针对因果常识推理任务的跨语言迁移学习系统

仇嘉星　熊光智
清华大学
{qiujx18,xgz18}@mails.tsinghua.edu.cn

摘要

基于常识的因果逻辑推理是自然语言处理中十分困难却又相当重要的任务之一，针对此任务近些年不断有新的数据集被公布。然而由于此类数据集的语言大多是英文，模型在语料相对稀缺的语言上难以得到充分的训练，最终效果都不大理想。近年来，迁移学习在自然语言处理领域的应用越加广泛，越来越多研究者提出了跨语言的模型或方法。本文中，我们提出了因果常识推理的跨语言迁移学习系统，将跨语言迁移学习应用于因果常识推理任务。针对预训练数据集噪声较大的问题，我们使用了基于 RoBERTa 的新颖降噪方法对其进行处理。除此以外，我们使用了多个预训练数据集进行多重预训练，并在有限的常识因果推理数据集中应用了数据增强，让模型的结果相比于前人的研究取得了一定的提升。

1. 引言

就自然语言处理而言，基于常识的因果逻辑推理是非常重要的任务之一。掌握常识因果推理的能力后，机器才能够更好地理解人类语言。和普通的文本包含关系不同，常识推理的过程中必须有文本之外的常识的参与 (Singer et al. 1992)。常识包括很多方面：时空关系、因果联系、自然法则、社会习俗等。

与此同时，在自然语言处理领域中，迁移学习的方法正处于快速发展的阶段，越来越受人们关注。自然语言处理中一直存在着一个问题：当今自然语言处理模型规模快速增长，对数据的需求也越来越大；但在全世界的诸多语言中，只有包括英语在内的少数热门语言拥有大量标注语料，很多较为冷门的语言标注数据很少，单语言的模型无法取得理想的效果。迁移学习就能有效解决该问题：人们设计了能够同时支持多种语言的模型，在大量的多语言无标注数据上进行预训练。最终针对具体任务时，只需使用一种语言的标注数据对模型进行微调，也能在其他语言上取得较好的效果。

不过，将跨语言迁移学习用于常识因果推理的研究并不多，常识因果推理的标注数据集更是屈指可数。因此，现有的研究 (Ponti et al. 2020) 将跨语言模型在普通的常识问答数据集上进行预训练，然后用有限的常识因果推理的数据集进行微调。这样的妥协会在预训练阶段带来很多噪声，降低训练的质量。

本文中，我们提出了针对因果常识推理的跨语言迁移学习系统。我们在Ponti et al. (2020) 的研究基础上，针对预训练的数据集提出了一种新颖的降噪方法，并使用了多个预训练数据集进行多重预训练，并在有限的常识因果推理数据集中应用了数据增强。最终，我们的系统相比于 Ponti 等人的模型在因果常识推理的 COPA 和 XCOPA 测试集上有着平均 2.06% 的提升。

2. 相关工作

因果关系的相关理论研究在很多领域都有涉及。在哲学领域中就有很多的理论来定义因果关系。有的学者 (Mackie & Mackie 1974, Trabasso et al. 1982) 提出了“在一定情况下的充分性”测试：事件 A 导致了事件 B，当且仅当在一定情况下，如果事件 A 发生了，并且各种事件正常地进行，那么事件 B 也会发生。不过，这样的定义比较模糊：“各种事件正常地进行”似乎暗示着常识的运用。于是，另外一些因果关系理论直截了当地指出，因果关系是和常识有关的。因果推理的 mechanism view 理论就认为，人们对因果关系的认知基于一些“基础知识”(Salmon 1984, Shultz 1982, Ahn

et al. 1995)。以表1中“小孩子放开了手中抓的气球线”和“气球飞走了”之间的因果关系为例，我们需要知道“气球会往天上飘”的“基础知识”。

常识因果关系	前提	假设
√	小孩放开了手中抓的气球	气球飞走了
×	泰迪是我的弟弟	泰迪是男性

表 1. 常识因果关系举例

Singer et al. (1992) 提出了“搭桥”的思想，其本质和前文的“基础知识”是类似的。他明确地区分文本包含关系和常识因果关系：因果关系之中必须有文本之外的常识的参与。“气球会往天上飘”就是文本之外的常识。而“泰迪是我的弟弟”和“泰迪是男性”之间不存在常识因果关系，因为从“弟弟”的文本含义中我们就能得知其所指对象泰迪为男性，而并不需要借助额外的常识。

我们最终采用了 Singer(Singer et al. 1992) 对于常识因果关系的划分。从上述讨论就可以看出，因果关系的定义和范围并不是定论。上述讨论中的因果关系等同于常识因果关系，两者不包括文本包含关系。但是在一些情景下，人们所谈论的“因果关系”的范围会不同于本文讨论的因果关系。为了避免歧义，本文使用了“常识因果关系”的术语。

在跨语言迁移学习模型中，掩码语言模型（MLM）是目前在跨语言理解任务中表现最佳的一类模型。常见的掩码语言模型包含 mBERT (Devlin et al. 2019)，XLM-Roberta (Conneau et al. 2020)，mT5 (Xue et al. 2020) 等。它们在跨语言推断、问答、命名实体识别等任务上都取得了最前沿的结果。部分语言模型加入了新的训练任务，比如 XLM 模型在预训练中引入了翻译语言模型 (TLM) 的训练任务，让模型通过不同语言的翻译信息来对被遮挡的词进行预测 (Conneau & Lample 2019)。还有一部分模型直接通过抽样的方式，用大量不同语言的语料进行 MLM 训练，最后也会有较好的结果 (Conneau et al. 2020)。

3. 方法

为了实现对于因果关系的常识推理，我们希望能通过一个句子的整体信息判断其是否蕴含了可用常识推理得出的的因果关系。为此，我们设计了一个完备的系统来实现对自然文本中因果关系的检测，并使用针对因果常识推理任务对其进行预训练及微调。我们的一共包含三个模块：文本编码模块会对输入的语句做编码处理，使其能被后续模块计算分析；接着是用于关系检测的模块，它接受经过处理后的句子，然后返还给我们该语句包含因果逻辑关系的概率；并且我们还设计了数据降噪模块，用于去除预训练所用的数据集中与任务不一致的噪音。接下来我们将具体介绍每个模块的实现原理。

3.1. 文本编码

对于任意文本，我们往往需要把它编码成一个向量或矩阵来对其含义进行解析，因此获取正确的向量化表示对于自然文本理解来说至关重要。在 BERT (Devlin et al. 2019) 横空出世后，越来越多的研究者投入到了预训练语言模型的研究之中，常见的有 RoBERTa (Liu et al. 2019)、T5(Raffel et al. 2020) 等等。此类模型能较好地将每个词的上下文信息融入词向量中，使得最终的编码结果能完整地提取出文本的含义。考虑到我们希望实现跨语言学习的目的，我们选择了在多种语言上进行了大规模预训练的 XLM-RoBERTa (Conneau et al. 2020) 模型作为我们编码器的内核。

我们首先用 XLM-RoBERTa 模型对应的分词器对输入文本进行分词并在前后插入特殊的标记符号，最后得到

$$S = \{[CLS], Tok_1, Tok_2, \cdots, Tok_n, [EOS]\}$$

接着我们将这串词输入到预训练好的 XLM-RoBERTa 当中，得到所有词对应的词向量

$$H = \{h_0, h_1, h_2, \cdots, h_n, h_{n+1}\}$$

其中每个词向量 h_i 的维度均为 768。对于每段文本前都插入的特殊标记符号 [CLS]，我们期望能用它对应的词向量作为整个文本的向量化表示，去处理后续的分析任务。该方法也反复被运用于自然语言理解任务 (NLU) 之中。

3.2. 关系检测

在得到文本对应的向量后，我们期望能从中提取出文本是否含有因果关系的信息。对此我们尝试了各种

Context	Alex was the bouncer at the bar the kai went to. Alex allowed Kai to enter even though she is underage.
Question	How would you describe Alex?
Answer A	breaking the law
Answer B	a good friend
Answer C	frustrated with Alex

图 1. 数据集噪音示例

组合来进行关系的检测任务。我们尝试了只用一个全连接层, 以及先用一个线性层对向量进行降维再用全连接的处理方式, 发现前者表现要好于后者, 这可能与模型复杂后易过拟合有关。我们还对比了让模型输出一个二维向量再使用 CrossEntropyLoss 进行计算, 和让其输出一个一维的值再通过 Binary Cross Entropy 计算 Loss 的效果。我们发现, 总的来说交叉熵损失函数的总体表现要更加稳定。最终我们输入一个二维的向量用来表示该文本有或无因果关系的概率分别为多少, 二者概率之和为 1。

3.3. 数据降噪

由于 COPA 本身数据集较小, 为了能让模型得到充分训练并实现跨语言迁移学习的效果, 我们使用了应用于常识推理任务的大规模数据集 SocialIQA 和 WinoGrande 进行预训练。但由于它们并非针对因果相关的常识推理来设计, 其中许多被标注为正样例的训练样本中并不含有实际的因果联系, 即在我们的下游任务中应被标记为不含因果逻辑。图1是一个 SocialIQA 数据集中的样本, 其中加粗的选项 A 是该问题的正确答案。根据我们的常识, 我们的确能确认答案就是 A, 但这种常识的推理并非是基于两个事件的潜在因果联系的。在该样例中, 背景知识可以说就是选项 A 本身, 它们在概念上并不构成因果逻辑。因此虽然在该数据集中它们的确是能通过常识推理得到, 但就因果常识推理任务而言, 我们希望将系统将此类语句标记为负样例而非正样例。因此针对预训练数据集中出现的大量伪正样例, 我们需要在正式训练前对其做降噪处理。

受Xiao et al. (2020) 启发, 我们希望用一个独立的神经网络对数据进行降噪。我们希望能让该模型对所有正样本含因果关系的可能性从大到小排序, 然后实验人员再截取出前 $k\%$ 个正样例作为真正的含有因果关系的训练样本。在这项任务里, 我们比较的是不同语句含有因果关系可能性的相对大小, 与我们的中心任务要简单一些。为了能训练出具有这样能力的模型, 我们设计了一个分类任务: 一次性向模型输入 m-1 个负样例和 1 个正样例, 让模型找到唯一的正样例在这 m 个样本中的位置。这里的正负样例都来自于我们主任务上最终用来微调和测试的 COPA 数据集, 这也是为了让我们的降噪过程更满足我们最终的任务需求。模型框架上我们同样也是采用预训练语言模型加全连接层, 再用交叉熵损失函数的方式进行训练。由于整个降噪模块的训练和运用都是在英文文本上的, 这里的预训练语言模型我们使用了在单语言任务上表现更好的 RoBERTa (Liu et al. 2019), 而非适用于跨语言学习任务的 XLM-RoBERTa (Conneau et al. 2020)。

最终整个系统的框架如图2所示。

4. 实验

4.1. 数据集

此次实验中我们总共用到了 4 个不同的数据集, 分别为 SocialIQA (Sap et al. 2019), WinoGrande (Sakaguchi et al. 2020), COPA (Roemmele et al. 2011) 和 XCOPA (Ponti et al. 2020), 各数据集的总结如表2所示。

SocialIQA 是面向社会场景中的常识推理任务而设计的数据集, 共包含 38000 个多项选择问题, 用于测试模型对于情绪和社会知识的掌握情况。每个问题有三个可选答案, 其中仅有一个是自然社会环境中人会

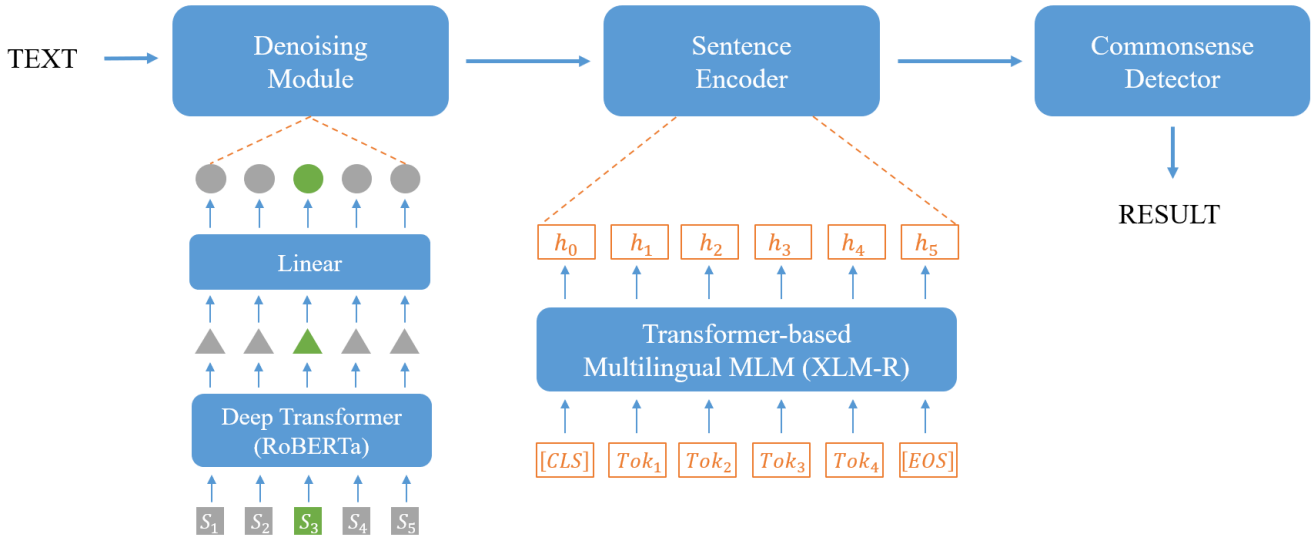

图 2. 系统框架

做出的反映。我们将每个正确选项与其对应的问题与语境拼接起来，作为一个能被需要被模型推理出因果关系的正样例；其余两个选项亦和问题、语境进行了拼接，作为训练的负样本。

WinoGrande 是一个受 The Winograd Schema Challenge (WSC) (Levesque et al. 2012) 启发，被创建出用来测试机器模型常识推理能力的数据集，其中共有 44k 个问题。在该数据集上目前最先进的模型也只能到达 80% 不到的准确率，远低于人类的表现 (94.40%) (Sakaguchi et al. 2020)。其问题的形式为完形填空，于是我们将正确答案填入空格中作为正样例，而将错误答案填入其中作为负样例。

COPA 是一个针对因果逻辑的常识推理任务而设计的英文数据集，一共包含了 1000 个多项选择题，每个问题中的正确选项为对应前提假设的原因或结果。与对 SocialIQA 的处理类似，我们将每个前提假设和对应的两个选项分别拼起来，形成一对正负样本。此外，由于每个问题中的前提假设和选项都是独立的事件，我们将其颠倒顺序进行拼接不会影响其是否含有因果逻辑关系，只是改变其因与果的出现先后次序。我们对划分出的训练集做了该数据增强操作，使可用于训练的数据增加了一倍。

XCOPA 是一个用于因果常识推理任务测试的多语言数据集，一共包含了 11 种语言，包括一些如东阿普里马魁楚亚语、海地克里奥尔语这样自然语言处理方面数据集比较稀缺的语言。XCOPA 是对 COPA 数据集验证集和测试集进行翻译和修订的结果，每种语言均有 100 个问题作为验证集，并包含 400 个问题来当做测试集。

4.2. 实验流程

我们首先以 RoBERTa-large 为核心，在 COPA 上训练出一个单语言的降噪模型 (排序模型)。随后提取出用作预训练的数据集 WinoGrade 和 SocialIQA 中全部的正样例，用我们训好的降噪模型对它们含有因果逻辑关系的可能性从大到小进行排序，只保留了根据人工的判定，保留了前一半样本用作预训练中的正样例。

我们选取 XLM-RoBERT-large 作为编码器来应用于我们的判别常识推理下的因果逻辑关系的模型。由于 SocialIQA 对应的论文中提到该数据集能被用于迁移学习到 COPA 的任务上 (Sap et al. 2019)，我们让模型先在 WinoGrade 进行预训练，随后让其在 SocialIQA 上进行第二轮预训练。这里的模型均指经过降噪处理后的模型。

最后我们让模型在 COPA 上进行微调。我们将 COPA 原本大小为 500 的训练集拆分成了 400 个训练样本和 100 个验证样本，并将这 100 个验证样本与 XCOPA 里 11 种语言的验证集合并在一起，作为我们微调过程中的验证集。在微调过程结束之后，我们将训好的模型在 COPA 和 XCOPA 的测试集上进行测试，

Dataset	Size	Target	Example
SocialIQA	38k	社会场景常识推理	context: Tracy protected her teammates from injury when she saw an accident about to happen prevented it. question: Why did Tracy do this? answerA: make a play answerB: prevent injuries answerC: injure them
WinoGrande	44k	常识推理	sentence: The GPS and map helped me navigate home. I got lost when the _____ got turned off. option1: GPS option2: map
COPA	1k	因果常识推理	premise: The cook's eyes watered. choice1: He ran out of onions. choice2: He cut an onion.
XCOPA	11*0.5k	因果常识推理	premise: 厨师的眼睛流泪了。 choice1: 他没有洋葱了。 choice2: 他切了洋葱。

表 2. 实验中使用的数据集

得到最终的评测结果。

我们采取的评测标准与Ponti et al. (2020) 的研究一致，选用的是全部问题回答的准确率。

4.3. 实验结果

我们的模型在 COPA-XCOPA 测试集上的平均准确率为 70.79%，Ponti et al. (2020) 的研究中平均准确率为 68.73%。具体的准确率对比请参照图3。可以看到，我们的模型在英语测试集上的效果不如论文中的模型，但在其他非英语的测试集上相比于该模型却有着明显的提升。我们对此的理解为，经过降噪后的预训练数据集使得模型吸收的英文预料信息减少，但其学习到的因果逻辑关系更加纯净，因此会有我们最终这样的结果。

5. 总结

在本文中，我们提出了因果常识推理的跨语言迁移学习系统。针对预训练数据集噪声较大的问题，我们使用了基于 RoBERTa 的降噪方法。为了提高模型的准确率，我们使用了多个预训练数据集进行多重预训练，并在有限的常识因果推理数据集中应用了数据增强，并最终取得了较好的结果。

参考文献

Ahn, W.-k., Kalish, C. W., Medin, D. L. & Gelman, S. A. (1995), 'The role of covariation versus mechanism information in causal attribution', *Cognition* **54**(3), 299–352.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020), Unsupervised cross-lingual representation learning at scale, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 8440–8451.
URL: *https://www.aclweb.org/anthology/2020.acl-main.747*

Conneau, A. & Lample, G. (2019), Cross-lingual language model pretraining, *in* 'Advances in Neural Information Processing Systems', pp. 7059–7069.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

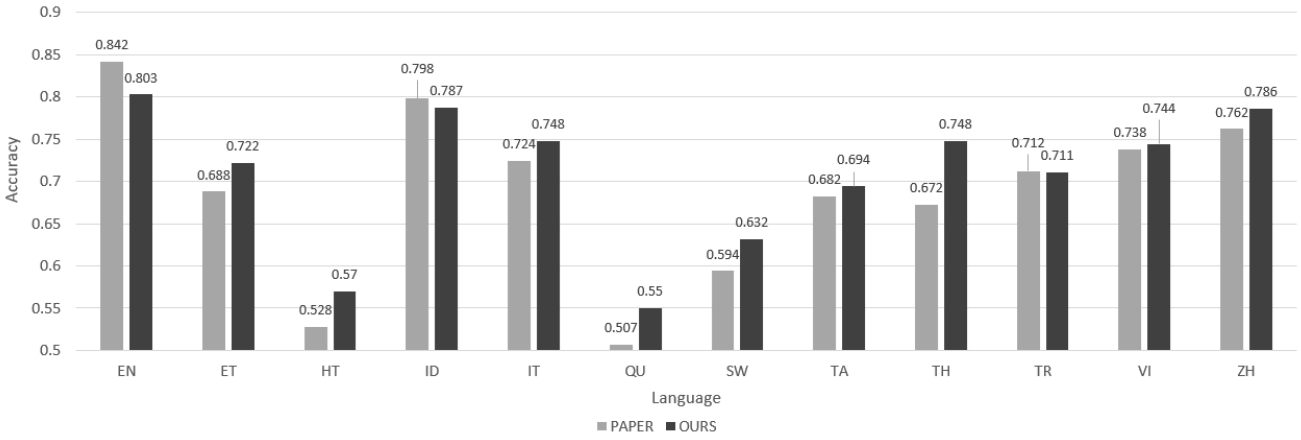

图 3. 系统在 COPA-XCOPA 测试集上的测试结果

Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: *https://www.aclweb.org/anthology/N19-1423*

Levesque, H., Davis, E. & Morgenstern, L. (2012), The winograd schema challenge, *in* 'Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning', Citeseer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Roberta: A robustly optimized bert pretraining approach'.

Mackie, J. L. & Mackie, J. L. (1974), *The cement of the universe: A study of causation*, Oxford: Clarendon Press.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I. & Korhonen, A. (2020), XCOPA: A multilingual dataset for causal commonsense reasoning, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 2362–2376.
URL: *https://www.aclweb.org/anthology/2020.emnlp-main.185*

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020), 'Exploring the limits of transfer learning with a unified text-to-text transformer', *Journal of Machine Learning Research* **21**(140), 1–67.

Roemmele, M., Bejan, C. A. & Gordon, A. S. (2011), Choice of plausible alternatives: An evaluation of commonsense causal reasoning., *in* 'AAAI spring symposium: logical formalizations of commonsense reasoning', pp. 90–95.

Sakaguchi, K., Le Bras, R., Bhagavatula, C. & Choi, Y. (2020), Winogrande: An adversarial winograd schema challenge at scale, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 8732–8740.

Salmon, W. C. (1984), *Scientific explanation and the causal structure of the world*, Princeton University Press.

Sap, M., Rashkin, H., Chen, D., Le Bras, R. & Choi, Y. (2019), Social iqa: Commonsense reasoning about social interactions, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', pp. 4453–4463.

Shultz, T. (1982), 'Rules of causal attribution. monographs of the society for research in child development'.

Singer, M., Halldorson, M., Lear, J. C. & Andrusiak, P. (1992), 'Validation of causal bridging inferences in discourse understanding', *Journal of Memory and Language* **31**(4), 507–524.

Trabasso, T. et al. (1982), 'Causal cohesion and story coherence.'.

Xiao, C., Yao, Y., Xie, R., Han, X., Liu, Z., Sun, M., Lin, F. & Lin, L. (2020), Denoising relation extraction from

document-level distant supervision, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 3683–3688.
URL: *https://www.aclweb.org/anthology/2020.emnlp-main.300*

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. & Raffel, C. (2020), 'mt5: A massively multilingual pre-trained text-to-text transformer'.