Event Extraction and Knowledge Discovery

——Analysis of Salsalate outcome texts and Viagra outcome texts

阳旭

2017/4/11

知识发现是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。知识发现将信息变为知识，从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展做出贡献。

目前已经出现了许多知识发现技术，分类方法也有很多种。
按被挖掘**对象**分有基于关系数据库、多媒体数据库；
按挖掘的**方法**分有数据驱动型、查询驱动型和交互型；
按**知识类型**分有关联规则、特征挖掘、分类、聚类、总结知识、趋势分析、偏差分析、文本采掘。
知识发现技术可分为两类：基于**算法**的方法和**基于可视化**的方法 。

材料： Salsalate and Viagra raw processed outcome texts

方法：MetaMap SemRep

MetaMap是由美国国立医学图书馆开发的高度可配置程序，用于将生物医学文本映射到UMLS（一体化医学语言系统），或者等效地发现文本中涉及到的超级叙词表概念。

MetaMap使用基于符号，自然语言处理和计算机语言技术的知识密集型方法。除了应用于IR和数据挖掘外，MetaMap是NLM的医学文本索引器（MIT）的基础之一，用于NLM生物医学文献的半自动和全自动索引。

Interactive MetaMap

We are currently using **MetaMap 2016 version 2**. The default knowledge source is **2016AB** with Data Version **USABase**, and our **Strict** Data Model. Please remember that users are responsible for compliance with the UMLS Metathesaurus License.

Text to be Processed (Single block of text, maximum 10,000 characters):

Batch MetaMap

Contact Us

User: tongkaisun: Indexing Initiative > Batch Access > UTS_Required > Batch MetaMap

FULL Email Address (Required):

File to Upload (Required):
选择文件 未选择任何文件

User Defined Acronyms File (--UDA) (optional): 选择文件 未选择任何文件

Batch Notes (Optional):

Data Options (select one of each):
Knowledge Source (-Z): 1617 (2016AB) **Data Version (-V):** NLM **Data Model:** Relaxed Model (-C)

Currently Using MetaMap 2016 V2

Submit Batch MetaMap

Reset Form

Reminder: Users are responsible for compliance with the UMLS hesaurus License

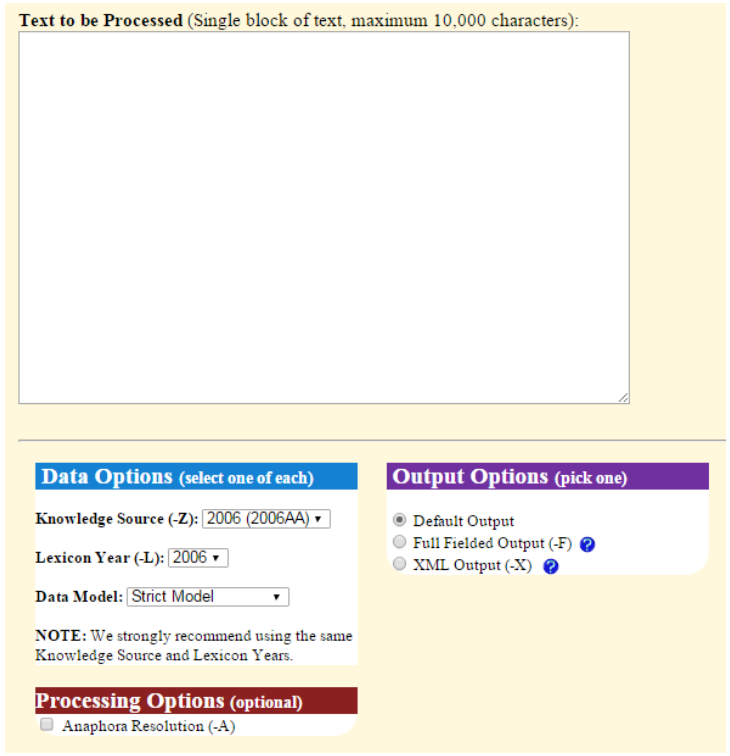
Release Notes (74 kb)

SemRep是从生物医学文本中提取语义预测（主题-关系-对象三元组）的程序。

例子：We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
SemRep提取：Hemofiltration-TREATS-Patients
Digoxin overdose-PROCESS_OF-Patients
hyperkalemia-COMPLICATES-Digoxin overdose
Hemofiltration-TREATS(INFER)-Digoxin overdose

每个预测的主体和对象参数是来自一体化医学语言系统(UMLS)的超级叙词表的概念，并且关系（大写）是来自UMLS语义网络中的关系。

SemRep也有交互模式和批处理模式。


Text to be Processed (Single block of text, maximum 10,000 characters):
Data Options (select one of each)
Knowledge Source (-Z): 2006 (2006AA)
Lexicon Year (-L): 2006
Data Model: Strict Model
NOTE: We strongly recommend using the same Knowledge Source and Lexicon Years.
Processing Options (optional)
Anaphora Resolution (-A)
Output Options (pick one)
Default Output
Full Fielded Output (-F)
XML Output (-X)

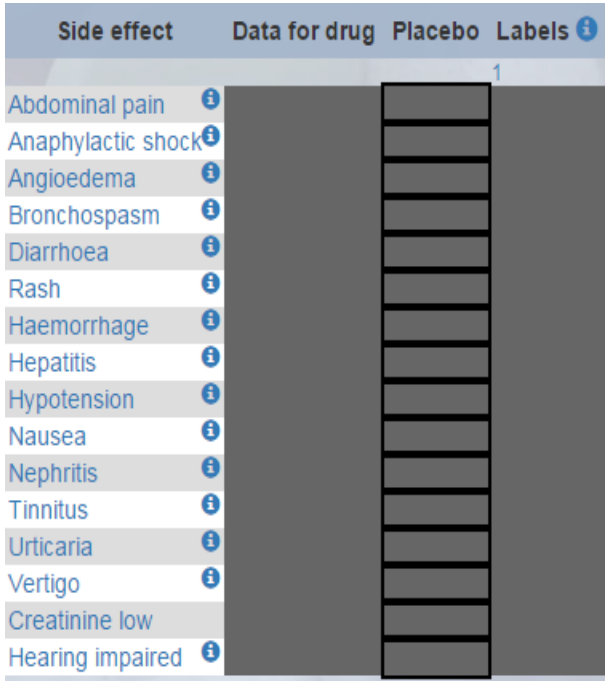
交互模式


FULL Email Address (Required):
File to Upload (Required):
选择文件 未选择任何文件
Batch Notes (Optional):
Data Options (select one of each)
Knowledge Source (-Z): 2006 (2006AA)
Lexicon Year (-L): 2006
Data Model: Strict Model
NOTE: We strongly recommend using the same Knowledge Source and Lexicon Years.
Processing Options (optional)
Anaphora Resolution (-A)
Output Options (pick one)
Default Output
Full Fielded Output (-F)
XML Output (-X)
Batch Specific (pick any or none)
Single Line Delimited Input
Single Line Delimited Input w/ ID
Silent on Errors
No Second Error Attempt
Individual Item Timeout:
900 seconds (15 min.)
Requested Run Priority:
Normal
Create dload.gz download file?: NO

批处理模式

Salsalate's **Pharmacology indication(药理说明)** in DRUGBANK:

For relief of the signs and symptoms of rheumatoid arthritis, osteoarthritis and related rheumatic disorders. （缓解类风湿性关节炎，骨关节炎和相关风湿性疾病的体征和症状）



Salsalate's **side effect** in SIDER

将Salsalate文本提交到SemRep中进行处理，处理完之后系统就会给你指定的邮箱发送一个处理完的邮件。

在邮件中可以得到系统处理后的连接，其中有一个为text.out的文件。

```
tx.54 To quantify the impact of anemia treatment by salsalate on self -reported outcomes measures by
      subjects answering 47 questions for patients with anemia and or fatigue.

tx.54|relation|C0002871|Anemia|dsyn|dsyn|||PROCESS_OF|C0030705|Patients|podg,humn|humn||
tx.54|relation|C0073983|Salsalate|orch,phsu|phsu|||TREATS|C0030705|Patients|podg,humn|podg||
tx.54|relation|C0073983|Salsalate|orch,phsu|phsu|||TREATS(INFER)|C0002871|Anemia|dsyn|dsyn||
tx.54|relation|C1825598|IMPACT gene|gngm,aapp|aapp|55364|IMPACT|TREATS|C0002871|Anemia|dsyn|dsyn||
```

unexplained anemia (UAE)	不明原因贫血
high interleukin	高白细胞介素
insulin sensitivity	胰岛素敏感性
postprandial lipemia	餐后脂血症
insulin clearance	胰岛素清除率
C-reactive protein	C-反应蛋白
anemia	贫血
serum hepcidin	血清铁调素
erythropoietin	促红细胞生成素
IL-6 and Tumor Necrosis Factor Receptor1	炎症标志物：IL-6、肿瘤坏死因子受体
Reactive hyperemia	反应性充血

从处理后的Salsalate文本中找到的相关症状及名词

Viagra，也称作Sildenafil(西地那非)，在DRUGBANK中的药理说明：For the treatment of erectile dysfunction and to relieve symptoms of pulmonary arterial hypertension (PAH).

Side effect
Headache
Bronchitis
Nausea
Upper respiratory tract infection
Oedema
Pharyngitis
Flushing
Diarrhoea
Epistaxis
Pneumonia
Cough
Dyspepsia
Myalgia
Dyspnoea exacerbated
Insomnia
Body temperature increased
Pain in extremity
Erythema
Back pain
Influenza
Disorder sight

Sildenafil's side effect in SIDER

Diabetic cardiomyopathy	糖尿病性心肌病
Red Blood Cell	红血细胞
Soluble platelet selectin	血小板凝集素
pulmonary hypertension	肺动脉高压
parenchymal lung disease	实质性肺病
hypersensitivity	超敏反应
atrial septostomy	心房间隔切开术

同样的，从处理后的Viagra文本中找到的相关症状及名词

总结

I. 对给出的未处理文本进行了预处理，去除其中重复的部分；

II. 了解和学习MetaMap和SemRep的使用；

III. 利用工具对处理过的文本进行处理；

IV. 在处理后的结果中找到一些相关的内容。

THANK YOU ！