文章编号: 1674-6864(2022)01-0065-06 DOI: 10.16508/j.cnki.11-5866/n.2022.01.011

基于迁移学习的实体关系抽取技术综述

郎春雨,侯 霞
(北京信息科技大学 计算机学院,北京 100101)

摘 要: 实体关系抽取任务的一大挑战是缺乏有效的训练语料,迁移学习可在一定程度上缓解其语料不足的问题。概述了迁移学习4种基本方法的原理及适用场景,分析总结了迁移学习在实体和关系抽取两方面的研究进展,最后总结展望了迁移学习技术在实体关系抽取领域的发展趋势。

关 键 词: 迁移学习; 实体抽取; 关系抽取; 机器学习
中图分类号: TP 391.1 **文献标志码**: A

Review of transfer learning technology for entity and relation extraction

LANG Chunyu, HOU Xia
(Computer School, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: A major challenge of entity and relation extraction tasks is the lack of effective training corpus. Transfer learning can alleviate the problem of insufficient corpus to a certain extent. The principles and applicable scenarios of the four basic methods of transfer learning were outlined, the current situation and research progress of transfer learning in entity and relation extraction were analyzed and summarized, and finally the development trend of transfer learning technology was summarized and prospected in the field of entity and relation extraction.

Keywords: transfer learning; entity extraction; relation extraction; machine learning

0 引言

在当今信息爆炸的背景下,如何从非结构化、复杂冗余的数据中获取有效的信息至关重要。信息抽取是从非结构化或半结构化数据中自动抽取信息的有效技术,在信息检索、问答系统等任务中有广泛应用。实体关系抽取则是信息抽取重要的子任务之一,其目的在于抽取出一对或多对实体并判断实体对之间是否存在某种语义关系。实体关系抽取分为流水式和联合式抽取。联合抽取[1-2]在一定程度上可以缓解流水式抽取的误差累积问题,但是其强行共享编码可能会导致实体抽取的特征与关系抽取的特征出现过于一致或者相互冲突等情况[3]。

近年来,借助于深度学习在特征提取和自动学习上的优势,基于深度学习的实体关系抽取研究取得了不少成果[4-5]。但是,深度学习在实体关系抽取任务中需要大量正确标注的语料进行训练,对数据的依赖性影响了其实际应用。迁移学习是将从相似领域学习到的知识应用到目标领域,可在一定程度上缓解实体关系抽取任务中训练数据缺乏的问题。

1 迁移学习的基本方法

迁移学习是机器学习的重要分支,它利用数据、任务或模型之间的相似性,让模型通过已有的源域标记数据向目标域未标记数据迁移,从而训练出适用于目标域的模型。迁移学习包括4种基本方法[6]:样本迁移,模型迁移,特征迁移和关系迁移。

1.1 基于样本的迁移

基于样本的迁移重复使用源域中的有标签数据,训练出一个在目标域中更准确的模型。其中存

收稿日期: 2021-11-05
基金项目: 国家自然科学基金资助项目(61672105)
第一作者简介: 郎春雨,女,硕士研究生;通信作者:侯 霞,女,博士,教授。

在两个关键问题: 一是如何从源域中筛选出与目标域有相似分布的有标签样本; 二是如何利用这些样本训练出准确的目标域上的理想模型。

第一种方法是基于样本的非归纳式迁移,它利用源域有标签数据和目标域无标签数据为目标域未见数据训练出预测模型。通过对源域和目标域的分布比值进行估计得到样本权重[7]。第二种是基于样本的归纳式迁移,利用源域的有标签数据和目标域一小部分有标签数据,为目标域训练预测模型。借鉴 AdaBoost 的思想,Dai 等[8]提出 TrAdaBoost,通过提高有利于目标分类任务的样本权重、降低不利于目标分类任务的样本权重,为目标域学习集成分类器。在实际场景中,基于样本方法的源域和目标域数据往往不重叠,而且某些特征只适用于源域,重新加权或采样的样本不能减少域间差异。为了解决这些问题,引入基于特征的迁移方法。

1.2 基于特征的迁移

基于特征的迁移将源域和目标域的数据特征变换到统一的特征空间,然后使用变换后的数据在新的特征空间中训练目标分类器。同时,需要将目标域未见数据映射到新的特征空间,然后进行预测。

第一种方法是最小化域间差异,识别不会导致域间差异的隐特征,并用它们表示源域数据,从而获得新特征训练目标分类器。如何学习域间隐特征十分重要,研究者们主要利用最大均值差异距离[9]最小化不同数据的分布差异,同时避免计算难和泛化难的问题。第二种方法是学习通用特征,从若干个源域的无标签数据学习通用的高级特征,用高级特征表示目标域有标签数据,然后利用这些有标签数据训练分类器。研究者们采用编码器[10]来学习通用特征并增强这些特征的可解释性。

1.3 基于模型的迁移

基于模型的迁移也称基于参数的迁移,其假设源域与目标域数据中存在一些可以共享的模型参数,它的核心目标是找到源域中哪部分有助于目标域学习。

第一种是基于共享模型成分的迁移。Williams 等[11]提出利用高斯过程在不同任务间共享知识,依靠训练数据间的相似性,预测未见数据标签。第二是基于正则化的迁移。Yang 等[12]提出的自适应支持向量机,成为后续研究的基础。

基于深度学习的迁移模型逐渐出现,参数微调是一种简单有效的模型参数的迁移方法。Long 等[13]改进了深度网络结构,通过在网络中加入概率分布适配层,进一步提高了深度迁移学习网络对于大数据的泛化能力。

1.4 基于关系的迁移

许多实际领域中存在样本间的关系结构,基于关系的迁移要构建源关系域和目标关系域之间关系知识的映射,其假设源域和目标域之间的关系具有共同的规律。Nickel 等[14]借助马尔科夫逻辑网络来发现不同领域之间的关系相似性,从而进行关系的迁移。

表 1 对迁移学习不同方法的适用场景进行了总结。

表 1 迁移学习方法的适用场景

方法名称	适用场景
基于样本的迁移	在域间分布差异较小时使用效果较好,如: 同一领域中的不同分支
基于特征的迁移	在域间分布差异较大时,找到域间可共享的特征
基于模型的迁移	在域间差异较小时使用效果较好,若差异较大,结合特征方法,将通用特征学习到的模型参数进行迁移,其他部分进行微调
基于关系的迁移	此方法关注样本间关系,如: 师生关系迁移到上下级关系

2 基于迁移学习的实体关系抽取

迁移学习最初应用在图像领域,近些年被应用到自然语言处理(natural language processing, NLP) 领域且逐渐获得了一些较好的成果。本节将主要总结迁移学习在实体抽取和关系抽取两方面的研究进展。

迁移学习在 NLP 领域通常被称为领域自适应。因为神经网络是领域自适应的基本模型,所以使用梯度下降法在源域和目标域进行模型优化,然后进行迁移是比较容易的。NLP 中的迁移主要有两种方法,分别是参数初始化和多任务学习,在某些情况下可以混合使用,先在源域参数初始化进行预训练,然后在源域和目标域同时进行多任务学习。其中参数初始化有两种方式: 参数冻结和参数微调。参数冻结是将源域训练的模型直接应用到目标域,不进行任何修改; 参数微调则将源域训练的模型部分层固定,目标域学习剩余的层。当目标数据集规模远小于源数据集时,参数冻结更优[15],反之微调方法更优[16]。

2.1 实体抽取

Qu 等[17]通过共享词汇和上下文特征,利用神经网络学习源标签和目标标签间的相关性,并对模

型微调以学习目标域特征的方式,在目标域与源域标签不匹配的情况下,将在大型医学源域训练的模型迁移至小型医学目标域。在强基线的基础上,仅基于 125 个目标域的训练句子,F_1 值提高了 160%。Giorgi 等[18] 基于长短时记忆网络(long short time memory,LSTM) + 条件随机场(conditional random fields,CRF) ,将在大型、嘈杂的数据集上训练的模型迁移到很小但由人工标注的数据集上,实体识别的错误平均减少约 11%,且 F_1 值有效提升,显著改善了生物医学实体抽取的最新结果,也证明了迁移学习对具有少量标签(约 6 000 或更少) 的目标数据集是非常有效的。电子健康记录大多以非结构化形式存在,对其进行实体抽取是 NLP 解决的典型问题之一。为了保护患者信息,相关机构在与研究者们共享信息前会去掉不同类型的个人信息,如姓名、地址和电话号码,这对实体抽取任务来说会更加困难。Lee 等[19] 利用 LSTM 获取字符特征,然后利用全连接网络在大型源域训练模型,最后将其迁移到较小的目标域,证明了对于标签数量较少的目标域,迁移学习是有效的。电子健康记录除了存在保密信息外,还存在格式错误的速记和非广泛使用的首字母缩略词,这使得实体识别难度更大,Gligic 等[20] 利用源域为目标域中未标注的电子健康记录提供预训练词嵌入表示,然后基于双向长短时记忆网络(bi-directional LSTM,BiLSTM) 、循环神经网络(recurrent neural network,RNN) 等模型进行迁移学习,在 I2b2 (2009) 数据集上算法的 F_1 值达到了 94.7%。

社交媒体上的用户生成文本同样存在数据缺失和语料少的问题。Von 等[21] 基于英文 Twitter 数据,通过合并句子级特征和利用不同于 Twitter 数据标签的数据,基于 BiLSTM + CRF 进行迁移学习。对于中文实体抽取任务,不仅只有很少的标注数据可用,而且语料处理时比英文更复杂。为了缓解 WeiboNER 数据集规模小、标注数据少的问题,Cao 等[22] 基于 BiLSTM + CRF + 对抗训练 + 自注意力机制进行迁移,采用多任务学习的方式将新闻领域的模型迁移至社交媒体领域。其中对抗迁移学习充分利用任务共享边界信息,自注意力机制捕获两个标记之间的长距离依赖关系。在两个公开数据集上的实验结果表明该模型显著优于其他模型。

近些年基于深度神经网络的预训练语言表示模型快速发展,如 ELMO、BERT 等。预训练的本质就是要进行迁移学习,对于实体任务来说,研究者们更倾向于利用源域获得预训练嵌入,然后对其他深度学习模型微调进行跨领域迁移。预训练模型的参数迁移使得模型训练更快,并且使用很少的训练样本就能达到特定的效果。

2.2　关系抽取

迁移学习在关系抽取方面获得了不少成果。因缺乏药物—疾病关系的标注数据集,张宏涛[23] 分别利用基于样本和特征组的方法进行关系抽取。基于样本的方法采用 TrAdaboost 算法,对样本权重进行学习调整; 基于特征组的方法,在特征级别上对源域中有利于目标域的多个特征进行学习并调整权重。以上两种方法在多个不同数据集上的召回率和 F_1 值相较于基线均有很大提升; 同时,基于特征组迁移比基于样本迁移在召回率方面提升了 10% 以上,这是因为基于特征组迁移选取了较为通用的特征,不需要更多领域性的知识,所以通用性更强。在不同领域间进行样本迁移时,由于样本差异,利用 TrAdaboost 算法容易出现负迁移。针对标注语料不足而导致蛋白质交互关系抽取性能较差的问题,李丽双等[24] 对 TrAdaboost 算法进行了改进,通过调整源域已标注数据集的样本权重,使得模型学习有利于目标域的样本特征,得到了改进算法 DisTrAdaboost,并验证了改进算法的收敛速度和抽取效果明显优于 TrAdaboost,且有效避免了负迁移。在公开数据集 20newsgroups 上的实验结果也证明了 DisTrAdaboost 能更好地使用源域数据辅助模型训练,加速收敛。

Di 等[25] 建立了领域感知的迁移方法,先提取目标域词汇特征,然后初始化实体关系的特征表示,再选取有利于目标域的源域知识库对实体关系表示进行规范、细化与推断,以 DBpedia 作为源域,Wiki-KBP 和 NYT 作为目标域,重新建立了新的知识库,并优于所有最先进的基线。Jiang[26] 利用源域有标签样本向目标域迁移,因域间关系类型不同,所以选择共享模型权重在域间提取通用特征,然后再通过人工加以实体类型约束信息,学习目标关系类型知识。在 ACE2004 数据集上的结果表明,将实体类型信息与自动选择通用特征相结合,多任务迁移方法达到了最佳性能。于海涛[27] 提出了一种基于 BERT 降噪的实体关系抽取模型: 为了解决因远程监督产生的噪声问题,通过在外部语料训练 BERT,然后将 BERT 迁移至目标任务进行微调; 在 BERT 输出后添加位置增强卷积层处理实体位置信息,弥补预训练任务与关系抽取任务的语义鸿沟,获取 BERT 的全局文本表示; 同时改进选择性注意力(selective

attention)机制,设计了时间衰减注意力机制,在训练的过程中按时间衰减机制避免低置信的样本,达到降噪效果,提升了模型的精度,在NYT-10和GIDS公开数据集上表现出优越的性能。

近年来,大多数基于模型迁移的关系抽取都与深度神经网络相结合,通过在神经网络中加入领域适配层,然后联合基于特征的迁移进行训练。其中在基于特征迁移时,大都采用特征选择法,从源域和目标域中,利用样本迁移估计数据分布,通过数据分布自适应来选择可共享的特征。在低资源条件下进行跨领域迁移时,根据实际情况,可以一对一迁移,也可将多源域迁移至单一目标域。

2.3 常用数据集

在实体和关系抽取研究中有一些常用数据集,表2对其中适合作为源领域的大型数据集进行了汇总。

表2 实体和关系抽取中的常用数据集

	一般领域	医学领域
实体抽取	CoNLL-2002、CoNLL-2003、CLUENER2020、人民日报语料2014版、WeiboNER、MSRA	MIMIC、I2b2、CCKS2019、DDIExtraction2013、CHEMPROT、MicrobiologyNER、PharmaCoNER
关系抽取	ACE04、ACE05、SemEval-2010 Task 8、Wikipedia、NYT、DBpedia	HPRD50、IEPA、LLL、AIMed、SemEval2013DDI、ADE-EXT、reACE

2.4 迁移学习的主要问题及措施

迁移学习的核心问题是找到两个领域的相似性。但是如果两个领域不相似或基本不相似,就会极大地影响迁移学习的效果,此种现象被称为负迁移。产生负迁移的原因主要有两点:首先是数据问题,源域和目标域数据不相似;其次是方法问题,源域和目标域数据相似,但是迁移方法不对。针对数据问题,Tan等[28]提出了传递迁移学习,其目标是在源域和目标域共享较少样本或特征的情况下,引入一个与源域和目标域都相似的领域作为中间域,从而实现3个领域间知识的迁移。Tan等[29]又提出了远域迁移学习,将其扩展到了多个领域,且极大地提升了算法的精度。针对方法问题,需要利用合适的方式找到可迁移的部分,如DisTrAdaboost通过调整样本权重有效地避免了负迁移。

3 结束语

在一般领域和医学领域的实体关系抽取任务中,使用迁移学习可以在一定程度上有效缓解标注语料不足的问题,但仍需研究者们在更多领域进行不断探索。通过对现有研究工作的探讨与总结,未来可从以下几方面展开研究:

1)深度迁移学习。利用深度神经网络进行迁移越来越受到研究者的关注。深度迁移学习[30]分为4类:基于实例、基于映射、基于网络和基于对抗的深度迁移。目前的研究主要集中在有监督学习上,如何利用深度神经网络在无监督或半监督学习中进行知识传递,将成为今后研究的热点。

2)强化迁移学习。Taylor和Stone[31]定义了强化迁移学习的问题,并将强化迁移学习分为3类:从单一源任务到目标任务的固定域迁移、跨多个源任务到目标任务的固定域迁移、源任务和目标任务不同域迁移。强化迁移学习已经在图像翻译[32]、知识图谱[33]等领域中获得较大成果,如何将强化迁移学习更好地应用在实体关系抽取任务中,还需要进行更深入的研究。

参考文献:

[1] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 1105-1116.

[2] Zheng S C, Hao Y X, Lu D Y, et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neurocomputing, 2017, 257: 59-66.

[3] Wang J, Lu W. Two are better than one: joint entity and relation extraction with table-sequence encoders [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 1706-1721.

[4] 孙镇,王惠临.命名实体识别研究进展综述[J].现代图书情报技术,2010(6):42-47.

[5] 王传栋,徐娇,张永.实体关系抽取综述[J].计算机工程与应用,2020,56(12):25-36.

[6] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[7] Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance

estimation [J]. The Journal of Machine Learning Research, 2009, 10: 1391 – 1445.

[8] Dai W Y, Yang Q, Xue G R, et al. Boosting for transfer learning [C]//Proceedings of the 24th International Conference on Machine Learning, Corvallis, Oregon, USA, 2007: 193 – 200.

[9] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis [J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199 – 210.

[10] Liao R J, Schwing A, Zemel R S, et al. Learning deep parsimonious representations [C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelone, Spain, 2016: 5083 – 5091.

[11] Bonilla E V, Chai K M. Williams C, Multi-task Gaussian process prediction [J]. Advances in Neural Information Processing Systems, 2007: 153 – 160.

[12] Yang J, Yan R, Hauptmann A G. Cross-domain video concept detection using adaptive SVMs [C]//Proceedings of the 15th ACM International Conference on Multimedia. 2007: 188 – 197.

[13] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks [C]// International Conference on Machine Learning. Sydney, NSW, Australia, 2017: 2208 – 2217.

[14] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs [J]. Proceedings of the IEEE, 2015, 104(1): 11 – 33.

[15] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension [J]. arXiv: 1611.01603v6, 2016.

[16] Kim Y. Convolutional neural networks for sentence classification [J]. CoRR, 2014, abs/1408.5882

[17] Qu L Z, Ferraro G, Zhou L Y, et al. Named entity recognition for novel types by transfer learning [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 899 – 905.

[18] Giorgi J M, Bader G D. Transfer learning for biomedical named entity recognition with neural networks [J]. Bioinformatics, 2018, 34(23): 4087 – 4094.

[19] Lee J Y, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with Neural Networks [C]//Proceedings of the 11 th International Conference on Language Resources and Evaluation, 2018.

[20] Gligic L, Kormilitzin A, Goldberg P, et al. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks [J]. Neural Networks, 2020, 121: 132 – 139.

[21] Von Däniken P, Cieliebak M. Transfer learning and sentence level features for named entity recognition on tweets [C]//3rd Workshop on Noisy User-generated Text (W-NUT). Copenhagen, Denmark. 2017, 3: 166 – 171.

[22] Cao P, Chen Y, Liu K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 182 – 192.

[23] 张宏涛. 面向生物文本的实体关系自动抽取问题研究 [D]. 北京: 清华大学, 2012.

[24] 李丽双, 郭瑞, 黄德根, 等. 基于迁移学习的蛋白质交互关系抽取 [J]. 中文信息学报, 2016, 30(2): 160 – 167.

[25] Di S, Shen Y, Chen L. Relation extraction via domain-aware transfer learning [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 1348 – 1357.

[26] Jiang J. Multi-task transfer learning for weakly-supervised relation extraction [C]. ACL, 2009.

[27] 于海涛. 基于深度迁移学习的实体关系抽取方法 [D]. 北京: 北京邮电大学, 2020.

[28] Tan B, Song Y, Zhong E, et al. Transitive transfer learning [C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1155 – 1164.

[29] Tan B, Zhang Y, Pan S, et al. Distant domain

transfer learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017,31(1).

[30] Tan C,Sun F,Kong T,et al. A survey on deep transfer learning [C]//International Conference on Artificial Neural Networks. Springer,Cham, 2018:270 – 279.

[31] Taylor M E, Stone P. Transfer learning for reinforcement learning domains: a survey [J]. Journal of Machine Learning Research,2009, 10:1633 – 1685.

[32] Gamrian S, Goldberg Y. Transfer learning for related reinforcement learning tasks via image-to-image translation [C]//International Conference on Machine Learning. PMLR, 2019:2063 – 2072.

[33] Ammanabrolu P, Riedl M O. Transfer in deep reinforcement learning using knowledge graphs [J]. EMNLP-IJCNLP 2019,2019:1.

(上接第64页)

[4] Bordoloi S,Kalita B. Designing graph database models from existing relational databases [J]. International Journal of Computer Applications, 2013,74(1):25 – 31.

[5] OQGRAPH Overview [DB/OL]. [2021 – 02 – 01]. https://mariadb.com/kb/en/oqgraph-overview/.

[6] Sun W, Fokoue A, Srinivas K, et al. SQLGraph: an efficient relational-based property graph store [C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,2015:1887 – 1901.

[7] Christopher J O L. Grapht: a hybrid database system for flexible retrieval of graph-structured data [D]. Cambridge: University of Cambridge,2016.

[8] Shute J, Vingralek R, Samwel B, et al. F1: a distributed SQL database that scales [J]. Proceedings of the VLDB Endowment,2013,6 (11):1068 – 1079.

[9] Grund M, Cudre-Mauroux P, Krueger J, et al. Hybrid graph and relational query processing inmain memory [C]//2013 IEEE 29th Internationl Conference on Data Engineering Workshops(ICDEW),2013:23 – 24.

[10] Concepts: relational to graph. [DB/OL]. [2021 – 01 – 27]. http://neo4j.com/developer/graph-db-vs-rdbms.

[11] Zhu Q, Larson P A. A query sampling method for estimating local cost parameters in a multidatabase system [C]//Proceedings of 1994 IEEE 10th International Conference on Data Engineering,1994:144 – 153.

[12] Zhu Q, Larson P A. Building regression cost models for multidatabase systems [C]//Fourth International Conference on Parallel and Distributed Information Systems,1996:220 – 231.

[13] Lu H J, Ooi B C, Goh C H. On global multidatabase query optimization [J]. ACM SIGMODRecord,1992,21(4):6 – 11.