关于短文本匹配的泛化性和迁移性的研究分析

马新宇　范意兴　郭嘉丰　张儒清　苏立新　程学旗
（中国科学院计算技术研究所 中国科学院网络数据科学与技术重点实验室，北京市 100190；
中国科学院大学，北京市 100190）

论文摘要

自然语言理解中的许多任务，比如自然语言推断任务，机器问答和复述问题，都可以看作或者转换为短文本匹配问题。大量数据集的涌现促进了短文本匹配任务的进步，但是很少有研究去分析这些数据集之间的泛化性和迁移性，以及如何将这些数据集应用到新的领域。该文使用深度学习模型ESIM和预训练语言模型BERT在十个通用的短文本匹配数据集上进行了详尽的实验。我们分析了数据集之间的泛化性和迁移性，通过可视化的方式展示了影响数据集之间泛化性的因素，并且我们发现即使是BERT这种在大规模语料预训练过的模型，合适的迁移仍能带来性能提升，最后我们发现在混合数据集预训练过的模型有较好的泛化能力和迁移能力，并且在新的领域和少量样本情况下，性能平均只降低了百分之五。

实验数据集

表一　相关数据集描述

数据集	规模	来源	类型
SNLI	570k	网页文档	自然语言推断
MNLI	433k	网页文档	自然语言推断
QNLI	130k	维基百科	自然语言推断
MSMARCO	120k	维基百科	问答
QQP	400k	网页文档	复述问题
WNLI	1k	书籍	自然语言推断
RTE	3k	网页新闻	自然语言推断
SciTail	27k	书籍	自然语言推断
MRPC	4k	网页新闻	复述问题
WikiQA	30k	维基百科	问答

论文简介

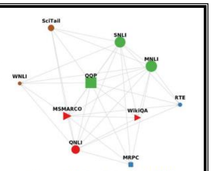
本文旨在分析不同类型、不同匹配任务数据集之间的泛化性和迁移性，并尝试探索如何利用这些大量的数据集到新的领域中；并且试图在低资源情况下取得不错的效果。

泛化性分析

表二　泛化实验结果

（表中值为准确率。其中上表是 ESIM 模型结果，下表是 BERT 模型结果，虚线右侧是大数据集，虚线左侧是小数据集，行代表源训练数据集，列代表目标测试数据集）

		Scitail	Wnli	Rte	Mrpc	WikiQA	Snli	Mnli	Qnli	Msmarco	QQP
ESIM	SNLI	62.8	46.5	54.9	46.7	86.5	–	**73.4**	50.9	50.0	67.5
	MNLI	67.6	43.7	**63.2**	58.0	**87.0**	**80.8**	–	51.2	49.9	66.8
	QNLI	60.7	53.5	47.7	69.5	76.7	62.8	63.6	–	**74.0**	64.0
	MSMARCO	**72.7**	**56.3**	52.0	**69.9**	81.5	64.1	63.9	**74.9**	–	**68.2**
	QQP	62.1	**56.3**	50.5	58.6	86.9	67.4	66.9	53.2	51.8	–
	MT100K	74.7	46.5	62.1	69.5	83.7	–	–	–	–	–
	SELF	82.6	57.8	57.0	69.0	88.0	90.4	81.8	82.4	83.8	88.1
BERT	SNLI	75.8	43.7	70.0	60.9	85.8	–	**82.5**	51.1	50.3	70.5
	MNLI	77.2	42.3	**73.3**	57.3	87.4	**86.7**	–	50.5	50.1	**72.5**
	QNLI	69.3	**57.8**	46.6	**71.3**	**87.6**	59.8	64.5	–	**82.3**	66.1
	MSMARCO	**82.7**	43.7	57.8	67.7	79.5	58.0	51.8	**77.1**	–	67.9
	QQP	73.9	47.9	53.5	65.6	86.8	68.8	69.5	54.0	52.3	–
	MT100K	79.4	52.3	70.8	65.4	82.2	–	–	–	–	–
	SELF	94.9	56.3	67.5	85.2	91.4	92.8	88.4	91.3	92.7	88.3



图一　数据集之间相似度的可视化

结论2：由图一可知，泛化性主要与数据集的任务类型和数据集来源有关，且受任务类型的关系更大一些

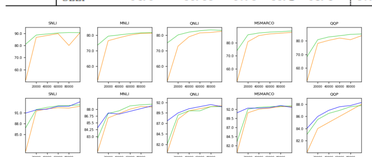
结论1：由表二可知，两种模型的泛化性在大数据集上表现较差，BERT模型要比ESIM的泛化性更强；MT100K指将大数据集取20K简单进行混合

迁移性分析

表四　迁移实验结果

（表中值为准确率。其中上表是 ESIM 模型结果，下表是 BERT 模型结果，虚线右侧是大数据集，虚线左侧是小数据集，行代表源数据集，列代表目标数据集）

		Sci-tail	Wnli	RTE	Mrpc	Wik-iQA	Snli	Mnli	Qnli	Msmarco	QQP
ESIM	SNLI	87.5	**59.2**	63.9	76.4	89.7	–	**72.0**	82.9	88.5	84.3
	MNLI	**88.6**	46.5	66.1	**79.5**	89.7	83.1	–	82.9	88.5	85.1
	QNLI	88.0	56.3	65.3	77.8	89.5	**83.7**	70.8	–	**89.3**	**85.2**
	MSMARCO	87.7	50.7	**66.8**	77.9	**90.0**	83.0	70.8	**83.0**	–	85.1
	QQP	85.2	54.9	62.1	75.7	89.6	81.6	68.8	82.2	88.3	–
	MT100K	87.9	56.3	64.6	76.6	89.8	–	–	–	–	–
	SELF	82.6	57.8	57.0	69.0	88.0	82.7	70.4	82.5	87.9	84.3
BERT	SNLI	95.3	56.3	**74.4**	85.7	91.8	–	81.6	90.0	92.7	88.1
	MNLI	95.3	56.3	72.9	**86.1**	91.1	**88.3**	–	90.0	**93.1**	87.8
	QNLI	95.7	56.3	72.9	85.2	92.2	87.9	81.5	–	92.9	88.2
	MSMARCO	94.9	56.3	70.8	85.3	**92.4**	87.8	**81.7**	**90.9**	–	**88.3**
	QQP	**95.8**	**62.0**	72.9	85.3	90.9	87.6	81.2	90.9	92.8	–
	MT100K	95.5	56.3	75.8	84.8	92.4	–	–	–	–	–
	SELF	94.9	56.33	67.5	85.2	91.4	87.7	81.0	91.3	92.7	88.3



图二　五个大数据集的学习曲线

（上为 ESIM 模型，下为 BERT 模型；横坐标代表目标数据集的数据量，纵坐标代表准确率；橘黄色的曲线代表只在目标数据集上训练，绿色曲线代表首先在迁移效果最好的数据集上预训练，然后在目标数据集上训练；蓝色曲线代表先在 MT100K 上训练，然后在目标数据集上训练）

结论1：由表四可知，即使在数据集很大的情况下，合适的预训练仍然能够有所提升。

结论2：由图二可知，在MT100K上进行预训练（蓝色曲线）的结果曲线要比在迁移效果最好的源数据集训练的曲线（绿线）不仅更加平滑，性能也有所提升，说明多个数据集的融合能提高模型的迁移能力。

低资源情况下

表五　MT100K在100个样本上的实验结果

	SciTail	WNLI	RTE	MRPC
MT100K	87.3	56.3	71.8	72.0
SELF	94.9	56.3	67.5	85.2

我们发现在少量样本（只有100个）情况下，MT100K的平均性能只比全量的小数据集低5%。因此在短文本匹配领域，对于新的领域的数据集，我们可以首先将其他领域的数据集等量混合，然后使用BERT模型在混合的数据集预训练之后，再进行少量的目标样本标注和训练，就可以达到不错的效果。

论文结论

本文在十个数据集上，使用两种深度学习模型对短文本匹配的泛化性和迁移性进行了详尽的实验。实验结果表明影响泛化能力的因素主要有匹配的类型和匹配数据集的来源，合适的迁移（在源数据集上进行训练），即使是BERT这种预训练模型，也能在目标数据集带来提升。实验结果显示不同数据集之间泛化性和迁移性趋势并不保持一致，将数据集进行简单的混合能带来更好的泛化能力和迁移能力，并且这种方法可以减少目标数据集的训练量。最后我们考虑到如何将这些现有的数据集应用到新的领域和少量样本的情况下，发现在目标数据集只有100个样本的情况下，使用混合的数据集在BERT模型上进行预训练，性能只下降了5%。

西安电子科技大学 XIDIAN UNIVERSITY