A Review of Recent Research in Extracting Information From Medical Textual Documents

AUTHOR
University of Florida

June 26, 2012

1 Automated identification of extreme-risk events in clinical incident reports

- Objectives: classification problem
- Methods: Naive Bayes and Support Vector Machine.
- Result assessments: precision,recall,F-measure,AUC.
- Feature extraction: Punctuation removed, converted to lower case,'bags of words'
- Input representation: binary,term-frequency, thresholding,tf-idf
- Feature selection: excluding words in similar frequency(dimensionality reduction) excluding determiners, prepositions, pronouns, conjunctions(pos tagging) stemming, bigrams.
- Training and validating the classifiers: Naive Bayes, SVM(linear,Polynomial,RBF), 10-fold cross-validation.

2 Extracting Information from Textual Documents in the Electronic Health Record:A Review of Recent Research.

- Spell checking, word sense disambiguation(?), POS, Contextual features like negation, temporarily, and subject identification.
- Automatic de-identification uses the extraction of personal information before its removal. Rule-based NER.
- Contextual Feature Detection and Analysis. Negation(e.g. "denies any chest pain") temporarily(e.g. "fracture of the tibia 2 years ago"). Event subject identification(e.g. "his mother has diabetes"). NegExpanding,NegEx,Negfinder: a program detecting negation terms TimeText detected temporal relations.

3 UMLS project

- The UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.
- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people. disparate databases and systems.

4 Automated systems to identify relevant documents in product risk management

- logistic regression, K-nearest neighbour, Naive Bayes, SVM
- Word occurrence, Binary frequence, TF-IDF.
- stemming, remove prepositions, indentify acronums, synonyms obtained from Omniviz.

5 De-Indentification

- Patient's and doctor's first and last names
- Id numbers, Phone, fax, pager numbers, Hospital names, Geographic locations, Dates, Ages above 90

features of the text used in de-identification

- Target word to be classified
- Words up to 2 words left/right of target
- Target part of speech
- Target capitalization
- Target length and others.

6 Coreference resolution: A review of general methodologies and applications in the clinical domain, Journal of Biomedical Informatics, 2011

- extract drug names and signatures. extract principal diagnosis from discharge summary. smoking status and medication dosage, frequency, and route.
-