thus improving the performance of ontology matching, and effectively reducing the cost of manual feature selection. The second is to resolve the problem of how to find better parameters to get the concept vector representation. In order to solve the problem of the ontology matching performance, this paper combines black box optimization and representation learning of knowledge graph for the first time. Through HORD algorithm, the parameters of the learning model can be adjusted automatically, thus the multi-view representation learning model can learn better features. In the end, through several experiments conducted on two medical ontology matching datasets, MA-NCI-FMA and MA-FMA-SNOMED, the method designed in this paper exceeds the existing research, and resolves the problems of the existing methods to a certain extent.

Keywords: ontology matching, multi-view learning, representation learning, black box optimization，HORD algorithm

目 录

1 绪 论

本章首先通过介绍知识图谱与本体匹配来说明本文的研究背景及研究意义，然后阐述本文的研究内容，最后告诉读者本文的论述结构与编排。

1.1 研究背景

自1989年Tim Berners Lee发明万维网后，人类便真正进入了信息爆炸式增长的时代。1.0时期的万维网由网页互相链接而成，但万维网中的网页内容都是通过文档呈现的。在万维网1.0时期，计算机将网页信息呈现给用户，但信息本身所包含的语义无法转化为计算机可理解的计算机语言，方便计算机进行理解，处理，比如："唐纳德特朗普"这一关键词，计算机只能将其识别为一个字符串，而无法理解"唐纳德特朗普"是现任美国总统，这类问题也影响了用户从海量的信息中筛选出自己需要的信息的效率。为解决万维网1.0中的存在的以上问题，万维网之父Tim Berners Lee在2001年提出语义网，语义网是一个能够让计算机理解互联网中数据的语义信息，以及数据之间的逻辑关系，使得网页中的数据实现互相关联的数据智能网络。语义网对网页中的文档添加计算机语言的释义，计算机可对用户输入内容所包含的语义，以及网页文档内容所包含的语义，作出理解和判断，进而使得基于万维网的应用更加智能，如对网页内容进行语义搜索，处理以及整合，最终将用户感兴趣的信息以知识卡片的形式呈现给用户，进一步提升搜索的准确率。随着语义网经过20年的发展，理论基础的不断丰富，科研成果的不断积累，语义网也逐渐被广泛应用于智能搜索，智能问答等方面。所以，对于语义网的研究，是非常重要，且有现实意义的研究。

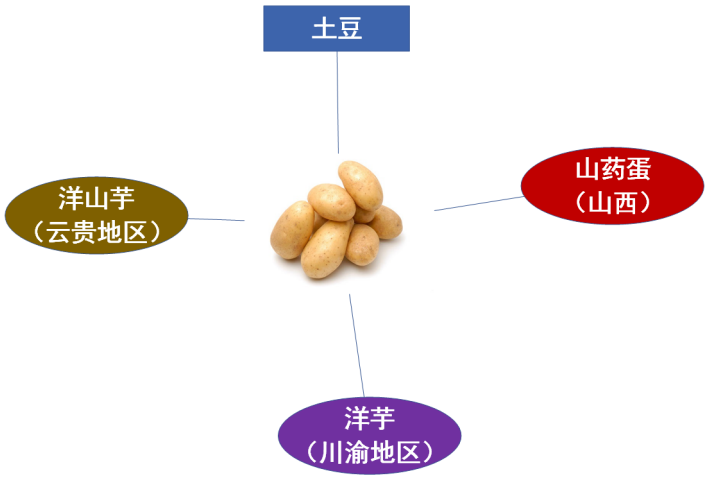

图 1.1 "土豆"及其别名

为了构建语义网，国际万维网组织(W3C)在2007年发起了开放链接数据项目[1]，该

[1]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

项目旨在将推动万维网由网页互联走向成知识互联，也有助于推动语义网的发展。项目发起后，研究者们可以对语义数据进行关联和独立发布，形成了众多的知识库，如：Freebase[2]、DBpedia[3]以及YAGO[4]等。本体可以看作知识库的结构框架，定义了概念与概念之间层次关系，形式化的定义了同一领域内共同认可的知识。由于知识库可以分布式的发布到万维网中，多个人对同一或相关联的领域知识构建的不同知识库，在发布数据时，个人的差异会导致对同一个概念命名不相同，取值范围不同等问题，即是本体的异构问题。例如：如图1.1在中国，对“土豆”这一蔬菜的叫法各不一样，在中国多数地区可被称作”土豆“，云贵地区被叫作”洋山芋“，山西称为”山药蛋“，川渝地区称作”洋芋“，对于同一个概念”土豆“有诸多的名称，这就是本体的异构问题。

1.2 研究现状

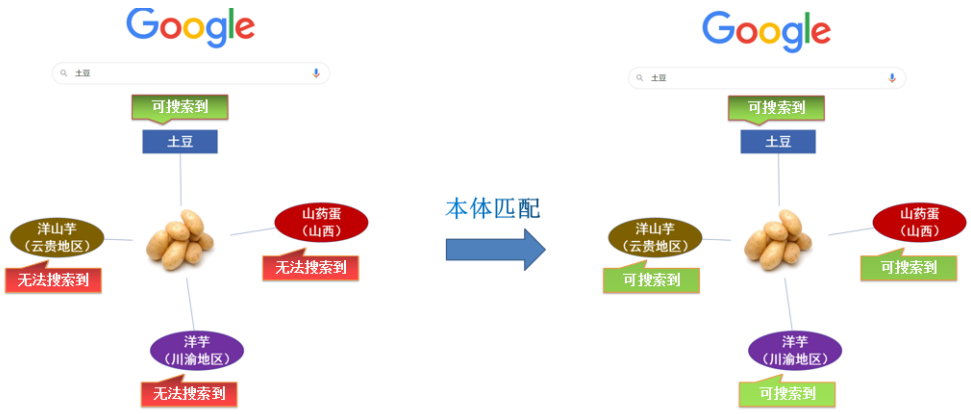

图 1.2 本体匹配在智能搜索中得应用

本体异构问题，在两个本体合并过程中，会导致语义网中许多知识互相关联的缺失并且产生冗余信息。若消除本体异构的问题，以智能搜索为例，则可以提升检索的准确率，举例说明：在开放链接数据项目中，若多个人对土豆相关的知识进行发布时，如图1.1，不同的人对”土豆“的命名也不一样，可能导致采用google搜索时，出现图1.2中左侧的情况，搜索”土豆“这一关键词，搜索引擎可以正确识别，而搜索”洋芋“，”山药蛋“等关键词时，却搜索不到；若借助本体匹配方法，则可以解决因个人的差异会导致对同一个概念命名等不相同的问题，如图1.2中右侧所示，经过本体匹配后，google可以根据”土豆“不同的名称进行正确搜索，提升了google智能搜索的准确率。本体异构问题是知识库关联过程中需要解决的重要问题，而本体匹配方法就是解决这一问题的方法之一，本体匹配方法就是用于寻找存在异构本体之间的语义映射关系，其中常用的做法是计算不同本体中两个概念之间的相似度，通过相似度的高

[2]https://developers.google.com/freebase/
[3]http://wiki.dbpedia.org/
[4]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/ yago/

低来判断两个概念之间的语义关系。

为解决本体异构的问题，自20世纪90年代末开始，国内外学者便开始了对本体匹配的研究，历经二十多年的积累，已有大量针对本体匹配的方法。这些方法大致可分为两类：基于统计学习的方法和基于深度学习的方法。基于统计学习的方法有较为经典的Jerome Euzenat等人针对OWL-Lite表示的本体提出的综合利用字符串距离和词汇距离的OLA[17]，Jayant Madhavan等人提出的使用了字符串和词典两种方式的Schema匹配工具Cupid[36]；Doan等人利用实例信息作为外部资源采用概念间联合概率分布计算相似度的GLUE系统[12]；Sergey Melnik等人提出的利用了本体结构特征的图模型匹配算法SF(Similarity Flooding)[38]；E.Jim enez-Ruiz等人的利用逻辑推理，采用了Horn命题逻辑表示本体层次结构与匹配的LogMap(Logic-based Methods for Ontology Mapping)方法[28]。以上方法都是只利用了本体某一方面（语言学特征，结构特征等）特征的匹配方法，除此之外，还有综合利用多种特征的本体匹配系统。比如：综合利用基于语言学特征与结构特征的基础匹配器，最后将二者线性加权求和的AML(Agreement Maker Light)系统[18]；蒙特利尔大学的在语言学特征模型得到的匹配的基础上，结合结构特征提高匹配质量的YAM++的方法[43]，以及2018年的M.Zhao等人的综合了基于语言学特征，结构特征，以及逻辑推理的FCA-Map方法[64]等。虽然这些方法在其给定的，与其方法特点相对应的数据集中可以取得比较好的效果，但在实际的本体匹配任务中，对于特征工程类本体匹配方法，人工的特征工程很难找到合适的特征，常常是非常耗时的工作[30]。而基于深度学习的本体匹配方法，通常不需要对特征进行人工选择，而是通过对模型参数的调节，实现模型自动选择特征，如：Prodromos Kolyvakis[30]的DeepAlignment，SCBOW[31]，以及Ernesto Jim[55]的Ontoemma，诸如以上三种基于深度学习的本体匹配方法不仅证明了模型自行选择特征是有效的，并且可以获得更适合的特征，提升本体匹配的效果。以上三种是较为典型的基于深度学习类本体匹配方法，这类方法大多借助字典或者上下文信息作为外部资源，借助深度学习模型，得到单词的词嵌入表示，再利用单词的词嵌入得到概念的向量表示，最终利用向量的语义相似度来判断概念的相似度。

所以，现有的本体匹配方法存在以下的问题：第一，基于统计学习的方法在其给定的，与其方法特点相对应的数据集中可以取得比较好的效果，但在实际的本体匹配任务中，对于特征工程类本体匹配方法，人工的特征工程很难找到合适的特征，常常是非常耗时的工作[30]；第二，虽然基于深度学习的本体匹配方法相比人工特征的方法，可以学到更加合适的特征，提升本体匹配的效果。但是，词向量无法准确表示概念的语义，而本体本身的结构信息也包含了大量的语义，忽略掉本体本身的结构信息会直接影响本体匹配方法的性能。以医学本体FMA[5]与NCI[6] 的匹配为例，”Blood

[5]http://owlapi.sourceforge.net/
[6]http://owlapi.sourceforge.net/

Capillary“（NCI中概念）与”Capillary“（FMA中概念）在以上几种基于深度学习本体匹配方法中会被认为是正确匹配（百度翻译也会翻译为同义词”毛细血管“），若考虑结构信息，NCI本体中包含关系”Blood Capillary subclass_of Capillary“，即是“儿子”与”父亲“的关系，则可避免这样的错误；第三，基于深度学习的本体匹配方法的性能依赖于模型所学特征，其常常使用随机向量或者基于文本的预训练模型来提取待匹配概念的特征或表示。但随机向量或者文本预训练模型得到的概念向量表示通常会将概念做为一个独立的单元忽略了概念之间的关联关系。这导致其表示不符合概念在知识图谱[7]中的表示。近年来知识图谱表示学习模型展现了其在知识表示的优势，但现有的工作证明，在知识图谱表示模型中，概念或者关系的表示非常依赖模型的参数。如何在知识图谱表示学习中找到更好的参数从而改善待匹配概念的表示并提高本体匹配的效果也是需要克服的问题。

1.3 研究内容

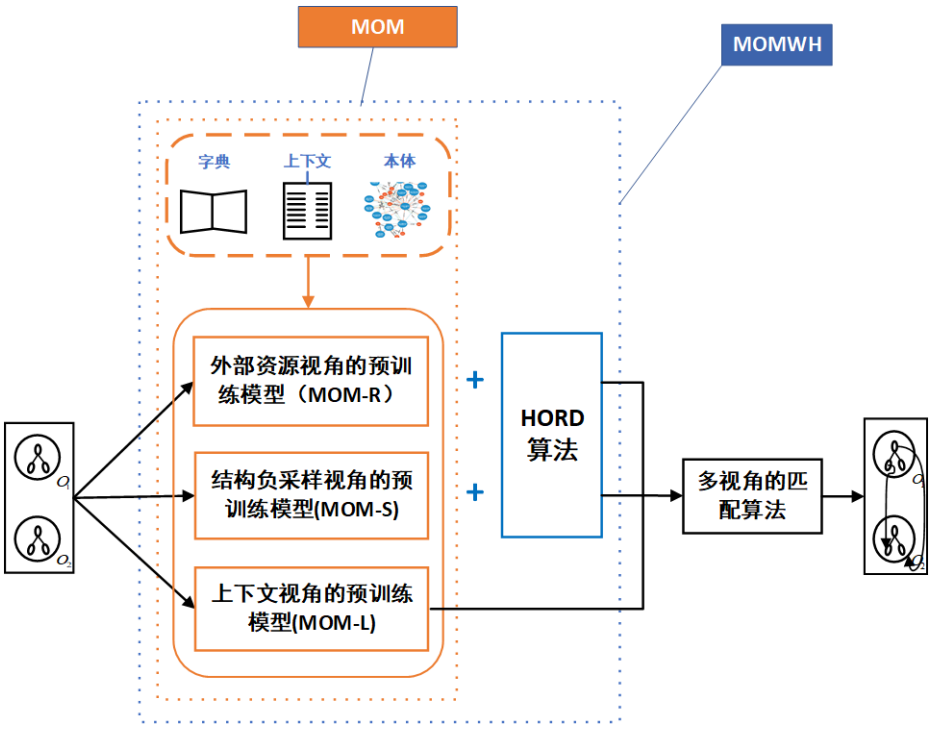

图 1.3 本文算法框架

本文针对以上现有基于深度学习的本体匹配方法存在的问题，设计了如图1.3所示的模型框架MOMWH，该模型分为两个子模型，第一部分是基于深度学习的多视角本体匹配模型MOM，MOM融入了同义词，上下文以及第三方本体的信息，由外部资源视角的预训练模型，结构视角的预训练模型，以及融入上下文视角的预训练模型组成，通过将由O_1和O_2两个本体中获得的概念对分别输入不同视角的模型中，经由模

[7]https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html

型处理，得到概念的向量表示，再利用多视角匹配算法基于概念对的相似度来筛选匹配。第二部分是在MOM的基础上，在外部资源视角的预训练模型，结构视角的预训练模型中分别加入带有全局思想的HORD算法，提出了采用HORD算法优化MOM中外部资源视角的预训练模型和结构视角的预训练模型的MOMWH模型框架，来改进模型最终MOM模型的性能。所以，本文的主要贡献包含以下三个方面：

1.为了更准确的表示知识图谱中的概念，提升本体匹配的性能，本文提出了一个利用多视角的方法来提取概念信息，该方法从概念的文本、概念在多个本体间的映射和概念所处的本体本身三个视角来计算概念的表示。相比与现有深度学习的匹配方法只关注本体概念的字面含义，MOM从多个视角来获取概念的表示，相对现有深度学习的方法，多视角可以从不同维度对待匹配的概念对进行计算，进而提升了本体匹配的性能。

2.MOM使用了多个表示学习模型通过多个视角来提取待匹配概念的特征，在该方法下人多个预训练模型得到的概念表示将影响本体匹配的性能，而预训练模型学习的概念表示对参数非常敏感。为解决该问题，本文首次将黑箱优化和知识图谱表示学习进行结合。在该方法下，通过黑箱优化HORD算法自动调节多个表示学习模型的参数，进一步使得多视角表示学习模型可以学习更好的特征。

3.在基准数据集上，利用常用的评估指标对模型效果进行评估，并将本文中的模型的效果与其他较新的模型结果进行对比，分析，验证本文工作的有效性。

1.4 文章结构安排

本文共分为六个章节，本文的结构编排如下：

第一章：一方面，简述本文的研究背景，介绍语义网，进而通过语义网中本体的异构问题，引出本体匹配；另一方面，介绍本体匹配研究的现状，根据目前的相关研究存在的问题，阐述本文的研究内容，并对本文的具体研究内容进行简述。

第二章：对本体匹配的相关工作，方法从基于统计学习类方法，以及基于深度学习的方法两个方面对已有工作进行综述，总结；对本文研究内容所采用的HORD算法，知识图谱表示学习及几类知识图谱表示学习模型，进行介绍。

第三章：介绍本文提出的提出基于多视角的本体匹配模型MOM，并通过丰富得实验验证MOM模型的有效性。。

第四章：介绍基于HORD算法对MOM模型的改进模型MOMWH，并通过实验验证MOMWH模型的有效性。

第五章：对本文工作的总结与对未来工作的展望。

2 相关工作

自20世纪末，国内外学者便开始了对本体匹配的研究，如今，已有大量用于解决本体异构问题的本体匹配方法，这些方法大致可以分为两类，基于特征工程的方法以及基于深度学习的方法。本章主要对现有的本体匹配方法及知识图谱表示学习模型，“黑箱”优化问题，HORD算法进行阐述。

2.1 基于统计学习的本体匹配方法

基于特征工程的方法有2001年Jayant Madhavan等人[36]提出的使用字符串和词典两种技术，利用本体结构信息，而未采用实例信息，通过键值约束，引用约束等信息对Schema进行匹配的方法Cupid；2002年，Sergey Melnik等人[38]提出了一种面向一般的图的匹配算法Similarity Flooding(SF)；Fausto Oiuachiglia等人[19]提出的一种采用多种自然语言处理技术，主要面向概念层次结构的本体的方法S-Match；2005年，P.Mitra等人[42]提出的利用贝叶斯网络，使用规则获取本体结构信息来生成贝叶斯网络条件概率表，通过阈值筛选匹配的的OMEN方法；2008年，有W.Hu等人[23]的方法和Wang等人的方法，W.Hu等人[23]提出了一种基于概念与属性之间的亲密度来分割本体，将大规模本体匹配任务转化为多个小规模匹配从而可以完成大规模本体匹配任务的本体匹配方法Falcon-AO；P.Wang等人[56]提出了一种能够处理大型本体映射任务，并且嵌入了匹配修复方法来进一步提高匹配的质量本体匹配方法Lily，2009年，J.Li等人[33]提出了一种基于贝叶斯理论，将本体匹配中的问题转换成经验风险最小化问题，先基于本体的概念的语义特征计算相似度，再对相似度聚类得到初始匹配，最后基于当前匹配挖掘潜在匹配的采用多策略的本体映射方法RiMOM；2011年，M.Niepert等人[44]提出的采用概率模型和马尔可夫逻辑网相结合的方式来评估概念对的相似度的CODI方法；Mao等人[37]的方法，将本体匹配任务转化为一个二分类的问题。通过生成各种与领域无关的与特征来描述实体的特征，并采用包含正例与负例概念对的训练集训练支持向量机分类器，最后用训练好的支持向量机来预测正确匹配。2012年，D.Ngo等人[43]提出了一个可以自行配置、灵活并且可扩展，即可根据本体匹配任务规模的大小选取不同匹配策略的的本体匹配方法YAM++；S.Albagli等人[1] 提出的基于马尔可夫网络，采用近似推理算法来计算相似度的交互式本体匹配方法iMatch；2013年Daniel Faria等人[18]提出了一种基于元素级匹配，以外部资源为背景知识，自动化的，且有效利用本体概念的字符串和结构信息的本体匹配方法AML；2014年，Djeddi 等人[15]提出的综合了概念的术语信息，结构信息，和上下文信息，然后通过相似度聚合得到匹配对的XMAP方法；以及2016年Gulic等人的方法[21] 和2018年M.Zhao等

人的工作[64]，Gulic等人[21]提出的在权重更新过程中引入新概念，基于本体结构特征的一对一匹配系统CroMatcher[21]，2018年M.Zhao提出了一种利用形式概念分析逐步来建模本体的各类信息，在大型医学本体的映射任务上表现不错的本体匹配方法FCA-Map[64]。以上方法都是从本体中概念的术语信息，结构相似性，结构约束，概念属性等信息中提取特征，基于所选特征，再构建各种图模型，概率模型等各种不同的模型，通过极大化，极小化，求均值，加权求和等方式来获得概念之间的相似度，通过相似度的高低来选取合适的匹配。而还有少许方法使用一些逻辑将本体匹配转化为可满足性问题，通过约束，来选择满足条件的匹配。这类工作有2004年，Jerome Euzenat等人[17]提出的一种综合使用了字符串距离和词汇距离来比较计算两个URIref的相似度，针对OWL-Lite所表示的本体进行匹配的方法OLA；2011年，E.Jim'enez-Ruiz等人[28]提出了一种具有高扩展性，支持推理并且能处理不协调本体匹配情况的方法LogMap。

2.2 基于深度学习的本体匹配方法

虽然以上给定的基于特征工程的方法，在其给定的，与其方法特点相对应的数据集中可以取得比较好的效果，但在实际的本体匹配任务中，对于特征工程类本体匹配方法，人工的特征工程很难找到合适的特征，常常是非常耗时的工作[30]。而基于深度学习的本体匹配方法，通常不需要对特征进行人工选择，而是通过对模型参数的调节，实现模型自动选择特征，基于深度学习的本体匹配方法不仅证明了模型自行选择特征是有效的，并且可以获得更适合的特征，提升本体匹配的效果。到目前为止，深度学习的医学本体匹配方法有：2005年，Chortaras等人[6]提出的一种基于递归神经网络模型利用本体实例学习本体概念之间相似性的本体自动对齐方法；2006年Hariri等人[7]提出的基于多种基于特征工程的基础匹配器得到的匹配结果构建标签数据集，采用监督学习构建神经网络模型进行灵敏度分析，选出合适的匹配器，最后再通过神经网络模型学习各个基础匹配器的权重的方法；2007年，Merlin等人[8]提出了一种灵活、可扩展的本体映射和集成工具，通过一个前馈神经网络将几种匹配算法结合起来本体匹配方法X-SOM；2008年Huang 等人[24]提出了一种将从实例，概念字符串，概念名等信息中获得的特征，采用神经网络模型学习各个特征权重来实现匹配任务的深度学习方法；Gracia等人[20]提出了一种基于模式的神经网络本体匹配方法CIDER，以及在CIDER基础上实现跨语言匹配的CIDER-CL；Ichise等人[25]提出了一种综合多个概念相似性度量来解决本体映射问题的深度学习模型；2010年Peng等人[45]提出了一种重在识别对于同一元素存在多种关系映射的神经网络模型OMNN；Esposito等人[16]提出了一种通过结构验证和聚合来筛选匹配，支持多种机器学习技术，应用神经网络模型的自动本体匹配系统MoTo；2012年，Rubiolo等人[50]提出了一种将基于代

理模型的分类器与基于人工神经网络分类器相结合的本体匹配方法；2013年Djeddi等人[11]提出的一种通过将多个基于统计学习的不同基础匹配器得到的语义相似度，采用神经网络模型训练得到各个相似度的权重，借助权重将多个相似性度组合成最终相似度的方法XMAP++；Shenoy等人[52]提出了一种基于元数据和实例信息计算不同类型相似性，并使用神经网络模型融合不同相似度的本体匹配方法。

随着表示学习的发展，表示学习展现了其对语义特征提取的优势。利用表示学习的方法有2014年Zhang等人[63]的方法是第一个使用词向量表示来解决本体匹配问题的方式，他们用word2vec[40]来训练维基百科预料，通过语义转换来补充词汇的信息，通过最大相似度来筛选合适的匹配。2015年Xiang 等人[59] 提出的是基于堆叠自动编码器的实体表示学习算法，他们利用本体的自身信息，采用定点算法计算实体的相似性，最后使用stable marriage算法进行匹配。2017年Qiu等人[47]提出的采用表示学习方法，通过词向量来等价表示实体各类信息的无监督自动编码网络；2018 年Wang等人[55]提出的是一种在本体匹配的监督学习框架Ontoemma，采用本上下文作为外部资源，来提高预测准确性的神经结构模型。对于给定的一个候选对，首先对待匹配概念对进行编码，然后进入一个多层感知机网络，最后训练模型并预测这两个实体等价的可能性。然而监督学习类方法在本体匹配任务中，依然无法克服类别不均衡的问题。而更多的是无监督的方法。Prodromos Kolyvakis 等人的Deepalignmeent方法[30]主要利用字典作为外部资源来改进预训练的词向量，从而得到更适合本体匹配任务的实体向量表示；SCBOW[31]与Deepalignmeent方法不同，利用字典和上下文同时作为外部资源，将实体嵌入高维欧几里得空间，采用一种新的短语改造策略，将语义相似信息嵌入到预训练好的词向量的字段中，提出了一种基于降噪自编码器的离群点检测机制的本体匹配方法。

虽然基于深度学习的本体匹配方法相比人工特征的方法，可以学到更加合适的特征，提升本体匹配的效果。但是，词向量无法准确表示概念的语义，而本体本身的结构信息也包含了大量的语义，忽略掉本体本身的结构信息会直接影响本体匹配方法的性能。同时，基于深度学习的本体匹配方法的性能依赖于模型所学特征，其常常使用随机向量或者基于文本的预训练模型来提取待匹配概念的特征或表示。但随机向量或者文本预训练模型得到的概念向量表示通常会将概念做为一个独立的单元忽略了概念之间的关联关系。这导致其表示不符合概念在知识图谱中的表示。近年来知识图谱表示学习模型展现了其在知识表示的优势，但现有的工作证明，在知识图谱表示模型中，概念或者关系的表示非常依赖模型的参数。所以，如何在知识图谱表示学习中找到更好的参数从而改善待匹配概念的表示并提高本体匹配的效果也是现有深度学习类方法的问题之一。

2.3 知识图谱表示学习模型

2.3.1 知识图谱表示学习

深度学习技术作为人工智能领域的核心技术之一，被广泛应用于机器视觉，语音识别，自然语言处理等众多领域。深度学习实现了数据的分布式表示，采用稠密的低维实值向量来表示数据的含义，而表示学习正是实现这类表示的方法。表示学习，是对数据表征的学习，它从数据中提取有用的信息，以便于构建机器学习模型[2]，表示学习的发展也影响了知识图谱的表示形式，从而产生了知识图谱表示学习（也可称为知识表示学习）。知识图谱表示学习旨在将知识图谱中的实体和关系借助模型的学习表示成低维实值向量，进而通过向量之间的关系来描述实体之间的语义关联。

2.3.2 知识图谱表示学习模型

本节重点介绍本文研究内容中涉及的若干种知识图谱表示学习方法。按是否引入额外信息为区分标准，知识图谱表示学习方法可分为仅利用三元组的方法和引入额外信息的方法[57]，表示学习模型众多，本节主要介绍本文用到的仅利用三元组的方法，仅利用三元组的方法又包括基于转移距离的方法和基于语义匹配的方法[57]，基于转移距离的方法有TransE[4],TransH[58],TransR[35],TransD[27] 等，基于语义匹配的方法有DistMult[60],ComplEx[54],ConvE[10]等。接下来以TransE,TransH,TransR为例，进行介绍。

TransE

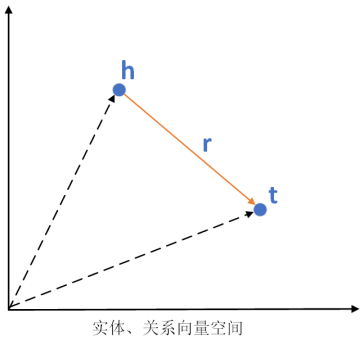

图 2.1 “TransE”模型示意图

TransE是由Bordes等人提出的将任意知识图谱三元组(h,r,t)中的关系r看作实体间的转移向量，实现头实体向量h和尾实体向量t之间的转移操作的方法。如图2.1所示：对于三元组(h,r,t)，模型尽可能使得$h+r\approx t$，所以模型的评分函数定义为：

$$f_{TransE}(h,r,t)=\|h+r-t\|$$

该函数意在表示$h+r$与t之间的距离，而这里的$\|\cdot\|$指1范数或2范数距离，即$\|h+r-t\|_1=\sum_{i=1}^{k}|h_i+r_i-t_i|$，$\|h+r-t\|_2=\sqrt{\sum_{i=1}^{k}|h_i+r_i-t_i|^2}$，距离越

小，则该三元组成立的可能性就越大，置信度越高。

构建模型过程中，为增强模型对于正确与错误三元组的区分能力，TransE采用了合页损失函数，定义了如下的损失函数：

$$L = \sum_{(h,r,t)\in\xi} \sum_{(h',r',t')\in\xi'} \max(0, f_{TransE}(h,r,t) + \gamma - f_{TransE}(h',r',t'))$$

其中，ξ是所有正确三元组(h,r,t)的集合，ξ'为错误三元组(h',r',t')的集合，两者构成模型的训练数据集，γ为正例评分与负例评分之间的间隔，该函数意在使得正例得分尽可能小，负例得分尽可能大，同时使得两者之间的间隔要大于γ，从而提升模型的判别能力。

负样本是正样本中的每个三元组(h,r,t)，任意替换头实体或者尾实体（不同时换）后构成的集合，符号语言为：$\xi' = \{(h',r,t)|h' \in E\} \cup \{(h,r,t')|t' \in E\}$，这里E表示所有实体的集合。

训练模型时，TransE需要对以下参数进行设定，训练轮数（Epoch），向量维数，批次数（nbatchs）,γ的大小，以及负采样率，即对每个正样本需构成的负样本数量，然后利用损失函数计算损失，采用随机梯度下降算法更新其参数。

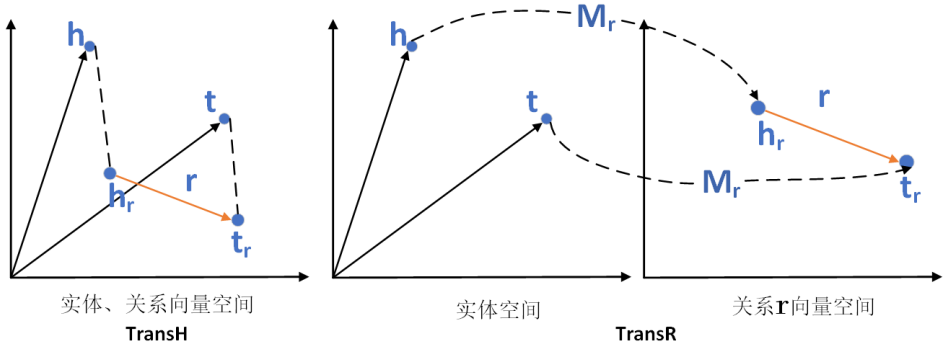

图 2.2 “TransH”和“TransR”模型示意图

TransH,TransR,TransD都是基于TransE的改进模型，与TransE拥有类似的损失函数及训练方式。而TransE模型在一对多或多对一的复杂三元组关系下，效果不理想，例如：给定任意一个头实体会对应多个尾实体的关系r，则存在(h,r,t_1)和(h,r,t_2)，因为$h + r \approx t_1$，$h + r \approx t_2$，所以有$t_1 \approx t_2$，这便无法区分t_1和t_1。为解决这一问题，TransH,TransR,TransD分别做出了不同程度的改进，此外，TransH将TransE的负采样方式改进为采用根据概率分布进行负采样的方式。下面逐一介绍这些模型的评分函数构造的思想。

TransH

Wang等人提出了TransH模型，为解决TransE在一对多或多对一的三元组关系下，效果不理想的问题，TransH规定实体的表示由不同的关系所决定，不同的关系对应不

同的超平面，如图2.2，对任意三元组(h,r,t)，该模型将头实体h和尾实体t沿关系r所在超平面的法向量w_r投影到关系r所在平面中，所以：

$$h_r = h - \mathrm{w}_r^T h \mathrm{w}_r, t_r = t - \mathrm{w}_r^T t \mathrm{w}_r$$

因此，TransH采用与TransE类似的思想，构建评分函数：

$$f_{TransH}(h,r,t) = \|h_r + r - t_r\|$$

TransR

TransH虽然实现了不同关系，实体的表示，但依然假设实体与关系处于同一向量空间，这也影响了TransH模型的表示能力，为进一步弱化该问题的影响，Lin等人提出了TransR模型，TransR认为实体和关系处于不同空间，在映射矩阵M_r作用下，将实体投影到关系空间中，从而有：

$$h_r = M_r h, t_r = M_r t$$

2.4 超参数优化与HORD算法

2.4.1 超参数优化问题

近年来，各类机器学习，深度学习技术被广泛应用于各种与人工智能相关的应用中，这类模型中参数规模较大，只知道其输入与输出没有显式表达式的模型就是黑箱函数。通常黑箱函数的性能会受到模型超参数的制约，所以通过算法进一步优化模型超参数来提升模型性能具有重要现实意义，对这样的黑箱超参数优化问题可以理解为将需要配置d个超参数看作一组向量$x \in R^d$，模型本身可以看作关于x 的“黑箱”函数$f(x)$，超参数优化可以看作解决以下的优化问题，在训练集Z_{train} 和测试集Z_{val}中，函数$f(x)$ 将d个可配置超参数中的超参数选择x 映射到具有学习参数 θ 的深度学习算法的验证误差。

$$\begin{aligned} &\min_{x \in R^d} \quad f(x, \theta; Z_{val}) \\ &s.t. \quad \theta = \arg\min \ f(x, \theta; Z_{train}) \end{aligned}$$

2.4.2 HORD算法

超参数调节较为流行的一类方法是基于贝叶斯的优化方法，例如：基于高斯过程的方法以及基于树结构的方法等，与这类算法相比，HORD算法能够同时优化连续超参数与整数超参数，并且可以以更少的估值次数找到较好的验证误差。表示学习模型的参数包括训练轮数（epoch），向量维数，批次数（nbatchs）,γ的大小，以及负采样率，学习率，既包含整数也包含连续实数，并且训练轮数常常设置比较大，无论是人工经验或者基于贝叶斯类优化方法调参都是一项十分耗时的任务。而HORD算法较可以较少的估值次数很好的实现既包含整数也包含连续数的超参数优化。

$f(x)$非常复杂并且没有显式表达式，所以求解大规模参数的问题5效率不高。为提高优化参数的效率，HORD算法[26]利用径向基函数插值模型构建响应面模型，通过动态超参数坐标搜索来找到整数或连续值超参数的近似最优配置。已知可行的n种超参数配置$x_i(i=1,2,3,\ldots,n)$，验证误差$f_i(i=1,2,3,\ldots,n)$，这里$f_i=f(x_i)$，径向基函数响应面模型具体构造如下：

$$S_n(x)=\sum_{i=1}^{n}\lambda_i\phi(\|x-x_i\|)+p(x)$$

这里$\phi(r)=r^3$，$p(x)=b^Tx+a$，$b=[b_1,\ldots,b_d]^T\in R^d$，$a\in R$是该插值模型的参数，通过求解以下线性系统得到:

$$\begin{bmatrix}\Phi & P\\ P^T & 0\end{bmatrix}\begin{bmatrix}\lambda\\ c\end{bmatrix}=\begin{bmatrix}F\\ 0\end{bmatrix}$$

其中，$\Phi\in R^{n\times n}$，$\Phi_{i,j}=\phi(\|x-x_i\|)$，$i,j=1,\ldots n$，

$$P=\begin{bmatrix}{x_1}^T & 1\\ \vdots & \\ {x_n}^T & 1\end{bmatrix}\lambda=\begin{bmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{bmatrix}c=\begin{bmatrix}b_1\\ \vdots\\ b_k\\ a\end{bmatrix}F=\begin{bmatrix}f(x_1)\\ \vdots\\ f(x_n)\end{bmatrix}$$

HORD算法的主要思想包含以下几步：第1步，借助径向基插值函数构建响应面模型；然后设置待优化参数的范围与初始采样点个数以及最大迭代步数，通过拉丁超立方体采样得到多组变量与函数值的初始点对$\mathrm{A}_{n_0}=\{(x_i,f(x_i))\}_{i=1}^{n_0}$；第2步，通过这些点对$\mathrm{A}_{n_0}$，来得到响应面模型的解析式；第3步，借助当前响应面模型，并结合动态坐标扰动[48]，求得最优值点x_*，同时计算出对应的函数值$f(x_*)$；第4步，将当前最优点对并入初始点对$\mathrm{A}_{n_0}=\mathrm{A}_{n_0}\cup\{(x_*,f(x_*))\}$，第5步，在未达到最大步数时，重复步骤2至步骤4。

3 基于多视角的本体匹配模型MOM

本章主要介绍针对现有方法本文提出的问题问题一和问题二，本文设计的基于多视角的本体匹配模型MOM，以下将从MOM模型的介绍，以及实验评估两个方面进行详细介绍。

3.1 基于多视角的MOM本体匹配模型介绍

3.1.1 问题定义

本体匹配是解决领域本体异构性问题的有效方法之一。为更好的阐述本文的研究内容，本文首先引入了本体匹配的形式化定义。

定义1：本文采用四元组(e_i, e_j, r, θ)来表示O_i，O_j的一组匹配，其中O_i和O_j表示两个本体，e_i，e_j分别代表O_i和O_j中的实体（实体可以是本体的概念，性质等），r表示两个实体e_i，e_j之间的关系，在多数本体中，r一般表示$\{\subseteq, \supseteq, \equiv\}$ 三类关系中的一种。

在很多本体匹配的实际任务中，匹配系统主要关注具有等价关系（等价关系可以理解为一般的同义词关系）的概念匹配。因此，在本文的剩余部分中，只针对需要获取等价关系的本体匹配任务，本文的模型设计也只注重挖掘领域本体中的存在等价关系的概念对。

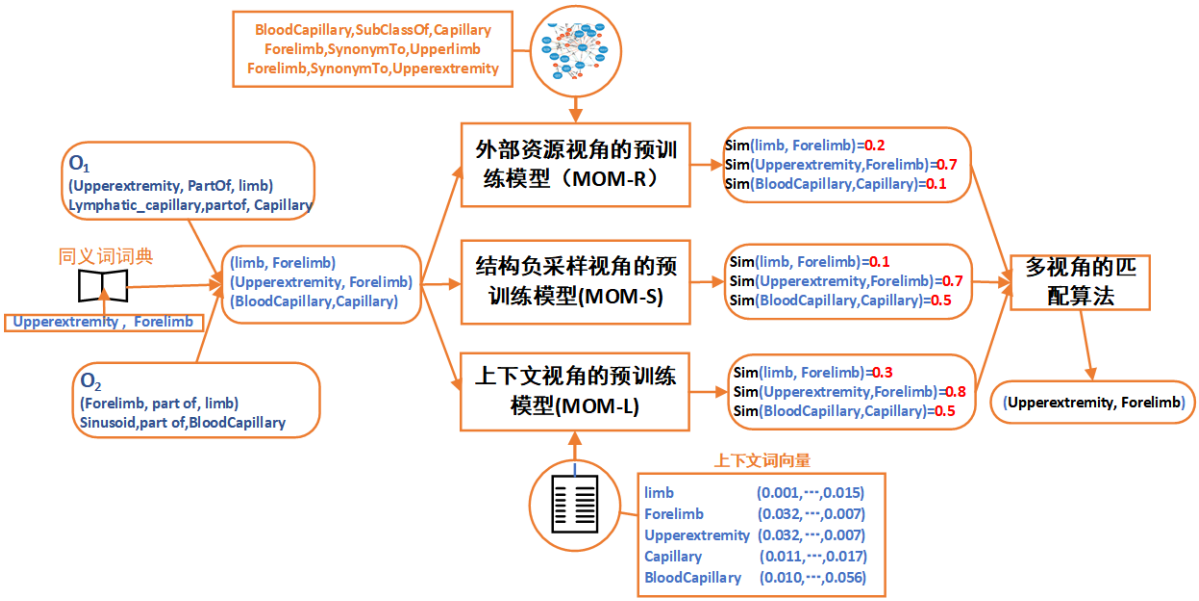

图 3.1 “MOM”模型工作流程图

3.1.2 MOM模型整体框架

现有的工作[29, 46, 61]表明，采用多视角的方式可以从不同角度获得本体更多的语义信息，并提升相应任务的准确性和稳定性。受以上工作启发，本文采用基于多视

角的方法来描述本体匹配的过程，并针对现有方法未充分利用本体结构信息的问题，尝试融入更多的结构信息，来提升匹配效果。

MOM模型的工作流程如图3.1所示，给定两个本体O_1和O_2，首先抽取本体中的概念，结构关系信息，得到的结构关系如O_2中的结构三元组如(Forelimb,PartOf,limb)；再对本体的概念进行规范化，并采用同义词词典挖掘同义概念，如图中采用同义词词典中的Forelimb与Upper extremity为同义词，所以可以找到同义词对(Forelimb,Upper extremity)，从而针对不同视角的模型构建相应概念对数据集；随后将对应的数据集输入到三个视角的预训练模型中，分别是基于外部资源视角，基于结构视角以及基于上下文视角的模型，基于外部资源的预训练模型主要关注第三方本体的结构信息，基于结构视角的预训练模型更加注重两个待匹配本体之间的结构信息，基于上下文视角的预训练模型注重利用本体概念本身的语义信息，采用多个视角旨在结合以上三个模型所学信息，来提升本体匹配的效果；同时为获得本体所有概念所涵盖单词的语义信息，MOM 在上下文视角模块引入了上下文信息，第三方本体的结构信息有助于提升匹配质量[34]，因此在外部资源视角模块中将第三方本体作为外部资源；然后不同视角的模型学习得到同一概念的三种不同表示，再采用各自不同的相似度计算方法得到概念对的相似度，MOM模型中不同视角的模型可以学得不同方面的特征，导致模型得到的概念对相似度也不一样，如图所示，如基于外部资源的模型根据第三方本体的结构信息可以得到匹配对(Blood Capillary,Capillary)的相似度只有0.1；基于结构信息的预训练模型得到(limb,Forelimb)的相似度只有0.1，和(Upperextremity,Forelimb) 的相似度有0.7；基于上下文的预训练模型采用词向量，可以确定(Upperextremity,Forelimb)的相似度有0.8。最后通过相互评价的多视角匹配算法来得到最终的匹配，所以最终正确的匹配为(Upperextremity,Forelimb)。

MOM通过表示学习技术得到概念的连续向量表示，避开了统计学习类方法人工构建特征的麻烦。与现有的基于深度学习的方法相比，MOM利用第三方本体，以及本体自身的结构信息，来进一步提升匹配的效果。在MOM中，对同一个概念存在不同粒度的向量表示，在外部资源视角，基于TransR构建的表示学习模型得到以概念整体为单位的向量表示；基于TransE构建表示学习模型的结构视角预训练模型，最终也得到概念整体为单位的向量表示，以上两个视角都是通过余弦相似度得到概念之间的相似值；上下文视角中，基于word2vec[41]训练上下文得到的本体概念中所有单词的向量表示，一般概念由多个单词的组成，将这些单词对应向量所构成的集合$\{\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_n}\}$ 来表示该概念，再利用这些单词表示，基于本文设计的算法计算概念之间的相似度。

3.1.3 基于上下文视角的预训练模型MOM-L

首先，通过MOM模型的工作流程图3.1中的例子来说明，基于上下文视角的模型采用预训练词向量，对于输入的概念对(limb,Forelimb)，(Blood Capillary,Capillary)以及(Upperextremity,Forelimb)，对照基于上下文训练Word2vec模型得到的词向量表，获得各个单词对应的词向量表示，再通过本文改进的TF-IDF算法求出每个单词对应的权重，再采用余弦距离求出两个概念的相似度。语义相近的概念相似度越高，而Upperextremity与Forelimb的相似度达到0.8，所以图中将(Upperextremity,Forelimb)正确匹配输出，至于其他概念对基于上下文视角无法确定，需结合其他视角，通过多视角匹配算法来确定。下面详细介绍本文提出的MOM-L模型。

基于上下文视角的预训练模型主要基于TF-IDF算法[51]，TF-IDF算法也是本体匹配中计算字符串相似度的非常有效的方法之一[5]。根据TF-IDF算法的假设，可将一个本体中所有概念涵盖的单词表示成一个由单词构成的词袋，每个概念C_i看作一个文档。单个概念所包含的单词$\{t_1, t_2, \ldots, t_n\}$看作术语。受软TF-IDF算法[5]的启发，本文提出了一种基于词向量嵌入的TF-IDF策略的改进算法来计算概念的相似性的方法MOM-L。首先通过word2vec模型[41]训练对应上下文，得到词袋中所有单词的向量表示；然后利用概念中的单词的表示，得到概念的表示$\{\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_l}\}$，这里$l$等于概念所包含的单词数量，最后基于概念表示（概念中的单词的表示构成的集合）利用本文设计的相似度计算方式得到概念对的相似度。这里的相似度计算方法不同于一般的字符串等价，相应的定义如公式：

$$Sim(C_1, C_2) = \sum_{i=1} w_i \cdot \arg\max_j \cos(\mathbf{t_{1i}}, \mathbf{t_{2j}}) \tag{3.1.1}$$

其中，C_1，C_2分别表示O_1和O_2中的概念，$\mathbf{t_{1i}}$，$\mathbf{t_{2j}}$表示C_1，C_2中两个单词t_{1i},t_{2j}的向量表示，w_i 表示t_{1i}在概念C_1中的权重，计算方式如公式：

$$w_i = \frac{TFIDF(t_{1i})}{\sum_{l=1}^{n} TFIDF(t_{1l})} \tag{3.1.2}$$

这里n表示概念C_1所包含的单词数，$TFIDF(\cdot)$表示每个单词的TFIDF函数值。

因为$\mathbf{t_{1i}}$，$\mathbf{t_{2j}}$向量包含了单词丰富的上下文信息，所以基于TFIDF算法的语义嵌入方法MOM-L能够更好的反映概念之间字面意思的相似情况，与字符串匹配相比，可以发现更多的概念同义词匹配。一方面，MOM-L依赖于词嵌入的质量，词嵌入的质量影响匹配的效果，因此这里采用了高质量的预训练向量来保证基于上下文视角的匹配效果（参见第5.2节）。另一方面，使用基于其他视角的预训练模型生成的匹配来评估基于上下文视角所得匹配的质量（参见第4.2.3 节）。

3.1.4 基于结构负采样视角的预训练模型MOM-S

首先，通过MOM模型的工作流程图3.1中的例子来说明，对于输入的概念对(limb,Forelimb)，(Blood Capillary,Capillary)以及(Upperextremity,Forelimb)，借助本节构建的表示学习模型MOM-S，再训练(Upperextremity,Forelimb)概念对表示时，需要对两个概念进行负采样，而在进行结构负采样时，可以发现两个本体中存在(Forelimb,SubClassOf,limb)，(Upperextremity,SubClassOf,limb)关系，对于存在这样关系的概念，不会是正确匹配，所以，模型计算得到的相似度也偏低。至于(Upperextremity,Forelimb)的相似度，需参照其他视角的模型，结合多视角匹配算法综合做出判断。下面重点介绍本节的结构负采样预训练模型。

现有的利用表示学习的本体匹配方法侧重于采用概念的名称，标签，定义等术语信息来学习的概念的词向量表示，却并没有充分利用本体中的结构关系，而有效的利用本体的结构信息，有助于提升本体匹配的效果。所以，本节将从结构视角构建表示学习模型来实现本体的匹配。首先假设由字符串等价或其同义词替换生成的待匹配对是正确的，将获得的待匹配对作为结构视角的表示学习模型模型训练的数据集，然后本文定义了基于交叉熵的损失函数来构建表示学习模型来得到概念的向量表示。损失函数定义如公式：

$$l_{SE} = -\sum_{(C_1,C_2,\equiv,1.0)\in M} \log f_{SE}(C_1,C_2) - \sum_{(C_1',C_2',\equiv,-1.0)\in M'} \log f_{SE}(C_1',C_2') \quad (3.1.3)$$

其中M表示根据假设得到的待匹配对$\{(C_1,C_2,\equiv,1.0)\}$构成的正例数据集，M'表示负例数据集，为生成负例训练集M'，MOM采用了新的负采用技术（参见第三段）。通过对每个正例$(C_1,C_2,\equiv,1.0)\in M$，根据正太分布函数得到的概率，在候选集中采用新的负采样策略（参见第三段）替换C_1或C_2得到负采样$(C_1',C_2,\equiv,-1.0)\in M'$或$(C_1,C_2',\equiv,-1.0)\in M's$。$f_{SE}(C_1,C_2)$表示由公式定义的计算待匹配对得分的评分函数，其中$\mathbf{C_1},\mathbf{C_2}\in\mathbf{R^d}$表示两个本体$O_1$和$O_2$中概念$C_1$，$C_2$的d维向量表示；$\|\cdot\|_2$表示向量的二范数，旨在使得对于两个相似概念的表示$C_1$，$C_2$，得分$f_{SE}(\mathbf{C_1},\mathbf{C_2})$尽可能大，相反，对于不相似的概念对，希望其得分尽可能小。基于TransE的距离转移的思想，将评分函数$f_{SE}(C_1,C_2)$定义如公式：

$$f_{SE}(C_1,C_2) = 2\cdot\frac{1}{1+e^{(\|\mathbf{C_1}-\mathbf{C_2}\|_2)}} \quad (3.1.4)$$

以下详细介绍本文提出的一种不同于TransE以及TransH等表示学习模型从所有概念中进行采样替换的负采样方式，本文的负采样方式充分利用了本体的结构关系，具体方式如下，如图3.2所示，被替换概念的候选集由所有与该概念不构成SubClassOf，PartOf关系的概念构成，并且替换时若候选集中存在与该概念构成disjointwith关系的其他概念，则优先将其作为采样对象。

例如：以医学本体中的SubClassOf关系为例，对于任意匹配对$(C_1, C_2, \equiv, 1.0)$，若C_1和C_2存在关系，即$(C_i, SubClassOf, C_1)$或$(C_j, SubClassOf, C_2)$，其中$C_i \in O_1$，$C_j \in O_2$，则对C_1负采样时，需保证候选集中没有C_i，对C_2负采样时候选集中没有C_j。此外，若C_1或C_2存在关系,即$(C_m, disjointwith, C_1)$或$(C_n, disjointwith, C_2)$，则对C_1负采样时，优先采样C_m，对C_2负采样时，优先采样C_n。利用这些本体的结构信息作为约束改进负采样，使得生成的概念的向量表示包含了更多的本体结构信息，有助于提升MOM的匹配性能。

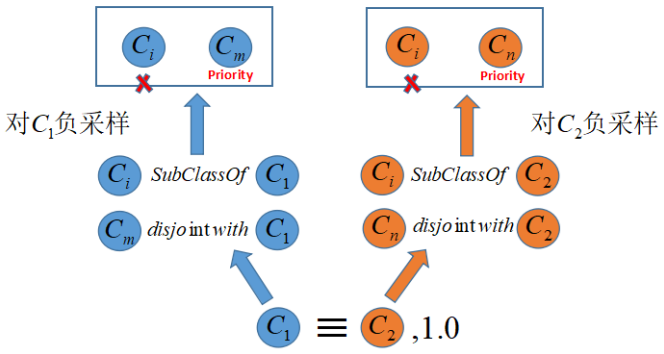

图 3.2 基于结构负采样示意图

3.1.5 基于外部资源视角的预训练模型MOM-R

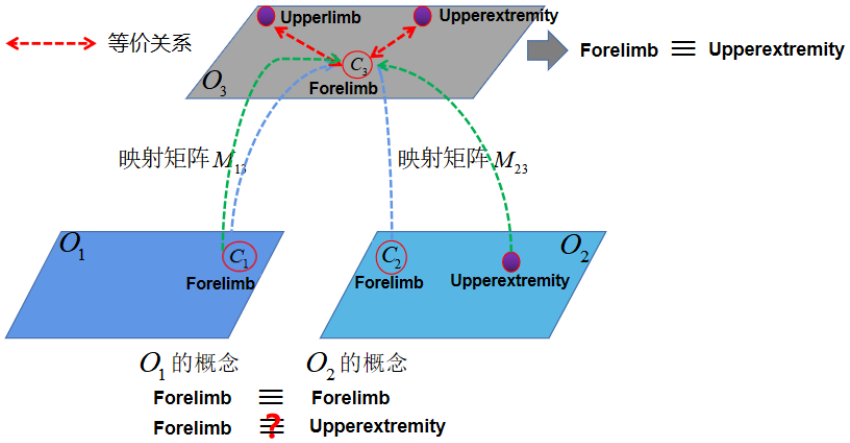

图 3.3 第三方本体作用示意图

首先，同样先通过图3.1与图3.3的例子进行说明。对于O_1中的概念Forelimb，O_2中的概念Forelimb，Upperextremity，两个本体可能的匹配对有(Forelimb,Forelimb)和(Forelimb,Upperextremity)，而比较显然容易判断(Forelimb,Forelimb)为正确匹配，只要通过字符串相等即可得到，而另一个待匹配对(Forelimb,Upperextremity)是不容易判断的。所以，本文借助了第三方本体O_3的信息，第三方本体中有Upperlimb与Forelimb是等价关系，Upperextremity与Forelimb是等价关系。通过Upperextremity与Forelimb的等价关系，可以进一步判定O_1和O_2中的匹配对(Forelimb,Upperextremity)是正确的。而对于模

型而言，对于O_1，O_2中分别与O_3中同一概念构成等价关系的概念，将这些概念的向量表示通过矩阵映射到O_3中，根据以上的关系，模型学得的向量表示会满足以上关系，在求解其语义距离时，若O_1和O_2中的待匹配对，同时与O_3同一概念满足等价关系，则两个概念通过矩阵映射后的语义距离相对较小，从而认为该匹配对为正确匹配；反之通过矩阵映射后的语义距离则较大，则认为其为错误匹配。

受文献[62]中方法的启发，本节采用第三方本体作为外部资源来实现两个本体之间的连接，现实应用中存在很多不同但存在交叉的本体（即两个本体中包含多个相同的概念），如：MA，NCI与FMA这三个本体中存在很多共有的概念和关系，FMA，NCI和SNOMED-CT这三个本体中，也有很多共有的概念和关系，以上两个例子也是本文实验部分的本体匹配数据集。与字典或上下文信息等外部资源相比，第三方本体作为外部资源可以提供更多的结构信息，这有助于改进本体匹配的质量[62]。然而，现有的利用第三方本体的方法主要是基于字符串的字面信息来实现匹配，这导致无法通过第三方本体结构信息挖掘更多的待匹配本体之间的语义信息，不利于本体匹配模型性能的提升[34]。因此，这里将第三方本体作为外部资源，使用表示学习技术，进一步通过第三方本体的结构信息来提升本体匹配的效果。

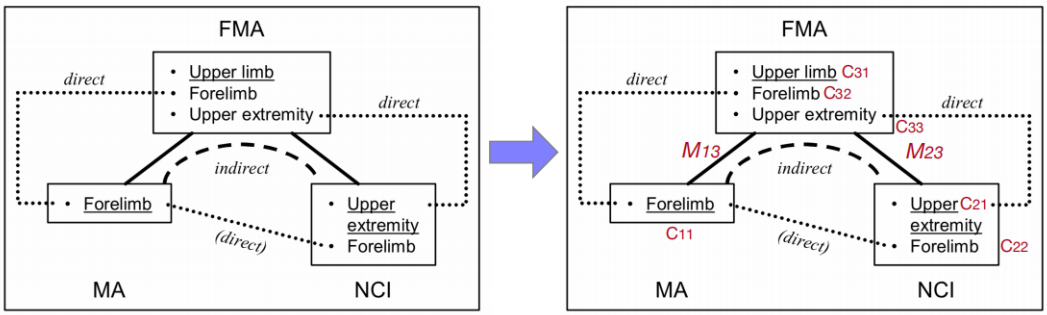

图 3.4 左：现有的利用第三方本体的模型框架；右图：本文使用第三方本体的表示学习模型框架

另一方面，如图3.4显示了现有的第三方本体类方法的模型框架到本文利用第三方本体作为外部信息的表示学习模型框架的演变。这里，采用$C \in R^d$向量来表示概念C，因为存在本体交叉的现象，所以不妨假设存在一些概念对(C_1, C_2)，并且这些概念或它们在O_1，O_2中的同义词出现在第三方本体O_3中，例如图中MA，NCI，FMA中都有“Forelimb”，并且“Forelimb”与“Upper limb”等都是同义词，所以第三方本体的结构信息也可以发现更多的匹配。将以上这样的三元组记为(C_1, C_2, C_3)。对于不同本体中的概念，这里将以上这些构成同义词的三元组(C_1, C_2, C_3) 作为模型的训练数据集，基于TransR的思想，不同本体的概念在不同的向量空间中，引入两个映射矩阵实现不同本体概念的向量空间之间的映射，设计表示学习模型，模型损失函数定义如

公式；

$$l_{RE} = -\sum_{(C_1,C_2,C_3)\in\gamma} \log f_{RE}(C_1,C_2,C_3)) - \sum_{(C_1{}',C_2{}',C_3)\in\gamma'} \log f_{RE}(C_1{}',C_2{}',C_3))) \quad (3.1.5)$$

这里γ表示由同义词三元组(C_1,C_2,C_3)所构成的集合，γ'表示随机替换C_1或者C_2后所构成的三元组集合$\{(C_1{}',C_2{}',C_3)\}$，$f_{RE}(C_1,C_2,C_3)$表示等式所示用于计算投影后概念间相似度的得分函数，这里，$\mathbf{C_1},\mathbf{C_2},\mathbf{C_3}\in\mathbf{R^d}$分别表示$O_1,O_2,O_3$中概念$C_1,C_2,C_3$的d维连续向量表示。$\mathbf{M_{13}}$和$\mathbf{M_{23}}$ 分别表示将概念C_1,C_2投影到第三方本体O_3中的映射矩阵，旨在实现将O_1和O_2中相似的概念投影到O_3中与他们相似的概念附近，相反，不相似的概念应该存在一定的距离，语义相反的词距离尽可能远。

$$f_{RE}(C_1,C_2,C_3) = 2\cdot\frac{1}{1+e^{(\|\mathbf{C_1}*\mathbf{M_{13}}-\mathbf{C_3}\|_2+\|\mathbf{C_2}*\mathbf{M_{23}}-\mathbf{C_3}\|_2)}} \quad (3.1.6)$$

两个映射矩阵非常重要，实现了不同本体间的语义关联，为了获得更好的映射矩阵，我们保持O_3中概念的向量表示不变，这样有助于缩减模型需学习的参数量，只更新两个映射矩阵以及O_1,O_2中概念的参数。这里需先将O_3中的结构关系作为数据集，利用预训练模型得到O_3中概念的向量表示，来保证O_3中概念的表示已包含其结构信息。因现有的预训练模型生成的表示稀疏性不好，无法体现O_3本体中的结构信息，所以，本文设计了如下的损失函数来构建表示学习模型以便获得更好的第三方本体的概念表示，如公式：

$$l_{RE} = -\sum_{(C_{31},r,C_{32})\in\lambda} \log f_r(C_{31},C_{32})) - \sum_{(C_1{}',r,C_3)\in\lambda'} \log f_r(C_{31}{}',C_{32}{}'))) \quad (3.1.7)$$

其中，评分函数f_r如公式所示：

$$f_r(C_{31},C_{32}) = 2\cdot\frac{1}{1+e^{(\|\mathbf{C_{31}}-\mathbf{C_{32}}\|_2-a)}} \quad (3.1.8)$$

这里λ表示由O_3中的结构关系三元组$(C_{31},SubClassOf,C_{32})$，$(C_{31},PartOf,C_{32})$构成的集合，$\lambda'$ 是随机替换C_{31}或C_{32}得到的负例样本的集合，得分函数$f_r(C_{31},C_{32})$表示在关系r中两个满足关系r的概念C_{31}和C_{32}之间的得分，$\mathbf{C_{31}}$，$\mathbf{C_{32}}$对应C_{31}和C_{32}的向量表示，值得注意额是这里的SubClassOf以及PartOf关系并不是等价关系，所以，这里利用超参数a来控制概念向量之间的语义距离。

3.1.6 多视角模型的匹配算法

在得到不同视角模型获得的匹配后，最终需要将它们进行组合。一个简单的策略是从这些模块中收集所有匹配，并用一个阈值或稳定的结合算法进行过滤[39]。

尽管这种策略最终可以获得很多正确匹配，但也可能引入许多错误的匹配以及容易忽略一对多，多对一或多对多的匹配[34]。因此，我们提出了一种基于相互评估的组合匹配算法。

算法 1 多视角匹配算法

输入:

参数$\delta_1, \delta_2, \delta_3, \delta_4$的值;

输出:

本体匹配的最终匹配对;

1: 通过相似度取极大来合并来自MOM-S和MOM-R的匹配。其合并结果被标记为MOM-SR;
2: 根据相应的阈值δ_1和δ_2选择MOM-L和MOM-SR的可靠匹配（相似度高于阈值即为可靠匹配）;
3: 如果一个可靠匹配属于MOM-L，且在MOM-SR中的相似性低于阈值δ_3，则将其删除;
4: 若一个可靠匹配属于MOM-SR，且其在MOM-L中的相似性低于阈值δ_4，则删除此匹配;
5: 对来自MOM-L和MOM-SR的匹配取并，生成最终匹配;
6: **return** 本体匹配的最终匹配对;

3.2 实验与评估

3.2.1 数据集

本节简要概述在本体匹配实验中使用的四种本体，其中两个本体FMA(Foundational Model of Anatomy)和MA(Adult Mouse anatomical)分别表示解剖学本体和成年小鼠解剖学本体，二者都是纯粹的解剖学本体论，而另外两个SNOMED C-T和NCI是涉及更广的生物医学本体，解剖学只是它们描述的一个子领域。虽然这些本体的最新版本是可用的，但是本文在整个本体匹配任务中参考了出现在OAEI(Ontology Alignment Evaluation Initiative) 中的版本，以便在本体匹配系统之间进行比较。

FMA(Foundational Model of Anatomy)：这是一个一直被更新的本体，自1994年由华盛顿大学开发并维护[49]，其目的是以机器可读的形式概念化人体的表型结构。

MA(Adult Mouse Anatomical Dictionary)：该本体是描述成年小鼠解剖结构的结构的词汇表[22]。

NCI(NCI Thesaurus)：该本体提供了癌症的标准词汇[9]，其解剖学子域描述人类自有的生物结构、液体和物质。

SNOMED(SNOMED Clinical Terms)：该本体是一个系统化，组织化的机器可读的医学术语集合，提供临床文档和报告中使用的代码、术语、同义词和定义[13]。

以上本体的概念及对应匹配数量具体如图4.2：

表 3.1 本体概念数及对应匹配数量

源本体	概念数	目标本体	概念数	匹配数
MA	2744	NCI	3304	1489
FMA	3696	NCI	6488	2504
FMA	10157	SNOMED	13412	7774

3.2.2 评估指标

OAEI为本体匹配提供了统一的评估标准，评估的指标主要是：匹配的准确率，召回率和F1值。为阐述以上三个指标含义，首先介绍以下概念：

TP(True Positives) ：表示模型将正例预测为正例的数据条数；

FP(False Positives)：表示模型将负例预测为正例的数据条数；

FN(False Negatives)：表示模型将正例预测为负例的数据条数；

TN(True Negatives) ：表示模型将负例预测为负例的数据条数；

所以，由以上四个参数值，可根据以下公式求得具体的指标值，其中total表示待预测数据的总数。

准确率(Acc)：$Acc = \frac{TP+TN}{\text{total}}$

召回率(Rec)：$Rec = \frac{TP}{TP+FN}$

F1 值：$F1 = \frac{2*Acc*Rec}{Acc+Rec}$

3.2.3 对比方法

表 3.2 实验对比方法表

模型名称	模型含义
MOM	本文提出的本体匹配模型（Multi-view Ontology Matching Model）效果
TFIDF	采用TFIDF算法计算单词权重的基于字符串等价的匹配效果
MOM-S	MOM模型中基于结构负采样视角的子模型效果
MOM-L	MOM模型中基于上下文视角的子模型效果
MOM-R	MOM模型中基于外部资源视角的子模型效果
MOM-SR	MOM基于外部资源模型的效果与结构负采样模型效果的融合效果
MOM$^-$	MOM模型未采用结构负采样的结果
StringEquiv	基于字符串等价的基础匹配器效果
StringEquiv-N	采用概念规范化的StringEquiv匹配效果
StringEquiv-S	采用同义词外部资源的StringEquiv匹配效果
StringEquiv-SR	采用同义词，第三方本体外部资源的StringEquiv匹配效果
StringEquiv-NS	采用字典，第三方本体外部资源的StringEquiv匹配效果
StringEquiv-NSR	采用同义词，字典，第三方本体外部资源的StringEquiv匹配效果

3.2.4 实验环境

为了验证本文工作的有效性，本文使用Python借助TensorFlow[1]深度学习框架来实现本文提出的方法。采用OWLAPI[2]（用于管理OWL本体的工具）本体解析工具来获得本体信息。实验是在具有64GB内存和TiTAN XP GPU 的Intel Xeon E5-2630 V4

[1]https://www.tensorflow.org/

[2]http://owlapi.sourceforge.net/

CPU的个人工作站上进行的。本文中MultiOM的源码和实验数据集可以由[3]下载得到。

3.2.5 实验结果

针对现有方法的两个问题，本文设计如下实验，逐步验证本文建模思想是否正确，是否可以有效解决现有方法的两个问题，具体实验步骤由两部分构成，分别是采用本体结构信息的MOM方法与基于字符串方法的对比实验，MOM多视角结合与单个预训练模型的结果对比实验。

本体概念中单词向量表示主要来自文献[32]的链接[4]，其维数设置为200。对于一些没有向量表示的单词，我们将其随机初始化，并满足$||t_{1i}||_2 \leq 1$和$||t_{2i}||_2 \leq 1$。

1.采用本体结构信息的MOM方法与基于字符串方法的对比实验

本实验采用MA-NCI-FMA和FMA-NCI-small-SNOMED两个数据集，将本文的MOM，MOM$^-$方法与StringEquiv，StringEquiv-N，StringEquiv-S，StringEquiv-SR，StringEquiv-NS，StringEquiv-NSR进行对比，以上方法中唯一的变量因素为有没有使用本体的结构信息，通过该实验来验证本体的结构信息是否有助于提升本体匹配的效果。其中，MOM和MOM$^-$方法训练时共享同组参数，具体参数如表3.2.5所示。

表 3.3 MOM模型训练时的参数配置表

参数	参数值
d	$\{50, 100\}$
d_M	$\{50 \times 50, 100 \times 100\}$
$nbatch$	$\{5, 10, 20, 50\}$
r	$\{0.01, 0.02, 0.001\}$
$nrate$	$\{1, 3, 5, 10\}$
$epoch$	1000
a	$\{0.01, 0.05, 0.10\}$
$\{\delta_1, \delta_2, \delta_3, \delta_4\}$	$\{0.8, 0.95, 0.65, 0.3\}$

实验结果如表3.4，3.5所示，表中列出了MOM以及多种字符串方法的结果对比表。通过对比可以发现,MOM的匹配效果要比基于字符串的基准匹配系统StringEquiv-NSR的效果无论是准确率，召回率还是F值都要高。两者的差别在于MOM采用了本体的结构信息，而基于字符串的匹配系统未采用结构信息。

同样，通过MOM与MOM$^-$的对比也可以发现,MOM较MOM$^-$加入了基于结构的负采样，表中MOM的匹配效果也优于MOM$^-$，所以，以上结果表明，充分考虑本体

[3]https://github.com/chunyedxx/MultiOM
[4]https://doi.org/10.5281/zenodo.1173936

的结构信息有助于提升本体匹配的效果。

表 3.4 MOM与基于字符串的基准匹配系统在MA-NCI中的结果对比表

Methods	MA-NCI				
	Number	Correct	P	R	F1
StringEquiv	935	932	0.997	0.615	0.761
StringEquiv-N	992	989	0.997	0.625	0.789
StringEquiv-S	1100	1057	0.961	0.697	0.808
StringEquiv-SR	1162	1094	0.941	0.722	0.817
StringEquiv-NS	1153	1109	0.962	0.732	0.831
StringEquiv-NSR	1211	1143	0.943	0.753	0.838
MOM$^-$	**1484**	**1296**	**0.873**	**0.855**	**0.864**
MOM	**1445**	**1287**	**0.891**	**0.849**	**0.869**

表 3.5 MOM与基于字符串的基准匹配系统在FMA-NCI-small中的结果对比表

Methods	FMA-NCI-small				
	Number	Correct	P	R	F1
StringEquiv	1501	1389	0.925	0.517	0.663
StringEquiv-N	1716	1598	0.931	0.595	0.726
StringEquiv-S	2343	2082	0.889	0.775	0.828
StringEquiv-SR	2343	2082	0.889	0.775	0.828
StringEquiv-NS	2464	2200	0.893	0.819	0.854
StringEquiv-NSR	2467	2203	0.893	0.820	0.855
MOM$^-$	**2471**	**2192**	**0.887**	**0.809**	**0.846**
MOM	**2470**	**2195**	**0.889**	**0.817**	**0.851**

2.MOM多视角结合与单个预训练模型的结果对比实验

表 3.6 MOM分别将不同视角的模型组合后的效果对比表

Methods	Number	Correct	P	R	F1
TFIDF (threshold= 0.8)	985	976	0.991	0.644	0.780
MOM-L (threshold= 0.8)	1286	1175	0.914	0.775	0.839
MOM-S^- (threshold= 0.95)	1836	1109	0.604	0.732	0.662
MOM-S (threshold= 0.95)	1189	1097	0.923	0.724	0.811
MOM-R (Random initialization, threshold= 0.95)	709	680	0.959	0.449	0.612
MOM-R (threshold= 0.95)	833	789	0.948	0.520	0.672
MOM-RS^- (threshold= 0.95)	1271	1147	0.902	0.757	0.823
MOM-RS (threshold= 0.95)	1237	1138	0.920	0.751	0.827
MOM-S^- (threshold= 0.95)	1821	1110	0.610	0.733	0.666
MOM$^-$	1484	1296	0.873	0.855	0.864
MOM	**1445**	**1287**	**0.891**	**0.849**	**0.869**

本实验针对MA-NCI-FMA数据集，将本文的MOM，MOM$^-$，MOM-S，MOM-L，MOM-R，MOM-RS方法进行对比，以上方法中唯一的变量因素为不同的方法对应的视角不同，通过该实验来验证不同的视角是否可以学习到不同方面的特征，多

视角的融合是否有助于提升本体匹配的效果。其中MOM和MOM⁻方法训练时共享同组参数，其他模型的参数也如表3.2.5所示。实验中也对比了对于同一视角的模型在不同的阈值下模型的匹配效果。

表 3.7 MOM不同视角的模型效果对比表

	MOM-L	MOM-S	MOM-R	MOM-LS	MOM-LR	MOM-SR
MOM-L	–	176	463	–	–	154
MOM-S	99	–	354	–	45	–
MOM-R	78	46	–	24	–	–

实验结果如表3.6，3.7所示，表3.6展示了MOM分别将不同视角的模型组合后的效果对比。对比数据可以发现，将任意的两个不同视角模型的匹配结果组合后，匹配效果会比单个视角的模型匹配效果要好，同时三个视角融合后的匹配效果即MOM 无论是正确率，召回率还是F 值都优于其他任意组合或模型的效果，这也说明不同视角的模型所学特征是存在差别的，不同方面的特征融合后，最终才有MOM 的整体效果的提升。

为进一步确定以上结论，本文通过表3.7中的结果进一步说明。表中的结果指将三个视角的模型匹配结果互相比较，并与他们的组合进行比较，可以发现，不同视角的模型总能找到不同于其他视角的匹配，这也进一步证明了不同视角的模型所学特征存在互补性，多视角的建模方式在本体匹配任务中是有效的。

4 基于HORD算法优化MOM超参数的MOMWH模型

本章主要介绍针对现有方法本文提出的问题问题三，本文设计的基于HORD算法对MOM的改进模型MOMWH，以下将从MOMWH模型的介绍，以及实验评估两个方面进行详细介绍。

4.1 基于HORD算法优化MOM超参数的MOMWH模型介绍

MOM模型中的子模型MOM-S是基于TransE构建的表示学习模型，MOM-R是基于TransR构建的表示学习模型，而无论是MOM-S，MOM-R模型，还是现有的TransE，TransR等表示学习模型得到的概念向量表示都非常依赖模型的参数。MOM中的MOM-S和MOM-R模型需调节的超参数包括训练轮数($epoch$)，学习率(r，Learning rate)，向量维数(d，Dimensions)，批次数($nbatch$)，以及负采样率($nrate$，Negative samping rate)，这些参数既包含整数也包含连续数；同时，模型本身的参数量较大，训练的时间成本也比较高，为进一步获得更好的参数，从而改进待匹配概念的表示并提高MOM本体匹配模型的效果，本节考虑采用HORD算法对MOM中的表示学习的超参数进行优化。

HORD算法较现有的贝叶斯类超参数优化方法，如GP[53],TPE[3]等方法，可以以更少的估值代价更加高效的同时优化整数或连续实数类超参数。而MOM中表示学习模型效果依赖于参数的选择，且参数既包含整数也包含连续实数，模型训练的时间成本也较高，而HORD算法较其他方法可以节省估值次数，同时优化整数以及连续实数超参，所以HORD算法是适合用于实现MOM模型超参数优化的方法。

为实现HORD算法对MOM中的表示学习模型参数的优化，进一步优化概念的向量表示，来提升MOM模型的性能，本节提出了如图1.3的MOMWH模型框架。因基于上下文视角的预训练模型采用预训练好的词向量，并没有涉及参数优化，所以HORD算法只针对基于MOM-R（外部资源视角的预训练模型）以及基于MOM-S （结构视角的预训练模型）进行参数优化。MOMWH与MOM的不同之处在于MOM采用经验调节参数，来得到概念的向量表示，从而实现本体的匹配，而MOMWH采用HORD算法对MOM-R模型与MOM-S模型的超参数进行优化，将模型采用优化后的参数训练得到的向量作为概念的表示，进行后续的本体匹配。

采用HORD优化MOM中表示学习超参进而实现本体匹配的大致步骤如下：以MOM-R为例，第一步，设置HORD算法的参数，初始采样点的个数n_0，最大迭代步数N_{max}，以及确定模型各个参数可行域D；第二步，通过拉丁超立方体采样获得初始点$\{x_i\}_{i=1}^{n_0}$列，这里x_i 表示一组向量，即一组参数；第三步，采用每一组参数x_i，以MOM中的数据集训练模型MOM-R，得到点对$\mathrm{A}_{n_0} = \{\{x_i, f(x_i)\}_{i=1}^{n_0}$，$f(x_i)$表

示模型训练后得到的验证误差值（实际可以是损失，错误率等）；第四步，采用A_{n_0}更新径向基响应面模型；第五步，通过当前响应面模型求得最优点x_*，再采用参数x_*训练MOM-R模型得到$f(x_*)$，更新$\mathrm{A}_{n_0} = \mathrm{A}_{n_0} \cup \{(x_*, f(x_*)\}$；第六步在未达到最大步数$N_{max}$时，重复步骤四至步骤五；最后得到改进后的参数以及概念的向量表示，再采用MOM中的多视角匹配算法得到最终的匹配。具体算法过程如算法2所示。

算法 2 使用HORD算法优化MOM中表示学习模型超参的算法框架

输入:

初始采样数$n_0 = 2(D+1)$;
候选点个数$m = 100D$;
最大迭代步$N_{\max}$;
初始迭代步$n = n_0$;

输出:

优化后的模型参数配置x_{better};
x_{better}参数配置下的概念表示;

1: 使用拉丁超立方体采样n_0个初始超参配置，记$\{x_i\}_{i=1}^{n_0}$;
2: 采用x_i训练模型，得到验证误差$f(x_i)$，$i = 1, \ldots, n_0$，记$\mathrm{A}_{n_0} = \{\{x_i, f(x_i)\}$;
3: **while** $n < N_{\max}$ **do**
4: 使用A_n构建径向基响应面模型$S_n(x)$;
5: 计算当前最优点$x_{better} = argmin\{f(x_i) : i = 1, ..., n\}$;
6: 计算坐标扰动的概率φ_n;
7: 构建由m个候选点$t_{n1:m}$组成的候选点集;
对$t_{n1:m}$中每个候选点y_i:
令$y_i = x_{better}$;
按扰动概率φ_n扰动y_i对应分量;
对被扰动分量添加扰动δ_i ($\delta_i \sim N(0, \sigma^2)$);
8: 根据候选点计算$V_n^{ev}(t_{n1:m})$，$V_n^{dn}(t_{n1:m})$，$W_n(t_{n1:m})$;
9: 根据当前$W_n(t_{n1:m})$，计算当前最优点$x_* = argmin\{W_n(t_{n1:m})\}$;
10: 根据当前最优参数配置x_*，训练模型，得到当前验证误差$f(x_*)$;
11: 调整方差δ_i;
12: 更新点值对$\mathrm{A}_{n_0} = \mathrm{A}_{n_0} \cup \{(x_*, f(x_*)\}$;
13: **end while**;
14: **return** x_*，x_*参数配置下，概念向量表示;

其中，扰动概率φ_n（对应步骤6）是关于迭代次数$n_0 \le n \le N_{\max} - 1$的严格递减函数，其计算表达式为：

$$\varphi_n = \varphi_0[1 - \frac{\ln(n - n_0 - 1)}{\ln(N_{\max} - n_0)}] n_0 \le n \le N_{\max}$$

在步骤7和步骤8旨在从候选点中选出最可能的极小点作为新的采样点，通过计算每个候选点的$V^{ev}(t)$和$V^{dm}(t)$作为其度量标准，函数定义如下：

$$V^{ev}(t) = \begin{cases} \frac{S(t) - S^{\min}}{S^{\max} - S^{\min}} & S^{\max} \neq S^{\min} \\ 1 & otherwise \end{cases}$$

$$V^{dm}(t) = \begin{cases} \frac{\Delta^{\max} - \Delta(t)}{\Delta^{\max} - \Delta^{\min}} & \Delta^{\max} \neq \Delta^{\min} \\ 1 & otherwise \end{cases}$$

其中$\Delta(t) = \min\limits_{1 \le i \le n} \|t - t_i\|$，候选点最终的得分为$V^{dm}(t)$和$V^{ev}(t)$的加权和，具体如公

式4.1.1所示,最后取得分最低的候选点为新的采样点。

$$W(t) = \omega V^{ev}(t) + (1 - \varpi)V^{dm}(t) \tag{4.1.1}$$

4.2 实验与评估

4.2.1 数据集

本节简要概述在本体匹配实验中使用的四种本体，其中两个本体FMA(Foundational Model of Anatomy)和MA(Adult Mouse anatomical)分别表示解剖学本体和成年小鼠解剖学本体，二者都是纯粹的解剖学本体论，而另外两个SNOMED C-T和NCI是涉及更广的生物医学本体，解剖学只是它们描述的一个子领域。虽然这些本体的最新版本是可用的，但是本文在整个本体匹配任务中参考了出现在OAEI(Ontology Alignment Evaluation Initiative) 中的版本，以便在本体匹配系统之间进行比较。

FMA(Foundational Model of Anatomy)：这是一个一直被更新的本体，自1994年由华盛顿大学开发并维护[49]，其目的是以机器可读的形式概念化人体的表型结构。

MA(Adult Mouse Anatomical Dictionary)：该本体是描述成年小鼠解剖结构的结构的词汇表[22]。

NCI(NCI Thesaurus)：该本体提供了癌症的标准词汇[9]，其解剖学子域描述人类自有的生物结构、液体和物质。

SNOMED(SNOMED Clinical Terms)：该本体是一个系统化，组织化的机器可读的医学术语集合，提供临床文档和报告中使用的代码、术语、同义词和定义[13]。

以上本体的概念及对应匹配数量具体如图4.2：

表 4.1 本体概念数及对应匹配数量

源本体	概念数	目标本体	概念数	匹配数
MA	2744	NCI	3304	1489
FMA	3696	NCI	6488	2504
FMA	10157	SNOMED	13412	7774

4.2.2 评估指标

OAEI为本体匹配提供了统一的评估标准，评估的指标主要是：匹配的准确率，召回率和F1值。为阐述以上三个指标含义，首先介绍以下概念：

TP(True Positives) ：表示模型将正例预测为正例的数据条数；

FP(False Positives)：表示模型将负例预测为正例的数据条数；

FN(False Negatives)：表示模型将正例预测为负例的数据条数；

TN(True Negatives) ：表示模型将负例预测为负例的数据条数；

所以，由以上四个参数值，可根据以下公式求得具体指标的值，其中total表示待预测数据的总数。

准确率(Acc)：$Acc = \frac{TP+TN}{\text{total}}$

召回率(Rec)：$Rec = \frac{TP}{TP+FN}$

F1 值：$F1 = \frac{2*Acc*Rec}{Acc+Rec}$

4.2.3 实验环境

为了验证本文工作的有效性，本文使用Python借助TensorFlow[1]深度学习框架来实现本文提出的方法。采用OWLAPI[2]（用于管理OWL本体的工具）本体解析工具来获得本体信息。实验是在具有64GB内存和TiTAN XP GPU 的Intel Xeon E5-2630 V4 CPU的个人工作站上进行的。

4.2.4 对比方法

表 4.2 实验对比方法表

模型名称	模型含义
MOM-S	MOM模型中基于结构负采样视角的子模型效果
MOM-L	MOM模型中基于上下文视角的子模型效果
MOM-R	MOM模型中基于外部资源视角的子模型效果
MOM-SR	MOM基于外部资源模型的效果与结构负采样模型效果的融合结果
MOMWH-S	MOMWH模型中基于结构负采样视角的子模型效果
MOMWH-L	MOMWH模型中基于上下文视角的子模型效果
MOMWH-R	MOMWH模型中基于外部资源视角的子模型效果
MOMWH-SR	MOMWH基于外部资源模型的效果与结构负采样模型效果的融合结果
MOM	本文提出的本体匹配模型（Multi-view Ontology Matching Model）效果
MOM$^-$	MOM模型未采用结构负采样的结果
MOM$^+$	在MOM基础上采用BERT模型改进其预训练词向量表示的模型效果
MOMWH$^-$	MOMWH模型未采用结构负采样的结果
MOMWH	采用HORD算法对MOMM超参进行优化的模型效果
MOMWH$^+$	在MOM基础上采用BERT模型改进其预训练词向量表示的模型效果

4.2.5 实验结果

针对现有方法的本文提出的第三个问题，本文设计如下实验，逐步验证MOMWH得建模思想是否正确，是否可以有效解决现有方法的这个问

[1]https://www.tensorflow.org/

[2]http://owlapi.sourceforge.net/

题，具体实验步骤由两部分构成，分别是MOMWH与MOM的效果对比实验和MOMWH$^+$与MOMWH，MOM及当前最好方法的效果对比实验。

本体概念中单词向量表示主要来自文献[32]的链接[3]，其维数设置为200。对于一些没有向量表示的单词，我们将其随机初始化，并满足$||t_{1i}||_2 \leq 1$和$||t_{2i}||_2 \leq 1$。

1.MOMWH与MOM的效果对比实验

本实验针对MA-NCI-FMA和FMA-NCI-small-SNOMED两个数据集，将本文的MOMWH，MOMWH$^-$，MOM，MOM$^-$以及MOMWH和MOM中多个不同视角的模型，将MOMWH与MOM中相对应的模型进行对比，以上方法中唯一的变量因素为MOM中的模型参数是人工调节的参数，MOMWH中模型参数一部分为采用HORD算法优化得到的参数，通过该实验来验证HORD算法是否可以改进表示学习模型学得的概念向量表示，是否可以因此提升本体匹配的效果。其中，MOM方法训练及其匹配算法的参数如图3.2.5，MOMWH模型训练时的参数如图4.2.5，实验中也对比了对于同一视角的模型在不同的阈值下模型的匹配效果。

该实验首先需采用HORD算法对MOM中的表示学习模型MOM-S，MOM-R进行超参数优化。HORD算法的执行与MOM-L或MOM-R训练的方式有所不同，需给出模型（MOM-S，MOM-R类模型）中待优化超参数的区间范围，而不用给定具体的值。对于模型中没有进行进一步优化的参数，则采用MOM中对应模型相同的参数。

表 4.3 MOMWH模型中HORD超参调节的参数设置表

参数	参数值
d	$\{50, 100\}$
d_M	$\{50 \times 50, 100 \times 100\}$
$nbatch$	$[10, 60]$
r	$[0.001, 0.05]$
$nrate$	$[3, 15]$
$epoch$	500
n_0	6
N_{max}	25

MOM-S和MOM-R的超参数包括训练轮数($epoch$)，学习率(r)，向量维数(d)，批次数($nbatchs$)，以及负采样率($nrate$)。为降低参数调节的计算量，提升MOMWH模型效率，本文未对以上六种参数都进行优化。首先，MOM中的表示学习模型的参数规模由向量维数大致确定，而参数的规模应与本体数据集中的概念以及关系的多少有关，在二者数量不变的前提下，可大致确定向量的维数(d)，保证可以适用于匹配任务即可。所以，MOM中未对这两个参数进行调节；其次，训练轮数($epoch$)根据人工经验

[3]https://doi.org/10.5281/zenodo.1173936

得到的值为1000，该参数决定了模型学习时间的长短，为保证参数优化后，可以提升MOM的的效率，我们将其固定为500，通过调节剩余的三个参数：学习率（r），批次数（$nbatchs$）和负采样率($nrate$)。所以MOMWH模型的参数设置如表4.2.5 所示。

MOMWH中通过HORD算法进行超参优化后，MOM-S与MOM-R的最优超参配置对比如表4.2.5所示：

表 4.4 HORD算法进行超参数优化前后的参数对比表

参数名	MOM-S		MOM-R	
	优化前	优化后	优化前	优化后
$nbatch$	5	8	10	12
r	0.0100	0.0131	0.0100	0.0107
$nrate$	3	6	5	8
$epoch$	1000	500	1000	500

除了以上经过HORD算法优化过的参数外，其他参数依然与原模型中参数相同。MOMWH与MOM的效果对比如表4.5，4.6，4.7所示。

表 4.5 MOMWH与MOM在MA-NCI下的效果对比表

Methods	MA-NCI				
	Number	Correct	P	R	F1
MOM$^-$	1484	1296	0.873	0.855	0.864
MOM	1445	1287	0.891	0.849	0.869
MOMWH$^-$	1462	1296	0.886	0.855	0.870
MOMWH	1431	1290	0.901	0.851	0.0.876

表 4.6 MOMWH与MOM在FMA-NCI-small下的效果对比表

Methods	FMA-NCI-small				
	Number	Correct	P	R	F1
MOM$^-$	2471	2192	0.887	0.809	0.846
MOM	2470	2195	0.889	0.817	0.0.851
MOMWH$^-$	2469	2192	0.888	0.809	0.847
MOMWH	2466	2200	0.892	0.819	0.854

如表4.2.5所示，采用HORD算法进行超参优化后与优化前的模型参数对比，对比可以发现，超参数优化后，对于同样的模型之前训练需要1000个$epoch$，而优化后只需要500个$epoch$，较之前只需要一半的训练次数，极大的提升了模型的效率。

对比表4.5，4.6中的数据，MOMWH无论是准确率，召回率还是F值都较之前的MOM都有所提升，MOMWH$^-$的性能也优于MOM$^-$。MOMWH是基于MOM采用HORD算法进行超参优化的模型，显然，经过优化后的参数，可以学到更加准确的特征；另外，无论是MOM与MOM$^-$还是MOMWH与MOMWH$^-$ 都说明同利用了改进

后的结构负采样，效果依然会有所提高，也进一步验证了本体的结构信息有助于本体匹配任务效果的提升。

表 4.7 MOMWH与MOM不同视角模型及其组合后在MA-NCI下的效果对比表

Methods	Number	Correct	P	R	F1
MOM-S^- (threshold= 0.95)	1836	1109	0.604	0.732	0.662
MOM-S (threshold= 0.95)	1189	1097	0.923	0.724	0.811
MOM-R (Random initialization, threshold= 0.95)	709	680	0.959	0.449	0.612
MOM-R (loss function, threshold= 0.95)	833	789	0.948	0.520	0.672
MOM-RS^- (threshold= 0.95)	1271	1147	0.902	0.757	0.823
MOM-RS (threshold= 0.95)	1237	1138	0.920	0.751	0.827
MOM$^-$	1484	1296	0.873	0.855	0.864
MOM	1445	1287	0.891	0.849	0.869
MOMWH-S^- (threshold= 0.95)	1821	1110	0.610	0.733	0.666
MOMWH-S (threshold= 0.95)	1170	1099	0.939	0.725	0.818
MOMWH-R (Random initialization, threshold= 0.95)	705	680	0.965	0.449	0.613
MOMWH-R (loss function, threshold= 0.95)	824	794	0.964	0.524	0.679
MOMWH-RS^- (threshold= 0.95)	1263	1147	0.908	0.757	0.826
MOMWH-RS (threshold= 0.95)	1221	1140	0.934	0.752	0.833
MOMWH$^-$	1462	1296	0.886	0.855	0.870
MOMWH	1431	1290	0.901	0.851	0.876

表4.7展示了MOMWH与MOM分别将不同视角的模型组合后的效果对比。通过各个视角模型的匹配结果详细对比可以发现，采用HORD算法进行参数优化后，基本上对各个视角的匹配结果都有所提升。众多的结果说明，经过参数优化后，不仅可以使得各个模型错误的匹配得以减少同时可以寻找到一些正确匹配，以上结果进一步说明参数优化后可以帮助表示学习模型学得更加合适得特征，获得得概念向量表示更加有效。

2.MOMWH$^+$与MOMWH，MOM及当前最好方法的效果对比实验

本实验依然针对实验一中的两个数据集，将本文的MOMWH$^+$等方法与当前最好的方法:SCBOW+DAE[31]，LogMapBio[14]，POMAP++[14]，XMap[15]，LogMap[28] FCAMapX[64]，SANO-M[14]，的匹配效果进行比较。根据之前的实验，发现利用改进后的TF-IDF算法的效果准确率有0.991, 其召回率不高，主要是因为我们利用的是预训练词向量，只包含其字面语义信息，并未对这些预训练词向量基于当前语境进行改进，所以而未包含该本体匹配任务下的语义信息。考虑到以上原因，本文基于MOM框架，提出采用BERT模型对其预训练词向量进行进一步改进的MOM$^+$，并采用HORD算法进行超参数优化的MOMWH$^+$模型，具体的对比结果如表4.8 所示。

通过比较表4.8，4.9中的数据可以发现，MOMWH$^+$的效果较MOMWH和MOM有了较大幅度的提升，并且从召回率与F值来综合比较，MOMWH$^+$的效果比LogMapBio，POMAP++等方法都要有优势。另外与第一名的SCBOW+DAE相比，

在FMA-NCI-small-SNOMED数据集中的效果，也与之可以相提并论。

表 4.8 本文模型与当前最好方法在MA-NCI下的效果对比表

Methods	MA-NCI				
	Number	Correct	P	R	F1
SCBOW + DAE	1399	1356	0.969	0.906	0.938
MOMWH$^+$	**1436**	**1363**	**0.949**	**0.891**	**0.919**
MOM$^+$	**1441**	**1363**	**0.946**	**0.891**	**0.918**
LogMapBio	1550	1376	0.888	0.908	0.898
POMAP++	1446	1329	0.919	0.877	0.897
XMap	1413	1312	0.929	0.865	0.896
LogMap	1387	1273	0.918	0.846	0.880
SANOM	1450	1287	0.888	0.844	0.865
FCAMapX	1274	1199	0.941	0.791	0.859
MOM	**1445**	**1287**	**0.891**	**0.849**	**0.869**

表 4.9 本文模型与当前最好方法在FMA-NCI-small下的对比表

Methods	FMA-NCI-small				
	Number	Correct	P	R	F1
SCBOW + DAE	2282	2227	0.976	0.889	0.930
MOMWH$^+$	**2787**	**2647**	**0.950**	**0.907**	**0.928**
MOM$^+$	**2793**	**2647**	**0.948**	**0.907**	**0.927**
LogMapBio	2776	2632	0.948	0.902	0.921
POMAP++	2414	2363	0.979	0.814	0.889
XMap	2315	2262	0.977	0.783	0.869
LogMap	2747	2593	0.944	0.897	0.920
SANOM	–	–	–	–	–
FCAMapX	2828	2681	0.948	0.911	0.929
MOM	**2330**	**2195**	**0.942**	**0.817**	**0.875**

综上实验，可以进一步验证说明本文提出的方法，可以在一定程度上解决现有基于深度学习的方法的问题，本文充分利用本体中的结构信息，并且对本文提出的基于表示学习的方法MOM等模型进行超参数优化，以四个不同的实验，进一步验证了充分利用本体的结构信息有助于提升本体匹配的效果，采用多视角的建模方式是实现本体匹配任务的有效方式之一，以及采用HORD算法对表示学习模型的超参数进行优化，可以改进其获得概念向量表示，从而提升本体匹配的效果，本文模型MOMWH$^+$，MOMWH也取得了比较好的效果。

5 结论及展望

5.1 本文工作总结

本文针对现有基于深度学习的本体匹配方法存在的两个问题，现有方法值利用了本体概念的字面含义，而未对本体的结构信息进行有效利用，从而影响了本体匹配的效果；现有方法中的表示学习方法获得的概念表示受到参数的影响，进而影响匹配的效果。未解决以上问题，本文采用多视角的建模思想，有效利用本体的结构信息提出了MOM 模型，同时，也利用HORD算法实现了对MOM中表示学习模型的超参数优化，进一步提升本体匹配的效果，并且，本文提出的方法与现有方法相比，也取得了非常好的效果。

5.2 未来工作展望

虽然本文提出的方法可以一定程度上解决现有方法的两个问题，也可以取得比较好的效果，但对于现有方法的两个问题，本文只是在一定程度上克服了，但对于结构较深的本体，可如何选择合适的本体的结构信息，也是有待解决的问题之一。同时，表示学习的超参数调节，本文采用HORD算法进行了参数由优化，但只是对其中三个参数的优化，而如何针对表示学习类模型，提出针对性的的，效率较高的超参数优化算法，也是未来值得继续开展研究的方向。

参考文献

[1] Sivan Albagli, Rachel Ben-Eliyahu-Zohary, and Solomon Eyal Shimony. Markov network based ontology matching[J]. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009.*, pages 1884–1889, 2009.

[2] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives[J]. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization[C]. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems,2011*, pages 2546–2554, 2011.

[4] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data[C]. In *Proceedings of the 27th Advances in Neural Information Processing Systems Annual Conference,2013*, pages 2787–2795, 2013.

[5] Michelle Cheatham and Pascal Hitzler. String similarity metrics for ontology alignment[C]. In *Proceedings of the 12th International Semantic Web Conference, 2013*, volume 8219, pages 294–309, 2013.

[6] Alexandros Chortaras, Giorgos B. Stamou, and Andreas Stafylopatis. Learning ontology alignments using recursive neural networks[C]. In *Proceedings of the 15th Artificial Neural Networks: Formal Models and Their Applications International Conference,2005.*, volume 3697, pages 811–816, 2005.

[7] Carlo Curino, Giorgio Orsi, and Letizia Tanca. X-SOM: A flexible ontology mapper[C]. In *International Conference 18th International Workshop on Database and Expert Systems Applications,2007.*, pages 424–428, 2007.

[8] Carlo Curino, Giorgio Orsi, and Letizia Tanca. X-SOM: A flexible ontology mapper[C]. In *Proceedings of the 18th International Workshop on Database and Expert Systems Applications,2007.*, pages 424–428, 2007.

[9] Sherri de Coronado, Margaret W. Haber, Nicholas Sioutos, Mark S. Tuttle, and Lawrence W. Wright. NCI thesaurus: Using science-based terminology to integrate cancer research results[C]. In *Proceedings of the 11th World Congress on Medical Informatics, 2004*, volume 107, pages 33–37, 2004.

[10] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings[C]. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence,2018*, pages 1811–1818, 2018.

[11] Warith Eddine Djeddi and Mohamed Tarek Khadir. Introducing artificial neural network in ontologies alignment process[J]. *Control Cybernetics*, 41(4):743–759.

[12] AnHai Doan, Jayant Madhavan, Pedro M. Domingos, and Alon Y. Halevy. Ontology matching: A machine learning approach[J]. In *Handbook on Ontologies*, pages 385–404. 2004.

[13] Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for e-health[J]. *Studies in health technology and informatics*, 121:279–90, 02 2006.

[14] Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative[J]. *Journal of Biomedical Semantics*, 8(1):56.

[15] Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative[J]. *J. Biomedical Semantics*, 8(1):56:1–56:28, 2017.

[16] Floriana Esposito, Nicola Fanizzi, and Claudia d'Amato. Recovering uncertain mappings through structural validation and aggregation with the moto system[C]. In *Proceedings of the 2010 ACM Symposium on Applied Computing ,2010.*, pages 1428–1432, 2010.

[17] Jérôme Euzenat and Petko Valtchev. Similarity-based ontology alignment in owl-lite[C]. In *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, 2004*, pages 333–337, 2004.

[18] Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F. Cruz, and Francisco M. Couto. The agreementmakerlight ontology matching system[C]. In *Proceedings of the On the Move to Meaningful Internet Systems,2013.*, volume 8185, pages 527–541, 2013.

[19] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-match: an algorithm and an implementation of semantic matching[J]. In *Semantic Interoperability and Integration*, volume 04391, 2005.

[20] Jorge Gracia and Kartik Asooja. Monolingual and cross-lingual ontology matching with CIDER-CL: evaluation report for OAEI 2013[C]. In *Proceedings of the 8th International*

Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference,2013., volume 1111, pages 109–116, 2013.

[21] Marko Gulic, Boris Vrdoljak, and Marko Banek. Cromatcher: An ontology matching system based on automated weighted aggregation and iterative final alignment[J]. *J. Web Semant.*, 41:50–71, 2016.

[22] Terry Hayamizu, Mary Mangan, John Corradi, James Kadin, and Martin Ringwald. The adult mouse anatomical dictionary: A tool for annotating and integrating data[J]. *Genome biology*, 6:R29, 02 2005.

[23] Wei Hu and Yuzhong Qu. Falcon-ao: A practical ontology matching system[J]. *J. Web Semant.*, 6(3):237–239, 2008.

[24] Jingshan Huang, Jiangbo Dang, Michael N. Huhns, and W.Jim Zheng. Use artificial neural network to align biological ontologies[J]. *Bmc Genomics*, 9(Suppl 2):S16–S16, 2008.

[25] Ryutaro Ichise. Machine learning approach for ontology mapping using multiple concept similarity measures[C]. In *Proceedings of the 7th International Conference on Computer and Information Science, 2008.*, pages 340–346, 2008.

[26] Ilija Ilievski, Taimoor Akhtar, Jiashi Feng, and Christine Annette Shoemaker. Efficient hyperparameter optimization for deep learning algorithms using deterministic RBF surrogates[C]. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence,2017*, pages 822–829, 2017.

[27] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix[C]. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing,2015*, pages 687–696, 2015.

[28] Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. Logmap: Logic-based and scalable ontology matching[C]. In *Proceedings of the International Semantic Web Conference,2011*, pages 273–288, 2011.

[29] Christopher D Manning Kevin Clark, Minh-Thang Luong and Quoc Le[J]. Semi-supervised sequence modeling with cross-view training. *In EMNLP*, 35(8):1914 - 1925, 2018.

[30] Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. Deepalignment: Unsupervised ontology matching with refined word vectors[C]. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2018.*, volume 1, pages 787–798, 2018.

[31] Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. Biomedical ontology alignment: an approach based on representation learning[J]. *J. Biomedical Semantics*, 9(1):21:1–21:20, 2018.

[32] Prodromos Kolyvakis, Alexandros Kalousis, Barry Smith, and Dimitris Kiritsis. Biomedical ontology alignment: an approach based on representation learning[J]. *J. Biomedical Semantics*, 9(1):21:1–21:20, 2018.

[33] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. Rimom: A dynamic multistrategy ontology alignment framework[J]. *IEEE Trans. Knowl. Data Eng.*, 21(8):1218–1232, 2009.

[34] Weizhuo Li, Xuxiang Duan, Meng Wang, Xiaoping Zhang, and Guilin Qi. Multi-view embedding for biomedical ontology matching[C]. In *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference,2019*, volume 2536, pages 13–24, 2019.

[35] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion[C]. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015*, pages 2181–2187, 2015.

[36] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. Generic schema matching with cupid[C]. In *Proceedings of the 27th International Conference on Very Large Data Bases, 2001*, pages 49–58, 2001.

[37] Ming Mao, Yefei Peng, and Michael Spring. Ontology mapping: as a binary classification problem[J]. *Concurr. Comput. Pract. Exp.*, 23(9):1010–1025, 2011.

[38] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching[C]. In *Proceedings of the 18th International Conference on Data Engineering,2002*, pages 117–128, 2002.

[39] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching[C]. In *Proceedings of the 18th International Conference on Data Engineering, 2002*, pages 117–128, 2002.

[40] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space[C]. In *Proceedings of the 1st International Conference on Learning Representations, 2013.*

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space[C]. In *Proceedings of the 1st International Conference on Learning Representations, 2013.*, 2013.

[42] Prasenjit Mitra, Natasha F. Noy, and Anuj R. Jaiswal. OMEN: A probabilistic ontology mapping tool[C]. In *Proceedings of the 4th International Semantic Web Conference,2005.*, volume 3729, pages 537–547, 2005.

[43] DuyHoa Ngo and Zohra Bellahsene. YAM++ : A multi-strategy based approach for ontology matching task[C]. In *Proceedings of the 18th International Conference Knowledge Engineering and Knowledge, 2012.*, volume 7603, pages 421–425, 2012.

[44] Mathias Niepert, Jan Noessner, Christian Meilicke, and Heiner Stuckenschmidt. Probabilistic-logical web data integration[C]. In *Proceedings of the 7th Reasoning Web.Semantic Technologies for the Web of Data,2011.*, volume 6848, pages 504–533, 2011.

[45] Yefei Peng, Paul W. Munro, and Ming Mao. Ontology mapping neural network: an approach to learning and inferring correspondences among ontologies[C]. In *Proceedings of the 5th International Workshop on Ontology Matching,2010.*, volume 689, 2010.

[46] Wei Hu Muhao Chen Lingbing Guo Qingheng Zhang, Zequn Sun and Yuzhong Qu[J]. Multiview knowledge graph embedding for entity alignment. *In Proceedings of IJCAI*, 2019.

[47] Lirong Qiu, Jia Yu, Qiumei Pu, and Chuncheng Xiang. Knowledge entity learning and representation for ontology matching based on deep neural networks[J]. *Cluster Computing*, 20(2):969–977, 2017.

[48] Rommel G. Regis and Christine A. Shoemaker. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization[J]. *Engineering Optimization*, 45(5):529–555, 2013.

[49] Cornelius Rosse and José L. V. Mejino Jr. A reference ontology for biomedical informatics: the foundational model of anatomy[J]. *J. Biomed. Informatics*, 36(6):478–500, 2003.

[50] Mariano Rubiolo, María Laura Caliusco, Georgina Stegmayer, Mauricio Coronel, and M. Gareli Fabrizi. Knowledge discovery through ontology matching: An approach based on an artificial neural network model[J]. *Inf. Sci.*, 194:107–119, 2012.

[51] G. Salton and Clement Yu. On the construction of effective vocabularies for information retrieval[J]. *ACM SIGIR Forum*, 9:48–60, 12 1974.

[52] Manjula Shenoy, Karthish Shet, and Dinesh Acharya. *NN Based Ontology Mapping[J]*, volume 296, pages 122–127. 01 2013.

[53] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms[C]. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems,2012*, pages 2960–2968, 2012.

[54] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction[C]. In *Proceedings of the 33nd International Conference on Machine Learning,2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080, 2016.

[55] Lucy Lu Wang, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, and Waleed Ammar. Ontology alignment in the biomedical domain using entity definitions and context[C]. In *Proceedings of the BioNLP workshop, 2018*, pages 47–55, 2018.

[56] Peng Wang, Yuming Zhou, and Baowen Xu. Matching large ontologies based on reduction anchors[C]. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011.*, pages 2343–2348, 2011.

[57] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications[J]. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.

[58] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes[C]. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014*, pages 1112–1119, 2014.

[59] Chuncheng Xiang, Tingsong Jiang, Baobao Chang, and Zhifang Sui. ERSOM: A structural ontology matching approach using automatically learned entity representation[C]. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,2015*, pages 2419–2429, 2015.

[60] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases[C]. In *Proceedings of the 3rd International Conference on Learning Representations, 2015.*, 2015.

[61] Meng Qu，Jian Tang，Jingbo Shang，Xiang Ren，Ming Zhang and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning[J]. *CoRR*, abs/1709.06636, 2017.

[62] Songmao Zhang and Olivier Bodenreider. Experience in aligning anatomical ontologies[J]. *Int. J. Semantic Web Inf. Syst.*, 3(2):1–26, 2007.

[63] Yuanzhe Zhang, Xuepeng Wang, Siwei Lai, Shizhu He, Kang Liu, Jun Zhao, and Xueqiang Lv. Ontology matching with word embeddings[C]. In *Proceedings of the 13th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data,2014.*, volume 8801, pages 34–45, 2014.

[64] Mengyi Zhao, Songmao Zhang, Weizhuo Li, and Guowei Chen. Matching biomedical ontologies based on formal concept analysis[J]. *J. Biomedical Semantics*, 9(1):11:1–11:27, 2018.

附录：作者攻读硕士学位期间发表论文及科研情况

发表的论文：

[1] G.Li, X.Duan and C.Wu. A New DC Algorithm for Sparse Optimal Scoring Problem[J]. IEEE Access, vol 8, pp 53962-53971, 2020.

[2] Li,W., Duan,X., Wang,M., Zhang,X., Qi,G. Multi-view Embedding for Biomedical Ontology Matching[C]. In Proceedings of the OM2019@ISWC, 2019.

参与的基金与课题：

1. 机器学习中大规模非凸优化问题的分布式并行算法研究及其应用、国家自然科学基金面上项目、参与、2019；
2. 具有可分结构非凸优化问题的理论与算法研究、参与、重庆市自然科学基金、2018-2021；
3. DC规划的理论和算法研究及其在机器学习中的应用、参与、国家自然科学基金面上项目、2018-2022；
4. 机器学习中几类稀疏优化问题的算法研究、参与、重庆市科委、2019-2022；
5. 用于人脸比对的身份证网纹照片修复技术、参与、横向项目、2018.6；
6. 基于RGBD传感器的人脸活体检测研究、参与、横向项目、2019.4.1；
7. 预测式自动外呼模型构建、参与、横向项目、2019.6；
8. BERT 预训练模型对问答机器人的效果提升、参与、横向项目、2019.6；

致 谢

值此论文完成之际，我由衷地感谢我的导师吴至友教授。从我读硕士开始已经将近三年时间，这期间吴老师在学习上给予了我极大的帮助和鼓励，在我的学业和论文的研究工作中无不倾注着导师辛勤的汗水和心血。吴老师渊博的知识、严谨的治学态度、一丝不苟的工作作风和无私的奉献精神都影响并激励着我。能够成为吴老师的学生是我人生中的一件幸事！同时也非常感谢杜学武教授对我的指导和帮助，杜老师锐意进取的精神、豁达的胸襟、乐观的心态都是我今后需要不断学习的。感谢白富生教授，白老师对人真诚、热情，我的进步也离不开白老师的关心和帮助。感谢李国权教授在我对硕士期间对我的帮助，不论是科研方面的合作还是生活上的相处都是非常愉快的事情。感谢院系领导老师对我的支持和帮助。

衷心感谢东南大学漆桂林教授，对我的关心和帮助。漆老师为人十分谦逊，对人热情。在我研究生期间漆老师给予了我很多的帮助，在此对他表示衷心的感谢！特别感谢高桓博士，高桓博士待人真诚，做事严谨。在我研究生期间给予我思想上，科研态度，以及方法上的指导，在此对他表示真心的感谢！也感谢李炜卓博士，李炜卓博士为人实在，踏实，在我研究生期间，引导我踏入了知识图谱相关研究，在此对他表示衷心的感谢。

感谢在学习和生活上给予我关心、鼓励的同门学们。

最后我要感谢我的父母、爱人和家人对我的理解和支持！

段旭祥

2020 年5 月6 日

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及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含他人已经发表或撰写过的研究成果，也不包含为获得重庆师范大学 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。

学位论文者签名：　　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重庆师范大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权重庆师范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，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

学位论文作者签名：　　签字日期：　年　月　日