

# 事理知识图谱

张 波

162010321

zb\_rule-ui@nuaa.edu.cn

岳亚高

182010326

saluzifengyueti@outlook.com

日期：2023 年 5 月 28 日

## 摘 要

事理知识图谱在处理高度复杂的动态系统方面显示出巨大的潜力，它对高效的关系查询提出了更迫切的要求。团队对豆瓣电影和 wikipedia 的数据进行了相应的爬虫和知识的抽取，图数据库的存储方式大大提升了团队的查询效率，团队构造了特定的专家规则，以实现意图识别。团队将在[GitHub::homework](https://github.com/zb_rule-ui/homework)上公布自己的代码。

**关键词：**知识表示作业，事理知识图谱，neo4j，word2vec，wikipedia

## 1 简介

### 1.1 知识图谱

知识图谱是一种用于描述物理世界中的概念及其相互关系的结构化数据模型，它由实体、属性、关系和事实四个基本要素组成。知识图谱的基本组成单位是“实体-关系-实体”三元组，以及实体及其相关属性-值对，实体之间通过关系相互联结，构成网状的知识结构。知识图谱可以用可视化技术展示知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系 [9]。

知识图谱是 Google 在 2012 年提出的一个概念，它是基于语义网络、本体论、语义网和链接数据等技术和理论，构建了一个开放、互联、可共享的知识资源，用于提供智能化、个性化、多模态的服务。知识图谱的发展历史可以追溯到 1960 年，从语义网络开始，经历了本体论、语义网、链接数据等几个阶段，逐渐形成了今天的知识图谱。

知识图谱在数据治理、机器语言认知、智能问答、推荐系统、知识管理和可解释人工智能等领域有很高的应用价值。知识图谱可以把领域中的异构知识结构化和关联起来，解决数据的分散、复杂和孤岛化的问题，提高数据的质量和值。知识图谱也可以利用丰富的语义关系，提高机器的语言理解能力，支持自然语言处理相关的任务，如分词、短语理解和文本理解等。知识图谱还可以为智能问答系统提供结构化的知识库，支持多种类型和推理的问题，提高问答的准确性和效率。知识图谱还可以为推荐系统提供多维度的用户、商品和兴趣信息，支持个性化、多模态和多场景的推荐，提高推荐的精准度和满意度 [8]。知识图谱还可以为知识管理系统提供统一的知识表示和存储方式，支持知识的获取、组织、存储、检索和共享等功能，提高知识的利用率和创新能力。知识图谱还可以为人工智能系统提供可解释的逻辑和证据，支持人工智能的决策过程和结果的解释，提高人工智能的可信度和可靠性 [2]。

知识图谱是一种采用结构化形式表示知识的方法，它能够为各种智能应用提供语义支持。然而，知识图谱也面临着诸多挑战，主要包括以下三个方面：第一，如何从海量的异构数据源中自动地抽取、整合和存储知识，这是构建知识图谱的难点。目前大部分的方法仍然依赖于人工或半自动的方式，导致构建过程耗时耗力。第二，如何保证知识图谱中的知识具有高质量，即准确、完整、一致和时效，这是评估和修复知识图谱的重点。目前的知识图谱常常存在错误、冗余、缺失和过时等问题，需要有效的机制来检测和纠正这些问题。第三，如何充分利用知识图谱中的丰富语义信息，以支持复杂的查询、推理和应用，这是发挥知识图谱价值的热点 [6]。

## 1.2 事实图谱

事理图谱是一种知识图谱的扩展，它能够表示事件及其逻辑关系。事件是指发生在时间和空间中的具有主体和客体的动作或状态变化，逻辑关系是指事件之间存在的因果、条件、并列等关系。传统的知识图谱往往只能表示实体及其属性和关系，事理图谱能够捕捉事件的动态性和不确定性，从而更好地支持事件演化和因果推理等任务。事理图谱和知识图谱都采用三元组的形式来表示知识，都可以用图结构来可视化知识，都可以用于回答常识问题。事理图谱的研究对象是谓词性事件，而知识图谱的研究对象是名词性实体；事理图谱中事件之间的逻辑关系是不确定的概率，而知识图谱中实体之间的关系是确定的 [13]。

构建事理知识图谱主要包含以下几个要点：

- 建立事件知识库。从各种文本中提取出事件的信息，再将这些事件的信息以合适的形式存储在知识库里。
- 抽取和表示事件关系。从各种文本中提取出事件之间的联系，再将这种联系以合适的形式存储在知识库里。
- 发现事件演化规律。利用知识库里的数据，分析事件之间的模式和规律。
- 实现事件后续预测。利用知识库里的数据，对未来可能发生的事件进行预测和推演。

当前构建事理知识图谱主要有以下几个难点 [12]：

- 事件表示。事件是发生的事情，它有很多种类，有的还会变化。表示好事件是事理图谱的基础。
- 事理融合。可以将其他知识图谱和事理图谱结合起来，一些静态的知识可以帮助理解事件和它们的关系。
- 知识表示学习：事理图谱中的知识表示学习要考虑事件和它们的关系，从而支持事件的变化和推理。
- 推理机制：推理机制是根据已有的知识和规则，得到新的知识或结论，这是事理图谱的难点。

## 1.3 团队工作

团队做了两方面的工作，一是使用爬虫爬取了豆瓣 250 的电影资料，直接解析获得了电影的结构化数据。然后将结构化的数据存入的 neo4j 图数据库当中，在使用 jieba 分词的方法进行实体提取，再使用 fuzzywuzzy 进行意图分析，最终找到合适的信息，这是团队构建的电影知识图谱。

此外，团队以大致同样的方法对 wikipedia 的数据进行操作，由于 wikipedia 是半结构化的数据，团队使用了 stanford nlp 进行了辅助抽取，使得抽取的效果更好。此外，为了能充分地在数据库查找到想要的信息，团队采用了 word2vec 的模型对识别出的实体分词进行了一定的扩充，这是团队构建的事理知识图谱。

团队工作主要有以下创新点：

- 构建了相应的事理知识图谱。
- 使用了 wikiExtractor 配合 stanford nlp 进行知识抽取。
- 使用了词向量的方法降低知识库缺失知识的可能。

## 2 方法

### 2.1 数据获取

知识图谱数据的获取方式有很多种，其中一些主要的方法包括：众包法、爬虫、机器学习、专家法。

众包法是指通过人工标注的方式来获取知识图谱数据，这种方法可以获得高质量的数据，但是成本较高。爬虫是指通过网络爬虫程序来抓取网络上的信息，这种方法可以获取大量的数据，但是数据质量不稳定。机器学习是指通过机器学习算法来自动抽取知识图谱数据，这种方法可以大大降低人工标注的成本，但是需要大量的训练数据。专家法是指通过专家来手动构建知识图谱，这种方法可以获得高质量的数据，但是成本较高。

对于豆瓣电影的数据，团队主要采用爬虫的方式进行数据的获取。爬虫是一种自动化程序，可以模拟人类在互联网上的行为，从而获取互联网上的信息。爬虫程序首先会访问一个网站，然后通过解析网页的 HTML 代码来提取出需要的信息。爬虫程序可以通过正则表达式、XPath 等方式来解析 HTML 代码，从而提取出需要的信息。爬虫程序还可以通过模拟登录、使用代理等方式来绕过一些反爬虫机制，从而获取更多的信息。

其中团队主要利用请求-响应的库 urllib.request 库对网页进行整体的爬取，然后利用 BeautifulSoup 对网页进行 lxml 方式的解析，接着利用正则表达式的库 re 对想要的数据进行筛选，团队构建正则表达式主要基于浏览器自带的开发者模式对网页结构进行观察进行的。在筛选出想要的之后，团队将数据存入了 csv 类型的表格文件中。

对于 wikipedia 的数据，由于 wikipedia 的数据规模太大，尽管团队采取了一些多进程的方法，利用了 multiprocessing 等可以开辟、管理多进程的库，但是爬取仍然需要大量的时间。wikipedia 事实上公布了官方的网页数据，即大量的 xml 格式的网页数据，团队采用了直接下载这些网页数据，省去了请求-相应流程需要的大量时间。在得到了 wikipedia 的 xml 的网页数据之后，团队使用了 WIKIExtractor 的工具包中的一些文件，其中使用正则表达式来解析 XML 文件，将 XML 文件转换为 HTML 格式，然后使用 BeautifulSoup 库来解析 HTML 文件，最后将 HTML 文件转换为纯文本格式，对于不能够识别的数据，团队采用了忽略的操作，这在一定程度上损失了一些数据，目前团队还没有发现更完善的方法 [1]。

## 2.2 知识抽取

知识抽取是指从非结构化或半结构化的文本中自动抽取实体、关系、事件等结构化的知识等。知识抽取的基本方法包括分词、词性标注、命名实体识别、关系抽取、事件抽取五个步骤。

分词就是指将文本分割成一个个单独的词语。词性标注是指对每个单词进行词性标注，一些常见的词性包括名词、动词、形容词等。命名实体识别是指从文本中自动抽取人名、地名、组织机构名等实体。关系抽取是指从文本中自动抽取实体之间的关系。事件抽取是指从文本中自动抽取事件及其相关信息。

由于要抽取的豆瓣数据实际上属于结构化的数据，团队并没有使用太多知识抽取相关的技术，只是使用了 `jieba` 库对抽取出来的数据进行了简单的实体识别，使得抽取出来的数据质量更高，结构化的属性对应的数据缺失值更少。

对于 `wikipedia` 的半结构化数据，团队使用了 `stanford` 的 NLP 工具包进行知识的抽取。`stanford` 的命名实体识别 (`stanford ner`) 主要使用的是基于条件随机场 (CRF) 的算法。[14] 条件随机场 (CRF) 是一种概率图模型，在给定观测序列的条件下，计算标注序列的概率。具体来说，CRF 模型将观测序列和标注序列看作联合随机变量，并通过最大化联合概率来确定标注序列。此外，规则匹配、字典匹配等工程技术同样是 `stanford ner` 强大功能的来源。通过 `stanford NLP` 对数据进行了处理之后，团队得到了一个结构化的数据集。

## 2.3 图谱存储

知识图谱的存储方式主要有基于 RDF 结构的存储方式和基于图数据库存储，团队采用了图数据库的方式对数据进行存储，采用的图数据库为 `neo4j`。

`Neo4j` 是一个基于 `Java` 的图数据库，它采用了基于节点和边的数据模型，支持 ACID 事务，具有高性能、高可扩展性、高可用性等特点。`neo4j` 的底层存储结构是基于 B+ 树的，每个节点都是一个 B+ 树节点，每个节点都有一个唯一的 ID，每个节点都可以包含多个属性和关系。`Cypher` 语言是 `neo4j` 的查询语言，它是一种类似 SQL 的声明式语言，可以方便地查询图数据库中的数据。

由于豆瓣电影的数据规模较小，可以很快的存储到本地的 `neo4j` 数据库当中去，因此团队采用了 `neo4j` 进行图谱存储。`wikipedia` 数据规模十分庞大，在使用 `python` 的 `py2neo` 进行数据的存储时需要花费大量的时间，因此团队暂时采用三元组的 `json` 文件格式进行存储。

## 2.4 图谱结构

知识图谱的常见结构是由节点和边组成的关系网络，每个节点表示现实世界中存在的实体，每条边为实体与实体之间的关系。知识图谱的基本单位是“实体-关系-实体”构成的三元组。

对于知识图谱的数据，团队以数据的事实为单位存储在 `neo4j` 图数据库，主要是使用“实体-属性-值”三元组作为事实的基本表达方式。

## 2.5 表示学习

表示学习将研究对象的语义信息表示为稠密低维的实体向量。而知识表示学习将知识库中的实体和关系表示为稠密低维的实值向量。知识表示学习的模型方法主要有基于距离、基于翻译、语义匹配、融合多源信息等方法。

针对于本项目，团队主要采用基于翻译的 TranE 模型。TransE 的核心思想是将每个三元组实例 (head, relation, tail) 中的关系 relation 看做从实体 head 到实体 tail 的翻译，通过不断调整 head、relation 和 tail，使得翻译后的头实体向量加上关系向量尽可能接近尾实体向量 [3]。具体来说，可以简单地表示为以下损失函数：

$$L = \sum_{\xi \in T} \sum_{\xi' \in T'} [\gamma + f(\xi) - f(\xi')]_+$$

---

### Algorithm 1 TranE

---

**Input:** Training set  $S = \{(h, l, t)\}$ , entities and rel. sets  $E$  and  $L$ , margin  $\gamma$ , embeddings dim.  $k$ .

- 1: **initialize:**  $\ell \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$  for each  $\ell \in L$   
 $\ell \leftarrow \ell / \|\ell\|$  for each  $\ell \in L$   
 $e \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$  for each entity  $e \in E$
  - 2: **loop**
  - 3:  $e \leftarrow e / \|e\|$  for each entity  $e \in E$
  - 4:  $S_{batch} \leftarrow \text{sample}(S, b)$
  - 5:  $T_{batch} \leftarrow \phi$
  - 6: **for**  $(h, \ell, t) \in S_{batch}$  **do**
  - 7:  $(h', \ell, t') \in S_{batch} \leftarrow \text{sample } S'_{(h, \ell, t)}$
  - 8:  $T_{batch} \leftarrow T_{batch} \cup \{(h, \ell, t), (h', \ell, t')\}$
  - 9: **end for**
  - 10: Update embedding w.r.t  $\sum_{((h, \ell, t), (h', \ell, t')) \in T_{batch}} \partial[\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+$
  - 11: **end loop**
- 

伪代码中 TransE 使用的边际损失函数为：

$$\mathcal{L} = \sum_{(h, \ell, t) \in S} \sum_{(h', \ell, t') \in S'_{(h, \ell, t)}} [\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+$$

其中  $[x]_+$  表示取  $x > 0$  的值， $\gamma > 0$  是一个超参数，对于  $S'_{h, \ell, t}$  有定义：

$$S'_{(h, \ell, t)} = \{(h', \ell, t) | h' \in E\} \cup \{(h, \ell, t') | t' \in E\}$$

TransE 模型的优点是简单易于实现，计算效率高，缺点是无法处理多对多关系，即 TransE 会将一些语义相近的词映射到一起，对于一些复杂的关系难以建模 [3]。对 TransE 做出改进的模型有很多，典型的有 TransH、TransR、TransD。TransH 通过对关系向量进行投影，将关系向量投影到一个超平面上，从而实现多对多关系的建模。TransR 通过将实体和关系分别映射到不同的向量空间中，从而实现复杂关系的建模。TransD 是在 TransR 的基础上提出的一种模型，主要是为了解决 TransR 存在的问题。TransD 通过对实体和关系进行分组，从而实现更好的建模效

果.

## 2.6 查询和推理

知识推理的方法主要包含基于逻辑、基于统计、基于图三种。基于逻辑的推理方法包含一阶谓词逻辑、描述逻辑等利用规则推理的方法，基于统计的推理方法包含机器学习、统计学习等利用概率推理的方法，基于图的推理方法包含 Path Ranking、神经网络 [4] 等在图上利用图的性质进行推理的方法 [5]。

团队主要使用了基于逻辑的推理方法。团队首先将 wikipedia 的实体名称载入 jieba 的词库，再将问题进行 jieba 分词，识别出对应的 wikipedia 实体后，再对 wikipedia 实体对应的属性进行意图分析，意图分析的过程主要是通过 jieba 中比对相似比率的方法实现的，然后将最有可能的属性对应的值作为答案。在这个过程中，为了保证能更好地找到对应的值，团队利用了词向量的方法对识别到的实体做了一定的扩充，以保证能再数据库找到对应的实体；又对意图分析的词进行了扩充，再对意图分析的结果取均值来减少错误的可能。团队采用的词向量模型是腾讯的 2021 年的中文词向量模型 (0.2.0 版本)，其中有超过 12 万个中文单词，该词向量模型主要是使用 word2vec 方法进行训练的 [18]。

word2vec 是由 Google 的 Tomas Mikolov 等人于 2013 年提出的，包含有两种模型：CBOW 和 skip-gram[16]。CBOW 模型是基于上下文预测目标词，而 skip-gram 模型则是基于目标词预测上下文，腾讯的词向量模型主要使用了 skip-gram 的模型方法。在训练过程中，word2vec 使用了负采样的技术来加速训练，并提高了模型的准确性 [7]。

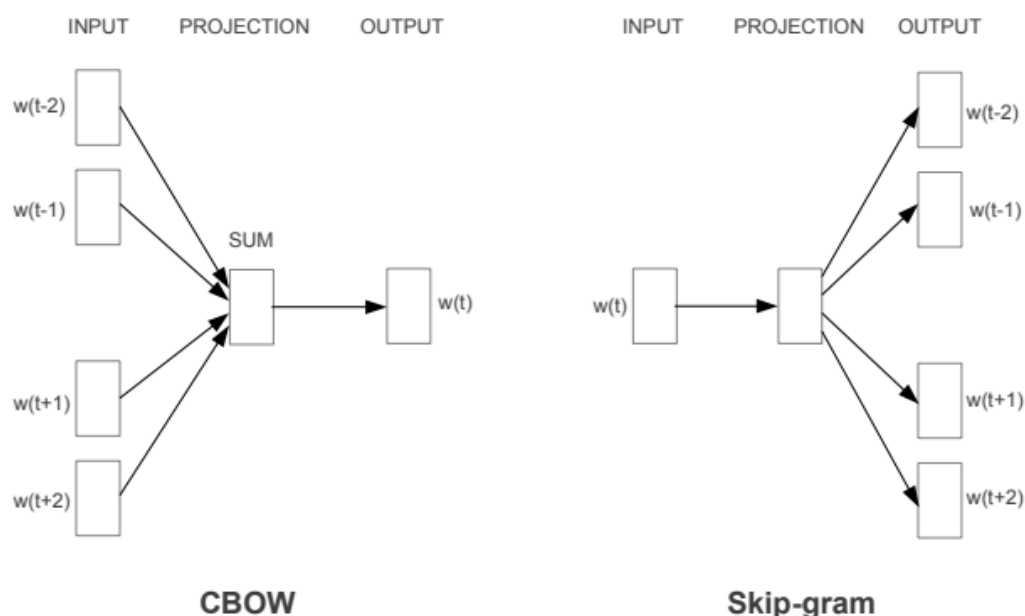


图 1: word2vec 的模型架构: CBOW 架构根据上下文预测当前单词, Skip-gram 根据当前单词预测周围单词

## 3 实验

### 3.1 图谱规模

表 1: 图谱规模

实体数目	属性数目
2356662	562292

### 3.2 表示学习

团队对豆瓣电影的信息进行了基于 TransE 的表示学习。

### 3.3 查询和推理

团队主要使用了基于逻辑的推理方法。首先，将维基百科实体名称载入 jieba 的词库，然后使用 jieba 对问题进行分词，识别出对应的维基百科实体。之后，分析维基百科实体对应属性的意图。意图分析过程主要是通过 jieba 中比对相似比率的方法实现的。最后，将最有可能的属性值作为答案。为了更容易地在数据库中找到相应的实体，使用采用 word2vec 方法进行训练的中文词向量模型来扩展已识别的实体。同时扩展意图分析的单词，并取意图分析结果的平均值以减少错误的可能性。

当查找失败时，采用粒度更小的 stanford nlp 来重新识别实体，重复之上的过程 [14]。

### 3.4 样例分析

经过大量的测试，团队发现查询时主要会出下以下几种问题：

- 分词粒度引起的问题

问题：道家思想简介

ChatBot：知识库未找到数据

问题：道家简介

ChatBot：道家是中国诸子百家中思想学派之一，也是九流十家之一，在春秋战国时期以老子、列子、庄子为代表。……

可以看出数据库含有道家的相关知识，但是在询问道家思想简介时，会识别出道家思想这个实体，导致查询失败。

- 意图识别引起的问题

问题：南京航空航天大学的学校特色

ChatBot：明故宫校区位于南京市秦淮区，在中山东路以南、瑞金路以北、御道街以东，原址为明代南……

可以看到在询问学校特色，模型选择回答了学校的地理环境。

- 难以处理多跳问题

问题：毛泽东领导的党派

ChatBot：自 1945 年起任中国共产党中央委员会主席和中国共产党中央军事委员会主席，成为中共中央主……

问题：中国共产党的理想

ChatBot：中国共产党始终宣称其理想与目标是共产主义，并将自身视为代表工人阶级领导工农联盟……

问题：毛泽东领导的党派理想

ChatBot：毛泽东思想有政治、经济、军事、外交、文艺、哲学等多方面内容，于 20 世纪广泛影响世界，如“枪杆……

难以处理多条问题在于团队只构建了单跳的查询，可以通过进一步完善模型解决这个问题。

## 4 结论

团队完整地构建了事理知识图谱和电影知识图谱，其中利用的技术方法包括 request 爬虫、stanford nlp、ner4j 图数据库、word2vec 词向量、tranE 模型。基于这些技术构建的知识图谱具有较好的查询性能，但也有一些不足，包括解析网页数据时忽略数据、意图识别错误、分词粒度不准确、不能处理多跳问题等典型的缺陷。总而言之，团队已经完成了图谱的结构构建，以后会通过追加规则和替换模型的方式进行图谱的优化。

## 参考文献

- [1] Giuseppe Attardi. *WikiExtractor*. <https://github.com/attardi/wikiextractor>. 2015.
- [2] Federico Bianchi et al. “Knowledge Graph Embeddings and Explainable AI”. In: *Knowledge Graphs for eXplainable Artificial Intelligence*. 2020.
- [3] Antoine Bordes et al. “Translating Embeddings for Modeling Multi-relational Data”. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013.
- [4] Huajun Chen et al. “Neural symbolic reasoning with knowledge graphs: Knowledge extraction, relational reasoning, and inconsistency checking”. In: *Fundamental Research* 1.5 (2021), pp. 565–573. ISSN: 2667-3258.



- [5] Qian Chen, Zhu Zhuo, and Wen Wang. “BERT for Joint Intent Classification and Slot Filling”. In: *ArXiv abs/1902.10909* (2019).
- [6] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.
- [7] Yoav Goldberg and Omer Levy. “word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method”. In: *ArXiv abs/1402.3722* (2014).
- [8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6645–6649.
- [9] Shaoxiong Ji et al. “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2 (2022), pp. 494–514.
- [10] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. ISBN: 1558607781.
- [11] Zhongyang Li, Xiao Ding, and Ting Liu. “Constructing Narrative Event Evolutionary Graph for Script Event Prediction”. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI’18. Stockholm, Sweden: AAAI Press, 2018, pp. 4201–4207. ISBN: 9780999241127.
- [12] Yankai Lin et al. “Learning Entity and Relation Embeddings for Knowledge Graph Completion”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 2181–2187. ISBN: 0262511290.
- [13] Xueyi Liu and Jie Tang. “Network representation learning: A macro and micro view”. In: *AI Open* 2 (2021), pp. 43–64. ISSN: 2666-6510.
- [14] Christopher Manning et al. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 55–60.
- [15] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *International Conference on Learning Representations*. 2013.
- [16] Xin Rong. “word2vec Parameter Learning Explained”. In: *ArXiv abs/1411.2738* (2014).
- [17] Shuming Shi et al. “Corpus-based Semantic Class Mining: Distributional vs. Pattern-Based Approaches”. In: *International Conference on Computational Linguistics*. 2010.
- [18] Yan Song et al. “Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 175–180.
- [19] Kaisheng Zeng et al. “A comprehensive survey of entity alignment for knowledge graphs”. In: *AI Open* 2 (2021), pp. 1–13. ISSN: 2666-6510.