専門分野の文書を対象としたキーワード抽出・類似文書検索と評価実験

吉田鑑地[1)]，中村明[2)]，川尻博光[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

言語処理技術を特定の専門分野の文書に適用する場合しばしば語彙の不足が問題となるが，専門性の高い知識が得られれば利用価値が高い．我々は，診療記録の電子化が進む医療分野において，文書に蓄積された知識を二次利用し，診療の質向上に役立てることを目指して研究を進めている．この分野では，カルテから診断に関する言語情報の抽出を行った事例 [5] 等が報告されているが，言語処理技術を適用した場合の課題や効果に関しては，まだ十分議論されていない．

本稿では，医療情報分野の文献を対象として類似文書検索・キーワード抽出を行い，その結果を 10 人の評価者の主観によって評価した内容を報告する．キーワード抽出では未知語・複合語・不要語を考慮することの有用性を確認した．また類似文書検索では，3 種類のタームの重み付け方法によって得られた検索結果の比較を試みた．

2　キーワード抽出と評価実験

文書の概要把握を支援することを目的として，単文書からのキーワード抽出を行った．第 24 回医療情報学連合大会論文集 [2]（320 文書）の各文書から Abstract，KeyWord を除いたものを処理対象とした．日本語文解析システム ibuki[1] を用いて形態素解析を行った後,「χ^2 値×出現頻度」を用いてタームの重み付けを行い，文書ごとに上位 10 位までの語をキーワードとして抽出した（抽出対象とした品詞については後述）．

$$\chi^2\text{値}\times\text{出現頻度} = \frac{(tf(t,d)-tf(t,d)\text{ の期待値})^2}{tf(t,d)\text{ の期待値}} * tf(t,d)$$

（$tf(t,d)$：文書 d におけるターム t の出現頻度 ）

全文書 (320 文書) の中から無作為に 10 文書を選んで評価対象とし，その 10 文書それぞれからのキーワード抽出結果を評価した．評価には”著者の付けたキーワード”と”10 人の評価者による主観評価”を用いた．

主観評価は，抽出したキーワード 1 つ 1 つについて，評価者に 1 点〜4 点の点数（高いほど適切なキーワード）を付けてもらう形で行った．10 人分の評価結果から点数の平均値を求め，1.0〜2.0 点を「適切でない」，2.1〜2.5 点を「どちらかといえば適切でない」，2.6〜3.0 点を「どちらかといえば適切」，3.1〜4.0 点を「適切」なキーワードとした．

2.1　キーワード抽出における未知語，複合語の有用性

形態素解析では，辞書に登録されていない語（未登録語）が現れた時，正しく解析できない．未登録語が辞書に登録されている語を部分的に含む複合語の場合，ibuki ではより短い単位に分割されて解析される．一方，未登録語がカタカナのみ・または英文字のみで構成される場合には「カタカナ」「英文字列」として解析される（本稿では以降，これらの語を未知語と呼ぶ）．

専門分野の文書では，通常の語彙に含まれない語が重要な意味を持つことが多い．今回の実験では，MEDIS（財団法人医療情報システム開発センター）が公開している標準病名マスター（19,599 語），標準手術・処置マスター（9,664 語），および標準臨床検査マスター（5,723 語）計 34,986 語を解析辞書に追加登録した．しかし未登録の複合語や略語，表記ゆれに起因する解析誤りは相当数発生する．対象文書の解析結果に基づいて未登録語を追加する方法では，未知の文書への対処の面であまり実用的ではない．そのため，専門分野の文書を対象としたキーワード抽出では未知語や複合語を用いることが有用と考え，検証を行った．

2.1.1　「一般名詞」「サ変名詞」による抽出実験

キーワード抽出対象の 320 文書を ibuki を用いて形態素解析したところ，延べ 141,823 語，異なり 8,965 語が名詞として解析された．名詞と解析された語には「形式名詞」や「固有名詞」などがあるが,「一般名詞」（延べ 75,992 語，異なり 5,175 語）と「サ変名詞」（延べ 50,933 語，異なり 1,649 語）の数が他に比べて非常に多かった．名詞と解析された語の情報は「一般名詞」,「サ変名詞」の 2 つを用いれば十分抽出できると考えられる．形態素解析結果から「一般名詞」,「サ変名詞」を用いてキーワードを抽出した.（図 1）

文書番号：30
タイトル：医療情報システムと医用電子機器の安定運用を目指した基盤構築
ー集中・分散併用UPSによるクリーン電源の提供システムー

キーワード	重み	
電源	24.109	
分散	16.6351	著者の付けたキーワード
ノイズ	15.7587	
停電	14.1933	電源ノイズ
電子機器	13.8837	UPS
電圧	12.4127	医用電子機器
医用	12.0238	電源
集中	11.3984	
制御	10.4716	
方式	9.50534	

図 1: 「一般名詞」「サ変名詞」による抽出結果

評価対象の 10 文書で抽出できた著者の付けたキーワードは 16.7 %（7/36）と非常に少なかった．また，10 人による主観評価の結果で「適切」,「どちらかといえば適切」と評価されるキーワードは，10 文書から抽出されたキーワード中 37.6 %（38/101）とあまり良い結果は得られなかった．名詞と解析された語を用いるだけでは十分でないことが伺えた．

2.1.2 未知語，複合語を含めたキーワード抽出実験

ibuki を用いて抽出対象の 320 文書を形態素解析したところ，延べ 11,255 語，異なり 3,640 語が「カタカナ」「英文字列」の未知語として解析された．2.1.1 で抽出できなかった「UPS」,「ゲノム」などのキーワードの抽出を期待して「カタカナ」,「英文字列」を用いた．

形態素解析結果から「一般名詞」,「サ変名詞」に加えて未知語「カタカナ」と「英文字列」を用い，ibuki の形態素解析結果の文節情報を基に同一文節内の連続する自立語を連結することにより複合語化を行い，キーワードを抽出した．(図 2)

文書番号：30
タイトル：医療情報システムと医用電子機器の安定運用を目指した基盤構築
ー集中・分散併用UPSによるクリーン電源の提供システムー

キーワード	重み	
ＵＰＳ	13.5032	
停電	12.0649	著者の付けたキーワード
（株）/Ｂｅｓｔ-ソリュ/ションズ	9.99334	
集中/ＵＰＳ/システム	9.99334	電源ノイズ
分散/ＵＰＳ/システム	9.99334	UPS
医用/電子機器	9.71293	医用電子機器
電子機器	9.28109	電源
分担	9.28109	
除去	9.02629	
電源	7.73751	

図 2: 未知語，複合語化を含めた抽出結果

評価対象の 10 文書で著者の付けたキーワードが 44.6 %（16/36），各文書最低 1 つ以上抽出された．また，10 人による主観評価の結果で「適切」,「どちらかといえば適切」と評価されるキーワードは 10 文書から抽出されたキーワード中 53.0 %（62/117）であり，未知語，複合語を用いることによる抽出結果の評価の向上が見られた．

未知語，複合語を用いて抽出したキーワードには複合語が 43，非複合語が 74 あった．図 3 に示すように，複合語では適切と評価されるキーワードが多く，非複合語では適切と評価されないキーワードが非常に多いことがわかる.「医用/電子機器」,「電子/署名」などの複合語のキーワードが新しく抽出されていた．

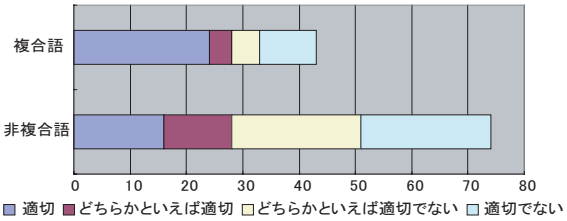

図 3: 複合語，非複合語キーワードの評価結果

未知語，複合語を用いて抽出したキーワードには，未知語を含むキーワードが全部で 16 あった．表 1 に示すように，低い評価を得ているものもあるが，多くが非常に高い評価を得ていた.「SNP」,「EHRS」などの未知語のキーワードが新しく抽出されていた．

表 1: 未知語を含むキーワードの評価結果

抽出されたキーワード	評価
（株）/Best/ソリューションズ	1.2
Flash	1.5
カレンダ/形式	2.2
クリーン	2.3
ケアクライアント	2.6
IT	3.1
SNP	3.7
ゲノム	3.8
EHRS/導入	3.8
ゲノム/情報	3.8
UPS	3.9
集中/UPS/システム	3.9
分散/UPS/システム	3.9
NCO	4.0
EHRS	4.0
PI	4.0

これらのことから，未知語，複合語にはキーワードとして評価されるものが多く，用いることで抽出結果の評価が向上することがわかった．キーワード抽出に未知語，複合語を用いることの有用性が確認できた．

2.2 キーワード抽出における不要語の扱い

一般に『どんな文書にも登場する語は文書の内容を表す語としてふさわしくない』と言われている．本稿ではタームの重み付けに「χ^2 値×出現頻度」を用いることでそのような語の抽出を抑えている．しかし，抽出対象の第 24 回医療情報学連合大会論文集（320 文書）のような専門分野の文書では，”特定の分野でよく使われる専門用語”が多くの文書に出現すること

があるため，そのような専門用語の抽出が抑えられてしまうことがある．2.1.2 での抽出では「電子カルテ」のような専門用語が抽出されにくい傾向にあった．また，320 文書程度の文書では，実際には”どんな文書にも登場する語”であっても特定の文書に偏って登場していることがある．2.1.2 での抽出では「このように」のような語が少なからず抽出されていた．

専門分野の文書を対象としたキーワード抽出では”どんな文書にも登場する語”の抽出を適切に抑え，”特定の分野でよく使われる専門用語”の抽出も行う必要がある．そのため，分野に偏りのない大量の文書から，”どんな文書にも登場する語”をキーワード抽出における不要語として選定し，不要語を考慮することで，不要語の抽出を抑制，これまで抽出できなかった新しいキーワードの抽出を試みた．

2.2.1 不要語の抽出を抑えた抽出実験

朝日新聞 2 年分の記事を ibuki で形態素解析し，2.1.2 と同様にして名詞，未知語，複合語を抽出した（延べ 84,025,001 語，異なり 871,752 語）．その出現頻度の上位 1 %の語（延べ 64,058,083 語，異なり 8,739 語）を不要語とみなした．2.1.2 で行った抽出を，不要語と一致するキーワードの重みを半分にして抽出しなおした．（図 4）

評価文書9
タイトル：電子記録の内容保障に向けた電子署名の実装と運用

	キーワード	重み	評価（×：評価平均値 1～2.5 ○：評価平均値 2.6～4.0）	
	電子-署名	21.1619	○	抽出されていたキーワード
	電子-保存	11.3225	○	
	病院-情報-システム	9.9713	○	
	電子-記録	8.98994	○	
不要語	署名	（半減した値）8.94245	○	
不要語	鍵	（半減した値）8.5516	○	
	医療-情報-連合-大会-論文-集	8.42152	×	新しく抽出されたキーワード
	実-装	8.16287	×	
不要語	運用	（半減した値）7.80885	×	抽出されていたキーワード
	ハッシュ値	6.99746	×	新しく抽出されたキーワード
	原本	6.99746	×	
	原本-記録	6.99746	×	

抽出されなくなったキーワード

不要語	担保	（半減した値）9.57647	×
不要語	このため	（半減した値）9.49336	×

図 4: 不要語の重みを半分にした抽出結果

10 人による主観評価の結果では，適切と評価されるキーワードは 10 文書から抽出されたキーワード中 52.0 %（65/125）であり，不要語の重みを半減しない場合よりも低下していた．著者の付けたキーワードの抽出数は同じだった．

抽出された 125 のキーワード中，不要語の重みを半減する前から抽出されていたものが 77，不要語の重みを半減することで新しく抽出されたものが 48，抽出されなくなったものが 30 あった．（図 5）

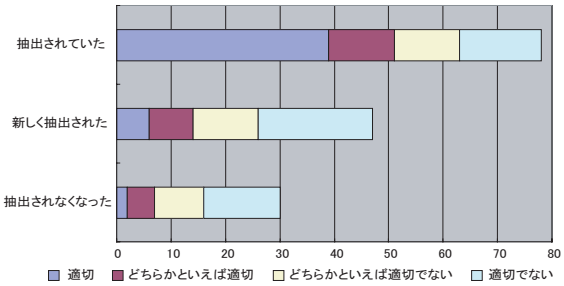

図 5: 不要語の重みを半減して抽出されたキーワードの評価結果

不要語の重みを半減する前から抽出されていたキーワードは多くが良い評価が得られた．反対に重みが半減して抽出されなくなったキーワードは大半が良い評価を得られなかった．不要語の重みを半減することで新しく抽出されたキーワードは良い評価が得られなかった．

不要語の抽出を抑えることで抽出結果の評価の向上が期待できることが確認できた．しかし，不要語の抽出を抑えることによるそれまで抽出できなかった重要なキーワードを新しく抽出することはできなかった．著者の付けたキーワードはまだ 44.6 %しか抽出できておらず,「融像域」,「PKI」などの抽出すべきキーワードが他にもあると思われるため，それらを抽出する方法を考える必要がある．

3 タームの重み付けの違いによる類似文書検索結果の評価実験

文書をタームの重みを要素値とするベクトルで表現し（ベクトル空間モデル [3]）ユークリッド距離や cos 類似度等の距離尺度を用いることによって，文書の分類や検索を行うことができる．文書間の距離が適切に求められれば妥当な分類・検索結果が得られる．ここでは，*cos* 類似度を用いて第 24 回医療情報学連合大会論文集（320 文書）を対象とした類似文書検索を行い，文書ベクトルのタームの重み付け方法の違いによる類似文書検索結果について評価，比較した．

日本語文解析システム ibuki を用いて形態素解析を行い「一般名詞」および「サ変名詞」を基底として各文書をベクトル化した後，主成分分析の近似計算方法である SimplePCA[4] によって次元圧縮を行った．ベクトル化の際のタームの重み付け尺度には「tf・idf」,「χ^2 値」,「χ^2 値× tf」の 3 通りを用いた．全 320 文書から評価対象として 10 文書を無作為に選び，これら各文書と残り 319 文書との類似度を求めた．各文書につき上位 10 位までを類似文書とし，10 人の評価者の主観により評価した．評価は類似文書１つ１つに

ついて，1 点〜4 点の点数（高いほど評価対象の文書と関連性がある）を付けてもらう形で行った．

tf・idf：$tf(t,d) * idf(t)$

$tf(t,d)$：文書 d での単語 t の出現頻度

$df(t)$：全文書中での単語 t の出現文書数

$idf(t)$：$log(\frac{\text{全文書の数}}{df(t)})$

χ^2 値：$\frac{(tf(t,d)-tf(t,d)\text{ の期待値})^2}{tf(t,d)\text{ の期待値}}$

「tf・idf」,「χ^2 値」,「χ^2 値×出現頻度」の 3 つの重み付けについて，評価対象の 10 文書の類似文書（各文書 10 文書ずつ）で 44 %（44/100）の文書が共通して抽出されていた．評価した類似文書の類似度と 10 人の評価の平均値を散布図に示す．(図 6)

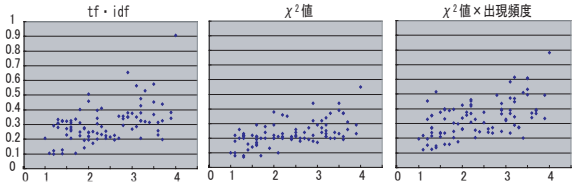

図 6: 各重み付けによる類似文書の類似度と評価

3 つの重み付け方法で類似度と 10 人の評価の平均値にはまずまずの相関関係が見られた．相関係数を求めたところ,「tf・idf」は 0.4841,「χ^2 値」は 0.5481,「χ^2 値×出現頻度」は 0.5375 だった．相関係数では「χ^2 値」が最も相関関係が高かったが，それぞれの相関係数にはあまり明確な差は見られなかった．

3 つの重み付けで共通して抽出された類似文書のそれぞれの重み付けによる類似度と 10 人の評価結果（平均値）を図 7 に示す．3 つの重み付けで共通して抽出された類似文書の類似度と評価の平均値による相関係数は「tf・idf」は 0.5192,「χ^2 値」は 0.5122,「χ^2 値×出現頻度」は 0.5249 であった．また，各重み付けで独立して抽出された類似文書の相関係数は「tf・idf」は 0.2134,「χ^2 値」は 0.4534,「χ^2 値×出現頻度」は 0.4374 だった.「tf・idf」で独立して抽出された類似文書の相関係数は，他の重み付けに比べて特に低く，文書間の距離を求めるには idf よりも χ^2値 のほうが適している可能性がある．

用いた「tf・idf」,「χ^2値」,「χ^2値 ×出現頻度」の 3 つの重み付けに関して，文書間の類似度を適切に求めることのできる特定の重み付けの有用性を見つけることはできなかった．idf よりも χ^2 値のほうが適している可能性があるが，今回は上位 10 位までの類似文書についてのみ評価を行っており，もっと下位の文書まで評価を行うことで重み付けの違いによる明確な差を確認できると思われる．また，ここでは ibuki の解析結果から「一般名詞」,「サ変名詞」を用いて類似文書検索を行ったが，2 で述べたように未知語，複合語を用いることでより明確な結果が得られるのではないかと思われる．

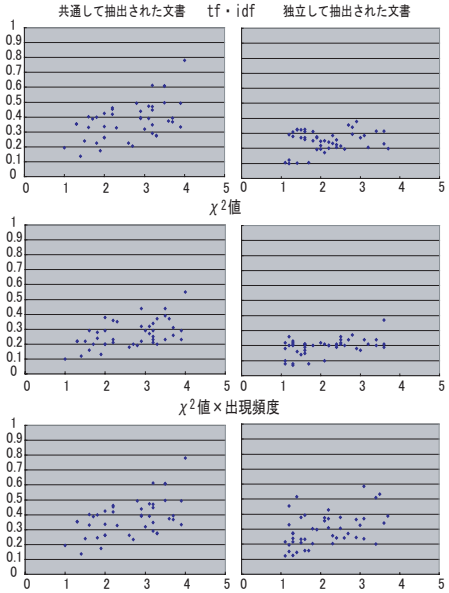

図 7: 各重み付けによる類似文書検索結果の共通部分と独立部分の類似度と評価

4 おわりに

医療情報分野の文献を対象として類似文書検索・キーワード抽出とこれらに対する評価を行った．キーワード抽出では，未知語を抽出対象とすること，複合語化を行うことの有用性を確認した．また分野に偏りのない文書集合から不要語を選定して抽出を抑制することの有効性を確認した．

類似文書検索では，3 種類のタームの重み付け方法のもとで得られた類似文書に対して評価を行った．3 種類の方法とも文書間類似度と評価者の主観との間に相関があることを確認できたが，重み付け方法の違いによる明確な差は見出せなかった．

本稿ではキーワード抽出の対象とする未知語として「カタカナ」「英文字列」を用いたが，複数の文字種からなる語は複合語化を行っても正しく切り出せない場合があるため，未知語領域の抽出方法は今後の検討課題である．さらに，複数の文書からのキーワード抽出に応用し，検索や分類によって得られる文書群の概要把握を支援することに役立てたい．

参考文献

[1] 日本語文解析システム ibukiC/S について　山田佳裕他，NLP2006

[2] 第 24 回医療情報学連合大会 第 5 回日本医療情報学会学術大会 論文集，2004

[3] 言語と心理の統計　金，村上，永田他，岩波書店，2003

[4] SimplePCA を用いたベクトル空間情報検索モデルの次元削減　黒石他，信学技報 NLC2001-17

[5] 退院サマリからの診断に関する言語情報の抽出　小野他，信学技報 PRMU2003-09