

基于异质信息网的短文本特征扩充方法

吕晓锋, 赵书良, 高恒达, 武永亮, 张宝奇

引用本文

吕晓锋, 赵书良, 高恒达, 武永亮, 张宝奇. [基于异质信息网的短文本特征扩充方法](#)[J]. 计算机科学, 2022, 49(9): 92-100.

LYU Xiao-feng, ZHAO Shu-liang, GAO Heng-da, WU Yong-liang, ZHANG Bao-qi. [Short Texts Feature Enrichment Method Based on Heterogeneous Information Network](#)[J]. Computer Science, 2022, 49(9): 92-100.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于异构网络表征学习的作者学术行为预测](#)

Author' s Academic Behavior Prediction Based on Heterogeneous Network Representation Learning
计算机科学, 2022, 49(9): 76-82. <https://doi.org/10.11896/jsjcx.210900078>

[一种面向电商网络的异常用户检测方法](#)

Method for Abnormal Users Detection Oriented to E-commerce Network
计算机科学, 2022, 49(7): 170-178. <https://doi.org/10.11896/jsjcx.210600092>

[基于生成对抗网络和元路径的异质网络表示学习](#)

Generative Adversarial Network and Meta-path Based Heterogeneous Network Representation Learning
计算机科学, 2022, 49(1): 133-139. <https://doi.org/10.11896/jsjcx.201000179>

[融合不完整多视图的异质信息网络嵌入方法](#)

Heterogeneous Information Network Embedding with Incomplete Multi-view Fusion
计算机科学, 2021, 48(9): 68-76. <https://doi.org/10.11896/jsjcx.210500203>

[基于异质信息网络表示学习与注意力神经网络的推荐算法](#)

Recommendation Algorithm Based on Heterogeneous Information Network Embedding and Attention Neural Network
计算机科学, 2021, 48(8): 72-79. <https://doi.org/10.11896/jsjcx.200800226>

基于异质信息网的短文本特征扩充方法

吕晓锋^{1,2,3} 赵书良^{1,2,3} 高恒达⁴ 武永亮⁵ 张宝奇^{1,2,3}

1 河北师范大学计算机与网络空间安全学院 石家庄 050024

2 供应链大数据分析 & 数据安全河北省工程研究中心 石家庄 050024

3 河北省网络与信息安全重点实验室 石家庄 050024

4 河北师范大学软件学院 石家庄 050024

5 石家庄铁道大学信息科学与技术学院 石家庄 050043

(1586821231@qq.com)

摘要 随着计算机技术深度融入社会生活,越来越多的短文本信息遍布在网络平台上。针对短文本的数据稀疏问题,文中构建了一个鲁棒的异质信息网框架(HTE)来建模短文本,该框架可集成任何类型的附加信息并捕获它们之间的关系,以解决数据稀疏问题。基于该框架利用不同外部知识设计了6种短文本扩充方法,引入Wikipedia知识库和Freebase知识库的实体、实体类别、实体间关系等实体信息和文本主题等文本信息,以丰富短文本特征。最后使用相似性度量结果来验证所提出的短文本特征扩充方法的效果。通过与传统的3种相似性度量方法的6种文本扩充方法以及目前主流的短文本匹配算法在两个短文本数据集上进行比较,结果表明,所提的6种短文本扩充方法均有所提升,最佳方法的相似性度量结果与BERT相比提升了5.97%,证明了所提框架具有鲁棒性,可以包含多种类型的外部知识,能够解决短文本的数据稀疏性问题,以无监督的方式高精度地对短文本进行相似性度量。

关键词: 异质信息网络;短文本扩充方法;短文本匹配;知识库;元路径

中图法分类号 TP391

Short Texts Feature Enrichment Method Based on Heterogeneous Information Network

LYU Xiao-feng^{1,2,3}, ZHAO Shu-liang^{1,2,3}, GAO Heng-da⁴, WU Yong-liang⁵ and ZHANG Bao-qi^{1,2,3}

1 College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang 050024, China

2 Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Data Security, Hebei Normal University, Shijiazhuang 050024, China

3 Hebei Provincial Key Laboratory of Network & Information Security, Hebei Normal University, Shijiazhuang 050024, China

4 Software College, Hebei Normal University, Shijiazhuang 050024, China

5 School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China

Abstract With the deep integration of computer technology into social life, more and more short text messages are spreaded all over the web platform. Aiming at the problem of data sparsity of short texts, a robust heterogeneous information network framework(HTE) for modeling short texts, which can integrate any type of additional information and capture the relationship between them to solve the data sparsity problem, is constructed. Based on this framework, six short text expansion methods are designed using different external knowledge, and the short text features are enriched by introducing entity information such as entities, entity categories, inter-entity relationships and textual information such as text topics from Wikipedia and Freebase knowledge bases. Finally, the similarity measurement result is used to verify the experimental effect. By comparing the six text expansion methods with the traditional three similarity measures on two short text datasets and the current mainstream short text matching algorithms, the results of the proposed six text expansion methods are improved. Compared with BERT, the similarity measurement results of the best method improves by 5.97%. The proposed framework is robust and can include any type of external knowledge, and the proposed method can overcome the data sparsity problem of short texts and can perform similarity metrics on

到稿日期:2021-07-26 返修日期:2021-10-17

基金项目:国家社会科学基金重大项目(13&ZD091,18ZDA200);河北省重点研发计划项目(20370301D);河北师范大学重大关键技术攻关项目(L2020K01)

This work was supported by the National Social Science Fund of China(13&ZD091,18ZDA200), Hebei Provincial Key Research and Development Project of China(20370301D) and Key Technology Development Project of Hebei Normal University(L2020K01).

通信作者:赵书良(zhaoshuliang@sina.com)

short texts with high accuracy in an unsupervised manner.

Keywords Heterogeneous information network, Short text enrichment method, Short text matching, Knowledge base, Meta-path

1 引言

随着计算机技术深度融入社会生活,越来越多的短文本信息遍布在网络平台上。短文本与一般文本有很大的不同,通常语义稀疏且篇幅较短,且存在句法不规范的问题。短文本在文本信息检索^[1]、智能问答系统^[2]等领域有着广泛的应用。如何解决短文本的数据稀疏问题,对短文本内容进行有效扩充,已成为近年来研究的重点。

对短文本进行有效扩充一般通过引入外部知识库或者借助短文本自身的内容特征实现。引入外部知识库指利用包含丰富语义信息的开源知识库(如 Wordnet 和维基百科)等扩充短文本^[3-4]。然而,目前这些研究没有充分考虑知识库中链接包含的结构信息以及文本主题等文本信息。许多知识库如 Wikipedia, Freebase^[5], DBpedia^[6] 等包含大量关于实体类型及其关系的世界知识,这为之后提出更好的方法以扩充短文本提供了丰富的资源。

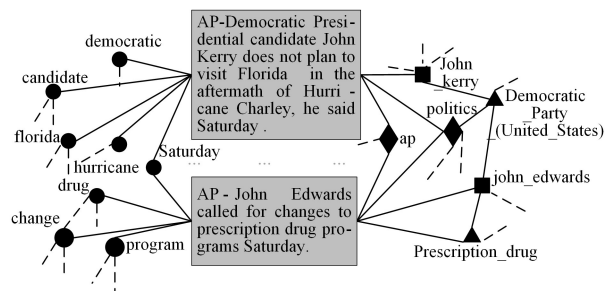
近年来,异质信息网(Heterogeneous Information Network, HIN)已经成为数据挖掘、数据库和信息检索等领域的研究热点^[7],在相似性度量、分类、聚类、推荐等任务上取得了良好的效果。异质网络可以通过融合丰富信息解决大数据的“多样性”难题^[8]。由于异质网络能够建模复杂交互系统,包含丰富的结构和语义信息,将短文本建模成异质信息网,可以包含任何附加信息(如实体和主题)并捕获文本和附加信息之间的丰富关系。

受上述思想的启发,本文提出了一种融合外部知识的短文本异质信息网构建算法,该算法构建了一个灵活的异质信息网框架建模短文本,可以集成任何类型的附加信息并捕获它们之间的关系。

图 1 给出了一个短文本异质信息网的例子。经过必要的去停用词等预处理操作后,本文先后加入了 Freebase 知识库中的实体(fb:en. john_kerry, fb:en. john_edwards)、短文本的主题词(ap, politics)以及 Wikipedia 实体(Associated_Press, Democratic_Party_(United_States), Prescription_drug),丰富了短文本的语义信息。与传统基于知识库的文本方法不同,本文还获取了文本主题等文本信息,并将其融入异质信息网中,形成了包含 5 种类型节点的短文本异质信息网,充分利用了文本的结构化信息及显式语义信息,从而丰富了文本特征。然后,使用基于短文本节点的元路径相似方法得到不同文本间的相似性。例如,短文本异质信息网中第一篇文本的文本主题和第二篇文本的文本主题均为 politics,则两篇文本通过主题词形成一条元路径实例 doc1-politics-doc2,进而建立语义上的连接。同时两篇文本均有 notable_person 类型的实体 john_kerry 和 john_edwards,那么这两份文件就变得相似。但如果不进行文本扩充,这两篇文本会由于包含的相同单词太少,导致相似度较低,但事实上这两篇文本是较为相似的。

本文提出了一种融合外部知识的短文本异质信息网构建模型 HTE(HIN Text Enrichment)。该模型充分利用了文本

的结构化信息及显式语义信息,同时结合了外部知识,以丰富短文本特征。首先,设计了 6 种不同方法,通过外部知识库和 LDA 模型向短文本添加额外的知识和文本信息;然后,提出了融合外部知识的短文本异质信息网构建算法,并构建了短文本异质信息网框架,将外部知识与文本信息充分融合,丰富短文本特征,缓解短文本的数据稀疏问题;最后,通过相似性度量来测试本文方法的有效性。



注:网络包含 5 个实体类型,即 doc, word, doc_topic, notable_person, Wikipedia_entity, 它们由矩形、圆形、菱形、正方形、三角形表示

图 1 异质信息网实例

Fig. 1 Example of heterogeneous information network

2 相关工作

目前研究者们提出了许多方法试图解决短文本的数据稀疏问题,主要包括基于外部知识库的文本特征扩展与基于短文本的自身内容特征扩展两个研究方向。使用外部知识库作为短文本内容扩展的来源在一定程度上可以丰富短文本特征,解决短文本内容稀疏问题。在文献[9]中, Yao 等提出利用词语分布式表示获得的词语语义相似度信息来丰富文本表示。基于词分布表示的短文本分类框架使用神经语言模型生成单词的分布式表示,从词的分布式表示中挖掘词的相似信息。文献[10]使用 DBpedia Spotlight 框架来识别短文本中的相关概念,并使用来自 DBpedia 本体的信息来丰富文本,从而改善了数据的稀疏性。

但是这些基于知识图谱的短文本扩充方法均是基于文本的 BOW 模型进行扩展,并不能完全有效地融合丰富信息。文献[11]集成了多语言 n -gram 信息,以丰富文本表示,但与本文方法相比,其需要一定的时间进行样本训练。

利用异质网络可以通过融合丰富信息的特性^[12-13],将文本建模为异质信息网,得到包含命名实体、文本和单词节点的异质信息网。这些方法与本文的方法最相似,都使用了异质信息网。但这些方法与本文方法也存在着巨大差异,它们没有充分考虑到文本的其他信息如文本主题等;同时,其只是从维基百科(Wikipedia)中提取的概念实体,而本文方法除了提取 Wikipedia 实体外,还设计不同方法引入了 Freebase 实体(下文简称实体)、实体类型、实体关系等信息,最大程度丰富了短文本信息。

短文本匹配是 NLP 和 IR 领域的重要问题,具有重要的研究价值。目前典型的文本匹配模型如 ARC-II^[14]、Match-

Pyramid^[15]等交互式模型在输入层进行词语间的匹配,并使用匹配的结果进行后续的建模,此方法虽然可以把握语义焦点,合理建模上下文重要性,但不能捕获句子内的语义关系(如实体关系),严重依赖于训练数据的数量,然而参数量大会使运算速度变慢。本文进行文本扩充后使用相似性度量(文本匹配)结果来验证文本扩充效果。

3 融合外部知识的短文本异质信息网框架及短文本特征扩充方法

本文提出了融合外部知识的短文本异质信息网构建算法,构建了异质信息网络框架(HTE),它可以集成任何类型的附加信息并捕获它们之间的关系。基于框架设计了6种不同方法,通过外部知识库和LDA模型向短文本添加额外的知识和文本信息,以改善短文本的稀疏性。通过添加外部知识,可以更好地表示短文本,更好地比较语义相似但没有共同词汇的短文本。

图2给出了完整的融合外部知识的短文本异质信息网框架。融合外部知识的短文本异质信息网框架是将Wikipedia知识库和Freebase知识库的实体、实体类别、实体间关系等实体信息和文本主题等文本信息融入短文本异质信息网,并应用于短文本相似性度量的过程。

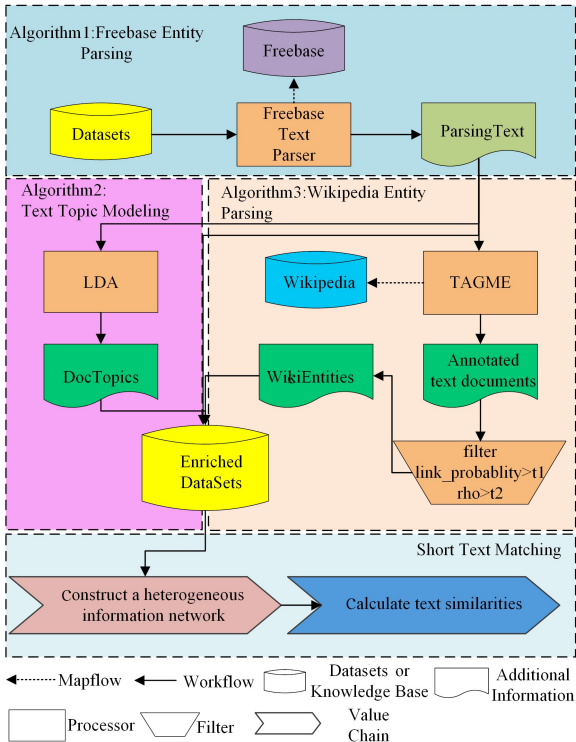


图2 融合外部知识的短文本异质信息网建模框架
Fig. 2 Short-text heterogeneous information network modeling framework incorporating external knowledge

本文的框架由以下几个步骤组成。第一步,使用算法1进行Freebase实体解析,使用算法2进行文本主题建模,使用算法3进行维基实体解析,获取来自Wikipedia知识库和Freebase知识库的实体、实体类别、实体间关系等实体信息和文本主题等文本信息。本文利用得到的附加信息提出了6种短文本扩充方法,这些方法将在3.1节中进行详细描述。

第二步,提出了融合外部知识的短文本异质信息网构建算法,并构建了短文本异质信息网框架,它可以集成任何类型的附加信息以及捕获它们之间的关系。利用提出的6种短文本扩充方法,生成包含doc, word, doc_topic, entity, Wikipedia_entity等类型节点的异质信息网。最后使用基于节点的元路径相似方法得到不同文本间的相似性。

3.1 问题定义

定义1(异质信息网络) 将异质信息网络定义为一个图 $G=(V,E,A,R)$ 。其中 V 是节点集合,图 G 中任意节点 $v \in V$; E 是边集合,图 G 中任意边 $e \in E$; A 是节点类型,映射 $\Phi: V \rightarrow A$; R 是边类型,映射 $\psi: E \rightarrow R$ 。图中的任意节点($v \in V$)都有预定义的节点类型, $\Phi(v) \rightarrow A$,任意边($e \in E$)都有预定义的边类型, $\psi(e) \rightarrow R$,且 $|A| > 1$ 或 $|R| > 1$ 。

定义2(网络模式) 网络模式是定义在对象类型集合 A 上的有向图,并以 R 上的关系为边^[16]。异质信息网络模式是构建异质信息网络所遵循的基本规则,利用异质信息网络具有融合丰富信息的特性,本文构建了包含doc_topic, doc, entity, word, Wikipedia_entity类型节点的短文本异质信息网。

3.2 短文本特征扩充方法

为了解决短文本数据稀疏的问题,本文设计了6种不同的文本扩充方法来引入外部知识。所有方法都使用相同的参数和设置来识别文本中的概念。它们之间的区别在于添加到短文本异质信息网中的文本信息不同。

由于单独的短文本特征过于稀疏,类型单一,不足以构建异质信息网,因此本文使用原始短文本与Freebase中的实体及实体间的关系(即下文的ER方法结果)作为异质信息网中的基础信息,考虑用ERT, ERW, ERCT, ERTW, ERCW, ERCTW这6种不同方法进行特征扩充,通过向短文本异质信息网中添加实体、实体类型、实体关系、文章主题、Wikipedia实体等特征信息来丰富短文本的语义。

ER(实体+实体关系):原始文本通过实体解析框架对文本中的概念进行扩展,同时解决了歧义问题。

本文使用 λ -DCS作为逻辑语言,从文本的每个句子中可以解析出4种可能的 λ -DCS逻辑形式。将文本中的某些词或短语映射到Freebase知识库中的相应类型的实体,得到具有不同实体类型的实体;同时结合上下文信息,对文本中的实体的含义进行消歧,最终从Freebase知识库查询并提取出输入文本的附加信息。

$$ER=[w_1, \dots, w_n, e_1, \dots, e_m]$$

其中, w_i 为短文本中的单词, e_j 为添加到原始文本中的已识别的Freebase实体, i 和 j 分别为单词数量和已识别的实体数量。

问题描述:

输入原始数据集DataSets,输出在Freebase知识库下扩充的语义解析数据集。

基本步骤:

- (1)通过Accurate Online Disambiguation of Entities工具进行实体识别;
- (2)利用Freebase文本解析器得到实体信息语料库corpus,其中包含每一个实体对应的实体类型与实体间关系等实体信息;

(3)计算候选实体类型在文本中出现的流行度得分,并利用流行度得分进行语义过滤;

(4)获取原始数据集中每一篇文本所包含的有效单词,并与前述步骤得到的实体及实体信息共同作为语义解析的结果 ParsingText 返回。

算法 1 Freebase 实体解析部分伪代码

```
输入:DataSets
输出:ParsingText
1. entities←NULL;
2. corpus←NULL;
3. ParsingText←NULL;
4. for doci∈DataSets
5.   for sentencei∈doci
6.     entities←NameEntityRecognition(sentencei);
7.     add entities to corpus;
8.   end for
9.   for entityi∈corpus
10.    EntityInfo←FreebaseTextParser(entityi);
11.    add EntityInfo to corpus;
12.   end for
13. for ParsingTexti∈corpus
14.   get PopularityScore;
15. end for
16. ParsingText ←filter(corpus);
17. for wordi∈sentencei
18.   if wordi is ValidWord
19.     add wordi to ParsingText;
20.   end if
21. end for
22. end for
23. return ParingText
```

Freebase 实体解析部分的算法以原始数据集合为输入,扩充得到解析后的具有实体、实体间关系与实体类型的数据集。算法 1 中第 5—8 步为进行实体识别,第 9—12 步为进行语义解析,通过 Freebase 文本解析器得到包含实体类型等的实体信息。第 13—16 步为结合上下文信息进行语义过滤,通过计算候选实体类型的得分,以在文本中流行度最高的类型为合法的实体类型,实现多义实体的消歧处理。第 17—21 步为获取文本中具有实际含义的有效单词,从而尽可能地保留原始文本的语义信息,最后返回扩充后的结果。

ERT(实体+实体关系+文本主题):首先使用 ER 方法扩展文本;随后基于 LDA 技术挖掘短文本的潜在主题。每个文本的主题 $t_i=(\theta_1,\cdots,\theta_w)$ (w 表示词汇量)由单词的概率分布来表示。根据概率分布将概率最大的 m 个主题分配给每篇文本,通过增加文本主题,可以解决文本之间的相关性问题,因为不同文本之间可能具有相同的主题。例如,一篇文本写的是“奥巴马”总统,另一篇文本中是“拜登”总统,那么这两份文件就变得相似,因为它们的主题都是“总统”,并与“美国”联系在一起。

$$ERT=ER+[t_1,\cdots,t_m]$$

其中, t_i 为短文本主题建模过程中产生的潜在主题, i 为短文

本主题数量。

问题描述:

输入算法 1 解析后的数据集 ParsingText,输出文本主题数据集 DocTopics。

基本步骤:

- (1)将算法扩充后的数据集 ParsingText 中的每一篇文本添加到语料库 corpus 中;
- (2)预处理得到的扩充语料库;
- (3)利用 CountVectorizer 类提取文本特征;
- (4)通过 LatentDirichletAllocation 类进行主题建模,得到主题单词的概率分布表示;
- (5)由 NaiveArgTopK 函数得到每篇文本对应的主题。

算法 2 文本主题建模部分伪代码

```
输入:ParsingText
输出:DocTopics
1. DocTopics←NULL;
2. corpus←NULL;
3. for doci∈ParsingText
4.   add doci to corpus;
5. end for
6. corpus←Preprocess(corpus);
7. BoWFeature ← CountVectorizer(corpus);
8. LdaFeature←LatentDirichletAllocation(BoWFeature);
9. DocTopics← NaiveArgTopK(TopK,LdaFeature);
10. return DocTopics.
```

主题建模部分的算法以算法 1 实体解析后的数据集作为输入,对每篇文本进行主题建模,得到文本主题数据集 DocTopics。算法 2 中第 3—5 步为将算法 1 解析后的文本添加到语料库训练的 corpus 中;第 6 步为对语料库 corpus 进行去停用词等的预处理,从而规范化文本;第 7 步为将语料库中的每篇文本转化为数字化的向量;第 8 步为进行 LDA 主题建模,得到每篇文本对应的主题单词的概率分布表示;第 9 步为根据概率分布将概率最大的 Top K 个主题分配给每篇文本;最后返回文本主题数据集。

ERW(实体+实体关系+Wikipedia 实体):首先使用 ER 方法扩展文本,随后使用实体链接工具 TAGME 识别每一篇文本中的实体,并将它们映射到 Wikipedia 知识库,得到每一篇文本在 Wikipedia 中的实体。为了防止一个知识库中的信息不够完善,需要综合使用 Wikipedia 知识库中的实体进行补充。有些单词在 Freebase 知识库中并不相关,但映射到 Wikipedia 知识库中却是相关的,从而最大限度地得到短文本的实体信息。

$$ERW=ER+[\omega_1,\cdots,\omega_m]$$

其中, ω_i 为实体解析过程中识别的 Wikipedia 知识库中的实体, i 为已识别的 Wikipedia 实体数量。

问题描述:

输入算法 1 解析后的数据集 ParsingText,输出 Wikipedia 实体数据集 WikiEntities。

基本步骤:

- (1)遍历数据集 ParsingText 得到每一遍文本,利用 Tag-

Me 工具得到解析后 Wikipedia 实体集合 entities;

(2) 过滤掉链接概率低于给定阈值 t_1 或者置信分数低于给定阈值 t_2 的 Wikipedia 实体;

(3) 将过滤后的 Wikipedia 实体添加到 WikiEntities 中, 并将 WikiEntities 返回。

算法 3 Wikipedia 实体解析部分伪代码

输入: ParsingText

输出: WikiEntities

```

1. WikiEntities ← NULL;
2. entities ← NULL;
3. for doci ∈ ParsingText
4.   entities ← TagMe(ParsingText.get(doci);
5.   for entityi ∈ entities
6.     if link_probability(entityi) ≤ t1 || rho(entityi) ≤ t2
7.       continue;
8.     else
9.       add entityi to WikiEntities;
10.    end for
11. end for
12. return WikiEntities

```

Wikipedia 实体解析部分的算法以算法 1 解析后的数据集为输入, 输出 Wikipedia 实体数据集 WikiEntities。算法 3 中第 4 步为利用 TagMe 工具发现文本中包含的 Wikipedia 实体的概念, 得到链接概率、置信分数等的概念信息; 第 5—10 步为利用概念信息进行语义过滤, 并将过滤后的 Wikipedia 实体添加到 WikiEntities 中, 最后返回 WikiEntities。

ERCT(实体+实体关系+实体类型+文本主题): 首先使用 ERT 方法扩展文本, 随后从 Freebase 知识库获取已识别概念的类型, 并将其添加到短文本异质信息网中。实体类型通常是已识别概念的名词, 如音乐家、总统等。由算法 1 与算法 2 可以得到实体、实体关系、实体类型和文本主题, 然后将实体、实体关系、实体类型和文本主题一起加入短文本异质信息网中, 融合更丰富的信息。

$$ERCT = ERT + [c_1, \dots, c_m]$$

其中, c_i 为 Freebase 实体解析过程识别的实体类型, i 为已识别的实体类型数量。

ERTW(实体+实体关系+文本主题+Wikipedia 实体): 首先使用 ERT 方法扩展文本, 随后使用实体链接工具 TAGME 识别每一篇文本中的 Wikipedia 实体。将实体、实体关系、文本主题与 Wikipedia 实体一起加入短文本异质信息网中, 得到更丰富的信息。

$$ERTW = ERT + [\omega_1, \dots, \omega_m]$$

其中, ω_i 为 Wikipedia 实体解析过程识别的 Wikipedia 知识库中的实体, i 为已识别的 Wikipedia 实体数量。

ERCW(实体+实体关系+实体类型+Wikipedia 实体): 首先使用 ERW 方法扩展文本, 随后从 Freebase 知识库获取已识别概念的类型, 并将其添加到短文本异质信息网中。实体类型通常是已识别概念的名词, 如音乐家、总统等。

$$ERCW = ERW + [c_1, \dots, c_m]$$

其中, c_i 为 Freebase 实体解析过程识别的实体类型, i 为已

识别的实体类型数量。

ERCTW(实体+实体关系+实体类型+文本主题+Wikipedia 实体): 首先使用 ERCW 方法扩展文本, 随后使用 LDA 技术挖掘短文本的文本主题。

$$ERCTW = ERCW + [t_1, \dots, t_m]$$

其中, t_i 为短文本主题建模过程产生的潜在主题, i 为短文本主题数量。

3.3 融合外部知识的短文本异质信息网构建算法

本文提出了一个融合外部知识的短文本异质信息网构建算法, 构建了一个灵活的异质信息网络框架来建模短文本, 它可以集成任何类型的附加信息并捕获它们之间的关系。结合上述的 6 种文本扩充方法, 生成包含 doc_topic, doc, entity, word, Wikipedia_entity 类型节点的短文本异质信息网。

首先, 按照 3.1 节中的方法识别文本中的实体, 进行语义解析, 文本中的某些单词或短语被映射到通用知识库中的相应类型的实体, 得到具有不同实体类型的实体。同时结合上下文信息, 对文本中的实体的含义进行消歧。如果文本包含实体, 则会构建文本和实体之间的边。值得注意的是, 本文使用“命名实体类型”来表示文本中提到的命名实体的节点类型, 例如, “拜登”的节点类型是“政治家”。因此, 不同命名实体类型的实体的节点类型是不相同的。两个实体间通过知识库中连接实体类型的关系来连接。实体集为 $E = \{E_1, E_2, \dots, E_{V_t}\}$, 其中 V_t 是实体的数量。 $t = 1, \dots, T$, 其中 T 是在 Freebase 知识库中找到的命名实体类型的总数。

随后, 通过 LDA 挖掘潜在主题来丰富文本的语义, 给文本分配最可能的多个主题, 构建文本和文本主题的边。文本的节点类型为“doc”, 文本主题的主题为“doc_topic”, 每篇文本对应概率最大的 P 个主题。文本集为 $D = \{D_1, D_2, \dots, D_M\}$, 其中 M 是文本的数量。

通过实体链接工具 TAGME 识别每篇文本的 Wikipedia 实体, 构建文本与 Wikipedia 实体的边, Wikipedia 实体的节点类型为“Wikipedia_entity”。Wikipedia 实体集为 $WE = \{WE_1, WE_2, \dots, WE_I\}$, 其中 I 是文本主题的数量。

若一个单词包含在某篇文本中, 则在单词与文本间构建边进行相连。单词的节点类型为“word”。文本主题集为 $W = \{W_1, W_2, \dots, W_J\}$, 其中 J 是文本主题的数量。

问题描述:

输入文本集合 Docs, 文本到单词的映射 DocWordMap, ParsingText, 文本到主题映射 DocTopicMap, 文本到 Wikipedia 实体的映射 DocWikiEntityMap, 文本到 Freebase 实体的映射 DocEntityMap, 输出异质信息网络 G 。

基本步骤:

(1) 遍历文本到单词的映射, 将文本与单词之间的关系以边的形式添加到 G 的边集合 E 当中;

(2) 遍历单词集合, 利用预训练的 Glove 向量计算单词之间的相似度, 若超过指定的输入阈值 SIM , 则将单词与单词之间的关系以边的形式添加到 G 的边集合 E 当中;

(3) 分别遍历文本到主题映射 DocTopicMap, 文本到 Wikipedia 实体的映射 DocWikiEntityMap, 文本到 Freebase

实体的映射 DocEntityMap,将文本与文本主题、文本与 Wikipedia 实体、文本与 Freebase 实体之间的关系以边的形式添加到 G 的边集 E 中;

(4)遍历实体类型的集合,若存在实体间的关系,则将实体之间关系到 G 的边集 E 中。

算法 4 短文本异质信息网构建部分伪代码

```
输入: Docs, DocWordMap, DocTopicMap, DocWikiEntityMap, DocEntityMap, SIM
输出: G=(V, E)
1. V←NULL;
2. E←NULL;
3. for doci ∈ Docs
4.   for wordi ∈ DocWordsMap. get(doci)
5.     add wordi to Words;
6.     E. add(doci, wordi);
7.   end for
8.   for wordi, wordj ∈ Words
9.     if wordi = wordj
10.      continue;
11.    end if
12.    if Sim(wordi, wordj) > SIM:
13.      E. add(wordi, wordj);
14.    end if
15.  end for
16. for topici ∈ DocTopicMap. get(doci)
17.   E. add(doci, topici);
18. end for
19. for wikiEntityi ∈ DocWikiEntityMap. get(doci)
20.   E. add(doci, wikiEntityi);
21. end for
22. for entityi ∈ DocEntityMap. get(doci)
23.   add entityi to entities;
24.   E. add(doci, entityi);
25. end for
26. for entityi, entityj ∈ entities
27.   if exists R(entityi, entityj)
28.     E. add(entityi, entityj);
29.   end if
30. end for
31. end for
32. return G=(V, E)
```

算法 4 为本文融合外部知识的短文本异质信息网构建算法伪代码,将短文本集合、短文本到单词集合的映射、短文本到主题集合的映射、短文本到 Wikipedia 实体集合的映射、短文本到 Freebase 实体的映射集合作为算法 4 的输入。算法 4 第 1—2 步为异质信息网的初始化;第 3—31 步为异质信息网中的不同类型的节点与边的添加过程,重复执行第 3—30 步,直到结束遍历短文本节点集合;第 4—7 步为生成所有单词节点的集合,并且将短文本节点与单词节点以边的形式

添加到异质信息网络的过程;第 8—15 步是在相似的不同单词之间加边的过程;重复执行第 9—15 步,如果是相同单词,执行下一次循环遍历,如果是不同单词对且之间的相似度超过给定阈值,则在这两个单词之间添加权重为相似度的边,直至两两单词组合遍历完全为止;第 16—18 步为在短文本与主题词间加边的过程;第 19—21 步为在短文本与 Wikipedia 实体间加边的过程;第 22—25 步为获取 Freebase 实体类型集合,并且在短文本与 Freebase 实体类型之间加边的过程;第 26—30 步为不同 Freebase 实体类型之间加边的过程,最后输出异质信息网络 G 。

3.4 基于元路径计算文本相似性

通过引入文章主题、实体、实体类型、实体关系、Wikipedia 实体等信息,两个文本可以通过多条元路径链接在一起。例如,如果两个文本通过元路径 Document→politics→Document 连接,则相应的元路径实例的数量可用于衡量两个文本之间的相似性。

首先采用 PageRank-Nibble 算法,寻找有用的子图来分配元路径。基于每一条特定的元路径,采用剪枝策略,对大规模异质信息网络进行修剪,从而得到包含给定文本节点集的局部图。根据局部图,计算并存储基于每一条特定元路径的交换矩阵。给定选定的元路径,使用 KnowSim^[17]方法进行文本相似性度量。

4 实验及评估

本文使用相似性度量结果来验证文本的扩充效果,与传统的 3 种相似性度量方法的 6 种文本扩充方法,以及与目前常见的短文本匹配算法 BERT, InferSent 和 Skip-thought 在两个短文本数据集上进行了比较。

4.1 数据集

本文在来自不同领域的两个不同数据集上进行了实验,第一个数据集是 Google Snippets¹⁾数据集,它包含 10 060 个训练样本和 2 280 个测试样本,分为 8 个类别,平均每个 snippet 有 18.07 个词。这些片段是使用不同域的预定义短语从网络搜索的结果中选择的。第二个数据集是 AG News²⁾数据集,它是由康奈尔大学发布的,从 2 000 多个不同的新闻来源搜集的超过一百万的新闻文章,其中包含 496 835 条 AG News 语料库中超过 2 000 个新闻来源的文章。该数据集仅采用了标题和描述字段,每种类别均拥有 30 000 个训练样本和 1 900 个测试样本。

4.2 基线方法

为了衡量所提方法的改进效果,本文将只添加了实体的短文本异质信息网作为基线方法进行比较。同时将目前代表性的 3 个相似性度量(余弦、Jaccard 和 dice 方法)也作为基线方法进行了对比。本文使用指定的世界知识作为特征来增强余弦、Jaccard 和 dice 方法。为了公平比较,3 种方法也都只添加了实体。

¹⁾ https://www.researchgate.net/figure/Google-snippets-as-training-test-data_tbl1_221023426

²⁾ <https://paperswithcode.com/dataset/ag-news>

4.3 评估方法

本文采用广泛使用的皮尔逊相关系数^[18]作为评价指标。该评估指标也被用作 SemEval STS^[19]中测试方法性能的官方指标,在文本相似性度量中被广泛使用。皮尔逊相关系数定义为:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

其中 $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, n 等于 X_i 或 Y_i 的总数。如果相似性结果与基本事实完全匹配,则相关性得分为 1;如果相似性结果是随机的,则相关性得分为 0。一般来说,得分越高相似性结果越好。

4.4 基于 4 种相似性度量方法的 6 种文本扩充方法比较

本文从两方面进行了对比,一方面,对两个数据集中的文本基于异质信息网使用了提出的 6 种文本扩充方法来对比不同的文本扩充方法的效果;另一方面,在 3 个代表性的相似性度量(余弦、Jaccard、dice 方法)和基于异质信息网方法上分别使用了 6 种文本扩充方法,比较了基于传统的 BOW 模型文本表示和基于异质信息网的文本表示的效果,最后综合比较了基于哪种文本表示方法的哪种文本扩充方法最好。在实验中,使用皮尔逊相关系数评估了相似性结果,结果如表 1 所列,每种方法中最优的皮尔逊相关系数以粗体显示。HTE 与 3 种代表性的相似性度量方法进行了对比。本文使用指定的外部知识作为特征来增强余弦、Jaccard 和 dice 方法。

特性设置的定义如下:

BOW+ TOPIC+ ENTITY:词袋模型中集成了来自 Freebase 知识库中的实体信息和通过 LDA 获得的主题词。

BOW+ ENTITY+ Wikipedia ENTITY:词袋模型中集成了来自 Freebase 知识库中的实体信息和 Wikipedia 知识库中的实体信息。

BOW+ TOPIC+ ENTITY+ Categories:词袋模型中集成了文本主题、来自 Freebase 知识库中的实体信息和实体精确类别。

BOW+ ENTITY+ TOPIC+ Wikipedia ENTITY:词袋模型中集成了文本主题、来自 Freebase 知识库中的实体信息和 Wikipedia 知识库中的实体信息。

BOW+ ENTITY+ Categories+ Wikipedia ENTITY:词袋模型中集成了来自 Freebase 知识库中的实体信息、实体精确类别和 Wikipedia 知识库中的实体信息。

BOW+ ENTITY+ TOPIC+ Categories+ Wikipedia ENT- TITY:词袋模型中集成了文本主题、来自 Freebase 知识库中的实体信息、实体精确类别和 Wikipedia 知识库中的实体信息。

表 1 列出了基于不同文本相似性度量方法的 6 种文本扩充方法在 Snippets 数据集上的结果。对于 Snippets 数据集,基于异质信息网使用 ERCTW 方法的效果最好。从图 3 可以看出,与基线方法相比,在 4 种文本相似性度量方法(余弦、Jaccard、dice 和基于异质信息网方法)上使用本文提出的 6 种

文本扩充方法后,相似性准确度均有所提升,提升幅度为 0.39%~5.89%,最佳方法的效果提高了 5.89%。这表明在 Snippets 数据集的短文本中,一方面,基于异质信息网的方法相比其他文本表示方法来说效果更好;另一方面,说明加入实体、实体类别、实体主题及文本主题等信息即本文提出的 6 种文本扩充方法在实验的所有文本表示方法上都是有效的。

表 1 不同文本扩充方法在 Snippets 数据集的相似性度量结果

Table 1 Similarity measurement results of different text augmentation methods in Snippets dataset				
Enrich method/ Similarity measurement method	Cosine	Jaccard	Dice	HTE
BASE	0.78966	0.72799	0.75087	0.84356
ERT	0.78998	0.75234	0.76075	0.84875
ERW	0.81030	0.76271	0.78102	0.84748
ERCT	0.86062	0.82589	0.82855	0.86735
ERTW	0.80452	0.82589	0.76433	0.85168
ERCW	0.86657	0.77472	0.87099	0.87241
ERCTW	0.87860	0.85916	0.84601	0.90255

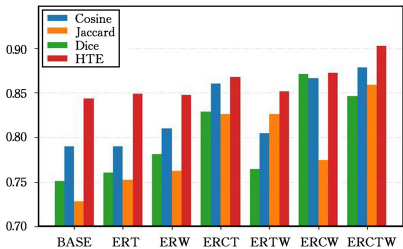


图 3 不同文本扩充方法在 Snippets 数据集上的相似性度量结果直方图

Fig. 3 Histogram of similarity measurement results of different text enrichment methods in Snippets dataset

表 2 和图 4 给出了基于不同文本相似性度量方法上 6 种文本扩充方法在 Agnews 数据集上的结果,与对比方法相比,在 4 种文本相似性度量方法(余弦、Jaccard、dice 和基于异质信息网方法)上使用本文提出的 6 种文本扩充方法后,相似性准确度均有所提升,提升幅度为 0.9%~34%,基于异质信息网络方法使用 ERCW 获得了最好的结果。综合 6 种文本扩充方法在两个数据集上的总体评价,本文所提 6 种方法均可以有效提高性能。同时可以看到,在添加相同的外部知识后,本文所使用的异质信息网络框架比其他 3 种方法更能有效融合信息,具有鲁棒性。

表 2 不同文本扩充方法在 Agnews 数据集上的相似性度量结果

Table 2 Similarity measurement results of different text augmentation				
Enrich method/ Similarity measurement method	Cosine	Jaccard	Dice	HTE
BASE	0.16974	0.15834	0.14740	0.28220
ERT	0.17221	0.16285	0.15099	0.29212
ERW	0.22629	0.21637	0.21333	0.32578
ERCT	0.36530	0.43543	0.36790	0.60919
ERTW	0.18451	0.15867	0.16902	0.34159
ERCW	0.40718	0.46003	0.40981	0.63039
ERCTW	0.27593	0.27015	0.26996	0.60622

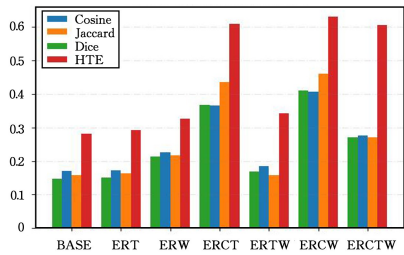


图 4 不同文本扩充方法在 Agnews 数据集上的相似性度量结果直方图

Fig. 4 Histogram of similarity measurement results of different text enrichment methods in Agnews dataset

4.5 与主流的短文本匹配方法比较

将本文方法与目前主流的效果较好的 BERT 方法、InferSent 方法、Skip-thought 方法进行对比。目前 BERT^[20] 模型在多个自然语言处理任务上都取得了突破,大幅度提升了实验效果,具有易于使用、稳定性强等诸多优点。由于其灵活的训练方式和不错的效果,在许多任务中得到了深入的研究和应用。

InferSent^[21]是一种基于双向 LSTM 的网络。它使用 SNLI(Stanford Natural Language Inference)数据集通过有监督学习进行训练,是一种可以在句子层面有效学习语句嵌入的模型,也是目前代表性的监督式学习模型。

Skip-thought^[22]通过给定中心句来预测前后两个句子,有效考虑了句子间的上下文关系,是一种传统的无监督句表示学习方法,也是 BERT 之前获得高质量句子嵌入的常用方法,可以有效解决情感分析、句对匹配等句子级别的任务。

因此,本文选择了这 3 种模型与 ERCTW 和 ERCW 方法进行比较。

在 BERT 对比实验中,为了具有更快的训练速度和更低的运算资源需求,本文采用了参数量相对较小的 BERT-Base (Uncased)模型 (12-layer, 768-hidden, 12-heads, 110 M parameters),之后采用了迁移学习中 fine-tune(微调)的方法来调整预训练模型,使得训练出来的模型可以更好地适应新闻等特定领域的短文本相似文本判定任务。

在 InferSent 对比实验中,使用在通过 PTB tokenizer 预处理后的文本上进行训练的 GLoVe 单词向量 (840 B tokens, 2.2 M vocab, cased, 300 d vectors),并加载与之对应的 InferSentV1 模型,之后构建单词向量对应的词汇表并对短文本句子进行编码,从而得到特定领域的短文本句子的嵌入式表示。

在 Skip-thoughts 对比实验中,使用在 BookCorpus 上训练的两个独立的模型 (2 400 维的单向编码模型 uni-skip 和向前与后向各 1 200 维的双向编码模型 bi-skip)得到短文本句子特征,最后依旧采用皮尔逊相关系数作为评价指标。

表 3 列出了不同模型在 Agnews 数据集和 Snippets 数据集上的结果。可以看出,本文提出的 ERCTW 方法在 Snippets 数据集上取得了最好的效果,ERCW 方法在 Agnews 数据集上取得了最好的效果。

表 3 不同模型在所有数据集上的结果
Table 3 Results of different models on all datasets

Enrich method /Dataset	Agnews	Snippets
ERCTW	0.60622	0.90255
ERCW	0.63039	0.87241
BERT	0.57082	0.63941
InferSent	0.29410	0.66420
skip-thought	0.20140	0.54940

从图 5 可以看出,在两个数据集上,本文的两种方法均优于目前主流的文本匹配算法。这是因为与现有文本匹配算法相比,本文使用文本扩充方法丰富了短文本特征,有效地解决了文本特征的稀疏性问题,从而提高了相似性度量结果的准确性。同时与 BERT, InferSent, Skip-thoughts 等模型相比,本文方法不需要任何标记数据、耗时的训练阶段或繁琐的参数调整步骤。

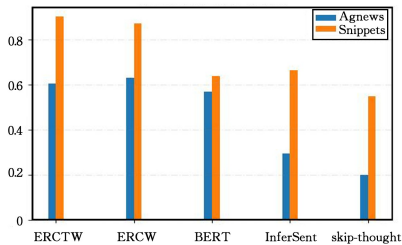


图 5 不同模型在所有数据集上的结果直方图

Fig. 5 Histograms of results of different models on all datasets

通过对基于 4 种相似性度量方法的 6 种文本扩充方法进行比较,一方面说明基于异质信息网的方法比其他文本表示方法效果更好,另一方面说明加入实体、实体类别、实体主题及文本主题等信息,即本文提出的 6 种文本扩充方法在实验的所有文本表示方法上都是有有效的。通过与主流的短文本匹配方法相比,验证了基于异质信息网的文本扩充方法的有效性和可行性。

结束语 本文提出了一种灵活的异质信息网络框架来建模短文本,利用异质网络可以融合丰富信息的特性,引入来自 Freebase 知识库和 Wikipedia 知识库的实体、实体类别、实体间关系等实体信息和文本主题等文本信息来丰富短文本特征以及捕获它们之间的关系。本文在两个开放数据集上对 6 种不同的文本扩充方法进行了全面的评价和比较。实验表明,本文提出的 6 种方法在两个数据集上性能更优。本文提出的框架具有鲁棒性,可以集成任何类型的附加信息,所提出的方法能够解决短文本的数据稀疏问题,可以以无监督的方式高精度地对短文本进行相似性度量。

由于短文本数量巨大,同时随着使用文本扩充方法引入越来越多的外部知识,导致在异质信息网中存在大量的实体类型,因此,如何更加快捷有效地找到异质信息网的元路径来进行计算变得十分重要。下一步将重点改进元路径的查询算法,降低实验的时间复杂度;同时,如何提取最重要的特征,剔除不相关的特征,也是下一步研究的重点。

参 考 文 献

[1] LI C L, CHEN S Q, XING J, et al. Seed-guided topic model for

document filtering and classification[J]. ACM Transactions on Information Systems(TOIS), 2018, 37(1): 1-37.

[2] NIE L Q, LI Y Q, FENG F L, et al. Large-scale question tagging via joint question-topic embedding learning[J]. ACM Transactions on Information Systems(TOIS), 2020, 38(2): 1-23.

[3] WANG X, CHEN R, JIA Y, et al. Short text classification using wikipedia concept based document representation[C]// 2013 International Conference on Information Technology and Applications. IEEE, 2013: 471-474.

[4] MOURINO-GARCIA M A, PEREZ-RODRIGUEZ R, ANIDO-RIFON L, et al. Wikipedia-based hybrid document representation for textual news classification[J]. Soft Computing, 2018, 22(18): 6047-6065.

[5] BOLLACKER K D, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008: 1247-1250.

[6] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[M]. Heidelberg: Springer, 2007: 722-735.

[7] SHI C, PHILIP S Y. Heterogeneous information network analysis and applications[M]. Cham: Springer International Publishing, 2017.

[8] SHI C, SUN Y, PHILIP S Y. Research status and future development of heterogeneous information networks[J]. Communications of the Computer Society, 2017, 11(13): 35-40.

[9] YAO D, BI J, HUANG J, et al. A word distributed representation based framework for large-scale short text classification [C]// 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, 2015: 1-7.

[10] FLISAR J, PODGORELEC V. Improving short text classification using information from DBpedia ontology[J]. Fundamenta Informaticae, 2020, 172(3): 261-297.

[11] GHADERY E, MOVAHEDI S, FAILI H, et al. A multilingual ngram-based convolutional network for aspect category detection in online reviews[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6441-6448.

[12] WANG C, SONG Y, LI H, et al. Text classification with heterogeneous information network kernels [C] // Proceedings of AAAI. 2016: 2130-2136.

[13] YANG T C, HU L M, SHI C, et al. HGAT: Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification [J]. ACM Transactions on Information Systems, 2021, 39(3): 1-29.

[14] HU B, LU Z, LI H, et al. Convolutional neural network architectures for matching natural language sentences[C]// Proceedings of the Conference on Neural Information Processing Systems. 2014: 2042-2050.

[15] PANG L, LAN Y, GUO J, et al. Text matching as image recognition[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016: 2793-2799.

[16] SUN Y, YU Y, HAN J. Ranking-based clustering of heterogeneous information networks with star network schema [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 797-806.

[17] WANG C G, SONG Y Q, LI H R, et al. KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks[C] // Proceedings of the 2015 IEEE International Conference on Data Mining(ICDM). 2015: 1015-1020.

[18] LI J, ZHANG X, ZHOU X. ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study [J]. JMIR Medical Informatics, 2021, 9(1): e23086.

[19] CER D, DIAB M, AGIRRE E, et al. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation[J]. arXiv:1708. 00055, 2017.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810. 04805, 2018.

[21] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[J]. arXiv:1705. 02364, 2017.

[22] KIROS R, ZHU Y, SALAKHUTDINOV R R, et al. Skip-thought vectors[C] // Advances in Neural Information Processing Systems. 2015: 3294-3302.



LYU Xiao-feng, born in 1996, postgraduate. His main research interests include machine learning and intelligent information processing.



ZHAO Shu-liang, born in 1967, Ph.D., professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning and intelligent information processing.

(责任编辑:何杨)