退院サマリ文章可視化システムの構築

荒牧英治† 三浦康秀‡ 外池昌嗣‡ 大熊智子‡ 増市博‡ 大江和彦+
†東京大学 知の構造化センター
‡富士ゼロックス (株) 研究技術開発本部
+ 東京大学 医学部附属病院

eiji.aramaki@gmail.com
{Yasuhide.Miura, masatsugu.tonoike, ohkuma.tomoko,
hiroshi.masuichi}@fujixerox.co.jp
kohe@hcc.h.u-tokyo.ac.jp

1 はじめに

平成 13 年度に政府が発表した「保健医療分野の情報化にむけてのグランドデザイン」にて，電子カルテシステムの普及が課題の一つとして掲げられて以降，我が国では急速に電子カルテが普及し，その結果，大量の臨床データが電子化された状態でストックされつつある．このデータをフル利用できれば，過去に類をみない大規模な統計的な臨床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，カルテ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．

このような背景から，本研究で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経過を要約した文書）の現病歴セクションを対象とし，そこから患者情報を抽出する手法を提案する．また，その応用例として，医師が一目で把握できるように退院サマリ文章を表形式に可視化するシステムを提案する．

本研究は，文章から何らかの情報を抽出するという観点からは，情報抽出の一種だと考えられるが，次の二つの特徴を持つ.（１）ドメイン：臨床医療文章を扱うためドメイン独自のアノテーションの枠組みが必要となる.（２）クローズ：対象となる退院サマリ文章は，医師間で閲覧されはするものの基本的にはクローズな文章であり，理解可能である場合は略語や箇条書きなど簡潔に記述されるため，これに対応する頑健な処理が要求される．

本稿では，本研究で扱う文章（退院サマリ）とそのアノテーションの枠組み，及び，提案するアプローチと予備実験結果を報告する．

2 コーパス

本研究で扱う退院サマリの現病歴セクションは患者の入院以前の経緯が簡潔に記された文章である．表 1 に一例を示す．例に示されるように，時間軸にそって，患者の状態，処置した内容，検査結果などのうち臨床的に重要だと判断されたものが記載される．我々は，現病歴文書において主要な情報は「いつ，何が起こったか」であると仮定し,（１）「いつ」に相当する時間情報,（２）「何が」に相当するイベント情報を分類／定義し，248 文章についてアノテショーンを行った．

2.1 時間情報アノテーション

時間情報については，汎用的な時間表現アノテーションの枠組みである TimeML (TIMEX3)*のサブセット（相対的な時間，繰り返し回数を除いた仕様）を用いた．

2.2 イベント情報アノテーション

イベント情報については，医師との議論を行い表 2 に示されるカテゴリに分けてアノテーションを行った．いくつかのカテゴリについて補足する．

病理所見 <P>: しばしば顕微鏡的検査結果である所見が記述される場合がある．多くの場合，ここま

*http://www.timeml.org/

で粒度の細かい情報は必要ないため，これらを区別した．

二つのレベルの検査 <TEST><T-NAME>: 「入院時検査」や「定期検診」といった複数の検査群からなるものと「ECG（心電図）」「心エコー」といった，具体的な検査値を持ちうるものの二種類がある．これらを区別するため，前者を<TEST>，後者を<T-NAME>とした．

匿名化 <DEID>: イベント表現とは別に，個人情報[†]に相当する表現をアノテートした．これにより自動匿名化を行うことが可能となる．

上記のカテゴリのどれに属すか曖昧な場合は複数のタグを許容した．例えば,「dynamic CT 施行したところ、富血性腫瘍の転移が疑われた。」という文における「富血性腫瘍の転移」は疾患表現でもあり，検査の結果でもある．このような場合，両方のタグを付与した．

また，各タグについて次の二つの属性を付与した．

1. **モダリティ**: この際,「リハビリを行った」といった事実と「リハビリの予定で」といった非事実（予定）などのモダリティを区別するため，表 4 のような属性を同時に付与した．

2. **レベル**: 明らかにあるカテゴリに属す用語と，文脈に依存する用語が存在する．これらを区別するためにレベルという属性を用意し，前者を level=1，後者を level=2 とした．例えば,「内視鏡検査で湿潤が疑われた」といった場合,「内視鏡検査」は明らかに検査であるので level=1 とした．一方,「湿潤」は，検査結果として解釈されるが，これは文脈に依存するため level=2 とした．

2.3 アノテーション例

表 3 に文「2001 年、舌癌手術時の腹直筋採取部位に疼痛を自覚し当院消化器内科に通院を開始。」に対して付与されるアノテーションを示す．表に示されるように，治療／処置表現「舌癌手術」と疾患表現「舌癌」，部位表現「癌」の入れ子を許した.「腹直筋採取部位」といった場合,「腹」「腹直筋」「腹直筋採取部位」いずれのレベルでも部位表現となるが，このような同

[†]個人情報の定義については HIPAA のガイドライン（http://www.hhs.gov/ocr/hipaa/）に沿った．

表 1: 退院サマリ（現病歴）の例.

2007 年 2 月 28 日〜 喉頭癌 T1N0 に対し RT66Gy(4/15 まで)。外来でフォローされていたが、6 月頃より披裂部の浮腫見られ、喉頭全体に発赤が見られるようになった。
2008 年 2 月 21 日 CT 撮影し、左仮声帯と声帯の腫脹あり、甲状軟骨の破壊（＋）。
生検勧められたが拒否していた。
4 月 14 日 近医に入院。翌日ラマによる生検施行し、左喉頭室、仮声帯、声門下より SCC(+)。
5 月 9 日 当院で喉頭全摘、lt.ND、気管前気管傍 ND、甲状腺亜全摘、永久気管孔作成術施行。その後咽頭皮膚瘻が形成され、6 月 20 日に左 DP 皮弁、大腿皮膚移植術施行。その後瘻孔は完全には閉鎖せず。
11 月 7 日 退院。
2009 年 1 月 9 日 再度近医に入院し、同日局麻下に咽頭皮膚瘻閉鎖術施行。1 月 27 日に退院している。以後外来フォローされていた。
5 月初め 気管孔右側の腫脹を自覚。
5 月 13 日 呼吸苦出現。近医受診し、CT にて気管孔周囲の再発が疑われた。手術できないと言われた。
5 月 25 日 当科受診。なお、この日より腫脹した気管孔右側より出血あり。嚥下困難、疼痛、発熱は見られていなかった。
5 月 31 日 当科入院。6 月 2 日〜7 日に CDDP120mg, 5-FU1000mgx4days の chemo 施行。Ccr の低下 (40 台)、低 K など見られたが腫瘍は著明な縮小をみた。
6 月 30 日 気管孔腫瘍切除、皮膚合併切除、ＤＰ皮弁による再建術施行。術後の経過は順調であり、7 月 13 日に全抜鉤、7 月 14 日に退院となった。

* この文章は（医師により作成された）ダミーのカルテ文章であり，現実の患者について述べられたものではない．

カテゴリで曖昧性がある場合は，最大範囲「腹直筋採取部位」をアノテートした．

他の例（ダミーカルテ）及びタグの詳細については http://lab0.com/med-nlp/guideline/を参照されたい．

3 提案システム

本システムは大きく４つの処理からなる．

STEP1: 表現の特定

まず，入力されたシステムの表現を特定する．このタスクは固有名詞特定と類似したタスクであるため，CRF [‡]を用いて行った．また，形態素解析エラーに対して頑健にするため，文字単位でラベル (IOB2 ラベル) を与えた．学習の素性には，医療用辞書を追加した形態素解析器（ChaSen）[§] の解析結果を用いた．

[‡]http://chasen.org/taku/software/CRF++/
[§]http://chasen-legacy.sourceforge.jp/

表 2: アノテーションされた表現カテゴリ.

M-NAME	医薬品名
M-NUM	医薬品の投与量および単位
TEST	検査群
T-NAME	具体的な検査名
T-NUM	検査値の検査値および単位
ACTION	退院，入院，転院など
A-LOC	ACTION の行われる場所
P	病理所見
CHANGE	変化に関する表現
R	治療処置表現. 治療，または治療行為を表す表現.（例）ギプス固定
D	疾患／症状表現: 疾患を表す表現.
DEID	匿名化すべき表現 （例： 医師名，施設名）
B	部位表現: 場所が特定できる部位を表す表現.（例）視神経

表 3: アノテーション例.

```
<TIMEX3 value=2001-XX-XX>2001　年</TIMEX3>、<R level=1>
<D level=1>　　<B level=1>舌</B>癌</D>手　術</R>時　の<B
level=1>腹直筋採取部位</B>に<D level=1>疼痛</D>　を自覚
し<A-LOC>当院消化器内科</A-LOC>に<ACTION>通院</ACTION>を
開始。
```

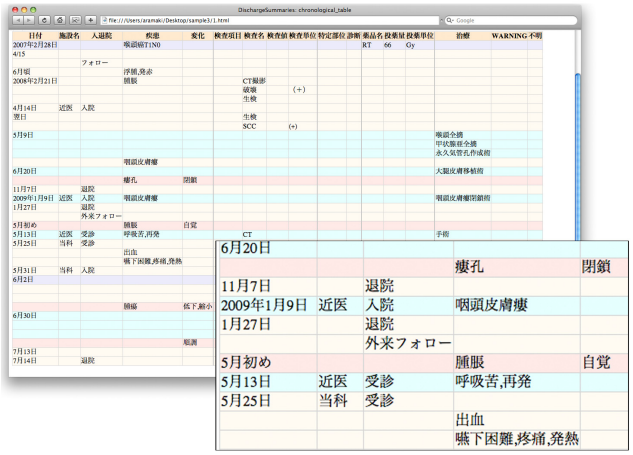

図 1: 可視化結果（全体図（左）と一部拡大図（右））.

STEP2: 正規化

次に，一部のカテゴリについて特定された表現を正規化する．現状のシステムは疾患表現<D>と日付表現<TIMEX3>について次の正規化を行っている．

疾患表現: 特定された表記と標準病名マスター [2] に記載された病名との表記ゆれを Aramaki 等 [1] の手法で吸収し，病名コード (ICD10[3]) を付与する．

時間表現: 省略された日付情報を補完する．現在は，人手で記述したルールによって行われる．

- 12 月 10 日 → 2008-12-10
- 昨年 12 月ごろ → 2007-12-XX

STEP3: TIME-EVENT 関係の特定

次に，特定された TIME 表現と EVENT 表現の間の関係を推定する．現在のシステムでは，i 番目に出現した時間表現が i+1 番目の時間表現までのイベントを支配する（スコープをもつ）といったアドホックなルールにより行っている．

STEP4: 可視化

最後に表形式の可視化（HTML 化）を行う．システムの出力例を図 1 に示す．

4 実験

4.1 実験設定と評価

今回，もっとも基礎となる処理である STEP1 の表現の特定について，予備的な実験を行った．実験は，10 交差検定で行い，Precision(P), Recall(R), F-measure(F) を調査した．

4.2 結果と今後の課題

実験の結果，表 5 の結果を得た．本システムの利用形態は二つある．（１）表形式に可視化することにより，臨床医が容易に患者の情報を把握する．（２）抽出されたデータをストックすることで，研究者がより大規模な統計的研究を行う，

（１）の観点からは，多くのカテゴリで 70-80 %の再現率を得ており，今後，コーパスを拡張することを考えると有望な精度だと考えられる．ただし，部位などバリエーションが多い一部のタグについては，低い精度にとどまり，外部リソースの利用が必要であろう．

（２）の観点からは，可視化した結果から，患者の経過が分かるかどうか（十分な情報を保存できているか）など実証的な評価が今後必要である．例えば，実際には以下のような文も出現する．

> **人工弁**を挿入していることもあり、**敗血症等細菌感染**を考慮し、**抗生剤投与**となった。

この文では,「**人工弁**」とそれが原因となってとられた処置「**抗生剤投与**」の関係を保存する必要があるが，現在の用語抽出のみに依存するシステムでは，これを表現できない．今後，この種のイベント間の関係（原因，結果など）についても整理が必要であろう．

また，今回の実験では，イベント表現の特定の部分のみの評価を行い，以降の処理については，評価を行っていない．現在の処理では，（１）時間表現の正規化，（２）時間表現とイベント表現の関係，などいくつかの部分で人手によるルールで行っているが，これらを緻密化することも今後の課題である．

5 まとめ

このような状況から，本研究ではカルテの一種である退院サマリを対象とし，そこから患者情報を抽出／可視化を行うステムを構築した．退院サマリをはじめ，カルテ文章は，医師間で閲覧されはするものの基本的にはクローズな文章であり，自然言語処理にとって，新しくチャレンジングな研究課題である．我々は，本研究をきっかけとし，今後，より多くの臨床文章が研究対象となることを望んでいる．

表 4: EVENT 表現のモダリティ．

事実	事実のとき（デフォルト）
半事実	「〜の疑いがある」「〜の可能性があり」「S/O〜」
未来	「〜予定で」「〜目的で」「〜のため」「〜の方針となる」
否定	「〜は認められず」
必要	「〜の必要性があり」
その他	「患者が〜を希望したため」

表 5: 結果.

tag	precision	recall	f-score
A-LOC	89.17	87.95	88.56
ACTION	94.63	91.04	92.80
B	68.87	59.39	63.78
CHANGE	84.64	74.89	79.47
DEID	82.76	77.13	79.84
D	85.56	80.24	82.82
M-NAME	86.99	81.34	84.07
M-NUM	95.99	93.81	94.88
R	84.50	76.36	80.22
T-NAME	84.74	76.68	80.51
T-NUM	80.76	76.34	78.49

参考文献

[1] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Orthographic disambiguation incorporating transliterated probability. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 48–55, 2008.

[2] K. Hatano and K. Ohe. Information retrieval system for japanese standard disease-code master using xml web service. In *American Medical Informatics Association (AMIA) Symposium*, pp. 597–602, 2003.

[3] WHO. *ICD10*. World Health Organization, 1992.