ACL 2019 | 图表示解决 长文本关系匹配 问题：腾讯提出概念交互图算法

[机器之心](javascript:void(0);) 1周前

机器之心发布

**作者：Bang Liu、Di Niu等**

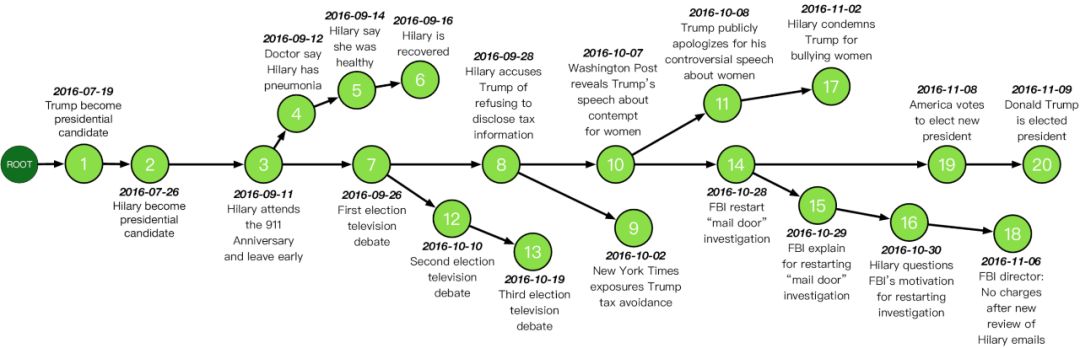
文章之间关系匹配是自然语言处理领域的重要问题。传统算法忽略了文本内部语义结构，而深度神经网络目前主要用于句子对之间的匹配。同时由于长文本对计算量需求较大，且目前缺乏训练数据集，因此长文本的匹配问题一直难以解决。对此，来自阿尔伯塔大学和腾讯 PCG 移动浏览器产品部的研究者提出了

概念交互图（Concept Interaction Graph）算法，

对比现有的文章关系匹配算法有明显的效果提升。该论文已被自然语言处理顶会 ACL 2019 接收，项目代码和数据集已开源。

**引言**

判断两篇文章之间的语义关系对于新闻系统等应用有着重要的意义。例如，通过对新闻文章之间的关系判断，一个新闻应用可以将讲述同样的事件的文章聚类在一起，去除冗余，并形成事件发展的脉络。在图 1 中，「2016 美国总统大选」这一故事的主要信息被组织成一条故事树。其中的每个节点，代表了讲述该故事中同样的一个子事件的文章集。这种文本组织方式，在信息爆炸的时代，能给人们带来极大的便利。



*图 1. 2016 美国总统大选故事脉络。（图源自：Growing Story Forest Online from Massive Breaking News）*

过去的研究工作，包括传统的 TF-IDF，BM25，LDA 等等，以及基于神经网络的 DSSM，C-DSSM，ARC-I，ARC-II 等等，不能很好地解决长文本的关系匹配问题。一方面，传统算法忽略了长文章内部的语义结构。另一方面，基于深度神经网络的模型目前主要用于句子对之间的匹配，难以处理文章长度的文本。因为随着长度的增加，模型计算量大大增加，并且缺乏训练数据。

针对长文本的匹配问题，我们利用「化整为零，分而治之」的思想，提出了一种文本表示形式。

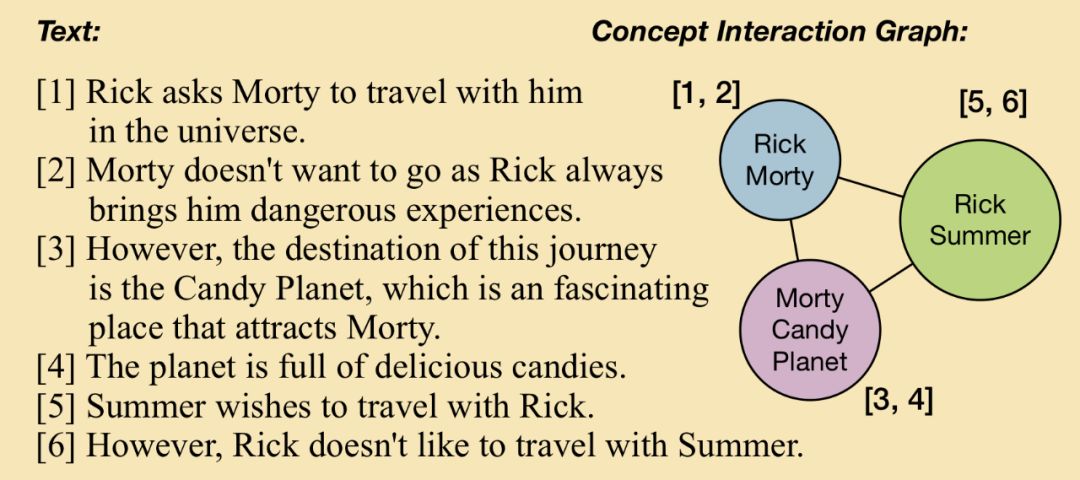
这种表示利用图结构（概念交互图）来分解长文章的内容，

将长文章的匹配变成图中每个节点上的短文本匹配，

并利用图卷积神经网络综合各个节点的匹配结果，得到最终的整体匹配结果。

我们把这种图叫做概念交互图（Concept Interaction Graph），图中每个节点称为一个概念，每个概念是一个关键字集合，集合中的关键字在文章中紧密相关。而所有的节点则把文章中所有的关键字划分成了多个子集，代表文章的多个子话题。

在得到节点之后，我们按照相似度将文章中的每一个句子，分配到不同的节点上。因此，每个节点上会包含两篇文章的一部分句子。节点之间的边则代表节点间的联系的紧密度。在这里，我们将两个节点的文本相似度作为边权重。如图 2 所示，该例子中包含的关键字被组织成三个节点，每个节点上有一部分句子集。



*图 2. 概念交互图示例*

在将长文本利用图分解之后，每个节点上的匹配可以利用句子匹配模型计算匹配特征，或者手动设计提取特征。为了充分利用文章内的语义结构，我们进一步利用图卷积神经网络（Graph Convolutional Network）来综合各个节点的匹配特征，以得到最终的结果。

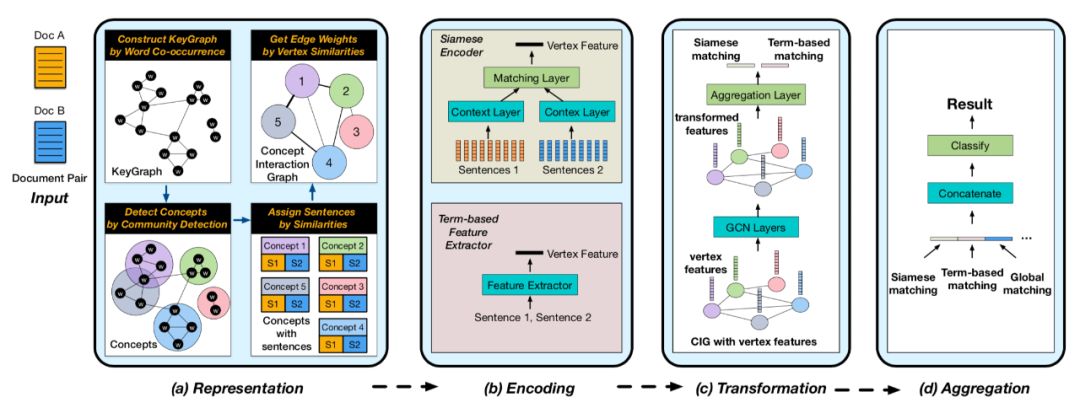
最后，目前的工作缺乏针对「长文本对关系」这个问题的实验数据。为了评估实验效果，我们构建了两个分别包含三万对新闻文章的二分类数据集。第一个数据集用于判断两个新闻是否讲述同一个子事件，第二个数据集用于判断两个新闻是否讲述同一个故事。例如，在图 1 中，任意两个文章都是讲述同一个故事「2016 美国总统大选」的，但是只有属于同一个节点的文章，才是讲述同一个事件的。

我们的主要贡献点包括：

* 我们提出了概念交互图（Concept Interaction Graph）用于表示长文本或者文本对；
* 结合概念交互图和图卷积神经网络，我们提出了针对长文本匹配的模型。实验证明该方法对比一系列已有的算法有明显的效果提升；
* 我们构建并开源了两个分别包含三万对文章的长文本对关系分类数据集，用于后续研究。实验代码也已开源。

值得注意的是，我们的算法并不局限于判断两篇新闻文章是否讲述同一个子事件或者同一个故事。只要有相应的训练集，它可应用于不同的长文本关系判断任务。同时，它也可以应用于英语等其他语言。

**方法**



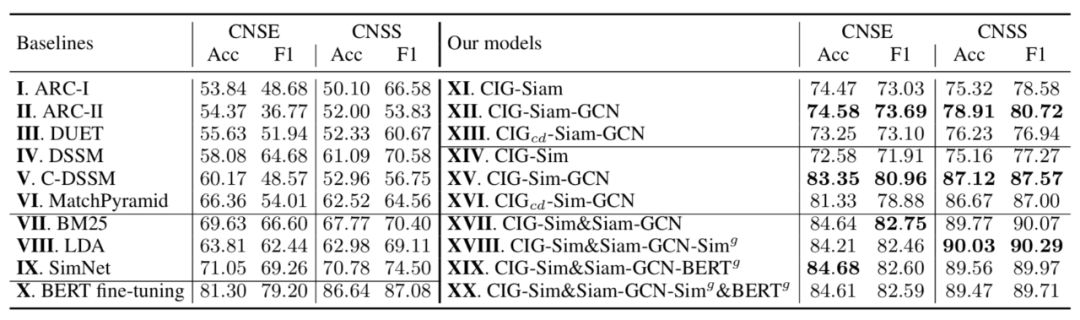
*图 3. 利用概念交互图表示文本以及利用 GCN 匹配的算法流程*

图 3 展示了长文本对关系匹配的流程，包括以下步骤:

* 文本图构建。在这一步中，我们利用关键字在句子中的共现构建 KeyGraph。之后，我们利用社区检测（community detection）算法将关键字聚类（注意，这一步是可选项），每个聚类是一个概念节点。在得到概念节点之后，将句子分配到各个节点上。同时，用节点文本的 TF-IDF 相似度表示边权重。
* 节点匹配特征编码。在这里，我们对每个节点上的文本对（来自两篇文章的句子集合分别拼接成一个文本）进行匹配，得到匹配特征。我们分别尝试了 Siamese Encoder 自动学习匹配特征，和计算各种 term-based 特征来作为节点特征向量。
* 节点特征转化。在得到每个节点的匹配特征向量之后，我们利用多层 GCN 进行特征转换。最后一层中，我们将所有节点的特征向量综合成一个向量（这里采用了 mean aggregation）。
* 综合匹配。最后，我们将得到的综合匹配向量，输入到一个多层神经网络分类器中进行关系分类。这里，我们可以拼接来自不同的编码器的匹配向量，以利用不同的特征。

**实验结果**

我们对比了本文的算法和一系列已有的文本匹配算法。同时，我们也对比了一系列本文算法的变种以分析不同部分的影响。表 1 展示了我们的实验结果。实验所用的两个数据集，Chinese News Same Event Dataset (CNSE), Chinese News Same Story Dataset (CNSS) 均已开源。

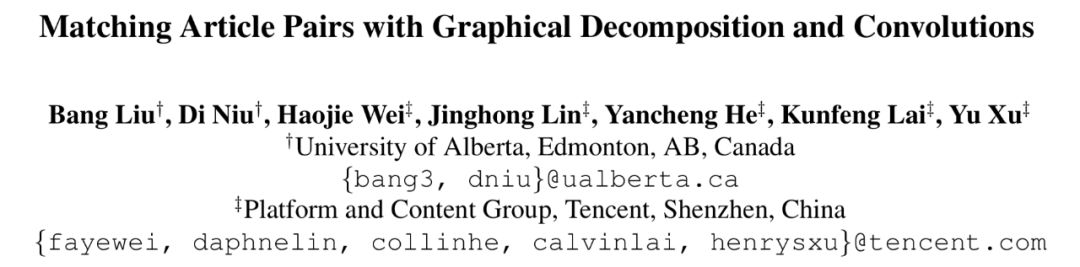


*表 1. 不同算法在 CNSE，CNSS 数据集上的分类效果对比*

从表 1 中，我们可以得到以下主要结论：

* 利用图分解，将文章化整为零，能明显提高匹配效果；
* 利用图卷积，将局部匹配综合，能明显提高匹配效果；

**论文：A User-Centered Concept Mining System for Query and Document Understanding at Tencent**



* 论文地址：https://arxiv.org/abs/1802.07459
* 相关数据资源：https://github.com/BangLiu/ArticlePairMatching

摘要：判断两个文章之间的关系，例如两个文章是否在讨论同一个事件，对于很多文本理解任务有重要意义。目前的算法较少处理长文本匹配的问题，也缺乏对长文本语义结构的充分利用。我们针对长文本匹配问题，提出了概念交互图文本表示。结合图卷积神经网络，我们提出了有效地处理长文本匹配问题的算法，将长文本化整为零，从局部匹配结果得到整体匹配关系的判断。同时，我们也构建了新的长文本匹配任务的数据集并开源。实验证明，我们的算法对比一系列现存的算法，在长文本匹配任务中效果提升明显。*https://mmbiz.qpic.cn/mmbiz_png/KmXPKA19gW8Zfpicd40EribGuaFicDBCRH6IOu1Rnc4T3W3J1wE0j6kQ6GorRSgicib0fmNrj3yzlokup2jia9Z0YVeA/640?wx_fmt=png&wxfrom=5&wx_lazy=1&wx_co=1*