# 基于事件本体的可用于虚假新闻识别的知识库构建方法

陈佩琳 学号：1120220668

**[摘要]**虚假新闻知识库可以用于虚假新闻识别等任务。近年来，研究人员开始借助外部知识库实现更加精准的虚假新闻识别。然而网络上的新闻是动态的，是不断更新、变化的，传统的知识库通常由结构化数据存储库提供支持，但是新闻中报道的最新事件通常会被这些数据库遗漏，因此传统的知识库存在遗漏最新信息的问题。以事件本体为中心构建知识库利用了自然语言处理技术，可以提取文本中提到的各种类型的事件，可以通过时间和地点锚定时间并把事件和实体进行连接，以表示整个事件的长期发展。这样的构建知识库的方式可以解决现有知识库构建方法的时效性、精度、普适性问题。

**[关键词]** 知识库；知识图谱；虚假新闻；事件本体

**1引言**

近年来，互联网上的新闻成为了各种事件的主要呈现形式。这些新闻是由自然语言构成的。由于自然语言具有非规范性、歧义性、个性化表达的特点；而且新闻是动态的，是不断在更新的，结构化的知识提取模式以及传统的以实体为中心的知识库难以适应虚假新闻的鉴别。因此，以事件为中心来构建知识图谱成为新的研究热点。事件是文本中包含的一种信息，其定义为在某个特定的时间以及特定的地点，由多个相关角色参与的一件事情或者一组事情。

知识图谱本质上是一种语义网络，一般的知识图谱包含实体以及实体之间的关系。实体是知识图谱中最基本的概念，一个实体代表了现实世界中的一个事物或者一个概念。关系代表的是实体和实体之间的关联。另外，知识图谱中的属性代表的是实体具有的某个特征，其描述了实体的相关信息。知识图谱刻画了现实世界中的事物以及事物之间的各种关系，其主要表示形式是三元组，包括属性三元组和关系三元组。知识图谱存储了结构化数据，适合对数据进行搜索和推理，已经在知识问答、智能搜索等方面发挥了重要的作用。随着知识图谱技术的发展，越来越多的研究者开始关注一种特殊的基于事件的知识图谱，即事件图谱。我们可以基于事件图谱构建关于某一事件的知识库，再将待鉴别的新闻与该知识库进行对比，判断其是否为虚假新闻。

文章“Content Based Fake News Detection Using Knowledge Graphs”提出从真新闻和假新闻分别构建一个知识图谱，并应用TransE学习三元组分数，从而进行虚假新闻检测。虚假新闻检测的效果很大程度上依赖于知识库的构建方法，以往的知识库大多是从结构化的百科数据库中获取的，倾向于表现世界当前的状态，忽略了事件的动态发展。

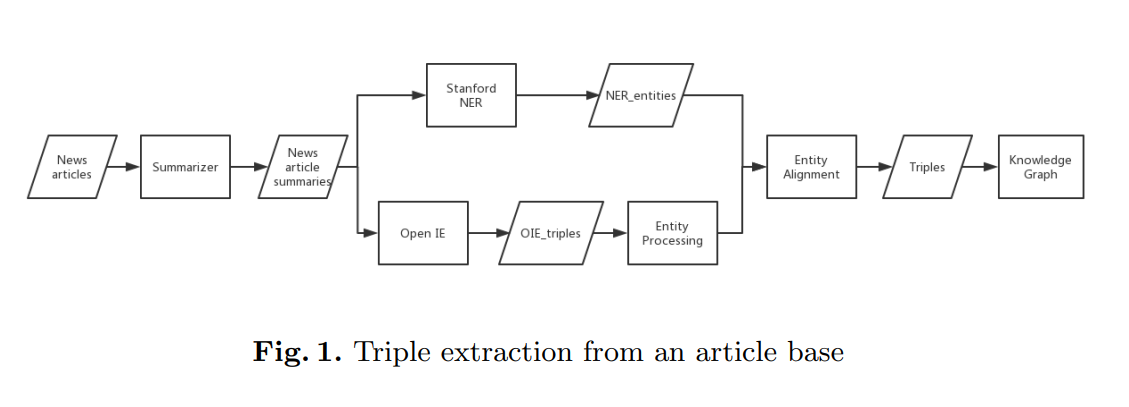
提高利用外部知识库进行虚假新闻核查的准确率需要像重建历史一样重建事件的全貌。因此本文将提出一种可能可行的改进知识库构建效果的方法：基于事件本体构建知识库。这样的方法可以让我们从新闻报道中重现一个新闻事件相关的“故事情节”，这种构建知识库的方式类似于人的大脑处理一个事件的来龙去脉的方式，有助于让机器对事件的全貌有充分地了解。

**2 现有模型介绍及其存在的问题**

文章“Content Based Fake News Detection Using Knowledge Graphs”提出通过文字内容，基于知识图谱来实现虚假新闻的检测。由于以往的虚假新闻监测大多数是通过写作风格实现的，不能直接辨别虚假新闻且准确率不够高，作者提出要构建知识库，从内容进行虚假新闻的检测，这样充分利用了事实数据，提高了虚假新闻检测的准确率。

作者提出要通过真新闻、假新闻、现有外部知识库（如DBPedia）分别建立知识库，检测过程中采用的B-TransE模型嵌入在一个低维向量空间的实体和关系。该模型通过将内容分别与以上建立的几个知识库进行比较来评估新闻文章的真实性。作者构建了3个知识库，分别是基于虚假新闻的“Fake News Knowledge Graph (FKG)”、基于可信新闻来源（例如BBC）的真实新闻知识库“True News Knowledge Graph(NKG)”、从包含事实性信息的DBPedia提取的事实知识库“Open Knowledge Graph（D4）”。知识库KG = {E, R, S}, 其中E是实体, R 是关系，S是“三元组”。这些知识库用图的形式存储从新闻文本中提取的“三元组”（即主谓宾）。作者这一步使用的工具包括OpenIE (Open Information Extraction)和 Stanford Named Entity Recognition (NER)。然后，通过应用诸如代词消歧义（pronoun disambiguation）、动词限定化（将动词转换为其基本形式）和命名实体识别（在文本中识别和分类适当的名词）等技术，提高提取的三元组的质量。

总的来说，作者构建知识库的流程可以用下图表示：



该方法使用知识图谱在向量空间中嵌入和表示新闻文章内容。通过比较这些向量，该模型可以有效地检测虚假新闻。该方法不仅能实现较高的准确性，而且可以通过识别新闻内容中的特定错误，提供可解释性。

虽然在虚假新闻检测中使用知识图谱提供了准确性和可解释性方面的显著优势，而且根据此研究，即使是不完整的知识库也能较好地辅助虚假新闻检测的任务。但这种方法也存在一些局限性和未解决的问题。这些问题包括：

1. 知识库本身：知识库的时效性（当前的知识库倾向于表示静态的事物，而虚假新闻是在随时间增长、变化的）、完整性（虽然不够完整和精确的知识库也能使用，但是知识库的质量仍然会影响模型的性能）和精度。
2. 其他方面：对数据来源的质量的依赖、三元组的提取和细化方面的困难、模型是否具有普适性（即在不同领域的虚假新闻中都适用）的问题等。

解决这些问题对于提高使用知识库的基于内容的虚假新闻检测方法的鲁棒性和可靠性至关重要。其次，作者研究的构建知识库的方法是适用于英文的，中文的语法规则与英文不同，对于英文适用的分词方式可能对中文不一定适用，为了适应中文虚假新闻知识库的构建，我们需要更换适用于中文的自然语言处理的工具。

**3 认知理论发现以及如何解决现有模型存在的问题**

知识库的构建和认知科学理论有着深刻而复杂的关系，因为这两个领域都涉及知识的表示、组织和处理。认知科学探索了人类如何在头脑中表征知识，语义网络、心理图式和框架等理论表明，知识是以相互连接的节点和关系组织起来的，类似于“图”的数据结构。认知科学中的关于认知模式的理论描述了知识如何根据过去的经验和背景组织成结构化的框架，这些结构有助于理解和预测事件。

一般的知识图谱都是以实体为中心的。使用知识图谱可以将信息组织成结构化的格式，使机器能够理解上下文、关系和层次结构。这种组织方式可以更有效地检索和推理信息。知识图谱由节点（实体）和边（关系）组成，形成一个图，直接反映了人类知识表示的互联本质。它们将这种组织形式形式化和系统化，用于问答、搜索、比较等目的。知识图谱的功能类似于人类记忆，因为它也可以存储相互关联的信息片段，用高效的算法可以遍历这些图来检索相关信息，类似于人类的回忆过程。

要建立具有时效性、全面性、准确性的用于虚假信息识别的知识库，应当以互联网开源新闻信息为数据源，采用文本分析以及事件本体技术，运用知识自动识别、提取和表示来制作基于事件本体的知识库。由于热点事件发生时互联网上的新闻是从不同角度聚焦同一个事件的，直接分析每一个新闻并且生成“三元组”再构建知识库可能会产生重复，导致效率降低。

文章“Building event-centric knowledge graphs from news”提出可以构建以事件本体为中心的新闻知识库。作者处理数据的方法有两个主要步骤：文档级的信息抽取和跨文档的事件共指。论文发现，通过将不同新闻文章中的提到的同一个事件进行共指消解，可以更准确地识别和合并同一事件的不同描述。通过识别实体和事件，可以显著提高事件表示的密度。采用严格的方法可以生成更为精确的表示，而宽松的方法则倾向于合并实例及其关系，这种严格性和宽松性的平衡允许根据上下文更有效地聚合事件。作者通过创建命名图，使该系统能够表达事件中关系的各种属性，例如来源、作者和数据所有者。这种灵活性使该系统能够更准确地存储和表达关系，支持更微妙和详细的事件表示。作者还实现了跨语言的事件抽取，实现了跨英语、西班牙语、荷兰语的语言处理，并根据经验研究了了合并事件要素的最佳粒度。

构建以事件本体为中心的知识库的大致流程可以分为如下几个步骤：1.用大数据平台获取互联网上的动态新闻数据。2.分析、识别文本，抽取事件关键要素，构建事件对象要素填充列表。3.结合事件关键特征设计事件本体知识模型，采用基于事件本体构建知识库的方法需要将事件按照人们认知一个事件的普遍方式结构化地描述出来，比如可以表示为“时间、地点、主体、谓词、客体”。4.将提取出的事件要素填充到事件本体知识模型中，并以三元组的形式生成知识库。

3.1 获取互联网上的新闻文本

这一步可以按照现有模型的方法进行，即对于真实新闻搜索官方媒体等可信新闻来源；对于虚假新闻搜索网络辟谣平台提供的虚假新闻数据库。

3.2识别并且抽取新闻文本中的事件关键要素

现有模型是针对英文的，为适应中文的语法规则，将自然语言处理的工具替换为适合处理中文的HanLP。可以使用Python调用HanLP（或者用Python调用Java的代码来实现）进行文本读取、文本分词、命名实体识别、词性标注、角色语义标注，并将角色语义标注作为确定事件要素的依据，以谓词为核心，识别其他词性的词语与谓词之间的关系来实现关键信息要素抽取，为以事件本体为中心的知识库构建提供基础。HanLP首先会对读取的文本进行分词处理，然后通过命名实体识别来提取文本中的机构名、人名、地名等命名实体集，再对分词得到的词语进行词性标注，作为后续按照主谓宾构建三元组和动词、名词匹配等匹配方式对词语进行重新组合的基础。最后以谓词（例如“保时捷家族从卡塔尔回购10%的股份”中的谓词为“回购”）为中心标注文本各个组成成分之间的关系（比如事件的主体与客体、事件的触发动作），并将各个成分按照“时间、地点、主体、谓词、客体”的排列方式进行重新组合，以按照人类大脑认知事件的方式准确描述客观事件。

在Python中调用Java代码使用HanLP进行实验，实验表明这个工具可以准确地提取命名实体集、标注词性和语义角色，适合用于虚假新闻知识库的构建。

3.3构建事件本体知识模型以及知识库

通过上文中对事件关键要素的识别和抽取，我们可以将事件以六元组的方式进行定义：事件E=（时间，地点，主体，谓词或者事件的触发方式，客体，状态），其中“状态”是由客体和作用在客体上的触发动作构成的描述当前时间发展状况的短语。构建知识库（知识图谱形式）时将事件要素列表中的相关要素转换为三元组形式，生成三元组图结构语义本体数据，基于新闻事件本体知识模型，使用知识库自动构建、填充算法将事件要素区分为实例、关系、属性及属性值，以三元组形式数据依次自动填充到事件本体知识模型的实例层中。

3.4总结

综合应用大数据技术、自然语言处理和事件本体知识建模的方法能够实现新闻事件信息（不论是真实新闻还是虚假新闻）的自动获得、事件要素提取、专题知识库自动构建，有效地保障了知识库的时效性和精确性。

**4 认知科学理论对人工智能实践研究的启发**

认知科学研究人类思考、学习、记忆的模式和规律，在理解、存储和检索知识的方面，人的大脑以结构化的形势组织知识（如语义网络）。受到这一点的启发，人工智能系统使用图的数据结构（即知识图谱）来构建知识库。这样存储实体、事件、属性、关系可以使机器更方便理解上下文和语义。

认知科学表明，人类大脑理解语言的过程与语法、语义、语用学有关，自然语言处理技术使用分词、解析和语义分析等技术处理和理解人类的语言以实现事件要素提取、情感分析、翻译等功能。在面对不断更新的动态知识和事件时，人类不断地将新信息与现有知识结合起来。为了确保认知的一致性和连贯性，这个过程涉及冲突的消解和“信念”的变化。知识图谱旨在整合来自不同来源的（这意味着可能出现不同表达方式的相同信息）信息，通过本体对齐和约束满足等机制确保一致可靠的知识表示。

知识库与认知科学理论的关系深深植根于知识表示、组织和处理的原则中。知识图谱可以被视为认知结构的计算模拟，利用认知科学的原则构建系统，以模仿人类认知过程的方式存储、检索和推理知识。这种协同作用增强了智能系统的发展，使得这些人工智能系统能够以类似人类的方式理解事物和与世界互动。

那么本文为什么选择以事件本体为中心构建虚假新闻知识库？这也要从对认知科学的研究说起。事件本体作为一种面向事件的知识表示方法, 更符合现实世界的存在规律和人类对现实世界的认知规律。过去的世界发生了许多的事件, 变成了今天的世界;今天的世界又将发生许多事件, 变成明天的世界。描述变化的历史, 就是描述这些发生的事件以及它们之间的关系。人类用话语文本表述,描述的只是事件及其关系的语言表现。要理解这些话语文本, 就必须知道这些事件类丰富的内容, 这些内容的绝大部分是不可能在话语文本中叙述的, 而是作为共同知识预先存在于每个交流者的头脑中。事件本体正是为计算机建造这样的共同知识。

通过认知科学的研究，我们可以对人类大脑表示、处理和应用知识的机制有更深入的了解。人工智能系统的研发者可以借助这些成果使开发的产品与人类大脑这一进化上的成功产物更加相似，处理问题更加高效、准确，有助于弥补机器智能和人类智能之间的差距，推动更符合人类需求的AI技术的发展。

**参考文献**：

1. Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, volume 11136, pages 669– 683
2. Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, Tessel Bogaard, Building event-centric knowledge graphs from news,Journal of Web Semantics,Volumes 37–38,2016,Pages 132-151
3. Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pages 312–320.
4. Sowa, J. F. (1991). Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann.
5. Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. Semantic Factors in Cognition, 37-53.
6. A. Fokkens, M. van Erp, P. Vossen, S. Tonelli, W. R. van Hage, L. Serafini, R. Sprugnoli, J. Hoeksema, Gaf: A grounded annotation framework for events, in: Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013), no. ISBN 978-1-937284-47-3, Association for Computational Linguistics, Atlanta, GA, USA, 2013.