**Medical information extraction**

Paper link: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fa63b19767b6024b6e5925139ca1d01626e319c0>

**Summary of the Paper:**

**Project Overview:**

The project focuses on extracting structured information (drug name, dosage, frequency, and method of delivery) from unstructured medical notes. This process is essential for large-scale analysis, such as data mining, as unstructured text cannot be directly analyzed.

**Methods Used:**

1. **Method 1: Basic Text Extraction**
   * Used a subset of diabetes drugs to trigger the extraction.
   * Verified the text processing pipeline.
2. **Method 2: Sentence-Level Analysis**
   * Targeted sentences containing keywords like "diabetes," common diabetes drugs, or "insulin."
   * Evaluated precision (61.33%) and recall (77.13%).
   * Challenges: Did not handle negation or differentiate family history from diagnosis.
3. **Method 3: Regular Expression for Dosage Extraction**
   * Used enhanced grep (egrep) and a C-shell script to extract drug dosages.
   * Identified patterns where a number followed a unit of measurement.
   * Precision: 100% for drug dosages.
   * Recall was affected by unexpected measurement units and multi-word drug names.
4. **Method 4: Improved Regular Expression**
   * Added support for floating-point numbers and expanded the list of measurement units.
   * Improved precision and recall for all extracted information:
     + **Precision (Average):** 96.74% (training) and 96.70% (test).
     + **Recall (Average):** 69.48% (training) and 79.72% (test).

**Dataset:**

* The dataset included transcribed and anonymized notes for **12,222 patients** from the University of North Carolina at Chapel Hill (UNC) hospital system.
* Training set: 30 randomly selected patient notes.
* Test set: 40 randomly selected patient notes.

**Conclusion:**

The study demonstrates that a small set of heuristics and regular expressions can effectively extract critical medical information from unstructured text with high precision. Further improvements are needed to address recall, especially for prescribed frequencies.

Paper link: <https://dsr.cise.ufl.edu/wp-content/uploads/2013/09/mlhealth13_submission_14.pdf>

**Summary of the Paper: Transforming Unstructured Data into Structured Data**

The paper focuses on methods to convert unstructured textual data in Electronic Health Records (EHR) into structured formats, facilitating efficient analysis and use in healthcare applications. The transformation process is critical for tasks such as risk stratification and patient management, as it enables the integration of textual narratives with numerical and categorical data.

**Key Methods and Approaches for Data Transformation**

1. **Feature Extraction from Unstructured Text:**
   * Bag-of-Words (BoW): A traditional method that converts text into a high-dimensional numerical representation based on word frequency.
   * Bag-of-Concepts: Builds on BoW by incorporating medical concepts extracted using tools like cTakes and UMLS (Unified Medical Language System), which map terms to predefined ontologies.
2. **Numeric Feature Extraction:**
   * Specialized modules in cTakes were utilized to extract discrete features such as vital signs (e.g., blood pressure, pulse) and lab results (e.g., glucose levels).
   * Numeric values were discretized (categorized into ranges) and appended to their corresponding medical concepts to create composite structured features.
3. **Section Identification in Medical Notes:**
   * Unstructured admission notes were partitioned into predefined sections, such as History of Present Illness, Physical Examination, and Assessment and Plan, using a rule-based system.
   * Each section was analyzed separately, and its content was passed to cTakes for extracting structured data.
4. **Addressing Sparsity in Feature Space:**
   * Sparse features resulting from unstructured data were mapped to a medical concept ontology (e.g., UMLS, ICD-9).
   * Concepts with insufficient sample instances were collapsed into higher-level parent concepts until a sufficient number of samples were achieved. This ensured statistically meaningful features for downstream tasks.
5. **Knowledge Exchange for Refining Extracted Data:**
   * A feedback mechanism was proposed where physicians validated extracted structured features.
   * Errors or sparse data issues in the structured output were iteratively corrected, integrating expert feedback into the data transformation process.
6. **Targeted Data Sources:**
   * Initially focused on History and Physical notes, with plans to expand to operating notes and discharge summaries for a broader scope of structured data.

**Impact and Utility of Transformation:**

* The structured data generated from unstructured text enables better integration with numerical EHR components for analysis.
* By transforming qualitative observations and textual narratives into structured numerical and categorical features, the system facilitates interoperability, statistical analysis, and predictive modeling.

**Tools and Techniques:**

* cTakes: Used for medical concept extraction and numeric data classification.
* UMLS and ICD-9: Ontologies employed to standardize and hierarchically organize medical concepts.

**Key Takeaways:**

* The paper emphasizes combining rule-based and machine learning approaches to extract structured data from medical narratives.
* The process enhances the usability of text-heavy EHRs by creating a seamless representation that integrates with existing structured data.

This focus on transforming unstructured text into structured data is foundational for improving healthcare workflows, enabling applications such as decision support systems and analytics.