**变更记录**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变更编号 | 版本号 | 日期 | 变更状态 | 变更简单描述 | 撰写人 | 审核人 |
| 1 | 1.0 | 20130904 |  | 初始版本 | 林基远 |  |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |
| 9 |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |
| 11 |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |
| 13 |  |  |  |  |  |  |
| 14 |  |  |  |  |  |  |
| 15 |  |  |  |  |  |  |
| 16 |  |  |  |  |  |  |
| 17 |  |  |  |  |  |  |
| 18 |  |  |  |  |  |  |

变更状态：**A** – 增加 **M** – 修改 **D** – 删除

**基于规则的关系抽取组件*drawline*使用说明**

目录

[1. 简介 3](#_Toc366092166)

[2. 规则 3](#_Toc366092167)

[2.1 定义 3](#_Toc366092168)

[2.2 注意事项 4](#_Toc366092169)

[3. 使用 5](#_Toc366092170)

[3.1 编译 5](#_Toc366092171)

[3.2 运行 6](#_Toc366092172)

[3.3 与*ICTLAP\_CORE\_NE*配合使用 6](#_Toc366092173)

# 简介

***drawline*** 是一个**基于规则的关系抽取组件**，它对***MTIE (Mengtao's information Extraction)*** 进行重写，目的是更好地维护以及减少抽取时间。 ***drawline***本身不会对文章进行命名实体抽取，它是对命名实体抽取结果按照规则抽出实体间的关系，比如

|  |  |
| --- | --- |
| 文章：  *玛依努尔接受了CNN的采访。热比娅的丈夫在接受采访的时候说. 热比娅的儿子呼吁美国议会.*  *热比娅受到了日本总理的邀请。奥巴马约见了热比娅。热比娅说要加强与藏青会的合作。* | 规则文件：  CONCEPT:NAME:玛依努尔  CONCEPT:LOCATION:美国  CONCEPT:ORGANIZATION:美国议会  CONCEPT:LOCATION:日本  CONCEPT:LOCATION:奥巴马  CONCEPT:ACTION:约见  CONCEPT:ACTION:接受  MCONCEPT\_RULE:PERSON\_ACTION(person, action):(SENT,(ORD,(DIST\_5, "\_person{NAME}", "\_action{ACTION}"))) |
| 结果：  164 15  person:奥巴马  action:约见  -----  1 18  person:玛依努尔  action:接受  -----  \*\*\*\*\*  NAME:玛依努尔  LOCATION:美国  ORGANIZATION:美国议会  LOCATION:日本  LOCATION:奥巴马  其中，在\*\*\*\*\*上的是抽出的关系，\*\*\*\*\*下的是概念。对于每个关系，第一行为offset和len，即关系对应字符串在原文本中的byet-offset，byte-len。接下来是关系，冒号前是规则的参数名，冒号后是实体实例。接着以一行-----结束该关系。 | |

# 规则

## 2.1 定义

目前，***drawline***支持:

* **定义概念**
  + **CONCEPT:concept\_name:sequence\_of\_words**

如 **CONCEPT**:COUNTRY:中国

* + **REGEX:UPSTR:[A-Z]{2,}**
  + **MCONCEPT:mconcept\_name:(word|CONCEPT|REGEX)**

如**MONCEPT**:SCHOOL:LOCNUM小学

**CONCEPT**:LOC:北京

**CONCEPT**:NUM:第[1-9]

由于定义了LOC，NUM， ***drawline*** 会自动将LOCNUM拆成LOC和NUM，不需在它们之间加空格（否则认为空格也需匹配）。 ***drawline***拆分单词时**优先考虑长单词**，如还定义了一个

**CONCEPT**:LOCNUM:北京第一

***drawline***就不会拆分LOCNUM，而是将LOCNUM当作一个整体， 因此不会用到LOC和NUM，这时不会匹配北京第2小学，而只会匹配北京第一小学。

* **定义规则**，如

**MCONCEPT\_RULE:obj\_pattern(per, loc, time, org):**

**(SENT, "\_per{PERSON}", "\_loc{LOCATION}", "\_time{TIME}", "\_org{ORG}")**

其中PERSON, LOCATION, TIME和ORG是定义的**CONCEPT**或**MCONCEPT**类型的概念。 而per, loc, time, org是参数名，将会在结果输出。模式规则可以任意**嵌套使用**(但不能递归), 如

**MCONCEPT\_RULE:obj\_pattern(per, loc, time, org): (SENT, (DIST\_2, "\_per{PERSON}", "\_loc{LOCATION}"), (OR, "\_time{TIME}", "\_org{ORG}"))**

目前支持的操作符有：

|  |  |
| --- | --- |
| 操作符 | 说明 |
| **AND** | 所有子句都出现的字符串才会被匹配。 |
| **OR** | 只要有一个子句出现，该字符串就会被匹配。 |
| **SENT** | 所有子句都出现在同一个句子，该字符串才会被匹配。  ***drawline*** 以".?!。？！"及换行符来**断句**（换行符是为了处理如表格型等无明确句号的文章）。 |
| **ORD** | 所有子句按规则定义的顺序同时出现，该字符串才会被匹配。 |
| **DIST\_n** | 所有子句同时出现在字符串，且**相邻**子句实例距离（**间隔距离**）不超过n个词时，该字符串会被匹配。  ***drawline*** 将**词**定义为：   * 单个中文符号（包括汉字，中文标点符号） * 断句符号（如上述） * 英文单词 * 数字串 * 连续的英文标点符号及空格 |

## 2.2 注意事项

* + 模板及文本文件都需要是UTF-8编码，结 果文件也是UTF-8编码
  + ***drawline*** 对输入规则的先后顺序没有要求
  + 定义概念及规则，严格按照附录的说明定义，如

1. 不要用中文冒号代替

**CONCEPT**：COUNTRY：中国

1. 不要用中文引号代替
   * 如果不需匹配，匹配串不要出现多余空格（特别是首尾），如

**CONCEPT**:NAME: xxx

这里xxx前的空格会被当作需要匹配的空格。

* + 如果满足条件的同一字符串在文本多个位置都可能出现，结果文件会把所有位置都输出
  + 同一名字可以有不同规则，同一名字的规则会被认为是同一类型的
  + 对于一个字符串同时满足多条规则的， *drawline* 的处理为：

1. 如果这多条规则是同一类型（名字相同），则只输出一次。如定义了

**MCONCEPT\_RULE**:NAME\_AGE(person, age):(SENT, "\_person{NAME}", "\_age{AGE}")))

**MCONCEPT\_RULE**:NAME\_AGE(person, age):(DIST\_5, "\_person{NAME}", "\_age{AGE}"))

**CONCEPT**:NAME:A

**CONCEPT**:AGE:[1-9] year

这两条规则会匹配"A is 5 years' old."，但只输出一次。

1. 如果这多条规则是不同类型，每种类型都会输出一次。如上面将第2个NAME\_AGE

修改为NAME\_AGE2。

* 对于**MCONCEPT\_RULE**只输出在有指定参数名

**MCONCEPT\_RULE**:NAME\_AGE(person, age):(SENT, "\_person{NAME}", "\_age{AGE}", "\_sex{SEX}")))

这里sex虽然是参数，但没有在NAME\_AGE(person, age)指定，因此不输出，只输出person, age。

* **MCONCEPT\_RULE**内的子句对于以下划线开头的变量认为是参数，如"\_person{NAME}"。

如果要在这个子句加入其他概念，概念名不要包含下划线。如

CONCEPT:ACTION\_ATTACK:攻击

MCONCEPT\_RULE::" ACTION\_ATTACK \_person{NAME}"

这时无法判断\_ATTACK \_person还是\_person才是参数。

* 由于有需求，对于**MCONCEPT\_RULE**如果参数为(person, person)，在最后的结果里会自动增加一个coexist参数，匹配的词为"同"（person匹配的词均为非英文时），"and" (person均为英文时)

# 使用

## 3.1 编译

* 编译成可执行的组件，在当前目录输入

$ make

* 编译成库

$ make lib

主要步骤为将规则push进 *drawline* ，然后匹配即可：

|  |
| --- |
| Drawline::push(const std::string &one\_rule);  Drawline::match(const std::string &text); |

具体请参考src/drawline\_driver.cpp，需要头文件src/drawline.h, src/drawline\_ds.h。

## 3.2 运行

编译后 *drawline* 将会在当前目录产生，使用方法如下，其中template.cfg是配置的模板文件，article.txt是要抽取的文本文件，res.txt为结果文件。

$ ./drawline template.cfg article.txt >res.txt 2>/dev/null

## 3.3 与*ICTLAP\_CORE\_NE*配合使用

这两个模块搭配的功能是对一篇文章，抽取其中的命名实体，然后对这些实体按照配置的模板进行关系抽取。其中

* *ICTLAP\_CORE\_NE* 是对文章进行命名实体识别（识别人名、组织名、 地名），识别出的实体将会合并到关系模板
* *drawline* 是对文章按照合并的实体、关系模板抽取这些关系实例

$ cp template.cfg combine.cfg

$ ./ne\_driver article.txt | grep CONCEPT >> combine.cfg

$ ./drawline combine.cfg article.txt > res.txt 2 > /dev/null