密级：（黑体三号）

“基于文本的知识自动抽取技术与知识图谱构造技术”研究项目

测试大纲

**项目名称：**

**项目编号：**

**合同编号：**

**委 托 方：**

**承 研 方：**

**项目负责人：**

**编制日期：**

基于文本的知识自动抽取技术与知识图谱构造技术

（三号宋体）

（三号宋体）

（三号宋体）

（三号宋体）

（三号宋体）

**32069部队装备部**

# 一、测试目的

本测试大纲是“基于文本的知识自动抽取技术与知识图谱构造技术”研究项目结题验收的测试依据，通过测试，检验到达“基于文本的知识自动抽取技术与知识图谱构造技术”项目研究合同、研究任务书的情况，为结题验收提供参考依据。

# 二、测试依据

列出拟定测试方案的依据，包括合同、国军标以及给测试提供依据的各类文献等。

# 三、测试场地及仪器设备

## 3.1测试场地和环境

### 3.1.1 测试场地

西安电子科技大学计算机科学与技术学院，北校区新科技楼1308实验室。

### 3.1.2测试环境

操作系统：windows7/8/10 64位系统；

处理器：Intel(R) Core(TM) i5-7500 CPU @ 3.40GHZ；

内存：8G。

# 四、指标要求

（1）支持从特定数据源以及公开知识网站文本中获取军事实体、属性、关系和事件等知识。

（2）多阶语义关系抽取技术能够实现面向军事文本的多阶关系抽取、支持并行、连接和嵌套三种模式下的多阶关系抽取，实体识别准确率大于90%，关系抽取准确率大于85%，事件识别准确率大于80%，知识抽取总体召回率不低于85%。

（3）自动的识别情报中所包含的知识特征模式，在保证准确率条件下减少70%以上的样本数量。

（4）语义关系推理规则自动构建能够实现基于既有多元、多重关系的自动推理，支持不少于三层关系的推理，推理规则能够覆盖等价、传递、自反、聚合、泛化、递进等六种关系模式，自动推理的关系准确达到80%以上。

（5）多源知识融合准确率大于85%，知识抽取速度大于5K字节/秒。

（6）自动构建知识图谱中上亿条事实在分布式查询模式下平均响应时间小于1秒。

（7）构建军事文本测试集对知识抽取算法进行验证。

# 五、测试内容、方法及步骤

## 5.1知识抽取数据源测试方法和测试步骤

### 5.1.1 测试方法和条件

在由特定数据源以及公开知识网站文本等构成的文档数据集中，随机抽取一定数量的数据样本，同时应用知识抽取模型对数据样本进行处理，得到军事实体、属性、关系和事件等知识并对得到的知识数据进行检验，以完成对该指标的评测。测试时需要一定数据规模的文档数据集。故需对例如已有的公开知识库等特定的数据源中检索适用的知识数据，需在公开的新闻网站或其他社交媒体中筛选出相关的报文，从而构成文档数据集。

### 5.1.2 测试步骤

（1）在文档数据集中，采取分层抽样的方式，分别在特定数据源和公开知识网站文本中随机抽取一定数量的知识数据和文本数据，得到测试的样本数据集。

（2）应用知识抽取模型对样本数据集中的文本数据所包含的军事实体、属性、关系和事件等知识进行提取。

（3）对提取到的知识数据与样本数据集中的知识数据进行检验。

（4）多次重复地在文档集中进行分层抽样，并检验各次抽样中得到的知识数据。

### 5.1.3 测试判据

若经过多次的抽样检测后，得到的知识数据的质量可达到预期，则判定本项测试满足指标要求。

## 5.2 多阶语义关系抽取测试方法和测试步骤

### 5.2.1 测试方法和条件

基于从特定数据源、公开知识网站文本以及自由文本中获取的军事实体、属性、关系和事件报文等知识，针对军事文本中并行、连接和嵌套三种句式，通过对每个句子使用知识抽取模型生成多个关系三元组，即多阶语义关系，具体的测试框图如图5.1所示。



图5.1 多阶语义关系抽取测试框图

### 5.2.2 测试步骤

（1）输入面向军事文本的测试报文。

（2）利用知识抽取模型对自然语言进行解析，识别出其中的军事实体、属性、关系和事件等知识，对不同句式的句子分别形成相应的实体关系三元组，得到多阶语义关系。

（3）最后保存并行、连接和嵌套三种句式的句子中的多阶语义关系到数据库中，完成多阶关系抽取的测试。

### 5.2.3 测试判据

在面向军事报文的多阶语义关系抽取结果中，若并行句式的抽取结果中拆分了并行实体，生成了相应的关系三元组，连接句式的抽取结果中识别出多个主语实体及相应关系，并且嵌套句式的抽取结果中识别出嵌入的句子和对应关系，即可判定本项测试满足指标要求。

## 5.3 实体识别准确率测试方法和测试步骤

### 5.3.1 测试方法和条件

基于从特定数据源、公开知识网站以及自由文本中获取的军事实体、属性和关系等知识，以及人工评测的真实、准确的实体和关系等知识，对面向军事文本抽取出的多阶语义关系进行评测分析，具体的测试框图如图5.2。



图5.2 实体识别准确率测试框图

### 5.3.2 测试步骤

（1）输入面向军事文本的测试报文，利用知识抽取模型对自然语言进行解析，识别出其中的军事实体、属性、关系和事件等知识。

（2）根据测试文本集人工标注相应的实体、关系以及事件等多元要素，其中总共包含的实体数量为EntityNum，其中EntityNum>10000。

（3）将知识抽取模型识别出的实体和人工标注的实体集进行比对，设置实体语义覆盖度阈值为。

（4）针对每条报文识别的实体e，计算人工标注的实体集中存在的实体和e的实体语义覆盖度s，若，则EntityScore加1，其中EntityScore为实体识别正确的分数。

（5）统计总共识别的实体的数量为EntityCount。

（6）根据公式 ，计算实体识别的准确率。

（7）根据公式 ，计算实体识别的召回率。

### 5.3.3 测试判据

若面向军事报文的多阶语义关系抽取结果中实体识别的准确率>90%，即可判定本项测试满足指标要求。

## 5.4关系抽取准确率测试方法和测试步骤

### 5.4.1 测试方法和条件

利用静态结构化数据源和动态数据源建立军事报文测试集，利用知识抽取模型进行实体、属性和关系等知识抽取。为了便于对关系抽取的准确率进行测试，人工对测试集进行实体关系三元组标注，同时与抽取结果进行比对，其中关系的准确率包括实体以及相应的关系的准确性。

### 5.4.2 测试步骤

（1）对测试集进行知识抽取，对每条报文抽取出相应的实体和关系三元组。

（2）人工对测试集进行标注，建立真实、正确的三元组集合。

（3）针对每条军事报文，将抽取结果与人工标注结果进行比对。其中，假定实体识别正确的情况下计算关系抽取的准确率，记为ReAcc 。

（4）设置关系语义覆盖度阈值为。针对每条报文计算抽取三元组和标注三元组的关系语义覆盖度s，若，则RelationScore加1，其中RelationScore为关系抽取正确的分数。

（5）统计总共抽取的关系的数量RelationCount和人工标注的关系的数目RelationNum。

（6）根据公式 ，计算关系抽取的相对准确率。

（7）根据公式 ，计算关系抽取的相对召回率。

（8）关系抽取的最终的准确率需要在实体识别的基础上计算关系抽取的准确率，因此根据公式，计算关系抽取的准确率。

（9）关系抽取的最终的召回率需要在实体识别的基础上计算关系抽取的召回率，因此根据公式，计算关系抽取的召回率。

### 5.4.3 测试判据

若面向军事报文的多阶语义关系抽取结果中关系抽取的准确率>85%，即可判定本项测试满足指标要求。

## 5.5 事件识别准确率测试方法和测试步骤

### 5.5.1 测试方法和条件

在假定从特定数据源中得到实体，属性和关系等数据都是正确的基础上，人工标注事件活动三元组得到事件测试数据集，然后验证事件识别模块的准确率。具体过程如图5.3所示。



图5.3 事件识别测试流程图

### 5.5.2 测试步骤

（1）输入从特定数据源抽取得到的军事实体、属性和关系三元组测试集。

（2）导入事件识别模块，自动判别测试集数据中包含的事件。

（3）对比人工标注测试集，得到判定正确的事件数量为EventScore。

（4）关系抽取作为事件识别的前提，事件识别的准确率是在假设关系抽取结果正确的条件下计算得到。根据公式，计算事件识别准确率，其中是输入三元组的总数量。

（5）根据公式，计算事件识别的召回率，其中是人工标注得到的事件总数量。

### 5.5.3 测试判据

若面向军事报文的多阶语义关系抽取结果中事件识别准确率大于80%，即可判定本项测试满足指标要求。

## 5.6 知识抽取召回率测试方法和测试步骤

### 5.6.1 测试方法和条件

从特定数据源和公开知识网站获取报文建立文本测试集，利用知识抽取模型抽取测试集中的报文抽取实体、属性、关系和事件等知识，同时人工标注测试集构建准确的三元组集合。

### 5.6.2 测试步骤

（1）针对军事领域的测试报文利用知识抽取模型对自然语言进行解析，识别出军事实体、属性、关系和事件等知识。

（2）人工标注测试集中报文的相应实体、关系以及事件等多元要素。

（3）将知识抽取结果与人工标注结果进行对比分别计算实体识别模块、关系抽取模块、事件识别模块相应的召回率，分别为、和。

（4）知识抽取模块由实体识别、关系抽取和事件识别组成，因此根据公式，计算知识抽取的召回率。

### 5.6.3 测试判据

若面向军事报文的多阶语义关系抽取结果中知识抽取的召回率>85%，即可判定本项测试满足指标要求。

## 5.7 特征模式识别测试方法和测试步骤

### 5.7.1 测试方法和条件

深度学习模型军事实体识别的准确率严重依赖于人工标记数据的质量与数量。人工标注训练数据量的减少会导致军事实体识别准确率的下降，因此我们设计特征模式识别模块对未标注的军事报文进行特征提取预处理，自动标注数据以避免人工标注数据量不足而导致的军事实体识别准确率下降的问题。

### 5.7.2 测试步骤

（1）测试方法分为两组对比实验，A组实验不经过特征提取预处理模块，将人工标注的数据作为训练集的唯一来源，每轮实验逐次减少训练集的训练样本数量。

（2）B组实验每轮采用与A组实验相同样本数量的人工标注数据作为训练集，并通过特征模式识别模块对定量未标注的军事报文进行特征提取预处理生成标注数据加入到训练集中。

（3）统计使用F1值作为评价指标。

（4）在两组间对比在训练集中添加特征模式识别模块自动生成的标注数据后，B组模型军事实体识别相较于A组的同一轮次准确率是否有所上升。

（5）在A组内测试随着训练集训练样本数量的减少，模型军事实体识别的准确率是否下降。

### 5.7.3 测试判据

若在A组内测试随着训练集训练样本数量的减少，当训练样本数据减少70%，模型军事实体识别的准确率下降。并且在B组实验中经过特征模式识别模块补充标注数据后准确率回升，回升后的F1值与A组样本数据减少前的F1值的差值在1%以下，即可判定本项测试满足指标要求。

## 5.8推理规则测试方法和测试步骤

### 5.8.1 测试方法和条件

基于知识抽取得到的军事实体、属性、关系以及关系类别进行关系模式推理。关系类别根据其概念含义分为等价，传递，自反，聚合，泛化，递进六类。根据不同的关系类别对实体之间的属性和关系进行推理，得到知识覆盖程度更广的知识图谱。

### 5.8.2 测试步骤

（1）输入军事领域的结构化知识，通过处理得到存储支持的数据格式。

（2）等价关系测试：输入数据关系类别为等价时，通过推理规则得到关系和属性的并集，对两个实体缺少的关系和属性进行补全，得到关系和属性更加丰富的实体信息。

（3）传递关系测试：输入数据关系类别为传递时，说明该关系可以进行传递，通过推理规则对该关系进行推理可以将原本实体间不存在该关系的实体联系起来，丰富实体之间的关系。

（4）自反关系测试：通过对数据中的关系进行提取，得到自反关系词典，根据词典对实体进行自反关系数据补全。

（5）聚合关系测试：通过对输入数据的分析，识别实体之间的构成关系是否满足部分与整体的概念关系，若满足则添加实体之间的构成信息。

（6）泛化关系测试：输入数据关系类别为泛化时，说明实体之间存在继承关系，通过推理规则得到实体所继承的实体的属性和关系信息，对该实体进行信息的补全。

（7）递进关系测试：通过对输入数据进行关系属性分析，识别实体属性之间是否可以互相进一步解释，若能够进一步解释该属性关系则对属性关系进行补全。

（8）当数据通过以上推理规则推理后，通过查看推理之前的实体关系和推理之后的实体关系得到关系推理层数。

### 5.8.3 测试判据

若面向军事领域的数据经过推理，每种推理关系都可以增加数据，且推理关系规则支持超过三层的推理，即可判定本项指标满足要求。

## 5.9自动推理关系准确率测试方法和测试步骤

### 5.9.1 测试方法和条件

基于知识抽取得到的军事实体、属性、关系以及关系类别进行关系模式推理后，得到经过推理规则完善后的数据。同时根据人工推理的真实准确的推理结果对推理得到的数据进行评测。

### 5.9.2 测试步骤

（1）输入经过推理规则得到的数据，将其中的原始数据和推理数据分开，对推理数据进行评测。

（2）根据测试数据集人工标注通过推理规则得到的推理数据，共涉及M条知识以及N条推理数据，其中M>1000，N>1500。

（3）设定语义相似度阈值α。

（4）用语义相似度计算模型计算通过推理规则得到的数据与人工推理得到的数据的语义相似度s，当s>α时，认为推理正确。

（5）根据公式，计算得到自动推理关系准确率准确性，其中，=N。

### 5.9.3 测试判据

若在军事领域的数据中，在大于1000条知识的推理测试上，自动推理关系准确率大于80%，即可判定本项测试满足指标要求。

## 5.10 多源知识融合测试方法和步骤

### 5.10.1 测试方法和条件

针对从多个知识来源抽取到的三元组，采用数据融合、信息融合、知识融合等进行融合，对比人工标注的融合后的三元组，计算融合的准确率。需要从特定数据源、公开知识网站以及自由文本中抽取到的三元组。

### 5.10.2 测试步骤

(1)读取待融合的三元组文件file\_1和人工标注的融合后的三元组文件file\_2;

(2)依次进行数据融合、信息融合和知识融合，得到融合后的三元组数量为N1，并写入文件file\_3保存;

(3)统计file\_2和file\_3中相同的三元组数量为N2;

(4)根据公式计算多源融合准确率，其中TP=N2，FP=N1-N2，TN=0，FN=0。

### 5.10.3 测试判据

若多源知识融合准确率大于85%，即可判定本项测试满足指标要求。

## 5.11 知识抽取速度测试方法和测试步骤

### 5.11.1 测试方法和条件

构造大于5K字节的文本数据，调用知识抽取模块处理，并计算总耗时。然后计算知识抽取速度。

### 5.11.2 测试步骤

（1）输入大小为的文本数据，导入知识抽取模块处理文本，计算耗时为秒。

（2）根据公式，计算得到知识抽取的速度。

### 5.11.3 测试判据

若测试集的知识抽取速度大于5K字节/秒，即可判定本项测试满足指标要求。

## 5.12 图谱自动构建事实性能测试方法和测试步骤

### 5.12.1 测试方法和条件

基于HugeGraph-Loader支持本地磁盘文件或目录的数据源，支持 TEXT、CSV 和 JSON等格式的文件等条件，借助HugeGragh的数据导入组件HugeGraph-Loader将数据源的数据转化为图的顶点和边并批量导入到图数据库中。

### 5.12.2 测试步骤

(1)根据HugeGraph-Loader要求的数据类型、支持的数据格式准备顶点数据和边数据;

(2)根据数据文件设计图模型，根据数据文件所涉及的属性、顶点类型、边类型等定义图模型文件需要的元数据；

(3)综合预构的图模型和定义的元数据，用groovy编写最终的图模型文件;

(4)根据图模型和数据文件输入源的映射关系编写格式为JSON的映射文件，其中顶点映射和边映射分别对应某类顶点/边的输入源映射;

(5)根据导入数据的不同选择合适的参数，运行 bin/HugeGraph-Loader 并传入参数，提交命令控制导入过程，将数据批量导入图谱；

(6)根据导入结果统计导入时间、导入数据的完整性。

### 5.12.3 测试判据

若测试结果导入速度与region大小相匹配、导入数据完整性达100%即可判定本项测试满足指标要求。

## 5.2 图谱构建事实数量测试方法和测试步骤

### 5.13.1测试方法和条件

基于HugeGraph支持分布式存储、基于Apache TinkerPop3框架实现、支持Gremlin图查询语言，HBase支持百亿以上的顶点和边快速导入等条件逐次导入多个数量级的数据，测试Hugegraph对不同量级数据的容纳性。

### 5.13.2测试步骤

(1)按照HugeGraph-Loader对数据的要求，准备十万、百万、千万、亿等多个数量级的事实，其中顶点的入度和出度均随机;

(2)调整region大小，使其与将要导入的数据量级相匹配；

(3)借助HugeGraph-Loader将不同数量级事实依次导入图谱;

(4)借助HugeGraph-Server提供的restful服务调用Gremlin查询语言查询相关事实。

### 5.13.3测试判据

若测试结果支持亿数量级事实，即可判定本项测试满足指标要求。

## 5.14 分布式查询响应时间测试方法和测试步骤

### 5.14.1测试方法和条件

根据HugeGraph支持Apache Gremlin标准图查询语言、支持基于图的OLTP方案、提供毫秒级的关联关系查询能力（OLTP）的特性，测试分布式查询模式下平均查询响应时间。

### 5.14.2测试步骤

(1) 按照HugeGraph-Loader对数据的要求，准备上亿条事实，其中点和边各半，顶点的入度和出度均随机；

(2)编写匹配数据的图模型文件和映射文件，并执行命令通过HugeGraph-Loader导入指定图谱；

(3)调用Gremlin查询语言按照id查询指定顶点和边并统计时间，并根据公式计算平均响应时间，其中IDTime为响应时间之和，N为查询次数；

(4)调用Gremlin查询语言按照属性及属性值查询指定顶点和边并统计时间，并根据公式计算平均响应时间，其中为响应时间之和；

(5)调用Gremlin查询语言进行两层及三层关系推理并统计时间，并根据公式计算平均响应时间，其中为响应时间之和；

(6)综合上述三类测试次数及测试时长，计算总平均响应时间。

### 5.14.3测试判据

若测试结果在≤1s范围内，即可判定本项测试满足指标要求。

## 5.15军事文本测试集测试方法和测试步骤

### 5.15.1 测试方法和条件

基于从静态结构化数据源、动态数据源以及自由文本中获取军事领域的报文，构建军事样本测试集对知识抽取模型进行整体测试。人工对测试集中的实体、关系和事件等要素进行标注从而构建一套标准的军事报文标注集合。

### 5.15.2 测试步骤

（1）输入军事文本测试集对文本进行过滤处理。

（2）将经过过滤之后的文本作为知识抽取模型的输入，抽取出相应的军事实体、属性、关系和事件等知识。

### 5.15.3 测试判据

若测试集是属于军事领域的文本，且可以抽取出知识三元组，则可判定本项测试满足指标要求。

# 六、测试记录

附录：测试报告（根据测试大纲内容，格式自拟）。