非文法的かつ断片化されたテキストの頑健な分類

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　〒 113–8655 東京都文京区本郷 7-3-1
E-mail: †{aramaki,ken}@hcc.h.u-tokyo.ac.jp, ††miyo-sup@h.u-tokyo.ac.jp, †††kohe@hcc.u-tokyo.ac.jp

あらまし　電子カルテの普及により，カルテに記載された情報を利用した大規模な統計的研究の可能性に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，カルテ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．情報抽出分野では，最近の構文解析精度の向上により，文を依存構造に変換して扱う手法が注目を集めている．しかし，先行研究の多くは，ニュース，論文や特許など比較的フォーマルなテキストを対象としており，カルテ中にしばしば含まれる非文法的かつ断片化された文に対して従来法が有効でないことも多い．そこで，我々は構文解析を行う深い手法と bag of words な浅い手法を併用し，入力文章の性質により両者をダイナミックに切り替える手法を提案する．実験の結果，個々の手法より高い精度（88.9%）を得たので報告する．
キーワード　文章分類, 情報抽出, 自然言語処理, 電子カルテ, 構文解析

Robast Classification for Ungrammatical and Fragmented Texts

Eiji ARAMAKI†, Takeshi IMAI†, Kengo MIYO†, and Kazuhiko OHE†

† University of Tokyo Hospital　7-3-1 Hongo, Bunkyo-ku, Tokyo, 113–8655 Japan
E-mail: †{aramaki,ken}@hcc.h.u-tokyo.ac.jp, ††miyo-sup@h.u-tokyo.ac.jp, †††kohe@hcc.u-tokyo.ac.jp

Abstract　Nowadays, a large scale statistical medical studies using electronic medical records have drawn a great deal of attention in the medical field. However, some parts of information in medical records are written in natural language, requiring an information extraction technique from texts. In the information extraction field, there are many previous studies starting from an early pattern match method to a recent tree-based pattern match method. However, most of previous studies handle formal texts, such as news papers, patents, and technical papers. On the other hand, texts in medical records includes ungrammatical or fragmented texts. To deal with such texts, we propose a robust method to select a suitable approach for each input. The experimental results showed high performance (88.9%), demonstrating the feasibility of the proposed method.
Key words　Document Classification, Information Extraction, Natural Language Proceesing, Medical Record, Parsing

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，自然言語処理技術が必須となる．

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の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（喫煙／非喫煙／不明）を推定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，文章からなんらかの情報を抽出するという観点からは，自然言語処理でこれまで扱われてきた情報抽出の一種だと考えられるが，次の 2 つの点（タスクの特殊性，対象文章の特殊性）で異なる性質を持つ．

【タスクの特殊性】

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

表 1　喫煙状態とその定義.

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.

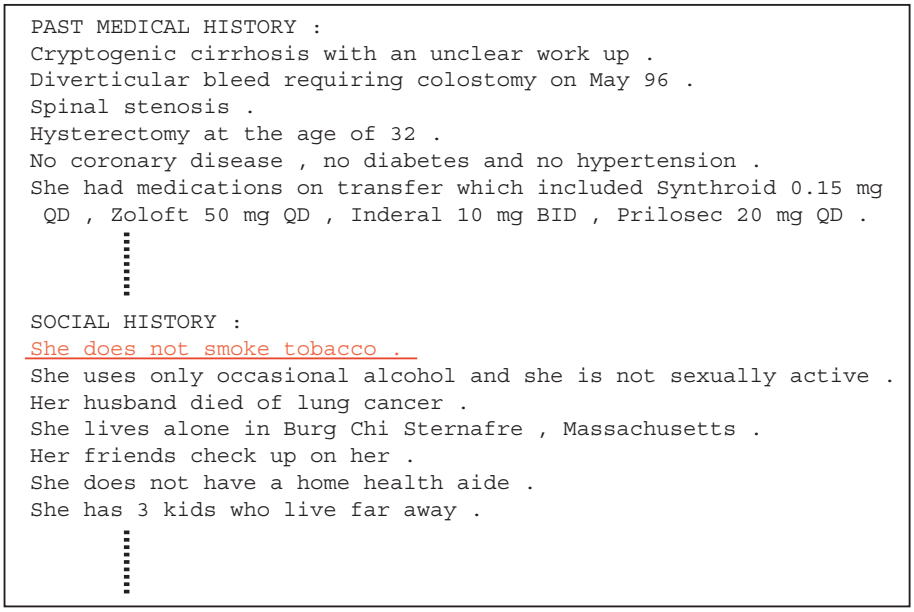

図 1　退院サマリの例.

* 下線部は喫煙関連文 (3.1 節で述べる) を示す.

のアプローチをとる.

【対象文章の特殊性】

もう一つの特徴は，対象となる文章の性質である．従来の情報抽出は論文 [25] やニュース [21], [26] など比較的フォーマルな文章を扱ってきた．一方，カルテ文章は，文献 [23] が指摘したように，次の 2 つの特徴がある．

（1） **断片化**: 以下の例のように，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い.

> Conditions , Infections , Complications , affecting Treatment / Stay Hypercholesterolemia , h / o Bell'palsy , smokin

（2） **非文法的**: カルテには，非文法的な表現がしばしば含まれる．例えば，上記の例では，“**smokin**” は正しくは，“**smoking**” が正確な記述である.

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，構文解析処理が有効でない場合も多い．そこで，我々は構文解析の結果得られる依存構造上での文の類似度と，表層的な語順の類似度の両方を併用することを考える.

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(注1)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（以下，**文章**）からなる．1 文章のあたりの平均の文数は 86.9 文である．また，1 文の平均語数は 8.85 語である.

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**）がアノテートされている．喫煙状態は，表 1 のように U,S,C,N,P の 5 つに分類されている.

これらのアノテーションは，医師がカルテに記述された情報にもとづいて行った．喫煙状態ごとの数を表 2 に示す.

（注1）: https://www.i2b2.org/NLP/

表 2　喫煙状態別の文章数.

喫煙状態	文章数
UNKNOWN	252
SMOKER	9
CURRENT SMOKER	35
NON SMOKER	66
PAST SMOKER	36

表 3　喫煙状態別の喫煙関連文を抽出できる割合.

喫煙状態	Ratio
UNKNOWN	1.1% (= 3/252)
UNKNOWN 以外 (C,P,S,N)	98.6% (=144/146)

表 4　喫煙関連文の統計量.

平均語数	18.7 語
中央値	11 語
最小語数	2 語
最大語数	187 語

3. 提 案 手 法

提案手法の流れを図 2 に示す．まず，最初にコーパス（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 (3.1 章)．次に，3 つの尺度（ED,NGRAM,TREE）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（3.2 章）．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（3.3 章）．

3.1 前処理：喫煙関連文抽出

まず最初に，コーパス（トレーニングセットと入力文章）から，喫煙と関連した文 (**喫煙関連文**) を抽出する．ここでいう，喫煙関連文は以下のキーワードを含む文とみなす：“*nicotine, smoker, smoke, tobacco, cigarette*”．また，これらの屈折形もキーワードとして扱った．つまり，“*smoked*” や”*smoking*” などもキーワードとして扱われる．例えば，図 1 では，下線部の文が喫煙関連文として抽出されることになる．

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．

表 3 にコーパスから喫煙関連文が抽出できた割合を喫煙状態別に示す．キーワードを用いた文の抽出手法は単純なものであるが，表 3 に示されるように UNKNOWN 以外の文章からは喫煙関連文を 98.6%の割合で抽出することができ，また，UNKNOWN からは，喫煙関連文は滅多に抽出されない（1.1%）．参考までに，抽出された喫煙関連文の統計を表 4 に，喫煙関連文の例を表 5 に示す．

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．

（1） **ED**: 文字単位の類似度 (3.2.1 節)

（2） **NGRAM**: 語 n-gram の類似度 (3.2.2 節)

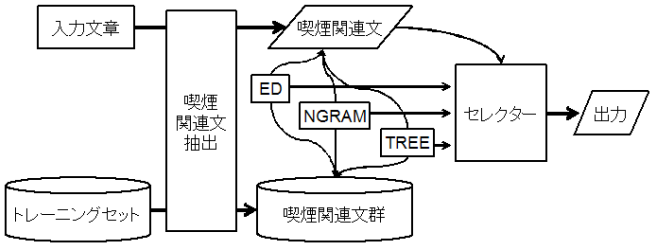

図 2　処理の流れ.

（3） **TREE**: 依存構造上での語 n-gram の類似度 (3.2.3 節)

これらは，(1)ED が語の切り分けさえも利用しない最も浅い尺度，(3)TREE は構文解析を行う深い尺度，(2)NGRAM がそれらの中間といった位置付けとなる．

3.2.1 ED: 編集距離を用いた尺度

編集距離 [13] は，文字単位の類似度であり，文字の挿入や削除、置換によって，一つの文字列を別の文字列に変形するのに必要な手順の最小回数として定義される．例えば，“smokin” と “smoking” の編集距離は，“g” を一度挿入するだけであるので，1 となる．類似度 ED は編集距離を以下のように文長で正規化して用いた：

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

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以下の式で重み付き投票し，最も高い確信度を持つものとする：

$$\text{確信度 } (S) = \sum_{S_t \in S} sim_{ED}(S_i, S_t), \quad (2)$$

ただし，S は，トレーニングセット中の同じ喫煙状態に属する S_t の集合である．

3.2.2 NGRAM: n-gram ベースの情報検索尺度

NGRAM は，文を n-gram 単位に分解して計る類似度である．まず，喫煙関連文を n-gram ($n = 1..4$) までの語（列）に分解する．例えば，以下の文は次のように n-gram に分解される：

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

1-gram: (He), (has),...,(use)

2-gram: (He has), (has a),...,(tobacco use)

3-gram: (He has a), (has a past),...,(and tobacco use)

4-gram: (He has a past), (has a past history),...,(hypertension and tobacco use)

次に，分解された語列間の類似度を計算する．これには，Okapi-BM25 [19] 尺度を用いた．Okapi-BM25 尺度の定義は表 6 に示す（詳細な定義は文献 [19] を参照のこと）．Okapi-BM25 尺度は NTCIR 特許検索タスク [14] や，病名の自動分類 [5] に

表 5　喫煙関連文の例.

喫煙状態	喫煙関連文
CURRENT-SMOKER	**Smoking** :
NON-SMOKER	**Tobacco** history :
NON-SMOKER	She is a non **smoker** .
CURRENT-SMOKER	Please attempt to quit **smoking** .
NON-SMOKER	The patient denied using tobacco . **smoking** .
PAST-SMOKER	The patient is an ex-**smoker** .
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 .
SMOKER	The patient is a 64-year-old male with a long standing history of peripheral vascular disease and **tobacco** use who has had multiple vascular procedures in the past including a fem-fem bypass , a left fem pop once above the knee with PTFE graph and then again below the knee with a refreshed saphenous vein graft as well as bilateral TMAs and a right fem pop bypass who presents with a nonhealing wound of his left TMA stump as well as a pretibial ulcer that is down to the bone .

表 6　Okapi-BM25 尺度 (sim_{BM25}).

$$sim_{BM25}(S_i, S_t) = \sum_{t \in T} (W_d \times W_q), \quad (3)$$

ただし，

$$W_d = \frac{(k_1+1)tf}{k_1((1-b)+b \times dl/avdl)}, \quad (4)$$

$$W_q = \log \frac{N-n+0.5}{n+0.5}. \quad (5)$$

T は両方（入力文章とトレーニングセットの文章）の喫煙関連文に共通して含まれる語の集合である．tf は，t の出現頻度である．dl は S_t の語数である．$avdl$ は S_t の平均の語数である．N は S_t の総数である．n は抽出された S_t の数である．k_1 と b は定数であり，予備実験の結果，$k_1 = 1.5$，$b = 0.75$ とした．

て高い精度を収めた情報検索尺度であり，提案手法はこれをそのまま用いた．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した．

3.2.3　TREE: 統語解析を用いた尺度

先の 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，TREE は依存構造上で文を n 語の組み合わせに分解する．すなわち，文を n 語の部分木に分解して扱う．図 3 に依存構造から生成される部分木の例を示す．また，n は NGRAM と同じく $n = 1..4$ とした．

部分木に分解された文同士の類似度は前節と同じく Okapi-BM25 [19] 尺度を用いて計算する．最終的な出力は，前節までの手法と同じく上位 k 個の重み付き投票によって決定する．

3.3　尺度のセレクター

最後に，前節で述べた 3 つの尺度（ED,NGRAM,TREE）で

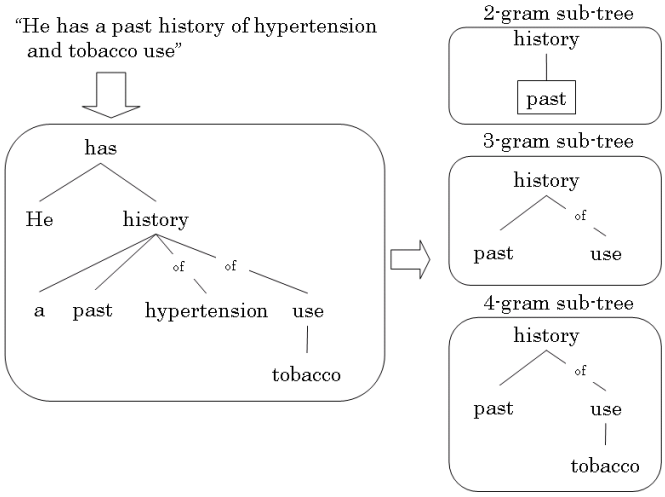

図 3　構文解析結果とそこから得られる部分木の例.

表 7　重要語ペアの例.

重要語ペア	Okapi 尺度
a former	663.8
denies any	587.9
she denies	579.8
years ago	561.4
not smoke	557.8
she does	553.3
does not	547.1
per packs	514.8
day x	464.4
(NUM) years	461.2
a remote	310.0
any use	309.0
she smokes	228.3
:	:

* (NUM) は数字（列）を示す.

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．

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，3.2.2 節で述べた NGRAM 尺度の 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．得られた重要語ペアの例を表 7 に示す.

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” が重要語ペアとなった場合，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 [17]）を用いて正解した尺度と以下の素性との関係を学習し構築した.

（1） ED，NGRAM，TREE の確信度の値,
（2） 入力文章の喫煙関連文の語数,
（3） 重要語ペアの距離.

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，どれを選んでよい場合は，ED ＞ TREE ＞ NGRAM という優先順位で学習を行った(注2)．例えば，ED,TREE が正解しており，NGRAM が不正解である場合，ED を使うことになる.

(注2)：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.

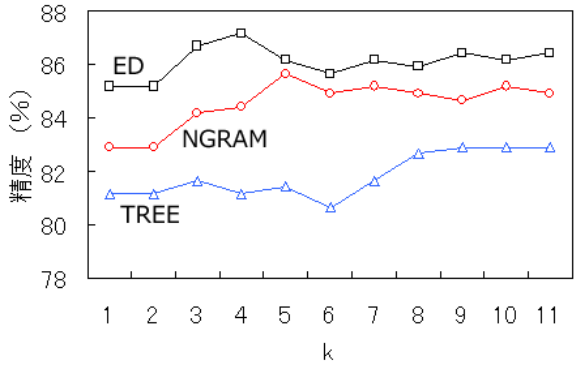

図 4　k の値と各尺度の精度.

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） **ED**: ED だけを用いた手法.
（3） **NGRAM**: NGRAM だけを用いた手法.
（4） **TREE**: 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，ED,NGRAM,TREE 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（図 4）の結果，それぞれ $k_{ED} = 4$; $k_{NGRAM} = 5$; $k_{TREE} = 8$ とした．また，TREE が用いる構文解析処理には Charniak の nlparser [7] を用いた(注3).

4.2　結　　果

結果を表 8 に示す．また，3 つの尺度の正解／不正解の頻度を表 9 に示す.

表 8 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 9 が示すように，TREE 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.

また，PROPOSED の精度は 3 つの尺度よりも高く，うまく各尺度を使い分けている．例えば，ED が間違う場合でも，他の尺度が正解する場合は 24 件あり（表 9），PROPOSED はその

(注3)：nlparser は句構造を出力するため，これを文献 [8] の手法にて依存構造に変換した.

表 8　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 9　各尺度の正解・不正解の頻度.

ED	NGRAM	TREE	頻度
×	×	×	27
×	×	○	10
×	○	×	8
○	×	×	7
○	○	×	24
○	×	○	13
×	○	○	6
○	○	○	303

* “○” は正解を示す. “×” は不正解を示す.

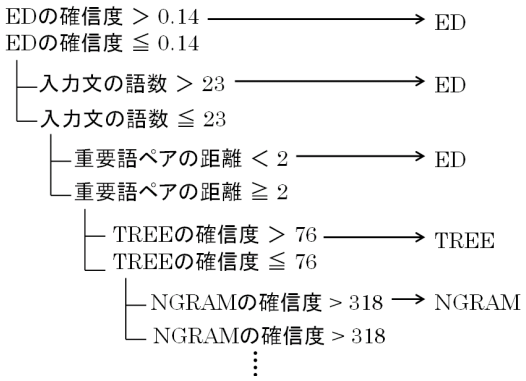

図 5　セレクターの決定木の例.

うちの 7 件について他の手法を用い正解し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). 有意差が得られない理由は，まず，実験のサンプルが少ないことが挙げられる. また，今回の 3 つの尺度のいずれを選んでも不正解となる場合が 27 件と比較的多く存在し（表 9），この結果，PROPOSED の上限は 93.21%となっている. この上限のもとで，PROPOSED が有意差を得るのは困難な状態だと言える.

PROPOSED において各尺度が採用された頻度に表 10 に示す. 単独の尺度としては ED が最も高精度のため，ED の利用回数が非常に高くなっている. 次に，尺度の判別にどのような素性が必要とされているのかをみるために，学習された決定木の一例を図 5 に示す(注4).

ED は，ED の確信度が高い場合や，文が長い場合，また，重要語ペアの距離が小さい場合に採用されていることが分かる. 一方，重要語ペアの距離が離れており，TREE の確信度が高い場合には TREE が採用されており，重要語ペアの距離が尺度の判別に貢献していることが分かる.

(注4)：実験は，5-fold で行ったため，実際には 5 つの決定木が得られている.

表 10　各尺度が選択された頻度.

尺度	頻度
ED	99
NGRAM	4
TREE	9

5. 関 連 研 究

電子カルテからの患者の情報抽出タスクは，新しいタスクであり，本コーパスを用いたワークショップである AMIA [3] i2b2-NLP [12] ワークショップ以外での本格的な先行研究は少ない. しかし，これを文章分類と情報抽出の両方にまたがるタスクと考えると関連研究を見つけることができる.

5.1 情 報 抽 出

一般の情報抽出と本タスクの違いは，カルテにおいては一人の患者が一つの文章に対応しているため，文章毎に必ず一つの喫煙状態を抽出しなくてはならない点である. そこで，提案手法は，トレーニングセットから情報抽出パターンを得るのではなく，トレーニングセットの喫煙関連文との類似度を計算し，もっとも近いものに分類するというアプローチをとった. ただし，どのように文を扱うかといった点では，情報抽出の先行研究が参考になる.

これまでの情報抽出の研究においては，表層的な語列のパターンを機械学習によって捉える手法が主流であった. 例えば，レジュメからの情報抽出 [27] や，セミナー・アナウンスメント [9]，求人情報 [1], [2]，住所 [15], [24] を対象とした研究がこれにあたる. これらは，表層上の単語列に基づくパターンを自動獲得しているにすぎず，多様な文の表現を扱えない.

そこで，構文解析技術の発展の影響を受けて，現在は，依存構造上でテキストを扱う手法に注目が寄せられている. 例えば, [25], [26] は述語-項構造で情報抽出パターンを扱った. [21] は部分木構造としてパターンを扱った. 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，現状の構文解析処理が割にあわないことも多い.

実際に，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. 例えば, [18] は人手で表層的なパターンを作成し分類を行った. [6] は n-gram 頻度を用いて分類を行った. [20], [22] は，語彙を素性として機械学習 (SVM [20]; C4.5+Ada-boost [22]) による分類を行った. 参加した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，n-gram 分布を情報検索尺度 (Okapi-BM25) 用いて分類する手法 [4] であった. そこで，本研究では [4] をもとに，構文解析とのハイブリッドなアプローチに挑戦した.

最後に, [11] [16] が指摘したように多くの研究は前処理とし

て，固有表現抽出を行っている．しかし，喫煙状態は固有表現と共起することはまれなため，固有表現抽出は行わなかった．

5.2 文章分類

本タスクは，患者の文章ごとに喫煙状態の分類を行うため，文章分類の一種だとも考えられる．文章分類手法は，TREC(注5)やNTCIR [10] など多くのワークショップが開かれて，盛んに研究されている．しかし，それらの多くは文章全体の内容によって分類するものが多い．

一方，本タスクは，喫煙の関連するたかだか数文によってのみ文章が分類される．このため，文章から重要な文を抽出し (3.1 章)，そこから分類を行う (3.2 章) という 2 段階のステップを踏んだ．

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．

文献

[1] A.D.Sitter and W.Daelemans. *Information extraction via double classification*. Proceedings of ATEM03, 2003.

[2] A.Finn and N.Kushmerick. *Multi-level boundary classification for information extraction*. Proceedings of ECML04, 2004.

[3] AMIA. American medical informatics association http://www.amia.org/, 2006.

[4] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Patient status classification by using rule based sentence extraction and BM25-kNN based classifier, 2006.

[5] Eiji Aramaki, Takeshi Imai, Masayuki Kajino Kengo Miyo, and Kazuhiko Ohe. A statistical selector of the best among multiple icd-coding methods. In *proceedings of MedInfo (to appear)*, 2007.

[6] Francisco M. Carrero, Jose M. Gomez-Hidalgo, Enrique Puertas, Manuel Mana, and Jacinto Mata. Quick prototyping of high performance text classifiers, 2006.

[7] 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.

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

[9] D.Freitag and A.McCallum. Information extraction with hmms and shrinkage. In *Proceedings of AAAI99 Workshop on Machine Learning for Information Extraction*, pp. 31–36, 1999.

[10] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at ntcir-4. In *Proceedings of the fourth NTCIR 4 workshop*, pp. 225–232, 2004.

[11] Ralph Grishman. Research in information extraction: 1996-98. In *Proceedings of a workshop on held at Baltimore*, pp. 57–60. Association for Computational Linguistics, 1996.

[12] i2b2. Informatics for integrating biology and the bedside https://www.i2b2.org/nlp/, 2006.

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

[14] Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, and Hitoshi Isahara. Using the k nearest neighbor method and bm25 in the patent document categorization subtask at ntcir-5. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 324–332, 2005.

[15] N.Kushmerick, E.Johnston, and S.McGuinness. *Information extraction by text classification*. IJCAI01 Workshop on Adaptive Text Extraction and Mining, 2001.

[16] Mary Ellen Okurowski. Information extraction overview. In *Proceedings of a workshop on held at Fredericksburg*, pp. 117–121. Association for Computational Linguistics, 1993.

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

[18] Magne Rekdal. Identifying smoking status using argus mlp, 2006.

[19] 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.

[20] Guergana K. Savova, Philip V. Ogren, Patrick H. Duffy, James D. Buntrock, and Christopher G. Chute. Mayo clinic nlp system for patient smoking status identification, 2006.

[21] 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.

[22] Gyorgy Szarvas, Richard Farkas, Szilard Ivan, Andras Kocsor, and Robert Busa-Kefete. Automatic extraction of semantic content from medical discharge records, 2006.

[23] 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.

[24] V.Borkar, K.Deshmukh, and S.Sarawagi. Automatic segmentation of text into structured records. In *Proceedings of ACM SIGMOD Conference*, pp. 175–186, 2001.

[25] 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.

[26] 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.

[27] Kun Yu, Gang Guan, and Ming Zhou. Resume information extraction with cascaded hybrid model. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2005)*, pp. 499–506, 2005.

(注5)：http://trec.nist.gov