医療相談テキストにおける文のタイプ判別

菅谷 哉　山田 寛康　島津 明
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

現在, 様々な電子化されたテキストが大量に存在し, 膨大な量のテキストから自分が必要とする情報を効率よく得ることは非常に困難である. そのため, これらテキストの情報を整理したり取り出したりする事が必要である.

そのような背景のもとで, 人間の生存健康に重要な医療に関する情報を扱うための試みが行われている [2]. web 上には様々な医療に関するテキストが存在し, 例えば医療相談テキストがある. これは医師が病気や症状に関する質問に対して答えたテキストで, 病気に関する対処法や症状等の情報が含まれており, 病気について知りたい読者にとって有用な情報源である. これらの必要な情報を質問応答 (QA) 等の技術により取り出すことが期待される. QA は, 自然言語の質問文を入力とし, ある与えられた文書集合中から回答を出力するタスクである. ユーザの質問に, より的確な回答を提示するには, 対象テキストの各文が何について書かれているかを自動的に判定する必要がある. 意味制約を用いた手法による質問応答システム [5][7] では, 求める情報と手がかりとなる語とが位置的に離れている場合や, 質問に対する回答が概念で求められるような場合は意味制約だけでは不十分である. 例えば, 「〜はどんな症状か?」等のように文としての回答が求められる場合には, 回答候補を得るための文書集合の各文が何を述べる役割を果たすかを判断しなくてはならない.

QA への利用だけではなく, 文のタイプを判別し, どこに何が述べられているかの情報を知ることが出来れば, テキストの整理や, 検索, 分類, 読解支援等にも役立つと考えられる. 例えば, 判別した文のタイプ毎にテキストを整理して示すだけで, 読み手に対して, 欲する内容を見つけるための支援になると考えられる.

以上の観点から, 本報告では医療相談テキストの文が果たす役割を文のタイプとして定義し, その判別を行う手法を述べる.

2 テキストの分析

読売新聞社のホームページ [4] で公開されている, 健康相談の記事を分析した. これらの記事は読者の質問に対して, 医師が解答する形式である.

2.1 判別する文のタイプ

分析にあたって, 5つの文のタイプを定義した.

病名 「病名」を示すことを意図した文
症状 病気の「症状」を述べている文
対処・治療法 患者が取るべき対応策, あるいは医師が治療のために行う事柄が述べられている文
検査・診断法 病気と判断するに当たり医師が行う診断法や, 検査方法に関して述べられている文
原因 病気の「原因」が述べられている文

これらの文のタイプは, 医療相談テキストに於いて主として述べられる事柄であり, 質問の回答あるいは読者が必要とする情報として適切である. 特に病名, 検査・診断法については, 「病名」, 「検査・診断法」を表した用語の出現する文が考えられるが, ここでは文が果たす役割としてタイプを定義しているため, それらの用語が出現する文全てにこのタイプが付与されるわけではない. 文のタイプは, 1文に対し一つとは限らずに複数与えることを許す. これは例えば, 以下のような文が考えられるからである.

> 極端に噛み合わせが悪い場合には、先の治療により症状がなくなった時点で噛み合わせの修正を行います。

この文に対して, 最もふさわしいと考えられるタイプは対処・治療法である. だが, "どういった症状の時にこの方法をとるか"での「症状」が前提として挙げられている. このためこの文のタイプとして症状も可能であると考えられる. しかしながら, 主の役割としては対処・治療法と考えるべきであるので, タイプ付けを行う際にはその順位も考慮に入れる.

手がかり語とパターンの獲得に先立って, 人手によるタイプを付与する. これは先の読売新聞社ホームページの記事302テキストに対して行った.

3 判別手法

タイプ判別は, 手がかり語とパターン規則によって行う.

3.1 手がかり語による判別

手がかり語による判別は, その語の出現によってその文のタイプが推定できると考えられる語を選択する.

表 1: 選択された手がかり語の例

病名 (3 語)	病名，呼ぶ，一種
症状 (74 語)	痛み，合併，変形，状態，腫れる，低下，疾患，異常，訴える，激しい，続く，障害，転移，出血，失う，損なう，萎縮，麻痺，不全
対処・治療法 (104 語)	受ける，有効，予防，大切，矯正，望ましい，勧める，なるべく，避ける，重要，服用，理解，検討，心がける，矯正，かかりつけ，利用，納得，克服，見極める，訓練，管理
検査・診断法 (10 語)	反応，確定，音波，測定，造影，区別，画像，脳波，判定，採取
原因 (10 語)	感染，きっかけ，契機，副作用，引き金，要因，伝染，先天，後天，病因

腫瘍が
大きくなって
視神経を
圧迫するようになると、
視力視野障害を
起こします。

図 1: 係り受けの解析結果

手がかり語の選択は，各タイプ毎に出現頻度の上位２割の語から，人手によって選択した語である．選択基準としては，固有な症状名，病名，検査法，機器等の語は除外し，症状や病気の種類に関わらず用いられ，そのタイプと特定できる語を判断し選択した．その例を表 1 に示す．

3.2 パターン規則による判別

各タイプの特徴的な表現をつかむことで文のタイプ判別を試みる．ここでは，格情報と動詞に着目した係り受けパターンに着目し，その特徴的な文の言い回しを推定し，それを判別規則として与えることで文のタイプ判別を行う．そのために特定の動詞と文末表現，それらに係る格として「が格」,「は格」,「を格」,「に格」そして「で格」に着目したパターンを主に着目している．例えば,「腫瘍が大きくなって視神経を圧迫するようになると、視力視野障害を起こします。」を例に挙げると，係り受けは図 1 のように解析される．

動詞節「起こします」に着目すると,「圧迫するようになると、」および「視力視野障害を」が係っていることが分かる．接続助詞「と」以上の文節が,「を格」の視力視野障害を起こすことが捉えられる．このように，特定の動詞において，何が係っているか (この場合は接続助詞「と」で終わる節と「を格」) の規則を与える．

表 2: 与えるパターン規則 (代表例)

病名	は 病気です， という 病気です 病気は です これが です，
症状	は を 起こす すると しやすい， する 病気です， が 見つかる
対処・治療法	すべき， すれば， すると よい， の場合 する
検査・診断法	する 検査， で 検査， という 検査
原因	が を 起こす， が 影響， すると 起きる， で 起きる， するため 起きる， 起こす～は

規則の適用は，この係り受け関係の有無によって判断される．つまり他の節が係っているかは判定に影響せず，対象とする係り関係が出現すれば，該当タイプと判別する．

そのパターン規則の代表例を表 2 に示す．□が一つの文節を示し，その前後や間には任意の節が入ることを許している．

4 評価実験

評価として，判別規則を求めた３０２テキストによって Closed テストを行い，新たにテキストを収集し，Open テストによる評価を行った．Open テストの対象テキストには，先の読売新聞社の記事での分析に用いていないデータ，朝日新聞社 HP，健康・医療コンテンツ内の健康相談のテキスト [1] および，保健同人社の J-Health 内コンテンツ“健康相談 Q&A”[3] を用いた．その結果を表 3 に示す．

テキストの形態素解析には「茶筌」[8] を用い，係り受け解析には「南瓜」[6] を用いた．その結果から，タイプによるそれぞれの判別の有効性を検討し，各タイプ毎にどのような規則を得ることが有効であるかを検討する．

4.1 手がかり語による判別

病名タイプにおいては，精度が Closed で 19.2%，Open で 36.1%の精度となった．再現率は Closed,Open でそれぞれ 11.2%，12.0%となっており有効には機能し

表 3: 二つの手法による判別結果

	手がかり語			
	Closed		OPEN	
文のタイプ	精度 (%)	再現率 (%)	精度 (%)	再現率 (%)
病名	16/83(19.2)	16/135(11.2)	13/36(36.1)	13/108(12.0)
症状	1389/1962(70.8)	1389/1959(70.9)	745/1044(71.4)	745/1270(58.7)
対処・治療法	1756/2135(82.2)	1756/2124(82.7)	787/1247(63.1)	787/1013(77.7)
検査・診断法	29/131(22.1)	29/55(52.7)	19/53(35.8)	19/52(36.5)
原因	109/233(46.8)	109/540(20.1)	75/202(37.1)	75/333(22.5)

	係り受け			
	Closed		OPEN	
	精度 (%)	再現率 (%)	精度 (%)	再現率 (%)
病名	77/135(57.0)	77/133(57.9)	25/60(41.7)	25/108(23.1)
症状	626/788(79.4)	626/1959(32.0)	272/367(74.1)	272/1270(21.4)
対処・治療法	548/677(80.9)	548/2124(25.8)	237/324(73.1)	237/1013(23.4)
検査・診断法	34/57(59.6)	34/55(61.8)	10/24(41.7)	10/52(19.2)
原因	277/392(70.7)	277/540(51.3)	137/201(68.2)	137/333(41.1)

ていない．これは,「病名」タイプの文は,「一番考えられるのは、「停留精巣」です。」等短い文が多く，さらに助詞や助動詞が出現上位に現れやすいため，病名以外の内容語を手がかり語とすることが難しい．手がかり語として 3 語しか選択できなかったのもこのためである．

症状タイプにおいては Closed,Open でそれぞれ精度が 70.8,71.4%，再現率が 70.9,58.7%と，Open においても同等な判別力を示しているが，再現率に若干の低下が見られる．

対処・治療法タイプにおいては Closed,Open でそれぞれ精度は 82.2,63.1%，再現率は 82.7,77.7%で判別出来ている．

検査・診断法タイプにおいては，Closed,Open でそれぞれ精度は 22.1,35.8%, 再現率は 52.7,36.5%となった．これは，頻繁に用いられる名詞は主にサ変名詞 (検査，診断，判定，判断など) であるが，これらは対処・治療法タイプにも良く現れ，なおかつ対処・治療法タイプの方がタイプ自体の出現も高い．このため手がかり語と特定することが難しく，かなり数を絞ったが対処・治療法タイプの文を誤判別し，精度が著しく落ちた．

原因タイプにおいては，Closed,Open でそれぞれ精度 46.8,37.1%，再現率 20.1,21.4%となっている．症状タイプの文の誤判別が多く，精度に響いている．手がかり語として選んだ「感染」「伝染」という語は，原因タイプの文に現れやすいが，症状タイプにも現れることがあり，誤判別が多いためである．

4.2 パターン規則による判別

病名タイプにおいては，Closed,Open でそれぞれ 57.0,41.7%という精度を得た．パターンにおける「病名」の語が入ることが期待される場所に対して，それが「病名」であるかの固有表現の情報があればさらに精度を上げることが出来ると考えられる．

症状タイプにおいては，Closed,Open でそれぞれ再現率は 32.0%，21.4%と低くパターン規則がまだ網羅されていないことが分かる．精度については若干向上している．症状のみを述べている文に対しては概ね判別できているが，症状を前提とした他のタイプ (症状によっての対処の仕方等) の文に対して，効果が現れた一方で，メインのタイプとはなっていないため，文によっては症状の語が出現していて，症状タイプと判断された文が多く見られた．

早めがよいと言っても、ひんぱんな服用は、薬
そのものが頭痛を誘発する原因にもなります。

この文は，対処法と原因タイプが人手では与えられている．確かに「頭痛」は症状であるが，この文で述べたいことは,「薬の服用は早めが良い」ということと「頻繁な服用が頭痛の原因になる」という事である．

対処・治療法タイプにおいては，再現率が著しく低い．これは，パターン規則がまだ網羅されていない事が原因である．精度に関しても，手がかり語からの向上はない．

検査・診断法タイプにおいては,「検査・診断・測定」等の語に対してそれに係る節を主にしたパターンであるが，着目した語は対処・治療法にも頻繁に現れ，誤判別が多かった．

原因タイプにおいては，主に「起こる」等の動詞に着目した係り受けパターンによって判別を試みた．精度は Closed,Open でそれぞれ 70.7,68.2%であった．病気に関わらず原因とそれに伴う結果を含んだ文が判別され，また，“起こる”が用いられる語義が異なる場合の文も見られ，誤判別に繋がった．

表 4: 係り受けの判別を優先した場合の結果

	Closed		Open	
文のタイプ	精度	再現率	精度	再現率
病名	50.0	60.0	38.0	27.8
症状	73.8	68.6	73.1	56.3
対処・治療法	83.3	78.0	64.2	76.5
検査・診断法	31.6	76.4	34.5	36.5
原因	64.1	59.4	53.9	54.1

4.3 2つの判別手法の結果

係り受けパターン規則は手がかり語に比べて，概して精度が高い．手がかり語は再現率が高い．よってまず係り受けパターン規則によって判別を行い，判別できなかった文に対してのみ，手がかり語による判別を行う．これによって精度の向上を図る．

精度向上のために，各タイプにおいて何を考慮して判別を行えば効果的かを考察する．

- 病名タイプ
  文のパターンで，ある程度の精度を出すことが出来るが，さらに精度を上げるためにはやはり「病名」の用語の情報が必要である．Open テストにおいて再現率が大きく低下した原因の一つは，Closed 及び Open テストで用いたテキスト毎に言い回しが大きく異なり，適用可能なパターン規則が不十分だったためである．
- 症状タイプ
  動詞や係り受けパターンによる精度が高かったが，手がかり語によっても 70%程度の精度を得ている．文末表現や係り受け等のパターン規則を充実させること，また手がかり語など内容語による判別も合わせて行うことで，さらなる精度の向上が期待できる．
- 対処・治療法タイプ
  手がかり語による判別で，係り受けと同等な精度を得ている．このため文の構造を考慮せずに手がかり語など内容語によってのみ判別を行うことでも十分判別可能と考えられる．
- 検査・診断法タイプ
  病名タイプと並び，良好な結果が得られなかった．手がかり語のみの考慮では，再現率の向上は期待できるが，対処・治療法タイプとの誤判別はやはり多い．それを避けるためには，内容語とその出現パターンを同時に考慮すればいいと考えられる．また，病名タイプと同様，対象テキストのサイトの違いに対する言い回しの違いが大きい．
- 原因タイプ
  文中の節と節の関係が重要な指標になっていると考えられる．係り受けパターンや特定の動詞に着目することは有用であるが，原因と結果の因果関係をつかむ事にこだわりすぎると，病気の原因以外の文に対しても原因タイプと判別されてしまう．

5 おわりに

文のタイプ判別として，本報告では手がかり語による判別と係り受けパターンによる判別によって判別を行った．病名と検査・診断法を除き 6 割前後の精度を得，また再現率では Closed において 6 割前後，Open においては病名，検査・診断法タイプが特に低く，その他のタイプについては Closed と同程度の値を得，対象に左右されないパターン，手がかり語を決定できた．しかし，精度，再現率共にまだまだ向上の余地が多分にあり，特に，パターン規則は，与える規則がまだ不足している．各タイプにおける指針を元にさらにテキストを分析する必要がある．

参考文献

[1] asahi.com：生活：健康・医療 http://www.asahi.com/life/health/index.html

[2] Hirst, Graeme, Chrysanne DiMarco, Eduard Hovy Authoring and generating health-education documents that are tailored to the needs of the individual patient, the Sixth International Conference, UM97, 1997, pp.107–118

[3] J-Health：健康相談Q＆A http://www.so-net.ne.jp/familyclinic/jhealth/

[4] Yomiuri On-Line/医療と介護 http://www.yomiuri.co.jp/iryou/index.htm

[5] 賀沢英人，加藤恒昭，意味制約を用いた日本語質問応答システム，情報処理学会 自然言語処理研究会 2000-NL-140,pp173–180，2000.

[6] 工藤 拓，松本 裕治 日本語係り受け解析器「南瓜」version 0.32 使用説明書

[7] 村田真樹，内山将夫，井佐原均，類似度に基づく推論を用いた質問応答システム，情報処理学会 自然言語処理研究会 2000-NL-135,pp181–188，2000.

[8] 松本裕治，北内啓，山下達雄，平野善隆，浅原正幸，松田寛 日本語形態素解析システム「茶筌」version 2.0 使用説明書 第二版.