中国石化
SINOPEC

化学品安全技术说明书知识图谱研究

中国石化工程建设有限公司
SINOPEC ENGINEERING INCORPORATION

日期: 2023.07

Contents
目 录

一、引言

- ✓ 近年来，我国化工园区发展迅速，化工园区的数量与日俱增。园区内危化企业众多，生产、储存的危险化学品种类多、数量大，安全问题成了化工园区发展的大患。“8·12天津滨海新区爆炸事故”、“3·21响水化工企业爆炸事故”等事故调查报告均指出危险化学品存在管理不到位、未按照其固有技术指导书进行安全使用等问题。
- ✓ **化学品安全技术说明书（MSDS）**是化学品生产、储存和销售企业按法律要求向政府、机构、客户或公众提供的有关化学品特征的一份综合性法律文件。
- ✓ 化学品生产商和进口商用MSDS文档来阐明化学品的理化特性、对使用者的健康可能产生的危害、包括安全使用化学品的指引、潜在的可能由化学品引发的危害，及给操作人员、运输/存储人员和应急救援人员的安全处理方法等。因此基于MSDS构建知识图谱，并为园区各种人员提供图形化展示方式，对落实新生产安全法等具有现实意义。

一、引言

- ✓ MSDS均包括以下十六项内容：化学品及企业标识、危险性概述、成分/组成信息、急救措施、消防措施、泄漏应急处理、操作处置与储存、接触控制/个体防护、理化特性、稳定性和反应性、毒理学资料、生态学资料、废弃处置、运输信息、法规信息以及其它信息，但这十六项内部的细节不完全相同。
- ✓ 由于化学品的纯物质有数百万种，其中常见的近3000种，混合物可能有上千万种，因此构建融合多种版本的MSDS知识库成为一种现实需要。
- ✓ 在构建MSDS知识库时必然会面临不同版本MSDS的知识图谱融合问题。不同版本MSDS知识图谱融合的关键问题是求解从局部模式知识图谱概念到MSDS全局模式知识图谱概念的一到多映射。
- ✓ 本文提出基于字符匹配的实体自动标注方法来求解该问题。实验结果显示本文提出的实体自动标注方法其性能一直优于基于深度学习的模型；在词袋数据量较小时，其准确率稍低于基于深度学习的模型；当词袋的数据达到一定量后，其准确率和基于深度学习的模型基本相等。

二、MSDS知识图谱构建技术

- 知识图谱是结构化的语义知识库，在逻辑结构上可分为**模式层**与**实例层**两个层次。

模式层

- 模式层构建在实例层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层。
- 本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小。
- 本文采用本体概念构建知识图谱的模式层：实体-关系-实体，实体-属性-属性值，使用这两种模式层来具体描述MSDS文档结构的知识图谱。

实例层

- 实例层主要是由一系列的事实组成，而知识将以事实为单位进行存储。
- 本文采用(头实体，关系，尾实体)和(实体、属性，属性值)这样的三元组来表达事实，并采用图数据库作为MSDS文档内容数据的存储介质。

二、MSDS知识图谱构建技术

- ✓ 图1是甲醇的知识图谱示例，可以看出其中心结点是“甲醇”，十六个小标题是一跳（one hop）结点，描述的是实体-关系-实体模式层。图2展开的是“急救措施”这个属性实体得到的知识图谱，可以看到它的属性值实体，描述的是实体-属性-属性值实例层。通过展开这些实体，可以看到甲醇MSDS的所有内容都呈现在知识图谱中，因此采用这两个层次的结构把一份MSDS文档构建为知识图谱应该是无损的。

甲醇知识图谱的模式层结点

急救措施实例层结点

三、MSDS知识图谱的融合——模式映射

- 基于本体的MSDS知识图谱全局模式集成分两种情况：

第一种情况：全局模式中没有相应的描述，此时需要对全局模式进行补充；

全局模式补充

- 先为各个版本的MSDS分别整理出它们的大标题和小标题。
- 只需要用其它版本的小标题对主版本相应大标题下面的小标题进行补充。

三、MSDS知识图谱的融合——模式映射

第二种情况：全局模式中有相应的描述，但是和局部模式之间是异构的，此时要找到局部模式向全局模式的映射；即其它版本MSDS的知识图谱向主版本知识图谱的标题概念映射，通过这个映射来把局部模式内容集成到全局模式下统一管理，从而完成多种版本的MSDS知识图谱集成。局部模式概念到全局模式概念的映射有如4种情况。

1、概念完全相同的融合

- 在局部模式KG_l中有一个属性实体e_l“相对分子质量”映射为全局模式KG_g中的属性实体e_g“相对分子质量”，可以看到不同版本中的这两个实体的字符串完全相同，并且它们的内涵也完全相同，其映射可以表示为：$E_l=E_g$。
- 这种情况在MSDS知识图谱融合的时候比较容易，因为这部分的概念完全相同，因此只需直接管理这种等值关系即可。

三、MSDS知识图谱的融合——模式映射

2、概念等价的一一映射

- 在局部模式KG_l中有一个属性实体e_l“危险性类别”映射为全局模式KG_g中的属性实体e_g“GHS危险性类别”，这两个实体概念内涵等价，但他们描述的字符串有区别的，虽然它们在两个模式中分别表示，但因表示相同的概念，其局部模式概念到全局模式概念映射可以表示为：$E_l \rightarrow E_g$。
- 在不同版本的MSDS中，其十六个大标题的描述一般完全相同，但有时可能有一些差别，此时就可以采用这种概念等价的一一映射来处理。

3、多概念合并的多到一映射

- 在局部模式KG_l中有两个属性实体e_{l1}“生物降解性”和e_{l2}“非生物降解性”映射为全局模式KG_g中的属性实体e_g“持久性和降解性”，这三个概念描述的字符串不同但在内涵上有交集。通过分析发现这三个实体有如下特点：$e_{l1} \subset e_g$和$e_{l2} \subset e_g$，并且有$e_{l1} \cup e_{l2} = e_g$。
- 因此多概念合并的多到一映射推广的形式化表示为：$e_{l1} \subset e_g, \ldots, e_{ln} \subset e_g$，并且有$e_{l1} \cup \ldots \cup e_{ln} = e_g$。映射可以表示为：$E_{l1} + \ldots + E_{ln} \rightarrow E_g$

三、MSDS知识图谱的融合——模式映射

4、复杂概念分解的一到多映射

- 比如在局部模式KG_l中有个属性实体e_l“刺激性”映射为全局模式KG_g中的两个属性实体e_{g1}“眼睛刺激或腐蚀”和e_{g2}“皮肤刺激或腐蚀”，这三个概念描述的字符串不同但内涵有交集。
- 通过分析发现这三个概念有如下特点：$e_{g1} \subset e_l$和$e_{g2} \subset e_l$，并且有$e_{g1} \cup e_{g2} = e_l$；因此局部模式概念到全局模式映射可以表示为：$E_l \rightarrow E_{g1} + E_{g2}$。此时要把一个局部模式的概念分解成两个概念，然后把它们分别映射到全局模式的一个概念。还比如局部模式KG_l中有个属性实体e_l“应急行动”映射为全局模式KG_g中的三个属性实体e_{g1}“作业人员防护措施、防护装备和应急处置程序”、e_{g2}“环境保护措施”和e_{g3}“泄漏化学品的收容、清除方法及所使用的处置材料”，可以发现$e_l \cap e_{g1}$=“应急”，$e_l \cap e_{g2} = \emptyset$，$e_l \cap e_{g3} = \emptyset$。通过分析发现这四个概念有如下特点：$e_{g1} \subset e_l$，$e_{g2} \subset e_l$，$e_{g3} \subset e_l$并且有$e_{g1} \cup e_{g2} \cup e_{g3} = e_l$；因此局部模式概念到全局模式映射可以表示为：$E_l \rightarrow E_{g1} + E_{g2} + E_{g3}$。
- 因此概念分解的一到多映射推广的形式化表示为：$E_{g1} \subset e_l, \ldots, e_{gn} \subset e_l$，并且有$e_{g1} \cup \ldots \cup e_{gn} = e_l$。
映射可以表示为：$E_l \rightarrow E_{g1} + \ldots + E_{gn}$

三、MSDS知识图谱的融合——实体自动标注

- 对于MSDS知识图谱的自动化标注，可通过Brat文本标注工具先把打标的结果存储到一个词袋entity_class，结合词袋entity_class就可以设计自动化标注算法。算法流程如下所示：

```
算法 1 autoLableTool
    读入词袋 entity_class //由 key 和 value 对组成，key 是词，value 是类型 enity_class[key]
    将原始的 MSDS 文档转换为 txt 文档；
    将 txt 文档转换成按小标题分段的文档 contents；
    for line in contents
        对文本采用后向最大匹配分词算法分词得到分词结果 segments；
        for token in segments
            if token in entity_class.keys()
                把该 token 自动标注为 enity_class[token]
    将自动标注的结果写入到.ann 文件中
```

- 本文设计了一个自动标注工具autoLabelTool，将算法标注的输出结果输入到Brat工具中，通过进一步的人工筛查，以有效的丰富MSDS实体识别所用的词袋。

三、MSDS知识图谱的融合——知识图谱融合

- 基于前面的MSDS文档的命名实体识别，本节主要研究根据实体标注的结果构建全局模式的知识图谱，即从局部模式到全局模式的四种映射，再加上采用局部模式对全局模式的补充，共五种情况的MSDS知识图谱的自动构建和融合。
- 根据3.1节的基于本体的MSDS知识图谱全局模式构建五种情况，本文提出的构建算法2 genChemKG对这五种情况分别进行处理。

- 如果是局部模式对全局模式的补充（即$E_g=E_g \cup \{e_l\}$），则在全局模式知识图谱中增加属性实体，然后把其对应的属性值实体也加入知识图谱，并加上属性实体到属性值实体之间的关系。
- 如果局部模式和全局模式是概念完全相同的映射（即$E_l=E_g$），直接构建属性实体和属性值实体。
- 如果局部模式到全局模式是概念等价的一一映射（即$E_l \rightarrow E_g$），则局部模式的属性实体换成全局模式的属性实体，而属性值实体不变。
- 如果局部模式到全局模式是多概念合并的多到一映射（即$E_{l1} + E_{l2} \rightarrow E_g$），则把局部模式的属性实体$E_{l1}$和$E_{l2}$删除，同时把它们对应的属性值实体连接到全局模式的属性实体E_g。
- 如果局部模式到全局模式是概念分解的一到多映射的情况（即$E_l \rightarrow E_{g1}+E_{g2}+\ldots E_{gn}\ (n \geqq 2)$），则采用的策略是把其文本输入提前训练好的BERT+BiLSTM+CRF模型进行标注，然后根据标注的结果把局部模式的属性值实体连接到对应的全局模式属性实体。

三、MSDS知识图谱的融合——知识图谱融合

- **基于算法2的genChemKG中，深度学习模型要预训练，然后把预训练的模型加载到genChemKG算法中来，以便根据该模型标注的结果构建和融合其它版本MSDS的知识图谱。在该知识图谱集成后，就可查询每个MSDS的各种特性和MSDS之间的各种关联关系；还可进行如知识图谱推理、面向知识图谱的图嵌入学习、构建化工园区相应的MSDS知识图谱管理系统和知识图谱问答系统等基于MSDS知识图谱的各种领域应用了。**

算法2 genChemKG

For 任意一个化学品安全技术说明书

将其原始MSDS文档转换为txt文档；

将原始的txt文档转换成按小标题分段的文档contents；

for line in contents

if $E_g=E_g \cup \{e_l\}$，then

在全局模式知识图谱中增加局部模式的属性实体;

把其对应的属性值实体也加入知识图谱;

并加上属性实体到属性值实体之间的关系;

else $E_l=E_g$, then

直接构建属性实体和属性值实体;

else if $E_l \rightarrow E_g$，then

局部模式的属性实体换成全局模式的属性实体而属性值实体不变;

else if $E_{l1} + E_{l2} \rightarrow E_g$，then

把局部模式的属性实体E_{l1}和E_{l2}删除;

而把它们对应的属性值实体连接到全局模式的属性实体E_g;

else

把其文本输入训练的BERT+BiLSTM+CRF模型进行标注;

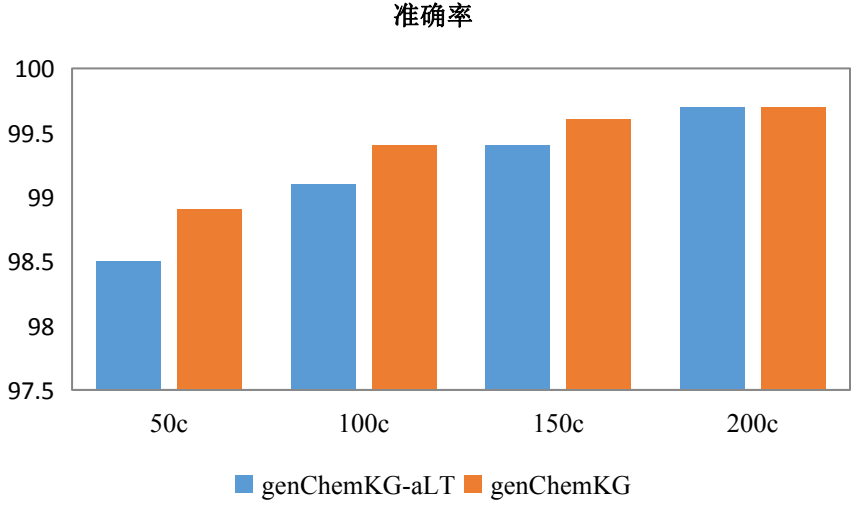
根据标注的结果把局部模式的属性值实体连接到相应的全局模式属性实体;

四、实验

- 本文需要解决的关键问题是MSDS知识图谱局部模式概念到全局模式概念的映射过程中的实体标注，对算法的比较要从两个方面进行，第一是算法的性能，第二是算法的准确率。
- 因此本实验要比较**genChemKG**和**genChemKG-aLT**在进行实体标注时的性能和准确率，实验中采用的数据集是上海有机所网站上面的MSDS，要把这些MSDS的知识图谱集成到中国MSDS注册中心版MSDS的知识图谱中。
- 算法genChemKG给出采用深度学习模型BERT+BiLSTM+CRF标注MSDS的方法，而算法genChemKG-aLT给出采用autoLabelTool标注MSDS的方法。
- 上海有机所网站上的MSDS一共有2000多份，本文一共整理了相对比较完整的1800份，其一部分用来标注、构建词袋或训练模型，其余的用来测试。

四、实验

- 比较分别采用两种算法进行标注的准确率：

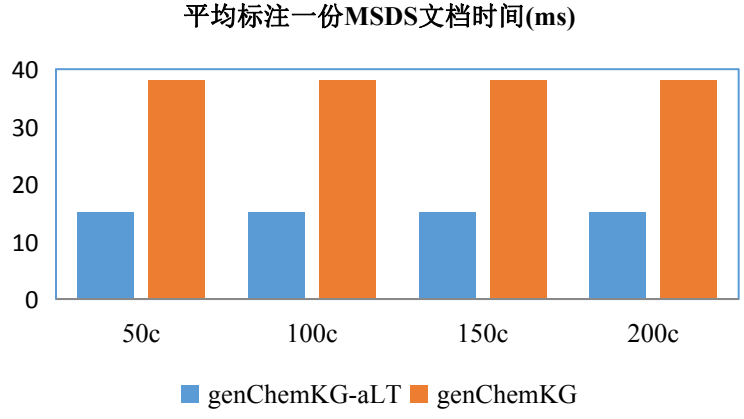


结论

- 当采用50份MSDS文档进行标注和构建词袋时，算法genChemKG的准确率是明显高于算法genChemKG-aLT的。
- 但随着采用的MSDS文档进行标注和构建词袋的份数增加，可以看到genChemKG-aLT和genChemKG的准确率差距逐步缩小；当采用200份标注MSDS进行词袋构建时，这两个算法的准确率基本相等。

四、实验

- 比较两种方法的性能：



结论

- 两个算法标注一份MSDS的时间都非常稳定。
- 原因：

1、每份MSDS需要标注的部分并不多；

2、对于genChemKG-aLT由50份文档组成的词袋和由200份文档组成的词袋相差并不大；

3、对于genChemKG由50份文档组成的训练数据和由200份文档组成的训练数据的模型实际上相差不大

- 算法genChemKG-aLT比算法genChemKG快的根本原因在于前者简单直接，后者相对来说要复杂很多，因此所需时间相对就多一些。

四、实验

实验小结

- 通过算法genChemKG-aLT和算法genChemKG在标注时间和准确率上的对比，本文发现对于特定领域的文档，由于专业术语比较固定，采用深度学习方法进行标注和采用字符匹配方法进行标注在准确率方面的差距其实不大，当然在标注数据量小的时候，深度学习方法是有一定优势的。
- 但在性能上，字符匹配方法由于相对简单所以运行的更快一些。

五、总结和展望

- 本文介绍了一种MSDS知识图谱全局模式构建和其他版本MSDS知识图谱与其融合和集成的方法，也就是说MSDS知识图谱的局部模式向其全局模式映射并进行补充和完善，最终形成一个集成的全局模式。
- 本文以在国家主管部门注册的MSDS为主版本，首先构建其知识图谱并作为为全局模式，再以上海有机所等MSDS知识图谱局部模式和全局模式的融合，验证了本文研发的一种实体标注工具实现局部模式向全局模式映射的有效性。
- 其中概念分解的一到多概念映射方法与CCF竞赛中表现很好且基于深度学习的实体标注工具进行了对比实验，实验结果显示本文提出的实体标注工具在性能方面一直比基于深度学习的实体标注工具好，且手工标注MSDS版本达到一定数量时，本文提出的实体标注工具和基于深度学习的实体标注工具在准确率上基本一致。

五、总结和展望

未来工作：

- 通过集成更多版本MSDS知识图谱，以完善不同MSDS知识图谱版本间的补充和映射方式及其算法，丰富MSDS知识图谱的局部模式到全局模式映射关系；
- 基于基础的MSDS本体，面向化工园区一线人员和专业人员从多种应用角度探讨关于MSDS知识图谱的应用模式及其在事故预防服务等领域的应用方法。

Contributing to society

乐业奉献 Devotion

竞争创新

Innovation

诚信为本 Harmony

宽松融洽

Lifelong responsible for engineering product

Sincerity

永久对工程负责

一起，做更好的！

Win-Win cooperation

合作共赢

Together，let's do better!

SINOPEC

[illegible]

Contributing to society and relief

Compliance

SEI

严格规范

A Provider of Quality Project

石化精品工程的创造者

谢谢

中国石化
SINOPEC