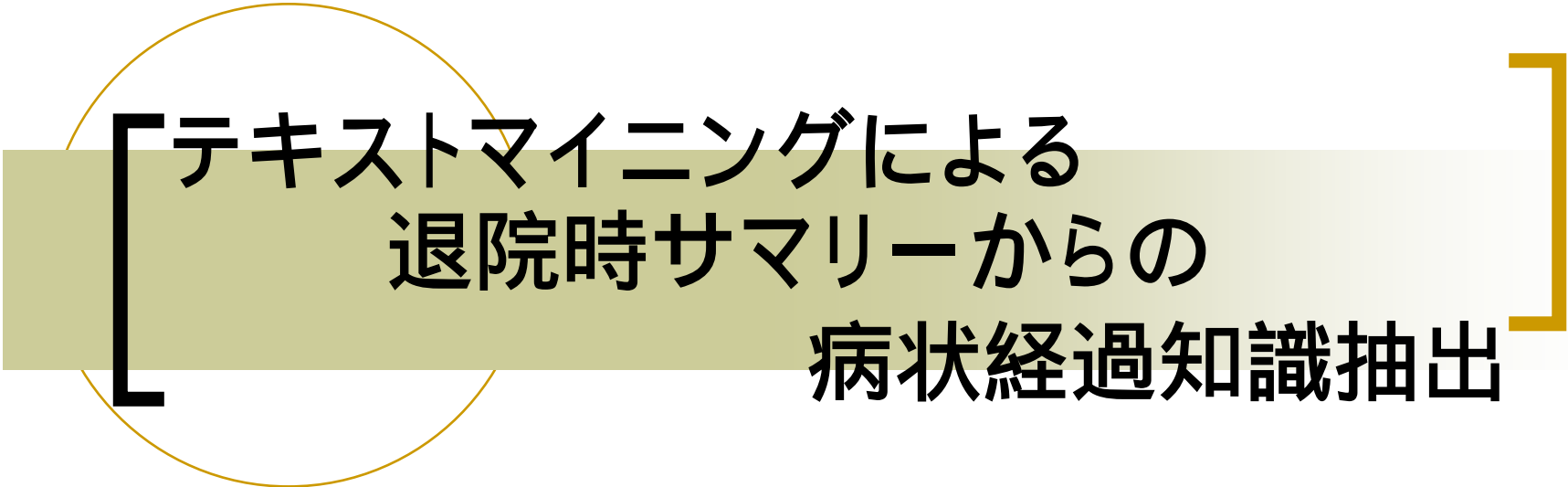




テキストマイニングによる 退院時サマリーからの 病状経過知識抽出

津本周作, 阿部秀尚, 平野章二

島根大学医学部医学科医療情報学講座

[研究背景]

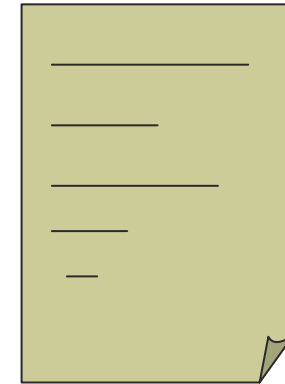
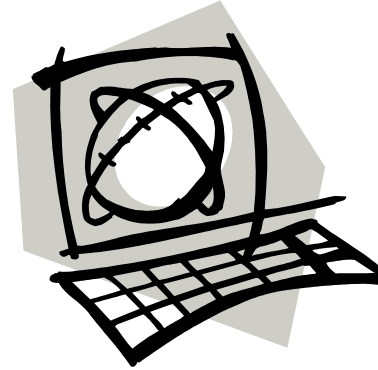
医師，実習生



記憶やキーワード
検索

診療に関する
根拠など

オンライン文書，蓄積された文書



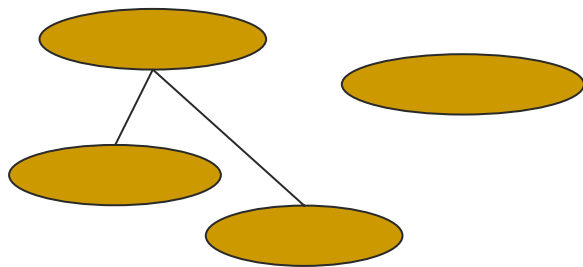
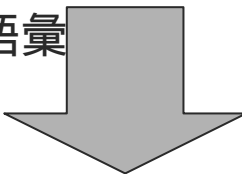
- オンラインや病院情報システムに蓄積されたテキストを効率よく利用するための手法が必要

[研究背景]

医師, 実習生

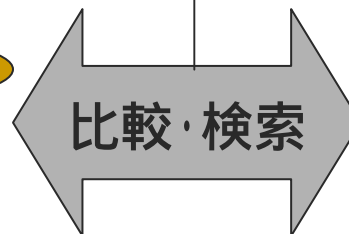
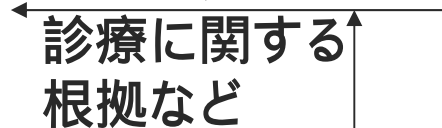


注目する語彙
疑問

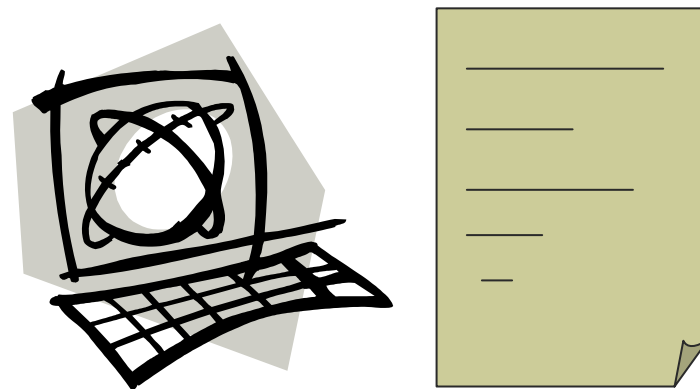


記憶やキーワード
検索

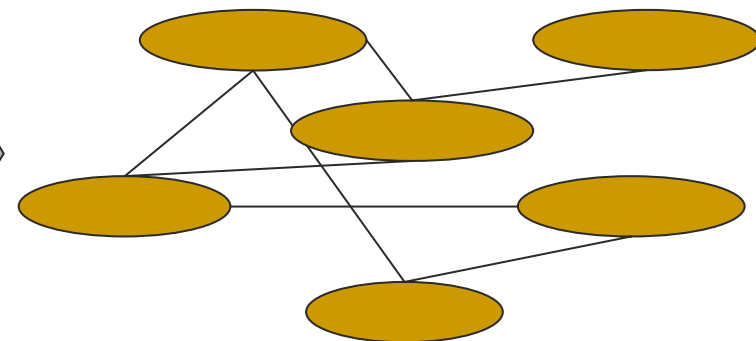
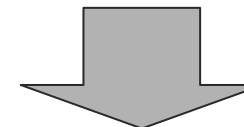
より適切な
診療に関する
根拠など



オンライン文書, 蓄積された文書



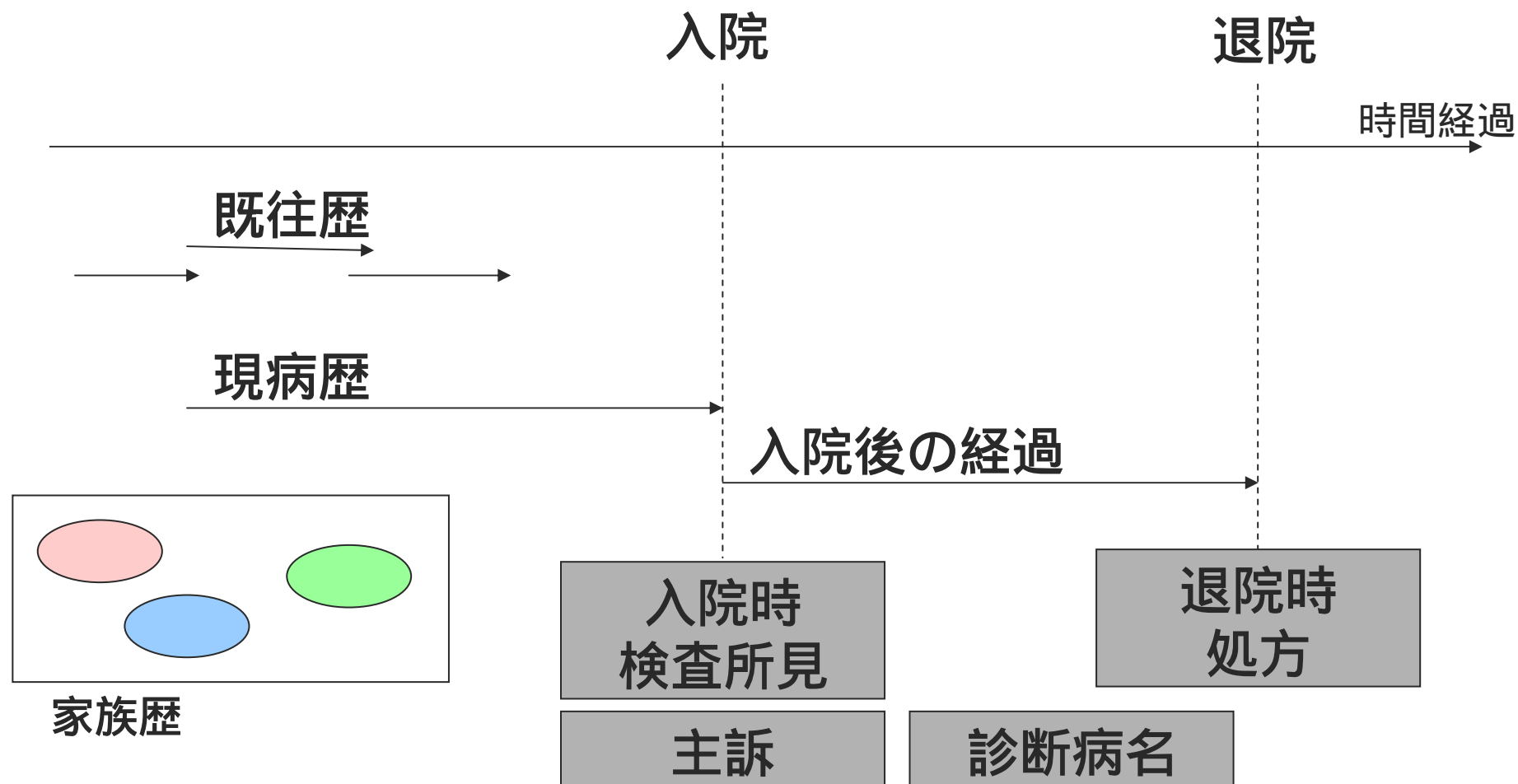
語彙間の関係を
抽出



退院時サマリーからの 病状経過知識抽出

- 入力：整備された退院時サマリー
 - 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

整備された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 神経内科から提供された退院時サマリー文書100症例分

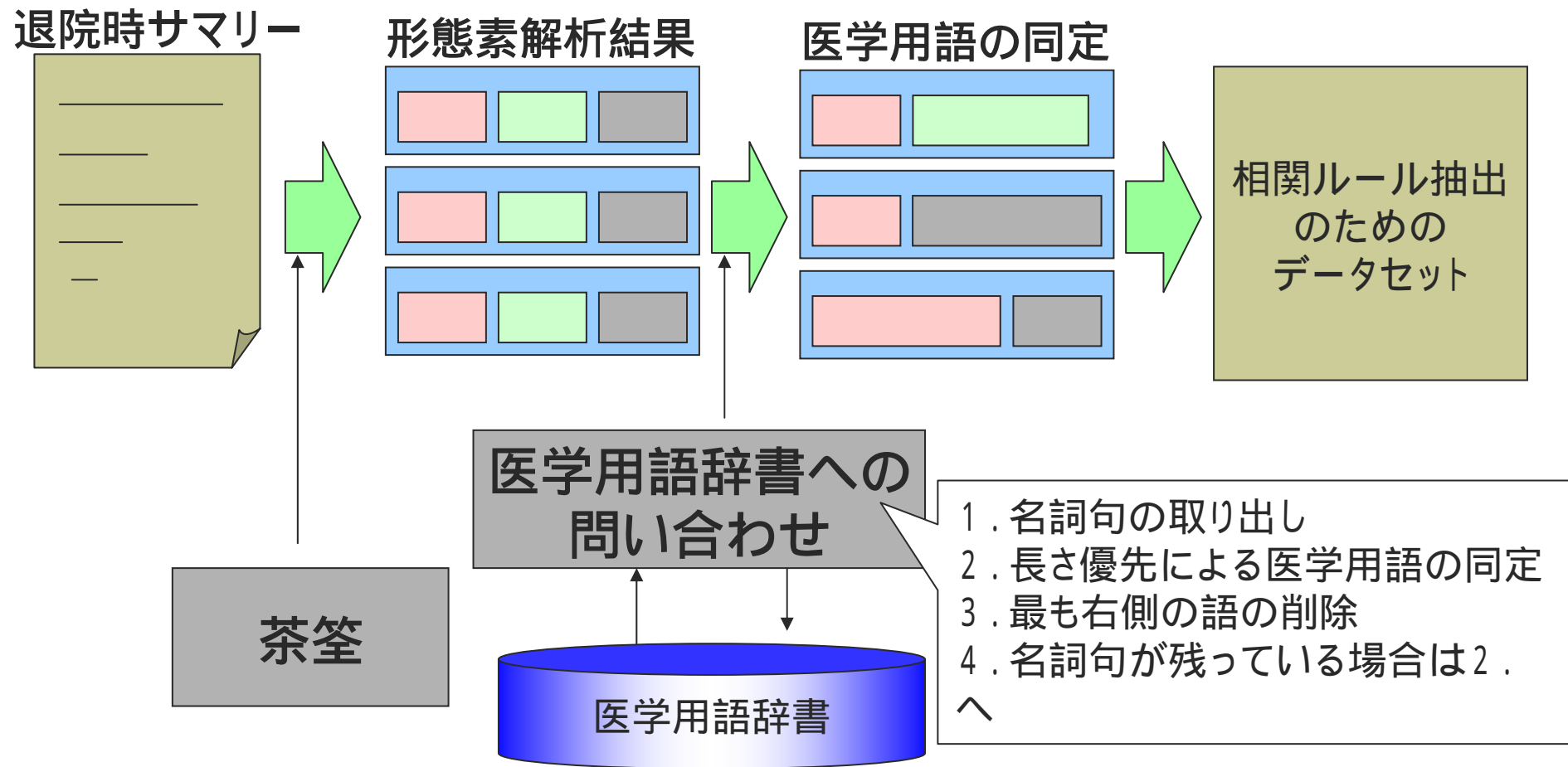
■ テキストマイニング手法

- 前処理: 医学用語の同定
- マイニング: 相関ルールによる各項目文章からの語彙の共起関係を抽出
- 結果の後処理: 出現頻度の差異によるルール選択 (各項目内での特徴的な語彙関係の抽出)

[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により構文解析が不可能になることを回避
 - 自然言語処理は形態素解析のみ
 - 対象語に対する否定表現の抽出は困難
- 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により同定
- 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]



予備実験： 時間経過に伴う使用語彙の変化

現病歴		入院後の経過	
語彙	頻度	語彙	頻度
出現	63	改善	86
入院	59	開始	37
受診	49	症状	30
診断	45	入院時	29
症状	36	退院	28
MRI	30	施行	26
改善	28	診断	23
めまい	24	入院後	23
頭痛	22	投与	22
自覚	20	脳梗塞	20
歩行	17	可能性	18
紹介	17	中止	18
異常	16	軽快	17
脳梗塞	16	内服	16
指摘	14	加療	16
CT	12	上昇	15
悪化	12	異常	15
点滴	11	予定	14
施行	11	MRI	14
軽快	11	経過観察	13
		所見	13

■ 「現病歴」「入院後の経過」で各語彙を含む文の数(上位20位まで)

■ 項目間の時間経過に伴う変化

- 使用される語彙
- 使用される頻度

相関ルールに基づく語彙間の関係抽出

- 入力: 各症例の「現病歴」「入院後の経過」についての文章を1データセット D_i 文を1データとする
 - 現病歴: 平均9.45文 (平均135.7語 / データセット)
 - 入院後の経過: 平均11.75文 (平均184.1語 / データセット)
- 相関ルール抽出アルゴリズム: Apriori
 - 最小支持度: $\min_support(D_i) \geq 2 / |D_i|$
 - 最小確信度: $\min_confidence(D_i) \geq 0.5$
- ルールの後処理
 - $P(Y | X) / P(X | Y) \geq 1$ となるルール対 $X \Rightarrow Y, Y \Rightarrow X$ を選択
 - 現病歴: 41データセットにルール対が存在
 - 入院後の経過: 45データセットにルール対が存在

[抽出されたルール対(例)]

特殊な関係にある語彙

[現病歴] 症例: 23

文章中で多く用いられる語彙

IF 起床 and 憎悪 and 構音障害 THEN 出現 確信度:(1.0)

IF 出現 THEN 起床 and 憎悪 and 構音障害 確信度:(0.5)

→ 出現したのは起床時のある症状の憎悪と構音障害だけではなかった

[入院後の経過] 症例: 23

IF 危険因子 and 糖尿病 THEN 脳梗塞 確信度:(1.0)

IF 脳梗塞 THEN 危険因子 and 糖尿病 確信度:(0.5)

→ サマリーを書いた医師は脳梗塞の危険因子として糖尿病に注目していた

考察

- 使用語彙の時間的变化
 - 項目間の時間的経過に伴って使用される語彙が変化
- 語彙の共起関係に基づく関係抽出
 - ルール対による特徴的な語彙関係の抽出
- 項目の時間情報を基にした経過的語彙関係の抽出
- 今回未使用の項目のデータ化
 - 診断, 主訴, 家族歴, 転帰などの利用
 - 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]

- 退院時サマリーに対するテキストマイニングの適用
 - 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
 - サマリーに記述された医師が注目した語彙間の関係を抽出
- 今後の課題
 - 退院時サマリー全体でのデータセットの構築
 - 同様の診断, 転帰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