**知识库构建与融合**

本方案通过自底向上的方法，采用“纵向构建、横向融合”的分层思想逐步构建指控系统领域适变知识体系。该知识体系主要从历史战例、活动规律、专家经验和样本数据等数据中提取实体、事件、模型三种知识。

首先，基于领域知识和通用知识构建实体层知识图谱；其次，通过实体与事件的关联构建事件层知识图谱；最后，面向事件服务响应构建模型层知识图谱，三层图谱之间层层递进、相互关联，构建“实体-事件-模型”驱动的多层次知识图谱总体架构框架。在每一层知识图谱内，采用智能算法、机器学习等方法实现不同知识类型的融合，包括实体链接、实体对齐和实体消歧三个方面。

# **知识库构建**

## 1、技术方法现状

### 1.1 实体知识图谱构建

知识库构建的主要任务是根据相关原始数据进行知识抽取。知识抽取通过自动化或半自动化的知识抽取技术获得实体、关系和属性等可用知识单元。知识抽取的主要三类任务为实体识别、关系抽取和属性抽取，对于每种不同的任务，有其对应的技术方法。

实体识别的目的是从海量的原始数据中提取知识图谱所需的特定实体信息。现阶段，实体识别方法主要为基于规则、基于统计模型及基于神经网络三类[1]。

基于规则的方法通过手工构建规则集，将纯文本等非结构化数据与规则集匹配来实现实体识别。该方法在处理小规模的知识图谱时精度较高，但是随着知识图谱规模的增大，规则构建困难，难以进行大规模扩展并应用于不同领域的知识图谱。

基于统计模型的方法将实体识别作为序列标注问题，以完全或部分标注的语料进行模型训练。常见的统计模型有条件马尔科夫模型、隐马尔科夫、条件随机场和最大熵等。基于统计模型的方法通用性强。但是状态搜索空间庞大、训练时间长、高度依赖特征选取和语料库；由于统计模型的状态空间爆炸问题，难以处理海量数据。

深度学习能够自动地从数据中学习复杂的隐藏特征，所需的领域专业知识和经验知识较少，基于神经网络的 NER 已成为目前 主流方法。主要模型有卷积神经网络CNN、循环神经网络RNN、长短期记忆网络、双向长短期记忆网络等。基于深度学习的实体识别算法有更好的通用性，一句对海量数据处理的能力。

关系抽取通过获取实体之间的某种语义关系或关系的类别，自动识实体对及联系这一对实体的关系所构成的三元组。早期的关系抽取研究方法主要是通过人工构造语法和语义规则进而采用模式匹配的方法来识别实体间的关系，此种方式要求指定规则的人对特定领域有深入的理解和认知且规则制定工作量大。近几年来，关系抽取研究大多是基于神经网络，主要包括卷积神经网络、循环神经网络、注意力机制、图卷积网络、对抗训练和强化学习等技术[2]。

属性抽取通过从不同信息源的原始数据中抽取实体的属性名和属性值，构建实体的属性列表，形成完整的实体概念。目前，属性抽取方法可分为传统的监督、无监督和半监督属性抽取以及基于神经网络的属性抽取和元模式的属性抽取。大部分属性抽取是从文本数据中抽取信息，与实体识别类似，传统方法中监督学习的属性抽取使用统计模型抽取属性。基于神经网络的属性抽取使用序列标注模型抽取文本中的属性值。基于元模式的属性抽取将类型化的文本模式结构命名为元结构，从而在海量语料库中发现元模式。

### 1.2 事件知识图谱构建

事件知识图谱多采用“自顶向下”的构建方法，先定义数据模式，再添加实例内容向下细化，构建技术包括事件知识表示、事件知识抽取、事件关系抽取等。事件知识图谱从自然语言文本中抽取事件和实体、属性、关系等并进行知识融合，然后通过本体构建体系框架，以结构化的三元组形式存储。

事件知识表示直接关系到知识推理、知识计算的应用，是事件知识图谱的核心部分。传统的事件表示模型主要基于5W ( Who，When，Where，What，Why) 构建，没有建立事件之间的关系连接。Silver等在2011年提出了一种离散事件表示本体模型，提供了事件之间的关系表示方法。Van Hage等2011年构建了简单事件模型( Simple Event Model，SEM) ，对不同领域中的事件进行建模，建立了事件与时间、地点和实体之间的链接，但无法表示事件之间的关系。Rospocher 等2016年和 Gottschalk 等2018 年在简单事件模型(SEM)基础上建立了事件之间的关联，构建事件知识图谱。

事件关系抽取以事件为基本的语义单元，自动抽取事件之间的逻辑关系，包括事件的共指关系、因果 关系、时序关系等。事件关系的抽取和构建可以揭示事件发展规律，厘清事件关联并全面了解事件，进而构建事件知识图谱。

因果关系指事件之间的作用关系，即某个事件是另一事件的结果。在事件知识图谱的推理应用中具有重大意义。杨竣辉等2016年提出了通过构建事件和事件元素的语义关联，进行事件因果关系的识别方法。付剑峰等2011年提出了一种基于层叠条件随机场的事件因果 关系抽取模型，将事件因果关系建模为序列标注问题。Sorgent 等2013 年通过制定规则进行事件因果关系抽取，并利用贝叶斯推理优化结果。

[1] Goyal, Archana, Vishal Gupta, and Manish Kumar. "Recent named entity recognition and classification techniques: a systematic review." Computer Science Review 29 (2018): 21-43.

[2] LiuQiao, LiYang, and LiuYao DuanHong. "Knowledge graph construction techniques." Journal of computer research and development 53.3 (2016): 582.

## 2、设计思路

指控系统需要敏捷地感知周围环境变化，并根据周围环境变化做出相应的自适应决策。为了更加迅速地感知外部环境变化并作出决策，需要构建相应的知识库。然而，单一的知识库构建不利于指控系统敏捷地做出反应，故针对指控系统，本设计采取了三层知识图谱的构建方式，即实体层—事件层—模型层知识库。~~（增加首段）~~

### 2.1 实体知识图谱构建

传统的实体识别任务主要分为实体类识别、时间类识别和数字类识别三个大类和人名、机构名、地点、时间、日期、货币和百分比识别七个小类。但是本文指控领域涉及的实体类别远远不止这些，因此本方法将带有领域特色的实体识别作为研究重点。

相较于传统的三大类和七小类实体识别，指控领域实体识别的特点是，领域中实体类别的粒度更细且普遍带有军事特征。军事实体识别的任务可以描述为，将指控信息文本语料进行预处理后，从中提取句子特征，将特征输入到实体识别模型中，从而识别出具有特定意义的军事实体。

军事领域关系抽取的任务可以描述为，将军事文本语料进行预处理后，依靠前文已经提取到的军事相关实体，利用神经网络模型对文本数据中实体之间的关系进行抽取，然后结合实体和抽取到的关系组成（实体 1，关系，实体 2）三元组数据，从而完善军事知识图谱的构建。

### 2.2事件知识图谱构建

事件是文本中包含的一种特殊信息，事件抽取就是从非结构化的文本数据中抽取与事件有关的各种角色，将信息用结构化数据表示。按照确定事件类别的方法，事件抽取可以被分为限定域事件抽取和开放域事件抽取。事件知识图谱构建与实体知识图谱类似，通常采用基于本体的构建方法进行知识建模，先构建事件知识图谱顶层表示模式，再向下细化补充实例。

如图1.2.1所示，军事领域实体标注的任务可以描述为，将军事文本语料进行预处理后，依靠前文已经提取到的军事相关实体，利用神经网络模型对文本数据中的实体进行标注，然后结合实体和抽取到的关系组成三元组数据，完善军事知识图谱的构建。



图1.2.1 事件知识图谱实体标注过程示意图

目前应用较为广泛的实体识别方法是BiLSTM-CRF，本项目拟通过该方法实现事件知识图谱的实体识别。如图1.2.2所示，该方法将句子中的词向量作为双向长短期记忆网络的输入。通过提取句子特征，结合条件随机场（CRF）完成实体的标注。



图1.2.2 BiLSTM-CRF方法流程示意图

预训练模型可以从大规模的语料中提取隐含的语义信息，学习到更好的通用语义表示向量，从而提高下游任务的表现。Yang S等人[2]针对现有远程监督事件抽取方法中存在的问题，将目光转向预训练的语言模型，希望利用从大规模语料库中学习到的知识表示向量来提高模型的性能。其设计了一种基于预训练语言模型的事件抽取（pre-trained language model based event extractor，PLMEE）模型。该模型的结构如图所示[2]。他们将事件抽取看作由两个子任务组成，两个子任务分别是触发词抽取和论元抽取，并提出了以预训练语言模型为基础的触发词抽取器和论元抽取器。基于预训练模型的方法会使事件的语义表示更加精确。

为了在神经网络中引入句法结构特征，Nguyen T等人[3]提出了一个基于句法依存树的GCN模型用于事件抽取。在GCN中，每个节点的卷积向量是由相邻节点的表示向量计算出来的，可以作为该节点的唯一特征进行分类。另外，模型中通过对当前单词的卷积向量以及句子中提到的实体进行池化操作，克服实体指称无法捕捉的问题。池化操作聚合了卷积向量，从而为事件类型预测生成了单个向量表示。该方法在事件抽取中引入了GCN模型，将句法依存树上的信息进行聚合，首次利用了句子中的结构信息。除此之外，Liu X等人[4]设计了一个新的联合多事件抽取（jointly multiple events extraction，JMEE）框架。该框架利用基于注意力机制的图卷积神经网络进行建模，并通过引入句法依存树中的句法捷径弧来增强信息流，以此来提升在一个句子中抽取多个事件的效果。使用句法捷径弧可以减少将信息流从一个节点转换。

### 2.3 模型知识图谱构建

模型类知识图谱的作用主要在于以实体知识图谱以及事件知识图谱为基础，并以既定的时间应对策略为知识图谱构建数据。对指控系统所发生的事件进行及时的决策响应。

类似于实体关系抽取以及时间关系抽取，使用与之相同的方法识别实体，如图1.2.3所示。

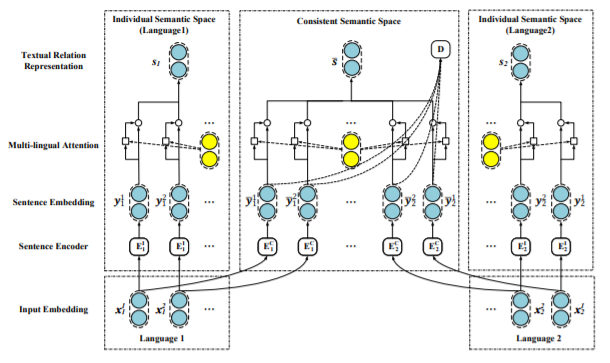


图1.2.3 模型知识图谱识别实体示意图

由于指控系统的专业性以及可靠性要求，以及指控系统的复杂性，常用的基于神经网络的模型通常忽略了输入数据中的噪声，导致模型效果不佳。将对抗训练AT引入关系抽取，对训练数据产生对抗噪声来优化分类算法，能够有效提升模型鲁棒性。Wu等[5]首次提出在关系抽取中应用AT策略，Wang等[6]在此基础上，提出了 AMNRE（Adversarial Multi-lingual Neural Relation Extraction），以 AT策略保证对多种语言的句子表征，提取出语义一致的关系。

# **知识库融合**

## 1、技术方法现状

知识图谱创建的信息抽取过程,实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标，然而，这些结果中可能包含大量的冗余和错误信息，数据之间的关系也是扁平化的，缺乏层次性和逻辑性，因此有必要对其进行清理和整合。通过知识融合,可以消除概念的歧义，剔除冗余和错误概念，从而确保知识的质量，帮助构建出高质量的知识图谱。

### 1.1 实体链接

实体链接(entity linking)是指对于从文本中抽取得到的实体对象, 将其链接到知识库中对应的正确实体对象的操作。它的基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象,然后通过相似度计算将指称项链接到正确的实体对象。

实体链接的一般流程是: 1)从文本中通过实体抽取得到实体指称项; 2)进行实体对齐、实体消歧，消除数据存在的实体冲突问题，判断知识库中的同名实体与之是否代表不同的含义以及知识库中是否存在其他命名实体与之表示相同的含义; 3)在确认知识库中对应的正确实体对象之后,将该实体指称项链接到知识库中对应实体。

### 1.2 实体对齐

实体对齐(entity alignment)，也称为实体匹配(entity matching)，目的是消除异构数据中存在的知识指代不明、实体冲突等问题。实体对齐本质上是在计算知识之间的相似度，通过相似度来决定知识对齐的操作。

实体对齐算法可以分为成对实体对齐、协同实体对齐以及基于表示学习的对齐算法三类。成对实体对齐主要考虑实例及其属性的相似程度，而协同实体对齐则是在成对对齐的基础上考虑不同实例之间的相互关系用以计算相似度，基于表示学习的实体对齐则主要通过对低纬度的空间映射来使用向量表示实体与关系，提供更丰富的语义信息。

1）成对实体对齐算法

成对实体对齐根据属性相似性评分来判断实体是否匹配，主要包括基于概率模型的对齐算法和基于机器学习的对齐算法。最早的基于概率的对齐算法是基于属性相似性评分方法[7]，该方法将实体匹配问题转换为分类问题(匹配、可能匹配和不匹配三种)，在此基础上建立起基于概率的实体链接模型，为每个匹配的属性分配不同的权重，提高准确性[8]。

基于机器学习的对齐算法则是将实体对齐问题转换为二分类问题进行处理。根据是否采用标注数据可以将该方法分为有监督学习和无监督学习两类。有监督学习采用一部分标注数据作为训练集来训练模型，然后将模型用于未标注数据的对齐。这一类算法集中于决策树、支持向量机、集成学习等[9][10][11][12]。无监督的学习方法则基于聚类的思想，将类似的实体聚在一起，然后通过少量的标记数据来推理实体的匹配情况[13]。

2）协同实体对齐算法

协同实体对齐算法为实体本身的属性以及与它相关的实体的属性分别设置不同的权重，并通过加权来计算总体的相似度，来实现实体的对齐。基于相似性传播的方法就是一种典型的协同实体对齐方法，主要考虑了需要匹配的两个实体及与他们直接关联的其他实体的相似性。而基于概率的协同实体对齐算法则采用 LDA模型、CRF模型、Markov逻辑网等[14][15][16]。

3）基于表示学习的实体对齐方法

表示学习方法来源于深度学习，主要思想是将知识图谱中的实体和关系映射到低维空间，学习得到实体和关系的向量表示。这种低维稠密的向量蕴涵了图谱的结构信息以及实体和关系的属性特征，具有丰富的语义信息。基于表示学习的知识图谱实体对齐算法由两部分组成：知识表示的学习和实体间映射关系的学习。首先，将待对齐知识图谱分别映射到低维空间得到对应的知识表；其次,基于新的知识表示和人工标注的实体对齐数据集来学习得到实体间的对应关系。

### 1.3 实体消歧

实体消歧(entity disambiguation)是专门用于解决同名实体产生歧义问题的技术，即经常会遇到的某个实体指称项对应于多个命名实体对象的问题。通过实体消歧，就可以根据当前的语境，准确建立实体链接。实体消歧主要采用聚类法，以实体对象为聚类中心，将所有指向同一目标实体对象的指称项聚集到以该对象为中心的类别下。聚类法消歧的关键问题是如何定义实体对象与指称项之间的相似度，常用方法模型有以下四种：

1. 空间向量模型(也称词袋模型)。典型的方法是取当前语料中实体指称项周边的词构成特征向量，然后利用向量的余弦相似度进行比较，将该指称项聚类到与之最相近的实体指称项集合中。

2. 语义模型。该模型与空间向量模型类似，区别在于特征向量的构造方法不同，语义模型的特征向量不仅包含词袋向量，而且包含一部分语义特征。

3. 社会网络模型。该模型的基本假设是物以类聚、人以群分。在建模时,首先利用实体间的关系将与之相关的指称项链接起来构成网络,然后利用社会网络分析技术计算该网络中节点之间的拓扑距离(网络中的节点即实体的指称项),以此来判定指称项之间的相似度。

4. 百科知识模型。百科类网站通常会为每个实体 (指称项) 分配一个单独页面，其中包括指向其他实体页面的超链接，百科知识模型正是利用百科类网站中实体页面的链接关系来计算实体指称项之间的相似度。

### 1.4共指消解

共指消解(entity resolution)技术主要用于解决多个指称项对应于同一实体对象的问题。共指消解的主要解决思路分为两种：基于自然语言处理和基于统计机器学习的方法。

基于自然语言处理的共指消解是以句法分析为基础的，代表性方法是Hobbs算法和向心理论(centering theory)。Hobbs算法主要思路是基于句法分析树进行搜索。向心理论则将表达模式 (utterance) 视为语篇(discourse)的基本组成单元，通过识别表达模式中的实体，可以获得当前和后续语篇中的关注中心(实体)，根据语义的局部连贯性和显著性，就可以在语篇中跟踪受关注的实体[17]。除此之外，统计机器学习方法也被引入该领域，并进入了快速发展阶段，通过将共指消解作为聚类问题来求解，以实体指称项为中心,通过实体聚类实现指称项与实体对象的匹配。

## 2、设计思路

根据上述技术方法现状的分析，可以看出，为了应对由于多源异构的信息提取导致的知识重复、质量差异和层次结构缺失等问题，目前针对适变知识融合的相关方法主要分为实体链接、实体对齐、实体消歧和共指消解四个部分。而在军事指控领域中，面向不同领域和数据模式的知识融合有了更高的要求。因此，为了解决上述构建提取的实体类知识、事件类知识和模型类知识的异构数据整合、消歧和语义融合问题，本方案提出了一种基于分层知识图谱的知识融合框架，主要从实体链接、实体对齐、实体消歧三个方面实现不同类型知识的融合。

### 2.1 实体链接

实体链接是把文档中给定的命名实体链接到知识库中一个无歧义实体的过程，是知识融合的重要手段。而在军事指控领域，面向实体、事件和模型三种知识类型的实体链接，上下文信息并不充分，具有更高的不确定性和难度。同时，不同结构和类型的实体指称项链接到知识库对应实体对象的准确性无法保障，也很难充分利用知识库中已有的知识体系。因此，本方法提出一种基于语义一致性的集成实体链接算法，通过利用知识库中实体间的结构化语义关系，提高对不同知识类型概念相似实体的区分度，从而更好的提高实体链接的准确性。



图2.1.1 实体链接结构图

如图2.1.1所示，本算法主要分为三个部分：候选实体集合生成、实体相关图构造以及集成实体链接实现。首先，对给定文本进行实体识别，使用基于规则的方法进行共指处理，得到该文本的实体指称项集合，然后根据该集合查找本地库，得到对应的候选实体集合。该本地库是根据军事指控系统的要求构建的知识库，包括实体、事件和模型三种。同时，还应该添加外部的维基百科知识库，利用该知识库中包含大量已经过人工消歧处理的同名实体，减少实体链接时选择实体的干扰项数量。

然后，需要对每次输入的文本构造一张实体相关图，利用共现实体间的语义相关性帮助提高实体链接的准确性并实现批量实体链接，我们通过对每个文本定点赋予一个先验置信度，表示实体指称项指向候选实体的可能性。先验置信度包括名字相似度和实体流行度两种；

（1）名字相似度的计算方法：

(1)

其中，表示实体指称项和候选实体名字间的编辑距离，即从字符串出发，通过字符替换转化成所需的最少编辑操作次数。表示在字符串和的长度中取较长者。以作为先验置信度的含义是：实体指称项与候选实体的名字相似度越大，则二者直接关联的可能性越大。

（2）实体流行度的计算方法：

(2)

其中，表示实体指称项的候选实体，表示对应的第ｋ个候选实体对象，表示维基百科中指向且锚文本内容为的超链接数目。实体流行度是与语料无关的测度，以作为先验置信度的含义是：候选实体的（实体）流行越大，则其作为目标链接对象的可能性越大。

在构造实体相关图的过程中，当顶点间存在直接关联关系时，则不考虑其间接关联关系，只有在顶点间不存在直接关联关系时，才进一步考虑其间接关联关系。此外，对于同一实体指称项对应的多个候选实体（顶点），不考虑其相互之间的关联关系，即实体相关图中同一实体指称项所对应的候选实体顶点间不存在关系边。

最后，考虑如何将候选实体的相关度和语义相关性结合起来得到一个实体链接的标准，我们提出了一个语义一致性判据，根据先验置信度计算候选实体和实体指称项的语义一致性程度，从而实现知识链接。

### 2.2 实体对齐

实体对齐也称为实体匹配或实体解析，是判断相同或不同数据集中的两个实体是否指向真实世界同一对象的过程。知识库实体对齐的目标是能够高质量链接多个现有知识库，并从顶层创建一个大规模的统一的知识库，从而帮助机器理解底层数据。这要求在军事指控领域，将实体、事件和模型三种类型的知识结合成一个统一整体，从而对军事知识图谱的一系列指令和事件进行分析决策。因此，我们提出一种基于网络语义标签的多源知识库实体对齐算法。该算法综合利用属性标签、类别标签和非结构化文本关键词，对齐维基百科实体知识库和前置的各个类型的军事领域的知识库。

该算法的核心是基于实体含有的一定量描述信息，结构化文本关键词，还有大量用户提供的语义标签，通过对两个实体语义相似度的计算，对实体进行对齐。

首先，我们需要对数据进行预处理，统计并记录了在各自知识库中拥有相同条目名称的实体；然后，分别计算实体间类别标签、属性标签以及非结构化文本关键词三方面特征的相似度，综合计算知识库实体与候选实体的相似性，取相似性最大且大于一定阈值的候选实体作为对齐的结果输出，视为语义指向相同，并将知识库实体作为与候选实体对齐的实体保存到体系中；若相似性最大的候选实体不满足阈值，则视为该知识库实体在体系中并无语义指向相同的实体存在，将其作为新建实体保存到体系中，并加入到拥有相同条目名称的实体记录中。

根据上述实体对齐过程，我们需要计算实体间的类别标签、属性标签和非结构化文本关键词三方面特征的相似度。

（1）基于属性标签匹配的语义相似度计算

属性标签是百科实体的重要特征，拥有很多个性化信息。我们结合了军事领域知识库和维基百科知识库，编订了若干属性映射规则，将通过标签特征将实体挂载到体系中，然后依据编定的属性映射规则，将挂载到体系中的百科实体属性映射到定义的属性上，从而克服不同百科知识库属性间的异构性。同时，我们使用三元组的形式表述属性标签，然后根据制定的好归一化规则对属性值进行归一化，利用编辑距离计算属性值的相似度。最后，将匹配属性的属性值相似度累加值作为实体的属性标签相似度。倘若两个实体匹配的属性名越多，两个实体越可能表示语义上的同一事物。

（2）基于类别标签匹配的语义相似度计算

为克服相关知识库缺乏完整的结构描述体系给知识库实体对齐带来的局限，我们从实体类别标签关于实体条目名称的共现信息入手，计算实体在潜在类别依赖关系中的相似度，即基于随机游走的类别标签语义相似度。

我们将实体的每个类别标签视为知识库中的概念，类别标签所标定的每个实体可视作概念的实例。将每个知识库实体用类别标签对应的概念向量来表示，利用式（3）来计算两个实体的类别标签语义相关度：

(3)

其中表示类别标签向量到类别标签向量的语义相关度。

同时，我们结合军事指控领域的知识库特点，采用随机游走的方法来衡量不同实体类别标签的语义相关度。首先，出于算法的时间和空间代价考虑，我们无法用上所有类别标签进行随机游走的计算，必须选取合适的类别标签数目；然后，建立类别标签间关于实体条目名称的共现矩阵，利用有重启的随机游走方法，计算类别标签间的语义相关度矩阵．公式如下：

(4)

其中，表示第ｉ步随机游走得到的类别标签间语义相关度矩阵；表示初始单位矩阵；表示归一化后的标签间关于实体条目名称的共现矩阵。

（3）基于非结构化文本关键词的语义相似度计算

我们利用词向量相似度的方法，计算不同知识库实体非结构化文本关键词之间的相似度．我们提取出非结构化文本中的关键词，建立关键词的词向量Ｓ，进而通过计算词向量的夹角余弦值作为相似度。

(5)

其中表示实体１的非结构化文本关键词向量；表示实体２的非结构化文本关键词向量；表示实体１与２的非结构化文本关键词向量的夹角余弦值。

### 2.3 实体消歧和共指消解

实体消歧是专门用于解决同名实体产生歧义问题的技术。共指消解(entity resolution)技术主要用于解决多个指称项对应于同一实体对象的问题。通过上述两种方法分别解决的不同类型知识融合中实体链接和实体对齐的问题，在此基础上，两种方法都采用了添加外部百科知识库以及相关军事指控领域的知识库。我们通过构建百科知识模型，在实现实体链接和实体对齐，计算相关实体类的相似度的同时，同时解决了实体消歧和共指消解问题。

1. 薛坤. 面向军事领域的知识图谱构建与应用研究[D].大连理工大学,2020.
2. Yang S, Feng D, Qiao L, et al. Exploring pre-trained language models for event extraction and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5284-5294.
3. Nguyen T H, Grishman R. Graph convolutional networks with argument-aware pooling for event detection[C]//Thirty-second AAAI conference on artificial intelligence. 2018.
4. Liu X, Luo Z, Huang H. Jointly multiple events extraction via attention-based graph information aggregation[J]. arXiv preprint arXiv:1809.09078, 2018.
5. Wu Y, Bamman D, Russell S. Adversarial training for relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1778-1783.
6. Wang X, Han X, Lin Y, et al. Adversarial multi-lingual neural relation extraction[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1156-1166.
7. Newcombe H B, Kennedy S J, Axford S J, et al. Automatic linkage of vital records [J]. Science, 1959, 130 (3381): 954-959.
8. Herzog T N, Scheuren F J, Winkler W E. Data quality and record linkage techniques [M]. Berlin: Springer, 2007.
9. Han J W, Kambe M. Data mining: Concepts and techniques [M]. San Francisco, CA: Morgan Kaufmann, 2006.
10. Vapnik V. The nature of statistical learning theory [M]. Berlin: Springer, 2000.
11. Kantardzic M. Data mining [M]. Hoboken, NJ: John Wiley & Sons, 2011.
12. Cohen W W, Richman J. Learning to match and cluster large highdimensional data sets for data integration [C]// Proc of the ACM SIGKDD Conf on Knowledge Discovery and Data Mining, New York: ACM Presser, 2002: 475-480.
13. Bhattacharya I, Getoor L. Alatent dirichlet allocation model for unsupervised entity resolution [C]// Proc of the 6th SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2006: 47-58
14. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.
15. Fisher J, Christen P, Wang Qing. Active learning based entity resolution using Markov logic [C]// Proc of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Auckland:Springer, 2016: 338–349.
16. 李春华. 基于机器学习模型与众包的知识融合方法研究 [D]. 苏州:苏州大学, 2017. (Li Chunhua. Knowledge fusion based on machine learning model and crowdsourcing [D]. Suzhou: Soochow University, 2017.)
17. Grosz B J, Weinstein S, Joshi A K.Centering:A framework for modeling the local coherence of discourse[J].Computational Linguistics, 1995, 21 (2) :203-225