電子カルテの入力支援を目的とした文書カテゴリ推定

津田裕亮[1)]，中村明[2)]，松本忠博[1)]，池田尚志[1)]
岐阜大学工学部[1)]，三洋電機（株）ヒューマンエコロジー研究所[2)]

1　はじめに

近年，医療の IT 化に伴って電子カルテが普及しつつある．電子カルテの利点として，情報検索が容易になる，カルテ保存に関して省スペース化が可能になるなどが挙げられる．しかし欠点として，データの入力作業により現場の医師や看護師が本来の医療行為に集中することを妨げられ，かえって臨床の業務効率の低下を招くことや，患者と医師とのコミュニケーションの低下など，診療の質の低下につながりかねないと懸念されている．入力の負担軽減を実現させることは，こうした欠点を解消し，電子カルテのさらなる普及，医療の進歩に貢献すると考えられる [1].

本稿では，文書カテゴリ推定とカテゴリ別言語モデルを用いた入力支援システムの構築を目的として，医療文書のカテゴリ（疾患群）推定を行った．カテゴリ推定結果に基づいてカテゴリ別言語モデルを動的に補間することにより言語モデルの精度向上を図る．またカテゴリ推定は入力支援に加え，診断支援も可能になると考える．実験では，カテゴリ識別アルゴリズムに k 最近傍法（以下，kNN）とカテゴリカル k 近傍法（以下，CAP）を使用し，両者の比較を行った．また，実際の入力支援を想定し文書の一部からの推定も試みた．

2　システム概要

カテゴリ推定とカテゴリ別言語モデルを用いた入力支援システムの概要は以下の通りである．

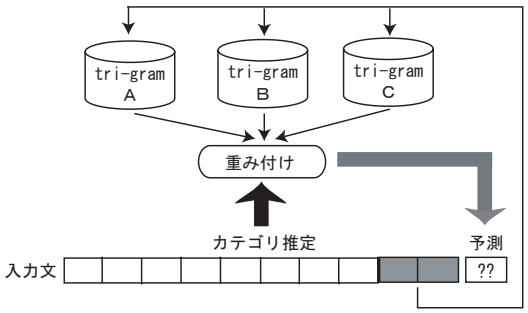

図 1: N-gram 混合モデル（tri-gram の場合）

(1) 各カテゴリの N-gram モデルを構築しておく．

(2) 入力文の直近の L 単語列から各カテゴリとの類似度を求め，カテゴリ推定を行う．

(3) 各カテゴリの N-gram 確率を，各カテゴリとの類似度に基づいて重み付けした確率を用いて入力予測を行い，予測候補を求める．

本稿ではカテゴリ推定の部分を報告する．

3　文書カテゴリ推定

3.1　文書カテゴリ推定の流れ

以下に，文書カテゴリを推定するまでの流れを示す．

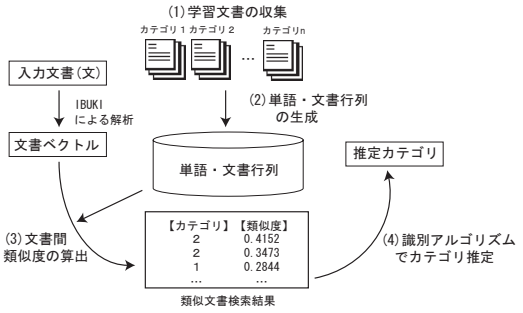

図 2: 文書カテゴリ推定までの流れ

(1) 学習文書の収集

あらかじめカテゴリ属性が付与されている文書群を収集しておく．

(2) 単語・文書行列の生成

収集した文書群から，日本語文解析システム ibuki[2] により形態素解析を行い，$tf \cdot idf$ により単語の重み付けを行い，単語・文書行列を生成，SimplePCA[3] で主成分分析を行い次元圧縮をする．

(3) 文書間類似度の算出

カテゴリが未知である文書（以下，未知文書）を検索質問文書とし，各学習文書との類似度をコサイン尺度により求める．

(4) 識別アルゴリズムによりカテゴリを推定

各学習文書との文書間類似度に基づき未知文書のカテゴリを推定する．本稿では，kNN と CAP のどちらかのアルゴリズムを使用してカテゴリを推定し，精度比較を行う．

3.2 識別アルゴリズム

3.2.1 kNN

k 最近傍法（k-nearest neighbor rule,kNN）は，未知パターンに近い上位 k 個の学習パターンによる多数決によって識別が行われる．本稿では多数決に，単純な投票式ではなく類似度を用いる．

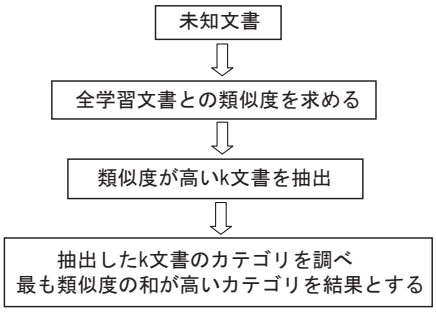

図 3: kNN の手順

3.2.2 CAP

カテゴリカル k 近傍法（classification using Categorical Average Patterns,CAP）は，カテゴリごとの k 近傍学習パターンからなる平均パターン（以下，カテゴリカル k 近傍平均パターン）と未知パターンとの類似度で識別が行われる [4].

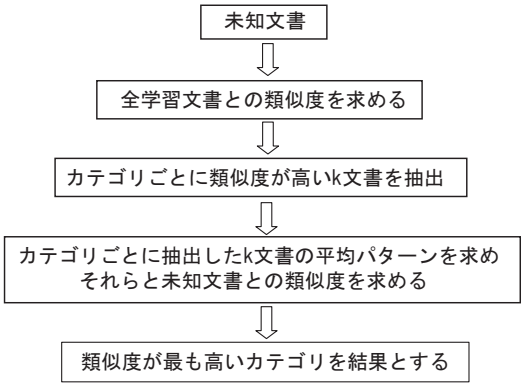

図 4: CAP の手順

4 評価実験

4.1 実験データ

「ロゴヴィスタ電子辞典シリーズ・南山堂医学大辞典第 18 版」[5] から選定した，ICD10（国際疾病分類第 10 版）[6] に含まれる病名を見出し語とする語句説明文を対象とする．図 5 に示す ICD10 の大分類（章）をカテゴリとし，実験では全 21 章の中から計 10 章を選び 10 個のカテゴリで実験を行う．各カテゴリの文書を学習用と評価用に分け，学習文書，未知文書とした．精度は未知文書にあらかじめ付与されているカテゴリを正解とし，全未知文書の正解率で表す．また，病名の中には ICD10 コードが 2 つ割り当てられているものがある．そのような文書は第 2 コードに対応するカテゴリも正解とした．文書ベクトル化する単語の品詞は「一般名詞」「サ変名詞」「固有名詞（地名）」とした．

〈章〉〈ICDコード〉〈分類見出し〉

1	A00-B99	感染症および寄生虫症
2	C00-D48	新生物
3	D50-D89	血液および造血器の疾患ならびに免疫機構の障害
4	E00-E90	内分泌，栄養および代謝疾患
5	F00-F99	精神および行動の障害
6	G00-G99	神経系の疾患
7	H00-H59	眼および付属器の疾患
8	H60-H95	耳および乳様突起の疾患
9	I00-I99	循環器系の疾患
10	J00-J99	呼吸器系の疾患
11	K00-K93	消化器系の疾患
12	L00-L99	皮膚および皮下組織の疾患
13	M00-M99	筋骨格系および結合組織の疾患
14	N00-N99	尿路性器系の疾患
15	O00-O99	妊娠，分娩および産じょく〈褥〉
16	P00-P96	周産期に発生した病態
17	Q00-Q99	先天奇形，変形および染色体異常
18	R00-R99	症状，徴候および異常臨床所見・異常検査所見で他に分類されないもの
19	S00-T98	損傷，中毒およびその他の外因の影響
20	V00-Y98	傷病および死亡の外因
21	Z00-Z99	健康状態に影響をおよぼす要因および保健サービスの利用

図 5: ICD10 国際疾病分類第 10 版

【実験で扱うカテゴリ】
- 感染症および寄生虫症
- 新生物
- 内分泌，栄養および代謝疾患
- 神経系の疾患
- 眼および付属器の疾患
- 循環器系の疾患
- 消化器系の疾患
- 皮膚および皮下組織の疾患
- 筋骨格系および結合組織の疾患
- 尿路性器系の疾患

【学習文書】
各カテゴリの文書それぞれ 100 文書，計 1000 文書.
【未知文書】
各カテゴリの文書それぞれ数文書〜100 文書，計 433 文書.

4.2 実験結果

4.2.1 文書カテゴリ推定

kNN と CAP の精度をそれぞれ示す.

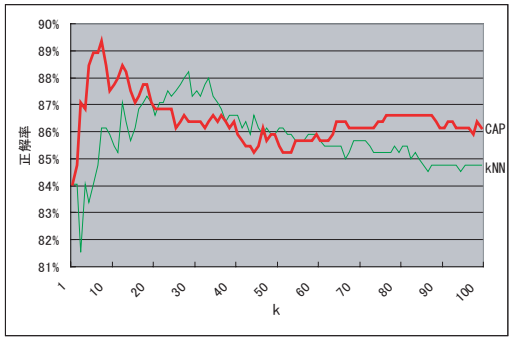

図 6: 各識別アルゴリズムでの精度

圧縮前の未知文書の未知語率を以下に示す．

表 1: 未知語率

未知文書単語異なり数	未知語数	未知語率
7340	2245	30.58 %

累積正解率（推定 5 位まで）を示す．

※ kNN → k=29 で推定 1 位が最高値

CAP → k=8 で推定 1 位が最高値

表 2: 各識別アルゴリズムでの累積正解率

kNN 累積正解率 (%)

推定＼k	1	8	10	29	50	100
1 位	84.06	86.14	85.91	88.22	85.91	84.75
～2 位	–	97.22	96.53	97.22	97.69	97.45
～3 位	–	98.84	98.84	98.61	98.38	98.15
～4 位	–	99.30	99.30	99.07	98.61	98.61
～5 位	–	99.30	99.30	99.53	99.07	99.30

CAP 累積正解率 (%)

推定＼k	1	8	10	29	50	100
1 位	84.06	89.37	87.52	86.37	85.91	86.14
～2 位	93.99	97.92	98.15	97.69	97.45	97.22
～3 位	98.84	98.61	98.61	98.38	98.61	98.38
～4 位	99.30	99.07	99.07	99.07	99.30	99.07
～5 位	99.53	99.07	99.07	99.07	99.30	99.30

kNN が k=29 のときに最高値 (88.22 %),CAP が k=8 に最高値 (89.37 %) となり，わずかではあるが CAP のほうが良い結果が得られた．k の値が小さい場合に kNN と CAP の違いが顕著に見られた．図 7 に，ある未知文書を CAP で推定した際の k とカテゴリカル k 近傍平均パターン (類似度上位の 3 カテゴリのみ) との類似度の変化を示す．

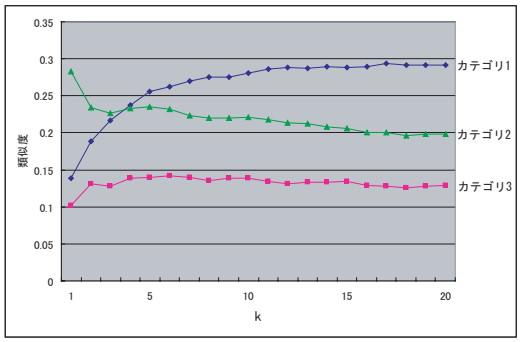

図 7: 近傍数 k と平均パターンとの類似度の関係

図 7 から，正解カテゴリ 1 は徐々に類似度が増加するが，他のカテゴリは少しずつ低下していっていることがわかる．これは，正解カテゴリ文書は未知文書の周りに均等に分布しており，平均パターンは常に未知文書の近くに生成されるが，他のカテゴリは k の値を増やすにつれて未知文書に類似していないものが多く平均に含まれるからだと考えられる．このことから kNN での例外的な文書の影響を抑えていると考えられる．しかし，k の値を大きくしすぎると類似度が全体的に低下していく傾向にあった．学習文書不足で類似する文書が少ししかない場合があることが原因と考えられる．また，文書によっては k の値を大きくすることで正解となったり，途中で不正解となったりと不安定なものがあり，文書の内容的な問題も考えられる．

未知語率が高いため，学習文書を増やすことが必要である．

4.2.2 文書の一部からの推定

実際の入力支援を想定し文書の一部からの推定を試みた．まず，未知文書の先頭 n 文を使い，n を変化させたときの各精度を以下に示す．未知文書の平均文数は 8～9 文であった．

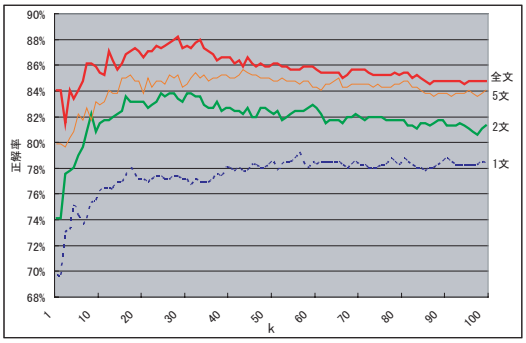

図 8: 文数での精度（kNN）

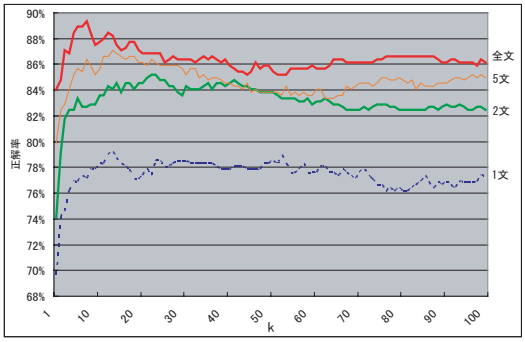

図 9: 文数での精度（CAP）

文書の一部からでもある程度の精度が得られた．特に 1 文と 2 文での差が大きかった．各識別アルゴリズムも，文書全体のときと同じような傾向であり，わずかではあるが CAP の方が良い結果になっていた．

文単位で精度を算出したが，文は長さがそれぞれ違うので正確な割合での精度がわかりくい．よって，次に単語数での精度実験を行った．

未知文書の先頭 n 単語を使い，n を変化させたときの各精度を以下に示す．未知文書の平均単語数は 70 語であった．

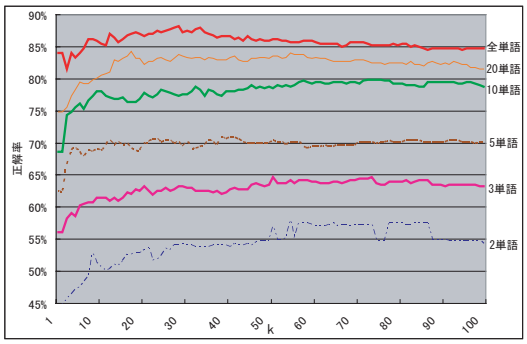

図 10: 単語数での精度（kNN）

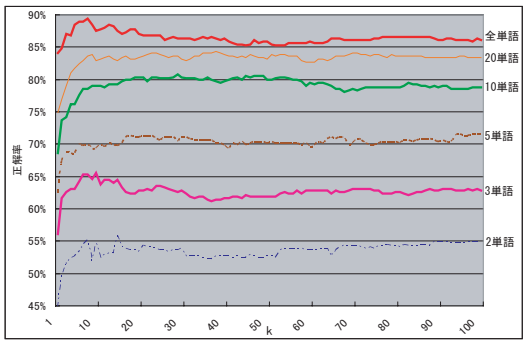

図 11: 単語数での精度（CAP）

単語数が少ないとやはり推定は難しくなる．10 単語くらいで文書全体の精度に近い値は出ている．また，10 単語で大体 1 文の精度と同じになっており，実際のカテゴリ推定をするときは 1 文〜2 文は必要であるといえる．

4.3 今後の課題・問題点

学習文書を増やして実験する

未知語率の高さや識別アルゴリズムでの問題の原因として学習文書の不足が懸念された．学習文書の不足は推定精度の低下につながるので，文書数を増やすことは重要である．また，今回の実験ではわずかに CAP の方が kNN よりも精度が良いという結果になったが，学習文書を増やすことで傾向が変わる可能性もある．

文書間類似度の精度向上

今回の実験では ibuki の解析結果から「一般名詞」,「サ変名詞」,「固有名詞（地名）」を用いて文書間類似度を算出したが，未知文書によっては関連のある類似文書がほとんどないことがあった．正確な文書間類似度はそのまま精度向上につながるため，基底となる単語の選定に未知語などを用いて，より正確な文書間類似度を算出する必要がある．また，識別に関係しないと思われる単語は使わないなど，重要語の取捨選択も必要であるのではないかと考える．

評価方法の見直し

学習文書と未知文書はそれぞれ共通のコーパスからランダムに抽出して使用したが，今回の実験では，学習文書と未知文書の組み合わせを固定して実験を行ったため，他の組み合わせでは異なる傾向をもった結果になる可能性もあり，精度に信頼性が欠けている．よって，今後は複数の異なる学習文書と未知文書の組み合わせで実験を行い，より信頼できる精度を得る必要がある．

5 おわりに

本稿では，電子カルテの入力支援を目的とした文書カテゴリ推定の実験・精度評価を行った．

実験では，識別アルゴリズムの比較を行い，わずかではあるが kNN よりも CAP の方が有用であることが確認できた．また，文書の一部からでもある程度の精度が確認できた．

しかし，今回の実験では小規模なデータで調査した結果にすぎない．今後はデータ量を豊富にして実験を行い，課題・問題点を見つけていく必要がある．そして，カテゴリ推定の精度向上に努め，最終的な目的である入力支援，または診断支援に役立てたい．

参考文献

[1] 予測入力による電子カルテ入力支援
川尻博光他，第 26 回医療情報学連合大会, 3-B-1-3,2006
[2] 日本語文解析システム ibukiC/S について
山田佳裕他，言語処理学会第 12 回年次大会, pp.185-187, 2006
[3] SimplePCA を用いたベクトル空間情報検索モデルの次元削減
黒岩他，NLC2001-17
[4] 未知パターンとカテゴリカル k 近傍平均パターンとの距離に基づくパターン識別
堀田政二他，電子情報通信学会論文誌, Vol.J88-D-II, No.8, pp.1357-1366,2005
[5] ロゴヴィスタ電子辞典シリーズ・南山堂医学大辞典第 18 版
http://www.logovista.co.jp/
[6] ICD10 国際疾病分類第 10 版
http://www.dis.h.u-tokyo.ac.jp/byomei/ICD10/
[7] 確率的言語モデル
北研二，東京大学出版会 1999
[8] 情報検索アルゴリズム
北研二他，共立出版 2002