

病状経過知識抽出のための テキストマイニング

阿部秀尚, 平野章二, 津本周作

島根大学医学部医学科医療情報学講座

[研究背景 (1)]

- 病院情報システムの普及
 - 蓄積される文書の増大
- EBMの実現
 - 科学的根拠に基づく質の高い医療の提供
 - 論文検索などの活用
- テキストを効率よく利用するための手法が必要

[研究背景 (2)]

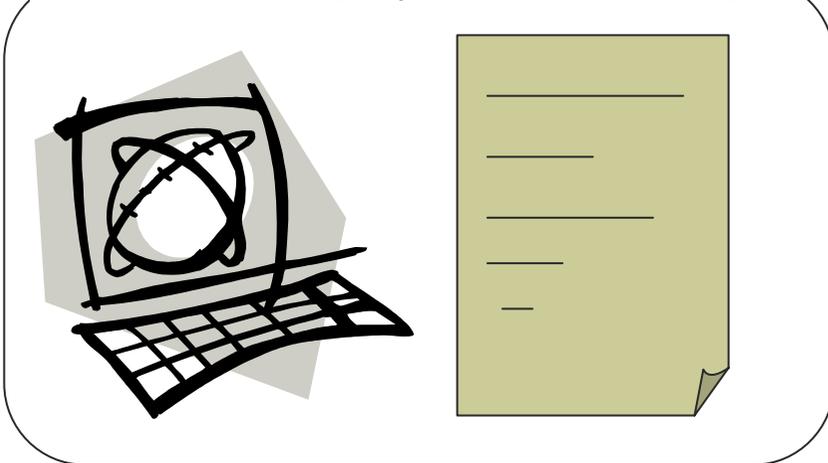
医師, 実習生



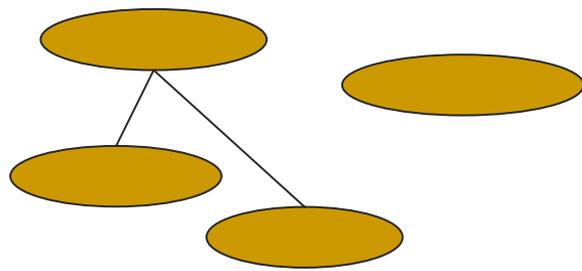
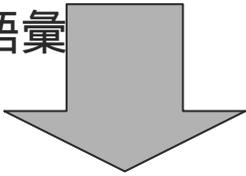
記憶やキーワード
検索

診療に関する
根拠

オンライン文書, 蓄積された文書

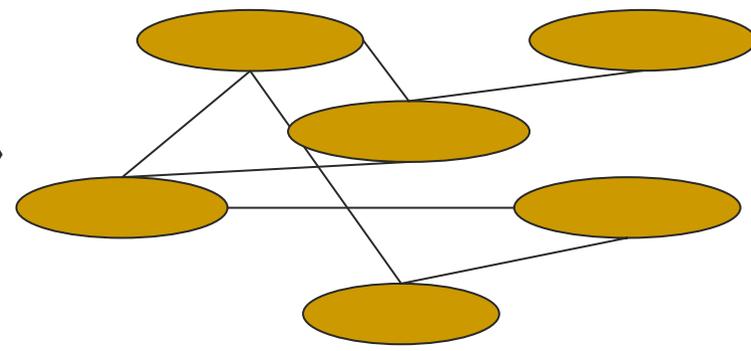
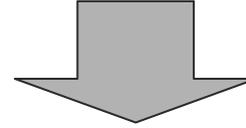


注目する語彙
疑問



比較・検索

語彙間の関係を
抽出



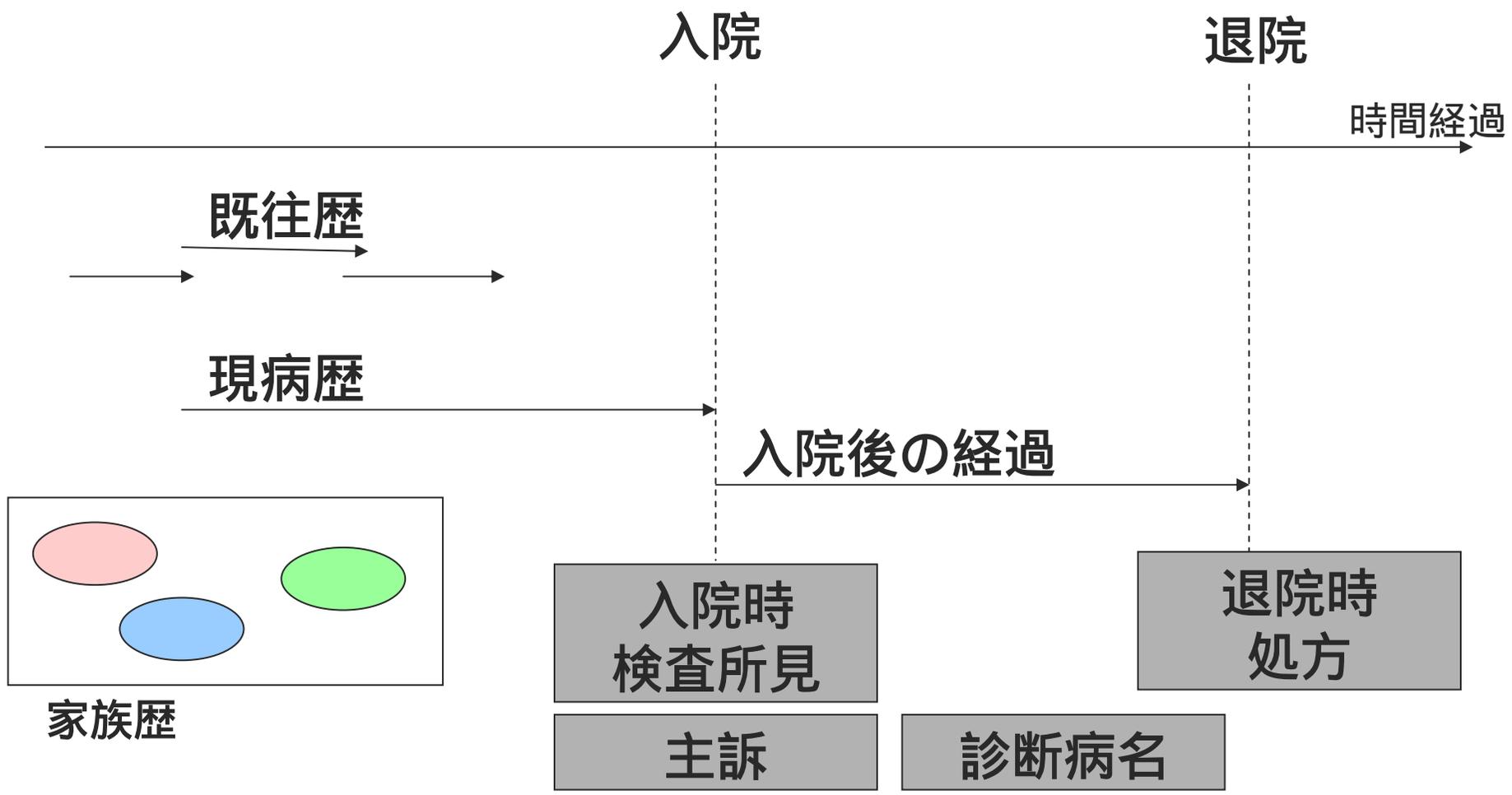
退院時サマリーからの病状経過 知識抽出

- 対象：退院時サマリー
 - 患者の退院時に入院中の経過などが自然言語でまとめられた文書
 - 時系列に沿って整理された要約
 - 経過に関して正確なタイムスタンプは記述されていない
- 出力：特徴的な語彙の関係
 - 疾患の進行に関する語彙の関係
 - 診療や検査に関する語彙の関係

退院時サマリー

- 整備された退院時サマリー
 - 診断病名: 診断された病名
 - 主訴: 入院の原因となった主な症状・病名
 - 家族歴: 近親の血縁者の疾病に関する記述
 - 既往歴: これまでに患った疾病に関する記述
 - 現病歴: 主訴について入院に至る経緯
 - 入院時検査所見
 - 入院後経過
 - 退院時処方
- 整備されていない退院時サマリー

[退院時サマリーの項目と時間経過]



[退院時サマリーの概要]

	内容の文体	
	不完全な文	完全な文
診断		—
主訴		—
家族歴		—
既往歴		—
現病歴・現症		
入院時検査所見		
入院後の経過		
退院時処方		—

[退院時サマリーの例]

【診断】

1 . # 2 . # 3 .

【主訴】

頭痛, めまい

【家族歴】

祖父母: 高血圧, 実父: 糖尿病

【既往歴】

20歳 虫垂炎

【現病歴】

昨夜, 9時過ぎテレビを見ていたところ.....

【入院時検査所見】

血液検査結果: ALB x.x, ALP x.x, ... 画像所見: 頭部CT...

【入院後の経過】

入院後, 3日間は を投与し, 経過を観察. の症状は徐々に改善した.

【退院時処方】

100mg, 60mg

退院時サマリーからの特徴的な語彙の関係の抽出実験

■ 対象データ

- 島根大学医学部附属病院 第三内科から提供された退院時サマリーの文書10症例分

■ テキストマイニング手法

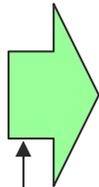
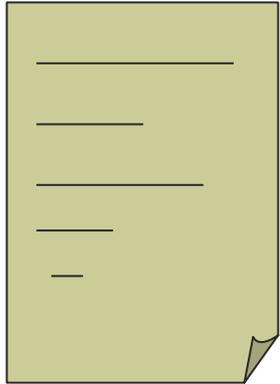
- 前処理: 医学用語の同定
- マイニング: 相関ルールによる語彙の共起関係を抽出
- 結果の後処理: 時間順序を反映したルールの選別

[データの前処理での課題]

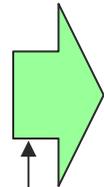
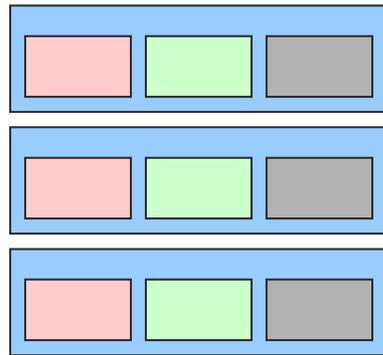
- 正しい文法で記述されていない文が存在
 - 構文解析から有効な情報の取出しが困難
- 特殊な用語(医学用語)の同定
 - 形態素解析と医学用語辞典の連携
- 対象とする語彙の品詞
 - 医学用語として用いられる名詞
 - 行為に関する名詞(名詞 + する)

[データの前処理手順]

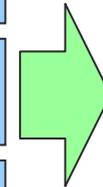
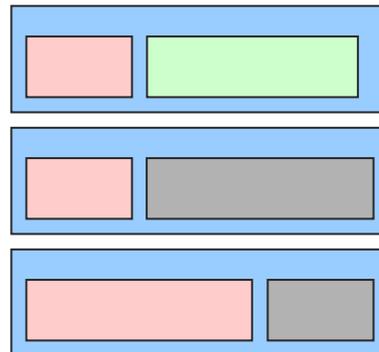
退院時サマリー



形態素解析結果



医学用語の同定

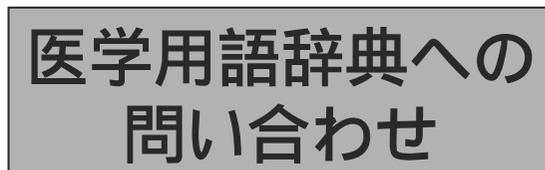


相関ルール抽出
のためのデータ

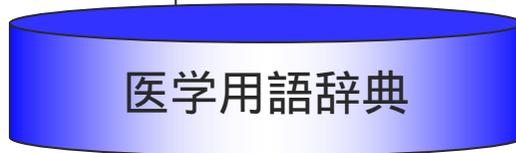
茶釜



医学用語辞典への
問い合わせ



医学用語辞典



1. 名詞句の取り出し
2. 長さ優先による医学用語の同定
3. 最右語の削除
4. 名詞句が残っている場合は2.へ

実験に使用する退院時サマリーの 前処理結果

症例No.	退院時	インスタンス数	語彙数
1	不変	5	14
2	不変	15	55
3	その他	10	44
4	その他	25	43
5	その他	12	47
6	軽快	17	55
7	その他	60	99
8	不変	12	50
9	軽快	6	23
10	軽快	7	31

相関ルールに基づく語彙間の関係抽出

- 各症例の「入院後の経過」についての文章
- 相関ルール抽出アルゴリズム: Apriori
 - 最小支持度: $\min_sup(D_i) \geq 2 / |D_i|$
 - 最小確信率: $\min_conf(D_i) \geq 0.6$
- ルールの後処理
 - $P(Y | X) / P(X | Y) \geq 1$ となるルール $X \Rightarrow Y$ を選択

[抽出されたルール]

症例6:

IF CT THEN 腹部大動脈瘤 (support(2/17), confidence(1.0))

IF 腹部大動脈瘤 THEN CT (support(2/17), confidence(0.67))

CTにより腹部大動脈瘤を検査した行為と一致

症例7:

IF 下痢 THEN WBC (support(2/60), confidence(1.0))

IF WBC THEN 下痢 (support(2/60), confidence(0.6))

下痢のためWBC(白血球数)の値を確認した行為と一致

考察

- 語彙の共起関係に基づく関係抽出
 - ルール対による語彙の時間的前後関係の同定
 - 時間間隔が不定な経過を表す語彙間の関係の抽出
- 入院時検査所見の利用
 - 検査項目を属性, 検査結果を属性値として利用
- 重要語彙の同定
 - 語彙の出現頻度による重要語彙の選択
 - 程度を表す語彙の活用

【まとめ・今後の課題】

- 退院時サマリーからの病状経過知識の抽出
 - 退院時サマリーでの医学用語の同定
 - 相関ルールを用いた特徴的な語彙間の関係抽出
- 退院時サマリー全体でのデータセットの構築
- 同様の症例, 経過の文書からの語彙間の関係抽出
- オンライン文書での語彙間の関係と退院時サマリーからの語彙間の関係の比較