医学用語属性と構文情報を用いた診断報告書からの重要所見情報の抽出

今井　健† 荒牧英治† 梶野正幸‡ 美代賢吾† 小野木雄三††† 大江和彦†
†東京大学医学部附属病院 ‡臨床医学オントロジー研究会 ††† 東京大学大学院医学系研究科

{ken, aramaki, yonogi, kohe}@hcc.h.u-tokyo.ac.jp
kajino@medical-ontology.jp, miyo-sup@h.u-tokyo.ac.jp

1　はじめに

近年、臨床現場における電子化の伸展に伴い、画像診断報告書を始めとする診断報告書が、電子的に蓄積されている。画像診断報告書とは、検査依頼医からのオーダーを受けた放射線画像診断の専門医が、「ある部位に画像診断上の異常が発見されたか否か、あるいは状態・属性がどうであったか」という所見情報を自然言語文にて記述し依頼医に報告するものである。この報告書中で、「悪性病変が存在する／可能性が高い」という旨の所見情報は重要度が高く、また画像検査では広範囲な部位を見るため、依頼医が問題としている部位以外でこれらの悪性病変が発見されることもある。

医療安全管理上、報告書中で重要度の高い所見情報を抽出して依頼医にその存在を警告するシステムは、人命にも関わる、依頼医による読み落としを未然に防ぐ診断支援として、非常に有用である。我々はこれら重要所見の自動抽出システムの構築を目指している。

しかし、文中から悪性病変を示す表現を端的に抽出し依頼医に提示するだけでは十分でない。例えば「骨腫瘍」という表現があっても以下の A) のような文脈においては「摘出術後である」という【患者状態】の説明の一部に過ぎず、抽出対象とするべきではない。同様に、「肝転移を疑わせる」という表現であっても、B) のような文脈であれば Negative Study となる。

A)	右眼窩転移性骨腫瘍摘出術後
B)	肝転移を疑わせるＳＯＬの出現を認めない

ここで、再現率を重視し、依頼医にスクリーニングを任せるという方針は、依頼医の負担が大きく現実的でない。従って、高い適合率を実現した状態で再現率を向上させるという方針が望ましい。

近年の自然言語処理の研究において、情報抽出 (IE) はアメリカで開催された MUC(Message Understanding Conference) で扱われたタスクを始めとし、多くの研究者が取り組んでいる課題であるが、臨床現場における医用文書を対象とした情報抽出研究は未だ十分に行われていない。画像診断報告書からの所見抽出を試みた例として、筆者らは形態素解析結果に人手で構築したルールによる Chunking を行い、得られた文節の連接パターンに基づいたパターンマッチングを行った [1] が、１つの文中に複数の所見情報が存在する場合や、近接でない係り受け構造を考慮できておらず、再現率の向上に課題を残している。

本研究ではこれを改善するために、係り受け構造の利用を考える。所見情報は「肝内【解剖学的部位】に｜肝細胞癌再発【病名 or 所見要素】を｜認めません【主張句】」のように、特定の医学的属性を持った構成要素と、それらの間の依存構造から成り立っていると考えることができる。そこで、報告書中の文の依存構造木の中で、これら所見情報を構成する要素で構成される部分木を抽出することを考える。ここで、考慮しなければならないのは以下の点である。

長い複合語の存在：
臨床医用文書においては、長い複合語が頻出し、この特定が問題となる。また Chunking により複合語が取り出されたとしても、次に解剖学的部位、病名、所見要素、検査手技、患者状態といった医学的属性を決定する必要がある。
(1)「右眼窩転移性骨腫瘍摘出術後」
→ 【患者状態】
(2)「右大腿骨断端付近左側」
→ 【部位】
またこれらは文中で臨時一語的に生成されること

もあり、全ての語を形態素辞書へ登録することは現実的でない。

助詞の欠損、数値・記号表現の頻出：

報告書中の文章は「１．肝;直径２．５ｃｍの～」など、数値・記号表現が多く見られ、記号表現が助詞に代替していることも多い。また助詞の欠損も見られ、

「上腸間膜動脈左側リンパ節転移疑い」

は構文解析と Chuking 処理により 1 文節となってしまう。そのため、構文解析を行う前の前処理と、解析結果に対する修正が必要となる。

そこで、本論文では構文解析と Chunking 処理により基本的な依存構造木を得た後、再度各文節の自立語に対して医学用語辞書を用いた形態素解析を行い*、医学的属性付与と構文解析結果の修正とを行うという手順を踏む。図 1 にその概略を示す。以下、提案手法の詳細について説明し、その後評価実験の結果を示す。

2　材料と提案手法

2.1　材料

画像診断報告書中には、検査内容、比較対象情報、患者の状態、所見情報、リコメンデーション、付加的な情報、を記述した文が混在している。本研究ではコーパスとして、東京大学医学部附属病院における 2005 年 2 月～7 月の CT, MRI に関する画像診断報告書 22,496 件を対象とした。またこの中の所見情報カテゴリ文のうち表 1 に示す悪性病変キーワードを含むものを選択、ルール構築用セット 10,000 文、評価用セット 100 文に分割した。

表 1: 悪性病変キーワード

[悪性病変を端的に表す用語]
malignant, malignancy, cancer, carcinoma, recurrence, recurrent, melanoma, metastasis, glioma, lymphoma, leukemia, sarcoma, blastoma, cytoma, 悪性, がん, 癌, 再発, 黒色腫, 転移, 脳腫瘍, リンパ腫, 芽腫, 肉腫
[悪性腫瘍を表す略語]
HCC, AML, CML

*構文解析の段階で医学用語形態素辞書を追加すると構文解析ルールに悪影響を及ぼす可能性があるためである

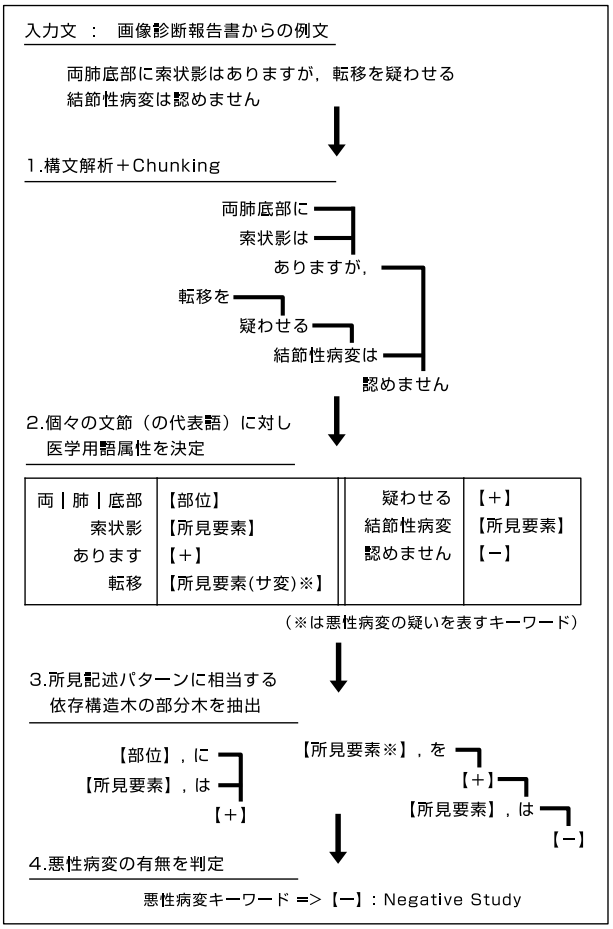

図 1: 処理の流れ

2.2　医学用語属性

文中の文節内自立語の医学用語属性を決定するため、既存の医学統制用語集を元に形態素辞書を整備した。使用したリソースは表 2 の通りである。また、これらの重複を整理し、さらに表 3 のような属性を付与した。ここで「所見要素」とは「画像診断上の発見物」(石灰化など) を指し、(1) の ICD コードで所見に分類されるもの、(4) の所見要素辞書を用いた。

表 2: 使用した医学用語リソース

(1)	ICD10 対応標準病名マスター v2.41	34,634 語
(2)	医学用語シソーラス第 5 版 内部作業データ	172,892 語
(3)	UMLS2005AA 中の日本語医学用語	54,274 語
(4)	過去のレポートから作成された画像診断領域辞書	2,773 語

*(1) http://www.dis.h.u-tokyo.ac.jp/byomei/
*(2) http://www.jamas.gr.jp/thesaurus.htm
*(3) http://www.nlm.nih.gov/research/umls/

表 3: 使用した医学用語属性

属性	(副)	語数	例
部位	2 種	9,043	肝臓
疾患	1 種	75,051	線維芽細胞性髄膜腫
所見要素	3 種	4,994	肺門部異常陰影, 胸水蛋白増加
検査手技	3 種	19,426	Ｔ２強調横断像, 二重造影像
修飾	8 種	777	右上, 肉芽性, 後部,
一般 (※)	5 種	106,140	βーアミロイド蛋白質

* (※) は今回の目的に対しては使用しない。(副) は副属性の種類で、「所見要素ーサ変」, 「所見要素ーナ形」などの細分類が存在する。

2.3 提案手法

Step1) 文節 Chunking と依存構造木の取得

まず、入力文に対して構文解析を行う。構文解析器としては KNP[3] を用いた。形態素辞書には、医療分野特有のサ変名詞 (137 語), ナ形容詞 (55 語) を追加し、KNP が並列解析のために内部で用いているシソーラスには、部位名、疾患名、所見要素名、検査名といった医学用語約 20 万語を追加した。

この段階で格助詞などの情報を用いた Chunking が成されるため、長い複合語であっても 1 語として特定することが可能となる。

Step2) 文節内代表語に対する医学属性の付与

次に、得られた文節中の代表語 (自立語) を特定し、JUMAN[2] を用いて再度形態素解析を行う。この際には医学用語特有の品詞設定を持つ医学用語約 20 万語を追加した形態素辞書を用いる。得られた形態素列から文節中の自立語の最終的な医学用語属性を決定する。基本的には Head Final なルールに従うが、最後が修飾語の連続の場合は、副属性とその直前の属性を用いて決定する。

(1) 右｜肺｜底部　　(修)(部)(修)　→　【部位】
(2) 右｜腎｜摘出｜後　(修)(部)(検)(修)　→　【患者状態】

次に、所見情報を構成する手がかり用言を考える。対象が用言の場合、所見の存在について肯定的（＋）否定的（－）かを決定した。手がかり用言として使用したものは 258 語であり、ルールセット中での頻度情報に従い、人手で作成した。

（＋）　165 語　認める, 認められる, 見られる, etc.
（－）　 93 語　認めません, 見られない, etc.

さらに、助詞の欠損に起因する、Step1) での過剰な Chunking に対し、用言となり得る箇所の分割を行う。文節末尾に上の手がかり用言が存在する場合、次の例のように分割した。

「古典的肝細胞癌あり」→ 【疾患】|【＋】

Step3) 所見情報に関する部分木の取得

次に、依存構造木を文末から逆に辿り、手がかりとなる主張句【＋／－】を起点とする部分木を全て抽出する。係り元文節が【＋／－】【部位】【疾患】【所見要素】【検査】で有る限り、再帰的に 1 つの部分木として結合する。その際に

1. 文末が【疾患】【所見要素】言い切りの場合は、【＋】をルートノードとして追加する。
2. 主張句から主張句に係る場合はそこで切断する。

図 1 の場合であれば、1 文から 2 つの所見部分木が得られることになる。

Step4) 悪性病変所見の存在の判定

最後に、各所見部分木の要素中に悪性病変キーワード (表 1) を含む場合に、これを悪性病変所見として抽出するかどうか決定する。

まず、悪性病変キーワード語が【疾患】【所見要素】である時のみ悪性病変所見の可能性があるとする。これにより例えば「癌」が【患者状態】や【検査】属性語に含まれていても抽出対象から除かれる。

次に、各々の悪性【疾患】【所見要素】の存在に関する肯定／否定の別は、意味的な反転を考えて付与する。すなわち、部分木の起点から到達するまでに経由した【＋】【－】の個数を用い下記の符号で考える。

$$\text{各要素の存在符号} = (+1)^{\#ofPlus} \times (-1)^{\#ofMinus}$$

図 1 の場合であれば、悪性キーワードを含む疾患・所見要素である「転移【所見】」は「疑わせる【＋】」に係っているが、部分木の起点から考えれば、＋, －と経由するため【－】属性で Negative Study となる。

3 評価実験

3.1 設定

提案手法の有効性を評価するため、所見抽出精度の評価実験を行った。評価用セット 100 文[†]中で、悪性病変キーワードにマッチした個々の語に対し、存在 (＋), 存在 (－), 抽出対象外 (0) の正解を人手で作成し、**(対象語) ＝「悪性病変キーワードを含む【疾患】or【所見要素】」**を肯定／否定の別まで含めて正しく抽出で

[†]今回は英文に対する処理を特に行っていないため、評価用セット 100 文は、日本語含む文のみを対象としている。

きるか判定する。ベースラインとして以下を用い、提案手法と比較した。

(BASE1)	**対象語**を全て肯定として抽出
(BASE2)	**対象語**＋直接の係り受けの符号
(PROP)	提案手法 (**対象語**を含む部分木と符号)

3.2 結果

得られた結果を表に示す。尚、正解とされた対象語, 肯定／否定のペアは 97 個である。

	抽出ペア	正答	Recall	Precision
(BASE1)	97	47	48.5%	48.5%
(BASE2)	74	58	60.0%	78.3%
(PROP)	81	74	76.2%	91.4%

提案手法の精度が上回っているものの高い再現率が実現できているわけではない。未抽出のものとしては「同定できない」など手がかりとなる主張句の整備が不十分であることが主な原因であった。また誤抽出は、「高分化ＨＣＣを疑われた結節は大きさに著変ありません」の係り受けが【高分化ＨＣＣ】→【＋ (疑われた)】→【結節】→【－ (ありません)】となることにより【－】と判定された事例などである。

3.3 考察

再現率の向上のためには手がかり主張句のさらなる整備が欠かせない。特に「否定する／ことは／困難である」等、複数文節にまたがる主張句を扱う手法が課題である。また今後は、少数であるが報告書のサマリーに出現する英文を扱う処理を構築し、これと併用する必要がある。本研究では試験的に悪性病変キーワードを設定したが、将来的には抽出された病名や所見要素を、国際的な疾病分類である ICD10 コードや大規模医療統制用語集である SNOMED-CT コードにマッピングすることで悪性の判定を行うことが望ましい。

4 関連研究

欧米では MedLEE[4] を始め診断報告書からの情報抽出研究が多数行われている [5][6] が、我が国で診断報告書を対象とした研究は分類や用語解析が主であり、所見抽出などの研究はほとんどない。特にその文体の特徴のため、構文情報を有効に活用したものは未だ存在しない。

本研究の特徴は医学用語属性に注目した依存構造木の部分木を用いたことと、さらに、助詞の欠損に起因する過剰な Chunking の修正により構文情報の効率的な利用を可能にしたことである。

5 まとめ

本研究では、画像診断報告書中の所見文から所見情報を抽出する際に、依存構造木から医学用語属性に基づく部分木に着目する手法を提案した。また予備実験によってその有効性が示された。現在我々は所見情報と医学用語属性を付与したタグ付き画像診断所見コーパスの作成に着手しており、今後より大規模な評価を行う予定である。

謝辞

医学中央雑誌刊行会に医学用語シソーラス構築作業用データを使わせて頂きました。また東京大学医学部附属病院放射線部の増本智彦氏には本研究の材料となる画像診断報告書の電子データを提供して頂きました。深くお礼を申し上げます。

参考文献

[1] 今井 健, 小野木雄三: 格フレームを用いた放射線読影レポートの文型分類と所見抽出. 医療情報学,24(Suppl.),pp.800-801,2004.

[2] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto and Makoto Nagao: Improvements of Japanese Morphological Analyzer JUMAN, In Proceedings of The International Workshop on Sharable Natural Language Resources, pp.22-28 (1994).

[3] Sadao Kurohashi and Makoto Nagao: KN Parser : Japanese Dependency/Case Structure Analyzer, In Proceedings of The International Workshop on Sharable Natural Language Resources, pp.48-55 (1994).

[4] Friedman C, Alderson P, Austin J, Cimino J, Johnson S (1994). A general natural language text processor for clinical radiology. J Am Med Inform Assoc 1(2), pp.161-174.

[5] Haug PJ, Ranum DL, Frederick PR (1990). Computerized extraction of coded findings from free-text radiologic report. Radiology 174, pp.543-48.

[6] Taira RK, Soderland SG, Jakobovits RM (2001). Automatic structuring of radiology free-text reports. Radiographics 21(1), pp.237-245.