Effective Medical Archives Processing Using Knowledge Graphs

Xiaoli Wang, Rongzhen Wang, Zhifeng Bao, Jiayin Liang, Wei Lu

Xiaofan Yan

Outline

- Abstract
- Introduction
- Methods
- Evaluation
- Conclusion

Abstract

- Medical archives processing (classification task)
  - Medical archives recognition
  - Feature extraction
  - Text classification
- Method
  - Build a semantic-rich medical knowledge graph
  - Recognize text using OCR and extract keywords using a KG based feature extraction algorithm from medical archives
  - Evaluate the similarity using a KG based semantic measure

Introduction

- Paper archives -- difficult to apply
- Meaningful to assign them with class labels
  - Typical classification techniques
  - Deep learning models
- Propose a novel KG-MDMF
  - Knowledge Graph based Medical Document Mining Framework
  - Medical KG construction
  - Medical archives processing
  - Text classification

Methods

- KG construction
  - Build a basic KG using entities and relationships from medical web resources and dictionaries
  - Extract entities and relationships from clinical data
  - Integrate expert Q&A results using majority vote algorithm into KG
  - 6 types of entities and 4 types of relationships

Methods

- Semantic Measures
  - Similarity score
    - In the corresponding conceptual hierarchy
    - In the KG (compute distance by path searching algorithm)
    - Of two medical documents
- Medical Archives Processing
  - OCR recognition
  - Preprocessing -- segment and remove stop words
  - Keyword extraction -- map each word to get entity set, and use semantic similarity to get edges' weight for ranking to extract keyword

Methods

- Improved Classification Algorithms
  - KNN
    - Typical: cosine similarity as the distance metric in tf-idf
    - Define: (two component)
      - Both tf-idf and word embedding
      - Text similarity and the weight
  - SVM
    - Using entity similarity to calculate the SVM based on a semantic-based Gaussian kernel

Evaluation

- Settings
  - Dataset: EMR & MRD
  - OCR: ABBYY & Baidu AI
  - Metrics: expert Q&A system
- Result on text classification

Table 3: The results on classification

Algorithms	*Precision*	*Recall*	*F-measure*
KNN	0.57	0.48	0.51
KG-KNN	0.58	0.55	0.56
CNN	0.72	0.71	0.71
SVM	0.71	0.6	0.65
KG-SVM	**0.78**	**0.75**	**0.76**

Conclusion

- Present a medical archives processing framework KG-MDMF, using dictionary matching, extraction and expert Q&A system to construct KG, and define semantic measures to calculate similarity score, then do medical archives processing and classification improving.