非文法的かつ断片化されたテキストからの頑健な情報抽出

荒牧英治† 今井健† 美代賢吾† 大江和彦†
†東京大学医学部附属病院

{aramaki, ken, kohe}@hcc.h.u-tokyo.ac.jp
miyo-tky@h.u-tokyo.ac.jp

1 はじめに

近年，電子カルテの普及により，大量の臨床データが収集されつつある．このデータを利用できれば，過去類をみない大規模な統計的研究が実現可能であり，大きな期待がよせられている．しかし，カルテ中の一部の情報は自然言語で記述されており，データをフルに利用するためには，テキストからの情報抽出技術が必須となる．

このような状況から，本研究はカルテの一種である退院サマリ（退院時に記述される患者の治療や経過をまとめた文書）を対象とし，そこから患者の喫煙状態を抽出するタスクに挑戦する．図 1 に退院サマリの例を示す．この例では，下線部の表現 “She does not smoke tobacco .” から，この患者が非喫煙者であることがわかる．

本タスクは，文章から情報を抽出するという観点からは情報抽出の一種である．しかし，カルテでは一文章が一患者に対応しているため，患者の喫煙状態の抽出はカルテ文章の分類タスクと考えることもできる．そこで，本研究では，文章分類と同様に類似度を用いたアプローチを採用した．すなわち，まず，入力文章とトレーニングセットから喫煙に関する文を抽出する．次に，それらの類似度を計算し，もっとも類似した喫煙状態へ分類するという二段階のアプローチをとる．

従来の情報抽出とのもう一つの差異は，対象となる文章の性質である．従来この分野では，論文 [9] やニュース [10, 7] など比較的フォーマルな文章が扱われてきた．一方，カルテ文章は，文献 [8] が指摘したように，文というよりも，短い名詞句の連続という形で記述されることが多く，さらには，誤字など非文法的な表現がしばしば含まれる．

このような非文法的かつ分断化されたテキストを扱う際には，構文解析など深い処理が有効でない場合も多い．そこで，我々は構文解析を用いる文の類似度と，表層的な語順の類似度の両方を併用することを考える．

```
PAST MEDICAL HISTORY :
Cryptogenic cirrhosis with an unclear work up .
Diverticular bleed requiring colostomy on May 96 .
Spinal stenosis .
Hysterectomy at the age of 32 .
No coronary disease , no diabetes and no hypertension .
She had medications on transfer which included Synthroid 0.15 mg
 QD , Zoloft 50 mg QD , Inderal 10 mg BID , Prilosec 20 mg QD .
        ⋮
SOCIAL HISTORY :
She does not smoke tobacco .
She uses only occasional alcohol and she is not sexually active .
Her husband died of lung cancer .
She lives alone in Burg Chi Sternafre , Massachusetts .
Her friends check up on her .
She does not have a home health aide .
She has 3 kids who live far away .
        ⋮
```

図 1: 退院サマリの例.
* 下線部は喫煙関連文 (3.1 節で述べる) を示す.

次に問題となるのは，どのような場合に構文解析が有効となり，どのような場合に有効でないのか，これらを判別する手がかりは何かということである．この問題は，次のようなトレードオフを抱えている．素朴には，文が長い（語数が多い）場合には，構文解析することで，文法的に意味のある形として文を取り扱いたい．しかし，文が長ければ長いほど構文解析が失敗してしまう可能性も増える．

そこで，提案システムは，文長や各尺度の確信度に加え，喫煙状況を左右する重要な情報となる語群をとらえ，それらの距離を手がかりとして，最適な尺度を判別することを試みる．

実験の結果，個々の尺度のみを用いた場合より高い精度（88.9%）を得たので報告する．

2　コーパス

本研究にあたっては，i2b2-NLP shared-task*で配布されたコーパスを用いた．このコーパスは，398 文章の英語の退院サマリ文章（以下，**文章**）からなる（1 文章のあたりの平均文数は 86.9 文,1 文の平均語数は 8.85 語）．

これらの文章には文書単位で患者の喫煙状態（以下，**喫煙状態**）が次の 5 つに区別されアノテートされている．

C (Current Smoker): 現在，喫煙している場合.
P (Past Smoker): 過去に喫煙履歴があった場合.
S (Smoker): 現在または過去かは曖昧であるが，喫煙していた／いることが明らかな場合.
N (Non-Smoker) : 現在喫煙しておらず，かつ，過去にも喫煙していない場合.
U (Unknown): カルテから喫煙履歴が判別できない場合.

3　提案手法

提案手法は，3 つの処理からなる．まず，最初にトレーニングセットと入力文章から喫煙関連文を抽出する．次に，3 つの尺度で，入力文とすべてのトレーニングセットの喫煙関連文の類似度を計算する．最後に，それぞれの手法の確信度と入力文の統計量により，適した尺度を選択する．

3.1　喫煙関連文抽出

まず最初に，コーパスから，喫煙と関連した文 (**喫煙関連文**) を抽出する．これには，キーワード “*nicotine, smoker, smoke, tobacco, cigarette*” を用意し，それらのうち一つでも含む文を喫煙関連文とみなした．

また，1 つの文章から複数の喫煙関連文が抽出できる場合には，最後に出現したものをその文章の喫煙関連文とみなした．また，入力文章から喫煙関連文が抽出されない場合は，その文章をただちに「UNKNOWN」へと分類した．詳細については文献 [1] を参照されたい．

3.2　文の類似度を計る尺度

次に文の類似度を計る以下の 3 つの尺度について述べる．

3.2.1　ED: 編集距離を用いた類似度

類似度 ED は編集距離 [4] を以下のように文長で正規化して用いた：

$$sim_{ED}(S_i, S_t) = \frac{\text{編集距離 } (S_i, S_t)}{|S_i| + |S_t|}. \quad (1)$$

*https://www.i2b2.org/NLP/

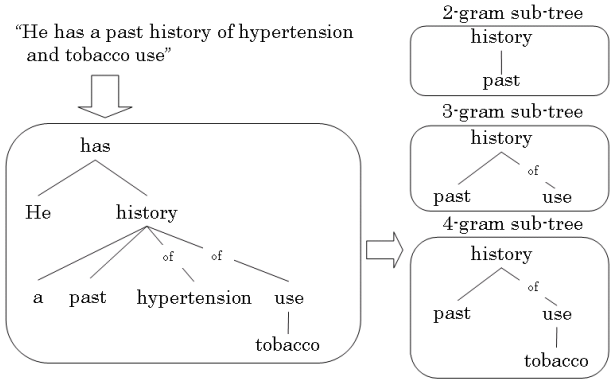

図 2: 構文解析結果とそこから得られる部分木の例.

ただし，S_i は入力文章の喫煙関連文，S_t はトレーニングセットの喫煙関連文，$|S_s|$ は S_S の文字数，$|S_t|$ は S_t の文字数とする．

最終的な喫煙状態は類似した上位 k 個の喫煙関連文が属する喫煙状態を類似度で重み付き投票し，最も高い確信度を持つものとする．

3.2.2　NGRAM: n-gram ベースの類似度

類似度 NGRAM は，文を n-gram 単位に分解して計る．まず，喫煙関連文を n-gram ($n = 1..4$) までの語（列）に分解する．

次に，分解された語列間の類似度を Okapi-BM25[6] 尺度を用いて計算する．また，最終的な出力は，前述の ED と同じく上位 k 個の類似度の重み付き投票によって決定する．

3.2.3　TREE: 統語解析を用いた類似度

先の類似度 NGRAM は表層的な語順で文を n 語の組み合わせに分解したが，この類似度 TREE は依存構造上で文を n 語の組み合わせに分解する点が異なる．すなわち，文を n 語の部分木に分解して扱う．図 2 に依存構造から生成される部分木の例を示す．以降の処理は類似度 NGRAM と同様である．

3.3　セレクター

最後に，前節で述べた 3 つの尺度で出力された喫煙状態のうち，どれを採用するかを選択する．このセレクターについて述べる前に，セレクターが用いる素性の一つである**重要語ペアの距離**について述べる．

表 1: 喫煙関連文の例.

喫煙状態	喫煙関連文
CURRENT-SMOKER	Please attempt to quit **smoking** .
NON-SMOKER	The patient denied using tobacco . **smoking** .
PAST-SMOKER	**Nicotine** abuse , quit in the 80s , and rare alcohol .
NON-SMOKER	She does not drink alcohol or **smoke tobacco** or take drugs .
CURRENT-SMOKER	**Smoker** for greater than 100 pack years (3-1/2 packs per day x 35 years) .
SMOKER	PAST MEDICAL HISTORY is remarkable for chronic lung disease due to **smoking** .
CURRENT-SMOKER	HPI. 51F w h / o **tobacco** and crack use p / w SOB / DOE worsening over the past 3 weeks and breast soreness , with troponin 0.15 in ED .

3.3.1 重要語ペアの距離

喫煙状態に大きく影響する語が隣接または非常に近い距離にある場合は，構文解析を行う必要性は少ない．そこで，Okapi-BM25 尺度で，もっとも高い情報量を持つ 2 語ペア（2-gram）を**重要語ペア**とみなし，それらの距離をセレクターの素性として利用した．

入力文章から抽出された喫煙関連文が重要語ペアを含んでいた場合は，それらの間の距離（語数）を調べる．複数の重要語が含まれる場合は，より高い情報量を持つ重要語ペアの間の距離を採用するものとする．例えば，以下の文において，“past” と “tobacco” が重要ペアである場合，重要語ペアの距離は，5 となる．

> He has a **past** history of hypertension and **tobacco** use.

3.3.2 機械学習によるセレクター

尺度のセレクターは，決定木（C4.5[5]）を用いて各手法の確信度の値，入力文章の喫煙関連文の語数，重要語ペアの距離を素性とし，正解した尺度を学習した．

この際，どの尺度を用いても不正解となるデータは学習に用いなかった．また，複数の尺度が正解している場合は，ED ＞ TREE ＞ NGRAM という優先順位で学習に用いた†．すなわち，ED と TREE の両方が正解している場合は，ED を使うようにトレーニングを行った．

†予備実験の結果，この優先順位の精度がもっとも高かった．

4 実験

4.1 実験設定

実験には 2 章で述べた i2b2 コーパスを利用し，交差検定法 (5-fold) にて，以下の 4 つの手法を比較した．ただし，セレクターの学習がクローズにならないよう PROPOSED に関してはトレーニングセットの 25%をセレクターの学習用に用いた．

1. **BASELINE**: マジョリティベースライン．喫煙関連を抜き出した場合は，UNKWON 以外の最頻値である NON-SMOKER とする．抽出できなかった場合は，UNKNOWN とする．
2. **E**: ED だけを用いた手法.
3. **N**: NGRAM だけを用いた手法.
4. **T**: TREE だけを用いた手法.
5. **PROPOSED**: 提案手法．ED，NGRAM，TREE をセレクターにより切り替える手法.

ここで，E,N,T の各尺度はいずれも重み付き投票に用いる定数 k を持つが，この値は予備実験の結果，それぞれ $k_E = 4$; $k_N = 5$; $k_T = 8$ とした．また，TREE が用いる構文解析処理には Charniak の nlparser[2] を用いた‡.

4.2 結果

結果を表 2 に示す．また，3 つの尺度の正解／不正解の頻度を表 3 に示す.

表 2 が示すように，3 つの尺度を比較すると ED が最も精度が高く，多くの文章は文字単位の編集距離という単純な方法で解けることが分かる．ただし，表 3 が示すように，T のみが正解する場合も 10 件あり，場合によっては構文解析は貢献することが分かる．

‡nlparser は句構造を出力するため，これを文献 [3] の手法にて依存構造に変換した．

表 2: 各手法の精度.

手法	精度（正解数）
BASELINE	77.94% (310)
ED	87.18% (347)
NGRAM	85.67% (341)
TREE	83.41% (332)
PROPOSED	88.94% (354)

表 3: 各尺度の正解・不正解の頻度.

E	N	T	頻度
×	×	×	27
×	×	○	10
×	○	×	8
○	×	×	7
○	○	×	24
○	×	○	13
×	○	○	6
○	○	○	303

* E=ED, N=NGRAM, T=TREE. ○は正解，×は不正解を示す.

ここで，PROPOSED の精度は 3 つの尺度よりも高く，うまく各尺度のよいところを利用できていることが分かる．ただし，ED 単独で用いた場合の精度との差はわずかであり，統計的有意は得られなかった (p=0.05)．この原因の一つには，実験サンプル数が少ないことも挙げられ，今後より大規模な実験を課題としたい.

5 関連研究

最近の情報抽出研究では，構文解析した上でテキストを扱う手法に期待が寄せられている．例えば，[10, 9] は述語-項構造で情報抽出パタンを扱った．[7] は部分木構造としてパタンを扱った．このような深い手法は，今後ますます，構文解析技術が向上することを考えれば有望であろう.

しかし，いずれの先行研究も本タスクと比較してフォーマルな文を扱っている点で本研究とは異なる．本研究のように対象となるテキストが，非文法的で断片化されている場合は，本実験が示すように，現状の構文解析処理が割にあわないことも多い.

このため，i2b2-NLP ワークショップに参加したシステムにおいては，構文解析を行わない手法がすべてを占めた．これらの参加システムのうち追加コーパスを用いないで，もっともよい成績であったのは [1] であった．そこで，本研究では [1] を拡張し，構文解析を行う手法とのハイブリッドなアプローチを試みた.

6 まとめ

本研究は複数の文の類似尺度を併用して電子カルテの文章から患者の喫煙状態（喫煙／非喫煙／不明）の抽出を試みた．複数の尺度を切り替える手がかりとして，分類上重要となる語同士の距離を用いることにより，高い精度（88.9%）で喫煙状態の抽出に成功した．今後は，より大規模なデータを用いて，実証的に精度を検証するとともに，喫煙以外の情報抽出にも提案手法が適応可能かどうか検証することを課題としたい.

参考文献

[1] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Patient status classification by using rule based sentence extraction and bm25-knn based classifier, 2006.

[2] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL 2000)*, pp. 132–139, 2000.

[3] M. Collins. Head-driven statistical models for natural language parsing, 1999.

[4] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, Vol. 163, No. 4, pp. 845–848, 1965.

[5] J.R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, Vol. 27, No. 1, pp. 221–234, 1987.

[6] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text Retrieval Conference*, pp. 109–126, 1995.

[7] Kiyoshi Sudo, Satoshi Sekine, and Ralph Grishman. An improved extraction pattern representation model for automatic ie pattern acquisition. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2003)*, pp. 224–231, 2003.

[8] Sibanda Tawanda and Uzuner Ozlem. Role of local context in automatic deidentification of ungrammatical, fragmented text. In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2006)*, pp. 65–73, 2006.

[9] Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, and Jun'ichi Tsujii. Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine*, pp. 60–69, 2005.

[10] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised discovery of scenario-level patterns for information extraction. In *Proceedings of International Conference on Computational Linguistics (COLING2000)*, pp. 940–946, 2000.