# 医療情報処理分野における自然言語処理

病院で自然言語処理？

ゲノム？

東京大学 医学部附属病院
荒牧英治

# 背景

- 最近よくある論文の導入

**インターネット**の急速な発展に伴い，利用可能なテキストの量が増加して．．．

爆発的に

- 医療分野では．．．

**電子カルテ**の急速な発展に伴い，利用可能なテキストの量が増加して．．．

# 目的と問題

- **目的：カルテの自動解析**
  - より大規模な統計的研究ができる
  - （例）喫煙と発癌率の相関関係は？
- **BUT：そのためには**
  - （1）カルテ中に含まれる個人情報の削除
  - （2）自然言語で書かれたカルテを処理する問題
    - （例）喫煙しているか否かをカルテ文章から分類できるかどうか？

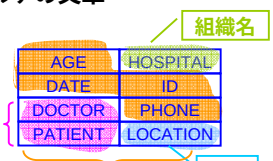
# 概要

- はじめに
- Challenge 1: 個人情報の匿名化
- Challenge 2: Smoking Challenge
- まとめ

AMIA-i2b2 Shared Task

# 個人情報(Personal Heath Information)の匿名化

- タスク
  - 入力：カルテの文章
  - 出力：PHIを削除したカルテの文章
- PHIとは
  - HIPAA ガイドライン
    (Health Information Portability and Accountability Act)




固有表現抽出（NER）

# 先行研究

赤色＝PHI

- 辞書＋規則ベース
  [Douglass2005]
  - 地名辞書・人名辞書
  - 規則「DR. XXX」→ XXX＝DOCTOR

| | |
|---|---|
| Precision | 88.9% |
| Recall | 67.6% |
| F（β=1） | 76.8 |

- 機械学習ベース
  [Sibanda HLTNAACL2006]
  - SVMで単語の近傍（前後2語）を素性

| | |
|---|---|
| Precision | 97.4% |
| Recall | 95.0% |
| F（β=1） | 96.2 |

070203832
DH 8446543;
4/2/2003 12:00:00 AM
ED DISCHARGE NOTIFICATION GUYNLUDZ , STASIERDI
MRN :8446543
REGISTRATION DATE : 04/02/2003 07:18 AM
PRELIMINARY REPORT
This is to notify you that your patient , GUYNLUDZ , STASIERDI arrived in the Emergency Department at Daughtersvillshamen's Hospital on 04/02/2003 07:18 AM .If you need additional information please call 613-870-0699 .
RCP Name : FYFE , PRERICK A
Provider Number : 06880

“Ungrammatical & fragmented”

# 提案手法

- Globalな情報を取り込む
  - (1) Non-local Features
    - Target Word (=current word) の文章中での位置
    - 周辺の文の語／語数

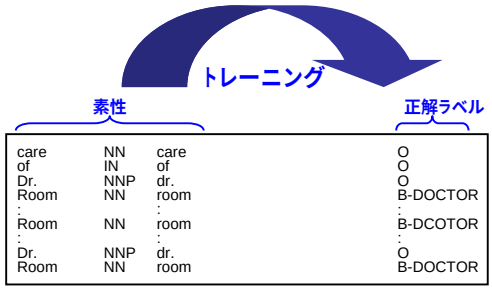
文章の最初／最後にPH!頻発

同じタイプのPH!が連続

213763231
CMC
07646518
10/3 /2002 12:00:00 AM ACUTE MYELOGENOUS LEUKEMIA
Date : 10/03/2002
Discharge : 11/03/2002
Report Status : Signed
ADDENDUM TO DISCHARGE SUMMARY : Please see prior admission summary for details of entire stay except for 10/31/02 to 11/03 /02 . Patient was admitted to Oncology B service .
ADMISSION DIAGNOSIS : ACUTE MYELOGENOUS LEUKEMIA .
DISCHARGE DIAGNOSIS : ACUTE MYELOGENOUS LEUKEMIA . :

# 2-Stage CRF [Krishnan2006]

- 1st CRF: 通常のラベリング




| | | | |
|---|---|---|---|
| care | NN | care | O |
| of | IN | of | O |
| Dr. | NNP | dr. | O |
| Room | NN | room | B-DOCTOR |
| : | : | : | : |
| Room | NN | room | B-DCOTOR |
| : | : | : | : |
| Dr. | NNP | dr. | O |
| Room | NN | room | B-DOCTOR |

# 2-Stage CRF [Krishnan2006]

- 1st CRF: 通常のラベリング

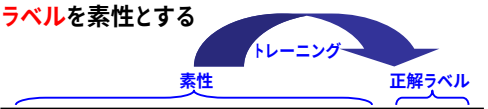

3回出現したROOMのうち
2回はDOCTOR
1回はPATIENT

| | | | |
|---|---|---|---|
| care | NN | care | O |
| of | IN | of | O |
| Dr. | NNP | dr. | O |
| Room | NN | room | B-DOCTOR |
| : | : | : | : |
| Room | NN | room | B-PATIENT |
| : | : | : | : |
| Dr. | NNP | dr. | O |
| Room | NN | room | B-DOCTOR |

ラベルの一貫性が保たれない

# 2-Stage CRF [Krishnan2006]

- 1st CRF: 通常のラベリング
- 2nd CRF: 1st CRFの結果, tokenが最もとりやすいラベルを素性とする




| | | | | |
|---|---|---|---|---|
| care | NN | care | O | O |
| of | IN | of | O | O |
| Dr. | NNP | dr. | O | O |
| Room | NN | room | DOCTOR | B-DOCTOR |
| : | : | : | : | : |
| Room | NN | room | DCOTOR | B-DCOTOR |
| : | : | : | : | : |
| Dr. | NNP | dr. | O | O |
| Room | NN | room | DOCTOR | B-DOCTOR |

# 実験

- コーパス: I2b2-shared task
  - 671 records & 14,309 PHIs
  - 10-fold cross-validation
- 結果:

| | Precision | Recall | F (β=1) |
|---|---|---|---|
| [Sibanda2006] | 97.4% | 95.0% | 96.2 |
| 1st CRF (BASELINE) | 98.0% | 95.2% | 96.5 |
| 1st CRF + Non-local | **98.4%** | 95.2% | 96.8 |
| 1st CRF + 2nd CRF | 97.4% | 95.8% | 96.5 |
| PROPOSED | 98.3% | **96.6%** | **97.5** |

+0.3 point

+1.3 point

# 実験

- コーパス: I2b2-shared task
  - 671 records & 14,309 PHIs
  - 10-fold cross-validation
- 結果:

| | Precision | Recall | F (β=1) |
|---|---|---|---|
| [Sibanda2006] | 97.4% | 95.0% | 96.2 |
| 1st CRF (BASELINE) | 98.0% | 95.2% | 96.5 |
| 1st CRF + Non-local | **98.4%** | 95.2% | 96.8 |
| 1st CRF + 2nd CRF | 97.4% | 95.8% | 96.5 |
| PROPOSED | 98.3% | **96.6%** | **97.5** |

+0.3 point

# 実験

- **コーパス**: I2b2-shared task
  - 671 records & 14,309 PHIs
  - 10-fold cross-validation
- **結果**:

| | Precision | Recall | F (β=1) |
|---|---|---|---|
| [Sibanda2006] | 97.4% | 95.0% | 96.2 |
| 1st CRF (BASELINE) | 98.0% | 95.2% | 96.5 |
| 1st CRF + Non-local | **98.4%** | 95.2% | 96.8 |
| 1st CRF + 2nd CRF | 97.4% | 95.8% | 96.5 |
| PROPOSED | 98.3% | **96.6%** | **97.5** |

+0.0 point

# PHIタイプ別の精度

- **先ほどの精度**＝PHI/nonPHI**の判定精度**
- PHI**タイプ別制度では**2nd CRF**の貢献が見られる**

| | 1st CRF | | | 1st CRF + 2nd CRF | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F (β=1) | Precision | Recall | F (β=1) |
| AGE | 33.3% | 7.69% | 12.5 | 33.3% | 7.69% | 12.5 |
| DATE | 98.3% | 94.5% | 96.4 | 98.0% | 95.1% | 96.5 |
| DOCTOR | 93.7% | 90.8% | 92.2 | 93.0% | 91.1% | 92.1 |
| HOSPITAL | 94.0% | 88.4% | 91.1 | 93.0% | 91.1% | 92.0 |
| ID | 96.8% | 98.2% | 97.5 | 96.8% | 98.2% | 97.5 |
| LOCATION | 69.1% | 45.1% | 54.6 | 54.6% | 45.1% | 49.4 |
| PATIENT | 84.8% | 83.6% | 84.2 | 84.0% | 84.8% | 84.4 |
| PHONE | 97.0% | 93.1% | 95.0 | 97.0% | 93.1% | 95.0 |
| ALL | 95.6% | 92.9% | 94.2 | 95.1% | 93.5% | 94.3 |

効果あり／効果なし

# 考察とまとめ

- **考察**
  - (**これまでカルテ匿名化では**GLOBAL**な情報は無益だとしてきた**)
  - BUT: **いくつかは貢献することが判った**
  - **人間の精度** (precision 98%; recall 95%) **＜提案手法**
    - → **これまで人間数人のユニオンをとっていたが，その中の一人として参加可能**
- **今後の課題**
  - GLOBAL**な素性をより多く／スマートに取り込みたい**
  - **例えば**: One person per record**の原則**
    - 「1 record **には** 1 **患者しかいない**」

BUT: **患者は人名**「PATIENT」**になることもあれば患者**「ID」**となることもある**

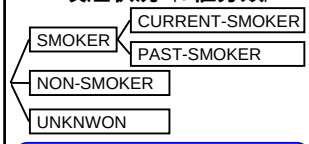
# 概要

- **はじめに**
- Challenge1: **個人情報の匿名化**
- Challenge2: Smoking Challenge
- **まとめ**

**喫煙履歴の自動分類**

# 喫煙履歴の自動分類

- **タスク**
  - **入力**: **カルテの文章**
  - **出力**: **患者の喫煙状況**
- **喫煙状況（5値分類）**

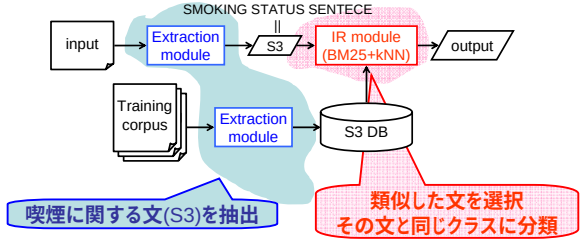



**問題**: **喫煙に関する文章はわずか数文**（**多くの場合，一文のみ**）

071962960 BH 4236518 417454
12/10/2001 12:00:00 AM
:
HISTORY OF INTRAVENOUS DRUG ABUSE ( HEROIN ) . PATIENT DENIES CURRENT USE . PATIENT REPORTS OCCASIONAL ALCOHOL USE . **HAS BEEN SMOKING APPROXIMATELY 10 CIGARETTES A DAY** . CLAIMS TO HAVE STOPPED A ; FEW WEEKS AGO ; . MARRIED WITH TWO CHILDREN .
:

# 提案手法

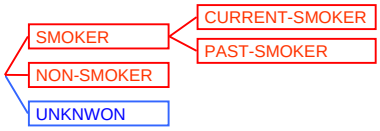
- **先行研究**: **なし!?**
- **提案手法のフレームワーク**



# Extraction モジュール

- キーワードを含む文を喫煙に関する文とみなす

キーワード={smoke, smoking, smoker, nicotine, tobacco, cigarette}




- キーワードを含むがなければ UNKNOWN とする

単純な方法だが，喫煙ドメインに関してはclear-cutできる

喫煙に関する文が抽出される割合

| | |
|---|---|
| S,N,C,P | 98.6% (=144/146) |
| UNKNOWN | 1.1% (=3/252) |

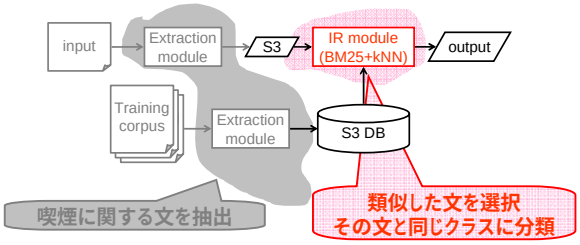
# 予備実験

- 抽出された喫煙に関する文

| | |
|---|---|
| NON-SMOKER | She does not **smoke tobacco** . |
| | The patient does not **smoke** . |
| SMOKERS | PAST MEDICAL HISTORY is remarkable for chronic lung disease due to **smoking .** |
| | 11. history of **cigarette smoking** , |
| PAST-SMOKER | He is not a current **smoker** . |
| | She quit **smoking** nine years ago . |
| CURRENT-SMOKER | **Please attempt to quit smoking** |
| | **Smokes** one pack per day x 40 years . |

# 提案手法

- 先行研究: なし!?
- 提案手法のフレームワーク




# BM25 [Robertson1995]+kNN[Cover1967]

- BM25で入力文とトレーニングセットの喫煙に関する文の類似度を計算

入力文の喫煙に関する文：He does not smoke .

| BM25類似度 | 文のクラス | |
|---|---|---|
| 480 | [NON-SMOKER] | The patient does not **smoke** . |
| 312 | [NON-SMOKER] | She does not **smoke tobacco** . |
| 252 | [PAST-SMOKER] | She quit **smoking** nine years ago . |
| : | : | : |

- 類似度の高い上位k個の重み付き和でクラスを決定

480＋312 [NON-SMOKER] > 252 [PAST-SMOKER]

# 結果＆考察

- 466 recordを用いて実験 (2-fold cross-validation)

| | |
|---|---|
| BASELINE 1 | 77.9% |
| BASELINE 2 | 86.0% |
| PROPOSED (k=1) | 81.6% |
| : | |
| **PROPOSED (k=10)** | **88.9%** |
| : | |
| PROPOSED (k=20) | 86.7% |

← S3 が抽出された場合すべてNとする

← BM25ではなく編集距離を用いる

今回の提案手法自体がBASELINEのようなもの

- 考察
  - BM25+KNNというBOWな手法で88.9%
  - これ以上の精度を求めるためには，より深い処理を行う必要
  - BUT: まともな文が少ないので構文解析など困難
  - リーズナブルな解決法を模索している段階

# 概要

- はじめに
- Challenge1: 個人情報の匿名化
- Challenge2: Smoking Challenge
- まとめ

## まとめ

- 医療分野における自然言語処理
  - 医療オントロジー
  - 個人情報の削除
  - カルテの自動分類

  } ∈ NLPタスク

- BUT: 研究者が少ない(日本では数人)

**医療分野はNLP研究者を**
**広く(切実に)募集しております**