《智能信息处理》课程考试

基于深度学习的知识图谱问答语义匹配研究

张建华

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 20 日

基于深度学习的知识图谱问答语义匹配研究

张建华
（大连海事大学　信息科学与技术学院　大连　116026）

摘　要：近年来，随着人工智能技术的热度持续上升，知识图谱的重视度不断上升，我们开始采用深度学习来处理知识图谱的相关知识。目前我们对单一事实类问答系统中问句和关系的语义匹配在小规模标注样本中难以获得较高准确率的问题，提出一种基于循环神经网络(RNN)的深度学习模型。首先，使用基于RNN的序列到序列无监督学习算法，通过序列重构的方式在大量无标注样本中学习问句的语义空间分布，即词向量和 RNN;通过使用问句特征和关系特征计算内积的方式，在有标注样本中训练并生成语义匹配模型。实验结果表明，在有标注数据量较少而无标注数据量较大的环境下，与有监督学习方法Embed-AVG和RNNrandom相比，所提模型的语义匹配准确率分别平均提高5.6和8.8个百分点。所提模型通过预学习大量无标注样本的语义空间分布可以明显提高在小规模标注样本环境下的语义匹配准确率。

关键词：语义匹配;；知识图谱；机器学习；深度学习

Research on Chinese Idiom knowledge graph based on deep learning

Zhang Jian Hua
(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract: In recent years, as the popularity of artificial intelligence technology continues to rise and the importance of knowledge graph continues to rise, we began to use deep learning to process the relevant knowledge of knowledge graph. At present, we propose a deep learning model based on recurrent neural network (RNN) to solve the problem that the semantic matching of questions and relations in single fact question answering system is difficult to obtain high accuracy in small-scale labeled samples. Firstly, a sequence to sequence unsupervised learning algorithm based on RNN is used to learn the semantic spatial distribution of questions in a large number of unlabeled samples by sequence reconstruction, namely, word vector and RNN. By using question feature and relation feature to calculate inner product, semantic matching model is trained and generated in labeled samples. Experimental results show that the semantic matching accuracy of the proposed model increases by an average of 5.6 and 8.8 percentage points, respectively, compared with the supervised learning methods Embed AVG and RNNrandom, when the amount of annotated data is small and the amount of no annotated data is large. By pre-learning the semantic spatial distribution of a large number of unlabeled samples, the proposed model can significantly improve the semantic matching accuracy in the context of small labeled samples.

Key words: semantic matching; metaphor knowledge; machine learning；deep learning

0 引言

随着人工智能的发展，智能的信息服务持续升级，在各种信息服务领域都能看到知识图谱的应用，如智能问答、个性化推送、信息检索等。知识图谱帮助计算机学习人的语言交流方式，使计算机像人类一样“思考”，使得各种信息服务反馈给用户更加智能的答案。可以说，知识图谱是传统行业和人工智能进行融合的方向，也是人工智能从研究走向落地应用过程中必不可少的环节。

深度学习是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。 深度学习是一个复杂的机器学习算法，在语音和图像识别方面取得的效果，远远超过先前相关技术。

在特定领域的问答系统中，例如石油勘探、医疗等领域，由于数据来源有限、人工标注样本的难度和成本过高，经常只能获取少量(数千条) 的有标注样本。而当前解决语义匹配任务的方法主要是有监督的学习方法，这些方法的语义匹配准确率依赖于标注样本的数据量，在标注样本数量较少的情况下难以学习到较高的语义匹配准确率，但是大部分情况下，通过互联网等途径容易获取大量通用领域的无标注样本，因此研究利用这些无标注样本来提高语义匹配的准确率具有重要意义。

通过引入大量通用领域(源领域) 的无标注样本来提升少量专业领域(目标领域) 有标注样本学习效果的方法，可以看作是一种迁移学习方法。已知的迁移学习方法通常假设源领域采用有标注样本进行训练，不能直接解决本文提出的问题。其中自学习方法跟本文提出的问题比较接近，但其源领域的无监督学习方法适合处理图像数据，无法直接应用于自然语言处理问题。至于自然语言处理方面的无监督学习方法主要有 word2vec、GloVe 等方法，但是这些方法只解决了词粒度的语义学习，不能学习到句子粒度的语义特征，因此需要设计一种能够解决源领域是无标注自然语言数据的迁移学习模型。

本文针对特定领域内有标注样本数量不足的问题，提出一种基于循环神经网络(Recurrent Neural Network，RNN) 的迁移学习模型，能够利用大量无标注样本学习问句的语义空间分布以提高语义匹配准确率，进而解决面向特定领域单一事实类问答系统中的语义匹配任务。该模型首先使用基于

RNN 的序列到序列(sequence-to-sequence，seq2seq)无监督学习算法，通过序列重构的方式在大量无标注样本中学习问句的语义空间分布，即词向量和 RNN，作为迁移学习模型中源领域的特征空间分布。然后，使用此语义空间分布作为有监督语义匹配算法相应的参数。最后，通过计算问句特征和关系特征内积的方式，在有标注样本下训练并生成语义匹配模型，其中通过线性函数来建立源领域语义空间分布到目标领域语义空间分布的映射。实验结果表明，与传统的有监督学习方法相比，本文提出的迁移学习模型在小规模有标注样本中取得较高语义匹配准确率。此模型在解决因特定领域标注样本不足而导致语义匹配准确率不高的问题有重要意义。此外，该模型还可以扩展应用到如情感分析等其他的语义匹配相关任务。

本文主要研究工作如下:

1) 提出了序列到序列无监督学习问句语义空间分布的算法，该算法使用大量无标注样本学习问句的语义空间分布，作为迁移学习模型中源领域的特征空间分布;

2) 提出了基于 RNN 的迁移学习模型，该模型通过将源领域的语义空间分布映射到目标领域的语义空间分布来提高语义匹配准确率;

3) 通过实验验证了基于迁移学习的语义匹配模型在小规模标注样本情况下可以提高语义匹配准确率。

1 相关工作

1.1 问答语义匹配

在基于知识图谱单一事实类问答系统的研究工作中，一条传统的研究线路是采用语义解析的方式，该方式针对特定领域设计语法解析规则，对于规则可以覆盖到的情况有较高的准确率，但是泛化能力较差，不易扩展。

除此之外，另一条重要的研究线路是词嵌入方式，这种方式主要使用基于深度学习的相似匹配来完成语义匹配任务。其核心思想是学习问句和知识

图谱中关系的语义特征向量表示，使相匹配的问句和关系在向量空间中距离最接近。目前已经提出很多基于神经网络学习问句和关系语义特征的方法。其中: Bordes 等使用相对浅层的词嵌入模型学习问句和关系的语义向量空间分布; Yih 等使用卷积神经网络(Convolutional Neural Network，CNN)生成问句和关系的语义向量空间分布; Dai 等使用 RNN 提取问句的语义向量空间分布。词嵌入方式具有很强的泛化能力，易于扩展，但是该方式采用有监督学习模型，依赖于有标注样本数量，在标注样本数量少的情况下准确率不高。

本文提出的方法接近第二条研究线路，应用神经网络学习问句和关系的语义特征向量表示; 但是有别于传统的有监督学习算法，本文提出迁移学习模型来解决标注样本不足的问题，通过无监督算法预学习问句的语义空间分布来提高有监督语义匹配算法的准确率。该方法解决了有监督学习算法在特定领域标注样本数量少而语义匹配准确率不高的问题，拥有更强的泛化能力。

1.2 迁移学习

迁移学习主要分为基于实例和基于特征的两种方式。其中基于实例的方法，具有代表性的工作有 Dai 等提出的提升方法，其通过改变源领域和目标领域中有标注数据样本的权值来达到迁移学习的目的; 但基于实例的方法通常假设源领域和目标领域都是有标注数据，而本文中源领域采用无标注数据，目标领域采用有标注数据，所以不适用于本文提出的问题。

此外，基于特征的方法对于源领域和目标领域是否有标签没有严格要求，其主要解决两个问题: 如何学习特征和如何迁移特征。例如 Oquab 等提出的基于特征迁移的方法与本文方法比较接近，其首先通过有监督训练方式学习得到源领域的分类器，然后将该分类器的模型参数作为目标领域分类器参数的初始值，最终通过训练目标领域的有标注数据学习得到目标领域的分类器。由于该方法假设源领域是有标注数据，所以具有一定局限性，不能完全适用本文所要解决的问题。在图像识别领域，Raina 等提出的自学习方法与本文方法更加接近，其首先通过源领域无标注数据学习得到表示图像的基特征; 然后根据上步的基来表示目标领域的有标注数据; 最终根据这些数据的特征进行有监督训练学习得到目标领域的分类器。本文与自学习方法的主要区别在于学习无标注数据特征的方法不同，并且特征迁移的方式也有所区别。近几年领域适应(domain adaptation)结合对抗训练思想取得了一些进展，但是由于这些方法假设源领域采用有标注数据，目标领域采用无标注数据，所以不能直接解决本文提出的问题。

2 问题的定义

本文假设 Dt ={〈qt，r〉| qt ∈ Qt，r ∈ K} 表示目标领域的有标注数据集，其中 Qt 是有标注问句集，关系 r 取自知识图谱 K; Ds ={〈qs〉| qs ∈ Qs} 表示源领域的无标注数据集，其中 Qs 是无标注问句集。需要强调的是有标注数据集的数据量远远少于无标注数据集的数据量，即 | Qt| | Qs|，并且无标注问句 Qs 不需要与有标注问句 Qt 符合同一分布。由于目标领域有标注数据量较少，难以直接学习得到准确率较高的语义匹配模型; 而源领域无标注数据量庞大，可以用来辅助学习语义匹配模型。本文通过将源领域中无标注问句的语义空间分布 Fs 映射到目标领域中有标注问句的语义空间分布 Ft 来提高语义匹配准确率，得到语义匹配模型 h:

$$r^{*} = \arg\max_{r \in K} h(q_t, r) \tag{2}$$

3 基于深度学习的语义匹配模型

3.1 基于 RNN 的深度学习模型

本文提出的迁移学习模型通过预学习大量无标注样本数据的语义空间分布来提高小规模标注样本环境下的语义匹配准确率。如图 1 所示，步骤①用于学习无标注问句的语义特征，本文使用基于 RNN 的序列到序列无监督学习算法，在大量无标注样本 Ds 中学习问句的语义空间分布 Fs，而此空间分布 Fs 由词向量 *E*s 和编码 RNN Encoders 组成，解码过程 d 由 Decoders 组成，该部分在 3．2 节详细描述; 步骤② 用于迁移参数，本文使用步骤 ① 中学习得到的语义空间分布作为有监督语义匹配算法中词向量 *E*t 和编码 RNN Encodert 的参数，即令 *E*t = *E*s，Encodert = Encoders; 在步骤 ③ 中，本文使用有标注样本 Dt 训练生成语义匹配模型 h，其中将线性函数 Linear 作为特征映射函数 φ 来建立源领域语义空间分布到目标领域语义空间分布的映射，语义

匹配函数 f 通过内积(Linear(*c*t) T *E*r(r)) 的方式计算问句与关系的语义匹配程度，该部分在 3．3 节详细描述。本文最后在 3．4 节详细阐述了迁移学习模型的整体训练算法。

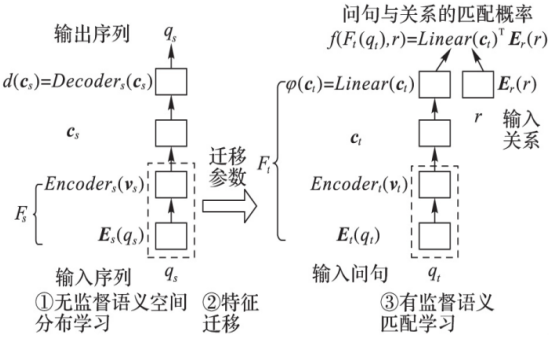

图 1 迁移学习模型框架

Fig. 1 Framework of transfer learning model

3.2 无监督语义空间分布学习算法

为结构体。本文通过设置序列 x = qs，并且令 y 重构 x 来学习源领域问句的语义空间分布。该算法主要分为编码 - 解码 (Encoder-Decoder) 两部分，具体表示如下:

$$vs = Es(x) \quad (4)$$

$$cs = \text{Encoders}(vs) \quad (5)$$

$$y = \text{Decoders}(cs) \quad (6)$$

其中编码部分首先通过一层词嵌入层 *E*s 将词序列 x 映射到分布式的向量表示 *v*s，然后经过由两层 BiLSTM 组成的编码器 Encoders 转化为固定长度的向量 *c*s，作为问句的语义特征向量表示，其中 *c*s 是 BiLSTM 的最终隐层状态; 解码器 Decoders 通过另外的一层 LSTM 进行解码得到输出序列 y。最终通过自编码的方式使词嵌入层 *E*s 和编码 RNN Encoders 学习到源领域问句的语义空间分布。

本文采用最大似然估计(Maximum Likelihood Estimation，MLE)和反向传播(Back Propagation，BP)算法训练序列到序列算法的参数，解码器 Decoders 以固定长度的向量 *c*s 为条件计算 y = (y1，y2 ，…，yT')的产生概率。

3.3 迁移学习模型的实现

综上所述， 此迁移学习模型的训练算法如下所示。

Input: D_s, D_t。

Output: $\boldsymbol{E}_t, Encoder_t, Linear, \boldsymbol{E}_r$。

1) Training semantic space distribution:

2) Randomly initialize the parameters of $\boldsymbol{E}_s, Encoder_s, Decoder_s$ layers

3) On the D_s data set, the parameters of each layer are unsupervised trained by MLE and BP algorithm according to formula (7)

4) Get the parameters of $\boldsymbol{E}_s, Encoder_s, Decoder_s$ layers

5) Training semantic matching:

6) Let $\boldsymbol{E}_t = \boldsymbol{E}_s$, $Encoder_t = Encoder_s$, and random initialize the parameters of *Linear* and $\boldsymbol{E}_r$ layers

7) On the D_t data set, the parameters of $\boldsymbol{E}_t, Encoder_t, Linear, \boldsymbol{E}_r$ layers are supervisedly trained by MLE and BP algorithm according to formula (14)

8) Get the parameters of $\boldsymbol{E}_t, Encoder_t, Linear, \boldsymbol{E}_r$ layers

该训练算法的输入是无标注数据集 Ds 和有标注数据集 Dt，输出是语义匹配算法中 Et 、Encodert 、Linear、Er 层的参数。其中第 1)～4)行用来训练源领域的语义空间分布，第 2)行是随机初始化 Es 、Encoders 、Decoders 层的参数，第 3)行是在无标注数据集 Ds 上根据式(7) 采用 MLE 和 BP 算法无监督训练各层的参数，第 4)行是得到训练完成后 Es 、Encoders 、Decoders 层的参数; 第 5)～8)行用来训练语义匹配算法，第 6)行是用第 4)行得到的 Es、Encoders 层的参数作为 Et、Encoder 层的初始值，并随机初始 Linear、Er 层的参数，第 7)行是在有标注数据集 Dt 上根据式(14) 采用 MLE 和 BP 算法有监督训练各层的参数，第 8）行是得到训练完成后 Et、Encoder 、Linear、Er 层的参数。

4 实验

4.1 实验数据和参数设置

语义匹配算法分别采用三套有标注数据集作为迁移学习的目标领域。为了验证本文提出的迁移学习算法在特定专业领域有标注数据量较少情况下的效果， 本文制定了一套中文有标注数据集 OIL 进行实验， 该数据集由 4465 条有标注数据组成， 主要涉及石油探勘领域， 并且该数据集的每条问句都对应同一个石油勘探领域知识图谱的实体-关系-客体三元组。除此之外， 为了验证算法的有效性和泛化能力， 本文采用两 套 英文数据集 WebQuestions (WBQ) 和 SIMPLEQUESTIONS(SQ)。WBQ 包含 5 810 条有标注数据，SQ 包含 108 442 条有标注数 WBQ

和 SQ 中的每条问句分别对应来自 Freebase 知识图谱中的实体-关系-客体三元组，主要涉及人物、地理等有限的几个领域。

本文将以上有标注数据集全部按照训练 70% 、验证 10%和测试 20% 进行划分。

序列到序列算法采用两套无标注数据作为迁移学习的源领域。中文采用 WebQA 作为训练集， 该数据集来源于某度知道的问答语料，包含 42 223 条无标注问句样本， 主要涉及人文、地理、影视、学科知识以及一些专业性的领域等多种领域。英文采用 WikiAnswers Paraphrase 数据集作为训练集，该数据集来源于 WikiAnswers 互动问答网站，用户可以在该网站提问或回答任何问题。该数据集包含 258 万条无标注问句样本，没有包含与之匹配的关系， 其中问句涉及科技、生活、人文、历史、地理等众多领域，问句描述较为复杂，有些问句可能涉及多种关系。本实验采用 Tensorflow 深度学习开发平台，其中问句词向量的维数设为 200，所有 RNN 结构体的隐层大小设为 256，关系词向量维数设为 256; 其中问句词向量和关系词向量参数的随机初始值满足均值为 0，标准差为 1E -4 的正态分布，RNN 参数的随机初始值都采用服从［-1，1］均匀分布的初始化方式; 所有合页损失函数的边界值 γ 设为 0.1，负采样数量设为 1 024; 所有实验采用基于 mini-batch 的 Adam 优化方式训练参数， 学习率设为 0.001。

4.2 对照基准和评估

本实验采用词嵌入平均模型作为对照基准，该模型的核心思想是采用问句中每个词向量的平均值作为其问句的向量表示本文称之为 Embed-AVG(Average Embeddings)。此外，本文将基于迁移学习的语义匹配模型称为 RNNpretrain，将基于 RNN 的完全随机初始化参数进行监督训练的语义匹配算法称为 RNNrandom(random initialization of RNN) 。为了验证迁移学习算法在小规模标注样本环境下的学习效果，本文分别设置了三组对照实验数据。如表 1 所示，语义匹配算法数据集表示 Embed-AVG、RNNrandom 和 RNNpretrain 语义匹配模型采用的训练集和测试集，序列到序列算法数据集表示 RNNpretrain 中序列到序列算法训练阶段的训练集。其中 SQ(5k)表示从 SQ 数据集中随机抽样 5 000 条样本，WAP(200k) 表示从 WAP 数据集中随机抽样 20 万条样本。

除此之外，为了验证有标注样本的数据量对实验的影响，本实验从 SQ 数据集的训练集中分别随机采样，产生多个不同数量的样本集，作为 Embed-AVG、RNNrandom 和 RNNpretrain 语义匹配模型的训练集.为了验证无标注样本的数据量对实验的影响，本实验从 WAP 数据集中分别随机采样，产生多个不同数量的样本集， 作为 RNNpretrain 中序列到序列算法训练阶段的训练集。最终所有语义匹配算法的评估全部采用 SQ 数据集的 21 687 条测试样本作为测试集。本文实验以目标关系得分最高视为匹配正确，以问句和关系的匹配正确率作为评估标准。

表 1 实验数据集设置

Tab. 1 Experimental data sets

实验数据对照组	语义匹配算法数据集	序列到序列算法数据集
data1	SQ(5k)	WAP(200k)
data2	WBQ	WAP(200k)
data3	OIL	WebQA

4.2 实验结果分析

图 2 展示了 Embed-AVG、RNNrandom 和 RNNpretrain 算法分别在 data1、data2 和 data3 数据集上的表现。从图 2 可以明显看出，在有标注数据量较少的情况下，RNNpretrain 在各个数据集上的准确率，与 RNNrandom 相比分别大约提高 8.6、9.1 和 8.8 个百分点,平均提高 8.8 个百分点; 与 Embed-AVG 相比分别大约提高 6.7、5.7 和 4.4 个百分点，平均提高 5.6 个百分点。甚至简单的 Embed-AVG 也比 RNNrandom 的学习效果好.从中可以看出 RNNrandom 在有标注样本规模很小的情况下学习效果比较差，因为越复杂的模型需要越多数据进行训练学习。而本文提出的 RNNpretrain 可以在有标注数据量较少的情况下， 通过学习大量无标注数据来明显提高语义匹配准确率。

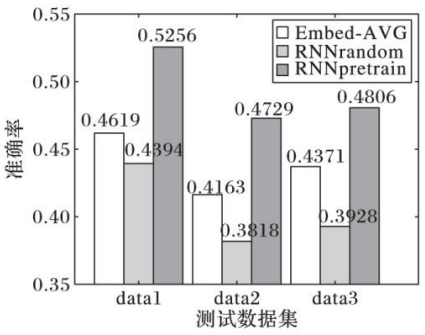

图 2 各语义匹配模型对不同测试数据集测试准确率

Fig. 2 Test accuracy comparison of semantic matching models under different test data sets

此外，为了展示有标注样本和无标注样本的数据量对实验的影响，表 2 列举了各个模型分别在不同有标注数据量和不同无标注数据量学习后的语义匹配准确率，其中 SQ 后面的括号内容表示语义匹配算法训练阶段在 SQ 数据集中随机采样的样本数量，WAP 后面的括号内容表示序列到序列算法训练阶段从 WAP 数据集中随机采样的样本数量。例如: SQ(20k) 情况下的 Embed-AVG 和 RNNrandom 分别表示从 SQ 数据集中随机采用两万条样本用来训练其语义匹配算法; SQ (5k) 情况下的 RNNpretrain(WAP100k) 表示 RNNpretrain 模型首先从 WAP 数据集中随机采样 10 万条样本用来训练序列到序列算法，然后从 SQ 数据集中随机采用 5 000 条样本用来训练语义匹配算法。图 3 以折线图的方式描绘了表 2 中的数据，通过图 3 可以更加明显地观察到所有语义匹配算法的准确率在不同数量训练样本上的变化趋势，以及其相互间的差距。

表 2　语义匹配模型测试准确率对比　%

Tab. 2　Test accuracy comparison of semantic matching models　%

Dataset	SQ(5k)	SQ(10k)	SQ(20k)
Embed-AVG	46.19	51.82	55.09
RNNrandom	43.94	56.45	63.69
RNNpretrain(WAP50k)	48.88	57.28	63.80
RNNpretrain(WAP100k)	49.78	57.80	63.90
RNNpretrain(WAP200k)	52.56	58.72	64.11

从表 2 可以看出在 SQ (5k)的情况下，RNNpretrain(WAP200k)的准确率分别比 RNNpretrain (WAP50k)和 RNNpretrain(WAP100k)大约提高 3.7 和 2.8 个百分点，说明在一定范围内使用相同数量的有标注样本，RNNpretain 会随着预训练无标签样本 WAP 数量的增加而提高语义匹配准确率，说明增加无标注样本数量可以提升语义特征提取效果，使源领域学习到的语义空间分布能更大程度地覆盖目标领域的语义空间分布。此外，RNNpretrain(WAP200k) 使用 5k 有标注样本训练得到的学习效果超过了 Embed-AVG 使用 10k 有标注样本训练得到的学习效果， 体现了迁移学习算法在小规模样本集上优越的性能。而且随着有标注样本数量的增加 RNNrandom 的学习效果逐渐超越了 Embed-AVG 的学习效果，证明在有监督学习中，越复杂的模型需要越多数据进行训练学习。从图 3 可以看出， 随着有标注样本数量的增加，RNNrandom 的学习效果将会缩小与 RNNpretrain 的差距.其中 RNNpretrain(WAP200k)的准确率相比 RNNrandom 在 SQ(5k) 、SQ(10k) 的情况下分别大约提升 8.6 和 2.3 个百分点。这主要有两点原因: 一个直观的原因是增加有标注样本的数量，可以学习到更加精准的目标领域语义空间分布，进而提升语义匹配准确率，但是在专业领域的实际开发中，难以获得数量足够庞大的有标注样本集; 另一个原因是当 RNNpretrain 使用相同数量的无标注本，在序列到序列算法训练阶段没有学习到更加通用的问句语义空间分布时，随着有标注样本的增加，将会减弱这些预训练得到的特征在有监督训练中发挥的作用。

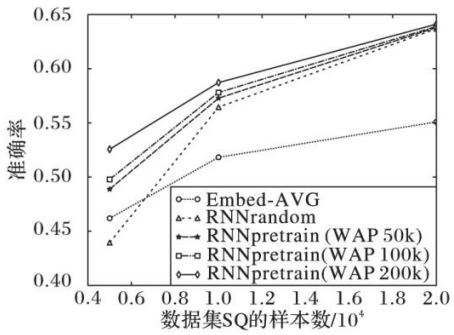

图 3　各语义匹配模型在不同数据集规模时测试准确率

Fig. 3　Test accuracy comparison of semantic matching models under different sizes of test data sets

综合以上实验结果表明，在有标注样本数量较少且无标注样本数量较大的情况，RNNpretrain 的语义匹配准确率相比 RNNrandom 和 Embed-AVG 有明显提升， 拥有更强的泛化能力。

5　结语

对于单一事实类问答系统中的语义匹配任务，传统的有监督学习方法需要依赖大量的有标注样本才能取得较高的语义匹配准确率。针对这种问题， 本文提出基于 RNN 的迁移学习模型，其通过预学习大量无标注样本数据的语义空间分布来提高小规模标注样本环境下的语义匹配准确率。实验结果表明，在小规模有标注样本上，该模型可以明显提高语义匹配准确率， 并且该准确率在一定范围内随着无标注样本数据量的增大而有所提升。本文提出的模型适用于特定领域因有标注数据量少而导致有监督学习模型准确率不高的情况， 此外这种迁移学习的方式也可以推广应用到其他与语义匹配相关的任务中。

下一步本文将探究预处理无标注样本的方法

来过滤更多的噪声，以及探索使用生成对抗网络改进序列到序列算法来提升学习问句语义空间分布的能力。

参 考 文 献

[1] 陈云,刘卫光.基于可分解注意力机制的医疗问句语义匹配研究[J].中原工学院学报,2020,31(01):74-79.

[2] 张卫,王昊,陈玥彤,范涛,邓三鸿.融合迁移学习与文本增强的中文成语隐喻知识识别与关联研究[J/OL].数据分析与知识发现

[3] 崔韬世,麦范金.词语相似度计算方法分析[J].网络安全技术与应用,2012,05:55-56+72.

[4] 孙海霞,钱庆,成颖.基于本体的语义相似度计算方法研究综述[J].现代图书情报技术,2010,01:51-56.

[5] 马良荔,孙煜飞,柳青.语义 Web 中的本体匹配研究[J].计算机应用研究,2017,05:1-3.