【读英文还是比较吃力，第一遍没有看太懂，第二遍才基本看明白，很多术语也查不到，还得麻烦刘老师解释一下几个术语的意思。。。】

【题目】Evaluation of text-processing algorithms for adverse drug event extraction from social media

【作者】Alejandro Metke-Jimenez, Sarvnaz Karimi Cecile Paris

【单位】The Australian e-Health Research Centre CSIRO, Australia, Computational Informatics CSIRO, Australia

【期刊】

【时间】2014

【笔记】

【摘要】

不良药物反应的发现不再局限从药品商和健康专家给监管部门的报告中挖掘。患者论坛和其他社会媒体也作为额外的信息来源来辅助不良反应的发现。药物、不良药物反应、疾病、症状和患者的统计特征等信息的抽取是很有必要的。该领域大部分研究都是使用的基于词法层的信息抽取方法，他们没有明确的评估文本处理步骤对最后结果的影响。本文通过实验定量的评估了比较流行的技术来完善信息抽取过程。

【本文贡献】

测试了不同的 tokenisation algorithms、stemming（词干提取？）、和去停用词方法对 ADR 相关术语抽取的影响；

测试了使用 medical controlled vocabulary（药物约束词表？）过滤非医药词语的效果；

测试了使用 consumer controlled vacabulary（这是什么词表？）过滤非医药词语的效果；

【背景】

目前，只有很少关于从社会媒体中进行 ADE 探测的研究工作。

【领域词表】CHV（http://www.consumerhealthvocab.org/）、MedDRA、SIDER、SNOMED CT、UMLS（http://www.nlm.nih.gov/research/umls）

【社会媒体】

DailyStrength：http://www.dailystrength.org/

Health and Wellness Yahoo! Groups

MedHelp

Bread cancer forums

Diabetes online community

【预处理：分词+过滤】

分词方法：（1）空格、（2）非字母字符、（3）基于语法的分词器：Unicode Text Segmentation algorithm（http://unicode.org/reports/tr29/）

过滤：（1）停用词（2）Porter stemmer（3）转换成小写

【匹配：controlled vocabulary and type of entity】

Controlled vocabulary 用来识别文本片段中的相关概念。

用到的 controlled vovabulary：CHV、UMLS、ALL

Table 2: Types of entities and the corresponding UMLS semantic types.

Entity Type	UMLS Semantic Type
Adverse drug reaction	Sign or symptom
Disease	Disease or syndrome
Medication	Organic chemical Pharmacological substance Clinical drug

【实验数据】

AskaPatient医药论坛提供的2001年至2013年Diclofenac相关的250个帖子和回复。涉及到Arthrotec, Cambia, Cataam, Diclofenac potassium,Diclofenac sodium, Flector, Pennsaid, Solaraze, Voltaren,and Zipsor。

4个医学学生使用Brat annotation tool对语料进行标注（将他们能够识别的症状、药物和不良药物反应标注出来）。表3是标签及释义。所有文档均分给四名标注者，此外，5篇文档分给所有人来计算inter-annotator agreement（标注者间的一致性？），用两个公式：strict agreement 和 relaxed agreement。

Table 3: Tags and their definitions for the annotation.

Tag	Definition	Example
Drug	Mentions of the name of a medicine or drug	Diclofenac
ADR	Mentions of adverse drug reactions	Dizziness
Disease	Name of a disease for which the patient takes the medicine	Anxiety
Symptom	Symptoms of a disease that led them taking to a drug	My heart was racing

一致性公式：

$$agreement(A_i, A_j) = \frac{max(count(A_i), count(A_j))}{count(match(A_i, A_j, \alpha, \beta))},$$

Strict：完全匹配；Relaxed：覆盖就可以

α 是跨度；

β 是内容；

【评估方法】和计算inter-annotator 一致性一样，有strict 和 relaxed 两种方法来评价算法。

Strict评估方法要求spans完全匹配；spans通过标注工具Brat生成的字符偏移量来计算；

Relaxed评估方法匹配两个能够覆盖的spans。例如，下边的也算匹配：

- ground truth = (182, 197) “aches and pains”
- actual result = (182, 187) “aches”, (192, 197) “pains”

【评估公式】

$$\text{Precision} = \frac{n_{TP}}{n_{TP} + n_{FP}},$$

$$\text{Recall} = \frac{n_{TP}}{n_{TP} + n_{FN}},$$

$$\text{F-Score} = 2 * \frac{precision * recall}{precision + recall},$$

nTP 是匹配的spans的数目；nFP是系统输出的且不在ground truth（这个是什么意思？）中的数目；nFN是在ground truth中但是没有被系统输出的数目；

【实验】

使用Lucene（indexer+searcher），indexer用来建立帖子的倒排索引（用不同的分词+过滤组合），searcher用来导入controlled vocabularies 和使用它们的entries作为queries。

每个query返回一个文档和span集合。

最终结果是所有queries查询结果的组合，然后系统就能够将最终结果和来自ground truth的结果进行对比。