病理診断報告書作成のための オントロジーを利用したテキスト入力支援

橋本 泰一[1)]　**TAM Wailok**[2)]　**鷹合 基行**[3)]　**荒牧 英治**[4)]　**宇於崎 宏**[5)]　**橋田 浩一**[6)]

[1)] **東京工業大学 総合プロジェクト支援センター**
[2)] **東京大学大学院 情報学環**
[3)] **富士ゼロックス株式会社 研究技術開発本部**
[4)] **東京大学 知の構造化センター**
[5)] **東京大学 医学部 人体病理学・病理診断学分野**
[6)] **産業技術総合研究所 社会知能技術研究ラボ**

[1)]**hashimoto.t.ab@m.titech.ac.jp**
[2)]**tam@is.s.u-tokyo.ac.jp**
[3)]**motoyuki.takaai@fujixerox.co.jp**
[4)]**aramaki@hcc.h.u-tokyo.ac.jp**
[5)]**uozaki@m.u-tokyo.ac.jp**
[6)]**hasida.k@aist.go.jp**

1 はじめに

医療への情報通信技術・知識処理技術の応用については学術的にも実践的にも様々な試みがなされている．近年は，Webやセマンティック Web [4] に基づく医療知識のモデリング [3,9] による知識情報サービスの開発が進んでいる．とりわけ，高度な医療情報を含む電子カルテなどから，大量のデータを持続的に収集・蓄積し，それを分析して得られる知見を活用して技術やサービスの高度化を図ることが重要である．医療に限らず，データの二次利用を普及させるためには，一次利用 (診断内容の確定や伝達) の生産性をも向上させ，データのライフサイクル全般にわたって価値を高めることが大きな技術的課題と考えられる．

我々は，医療に関するデータとして病理診断報告書に着目した．病理診断は多くの疾患における最終診断として重要な役割を担っている．しかし，病理診断報告書はテキストの手入力により作成されているため，表記の揺れ，入力ミス，意味的に曖昧な表現などを多く含む．病理診断報告書のテキストを入力する際に，計算機がその入力を動的に解析し，入力作業を支援 (修正，補完，予測) することができれば，入力ミス等を減らすことができ，かつ解析済のデータを作ることができ，データの一次利用と二次利用の両面において価値を向上させることができるだろう．加えて，テキストデータの構造化と画像データの構造化との連携によって診断の精度を向上させられる等の可能性もある．

これまでにオントロジーを用いた医療報告書の作成支援 [6,8] に関する報告では，テンプレート形式を採用している．しかし，テンプレート形式では多様な診断結果に対応した報告書を作成することが困難であると考えられる．本論文では，オントロジーと構文解析器および意味解析器を用いて，胃癌に関する病理診断報告書の作成を支援する入力支援に関して報告する．

2 病理診断報告書

本研究では，連結不可能匿名化された病理診断報告書の「診断名」と「所見」を対象とする．

診断名は，一般的には第1文で診断の結果すなわち病名を簡潔に記述する．次にその診断内容の詳細情報を記述する．診断結果が胃癌の場合はこの詳細情報は胃癌取扱い規約 [7] にしたがって記号や数値などを用いて表現する．最後に，備考を記述する．

診断名

Early cancer of the stomach, subtotal gastrectomy.
- Adenocarcinoma (tub1), type 0-IIc(5x4mm), pT1a(M), ly0, v0, pN0(0/8), pPM0(40mm), pDM0(38mm).
- ...

所見は，病理検体を観察し，その様子を詳細に記述する．一般的には，手術によって摘出された検体の場合，まず最初に摘出された検体全体の基本情報（部位，大きさ，質量など）を記述する．それに続いて肉眼で観察した際に発見された異常（病変）の個数や位置，状態を記述する．次に，顕微鏡を用いて細胞レベルで検体を観察した内容の記述を行う．所見には検体の詳細な状態を忠実に記述する必要があるため，自由入力を採用している．

所見

胃 ESD 検体．2.6x2.0 x 0.3cm 大．肉眼的に中心部に5x6mm 大のやや境界不明瞭な浅い陥凹病変を認める．

組織学的には...

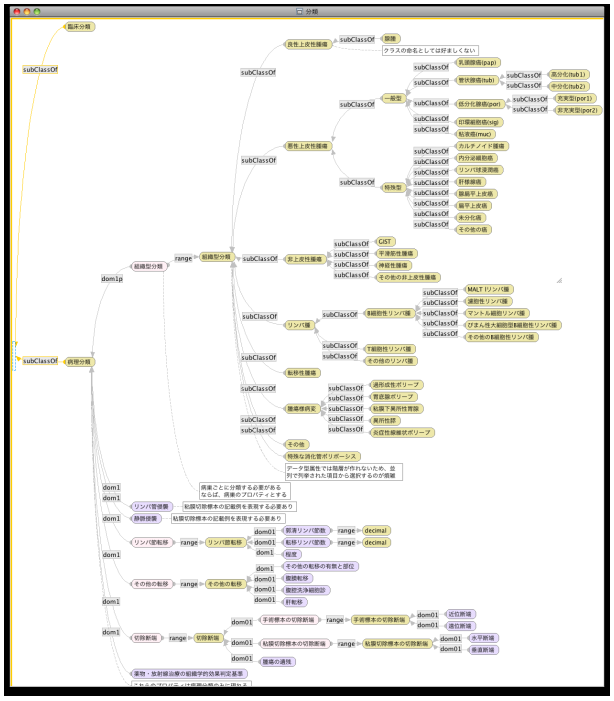

図 1: 胃癌取り扱い規約オントロジーの例

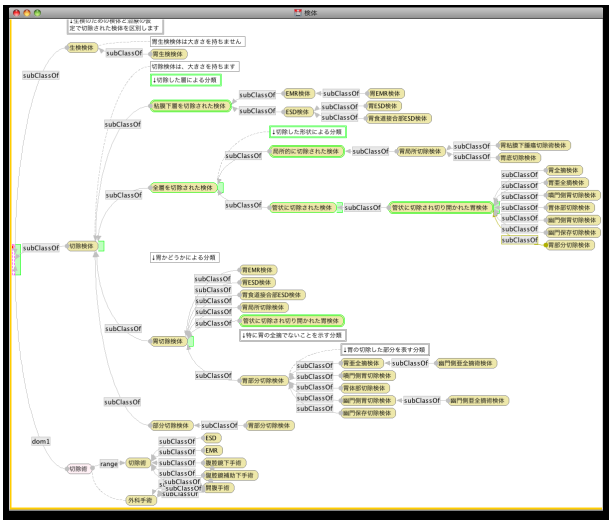

図 2: 胃癌病理所見オントロジーの例

3 病理診断報告書のためのオントロジー

病理診断報告書のためのオントロジーは，セマンティックオーサリングツールの一つであるセマンティックエディタ [2] を用いて作成している．セマンティックエディタとは，産業技術総合研究所社会知能技術研究ラボが開発しているツールであり，意味構造を明示できるコンテンツ作成を支援するグループウェアでもある．このツールを使ってオントロジーを作成できるだけでなく，産業技術総合研究所のデータベースサーバと通信することで作成したオントロジーをネットワーク共有することができる．これまでに絵文字の作成を支援するオントロジーの作成 [1] などで採用されている．

2010 年 1 月現在，下記の 2 種類のオントロジーを作成している．

- 胃癌取り扱い規約オントロジー（図 1）
- 胃癌病理所見オントロジー（図 2）

胃癌取り扱い規約オントロジーは，胃癌取扱い規約 [7] をベースとした胃癌の状態や治療記録に関するオントロジーである．胃癌取り扱い規約は，胃癌の進行度や治療・診断を記録するための，原発巣，転移や進行度といった腫瘍の状態と，手術や内視鏡治療の根治性や薬物の効果などの治療の評価を記録するための基本となる規則を示している．医学分野では国内・国際的な記録の共有化が重要であり，このような記録規則の標準化が積極的に進められている．

胃癌病理所見オントロジーは，病理診断報告書の所見に記述されたテキストの意味構造を表現するために，我々が人手で作成したオントロジーである．所見テキスト内には，胃癌取り扱い規約で定義された概念も含まれるため，胃癌病理所見オントロジーは，胃癌取り扱い規約オントロジーを内包している．

4 オントロジーを利用したテキスト入力支援

提案システムは，Web アプリケーションとして構成され，Web ブラウザ上でのテキスト入力の支援を行う．システムの概念図を図 3 に示す．本システムは，4 つのモジュールと 3 つの知識で構成され，外部のデータベースと知識マイニングモジュールと連携することを想定している．

ユーザがブラウザ上でテキストを入力すると，入力文字列をシステムへ送信する．システムは受信したテキストから辞書，文法，オントロジーを参照して，次に入力補完候補をユーザへフィードバックする．

4.1 Suggester

Suggester は入力テキストとブラウザ上でのユーザのアクションから入力予測に利用するモジュールを判断し，データを送信する．そして，モジュールから返ってきた入力補完候補をユーザに提示する．

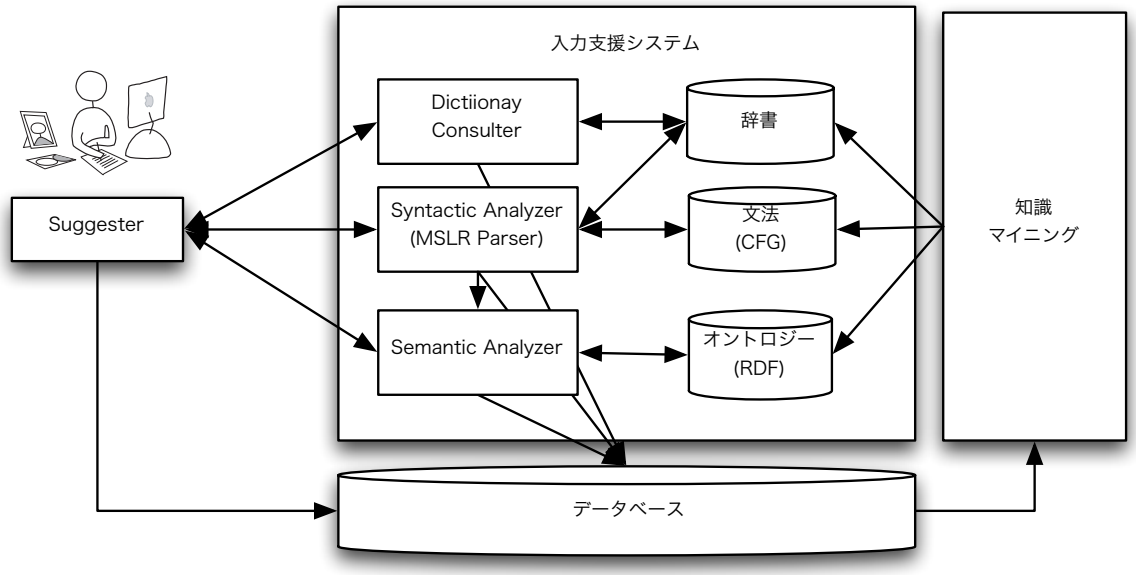

図 3: システム概要図

構文解析や意味解析といった高度な解析は大きな解析時間を必要とする．そのために，入力支援のような短時間での応答を期待されるシステムにおいて，常に高度な解析を利用することは難しい．また，構文解析や意味解析は，単語，節，文などの文法的な区切りで解析を実行しなければ，誤解析をしやすい．

そのため，Suggester では，

- 空白が入力された
- ユーザが入力補完候補を選択してテキストを入力した

などの条件を満たす場合には，構文解析器や意味解析器へ入力補完候補の推定を依頼し，

- 入力テキストを削除した
- 空白以外の文字列を入力した

などの条件を満たす場合には，Word Searcher へ入力補完候補の推定を依頼する．

4.2 Dictionary Consulter

Consulter は，ユーザの入力テキストの末尾と辞書登録されている単語の先頭の文字列マッチングをもとに入力補完候補を選定する．

4.3 Syntactic Analyzer

Syntactic Analyzer は，文脈自由文法（Context Free Grammar, CFG）をベースに形態素解析および構文解析を統合的に行う MSLR パーザ [5] を用いる．従来の MSLR パーザの違いとして，

- 正規表現を辞書に登録可能
- 空白などの規則ベースのスキップ処理の追加

が挙げられる．数字など辞書に単語登録しにくい表現の登録を容易にするために，辞書に正規表現を登録できるように拡張した．

実際の病理診断報告書を分析した結果，英語と日本語混じりのテキストが多く，文中に空白が挿入される場合が多い．また，空白の有無などによる表記ゆれも比較的多い．そのため，ステミング処理や形態素解析器などの処理を介して構文解析を行うよりも，余分な文字をスキップし，英語や日本語に関係なく分かち書きされてていない文として構文解析を行った方が効率的であると考え，MSLR パーザを採用した．

加えて，MSLR パーザは一般化 LR 法をベースにしており，文法を LR 表にコンパイルする．そのため，文脈に依存した先読み（入力補完候補）の予測が可能で，他の CFG ベースのパーザに比べ予測精度が高い．

4.4 Semantic Analyzer

Semantic Analyzer は Syntactic Analyzer によって解析された構文木とオントロジーをもとにして，文章の意味構造を構築する．次の「L,Less,Type 2, 50 × 20 mm, tub1 > tub2, pT2, int, INFb, ly1, v1, pN1(2/13),

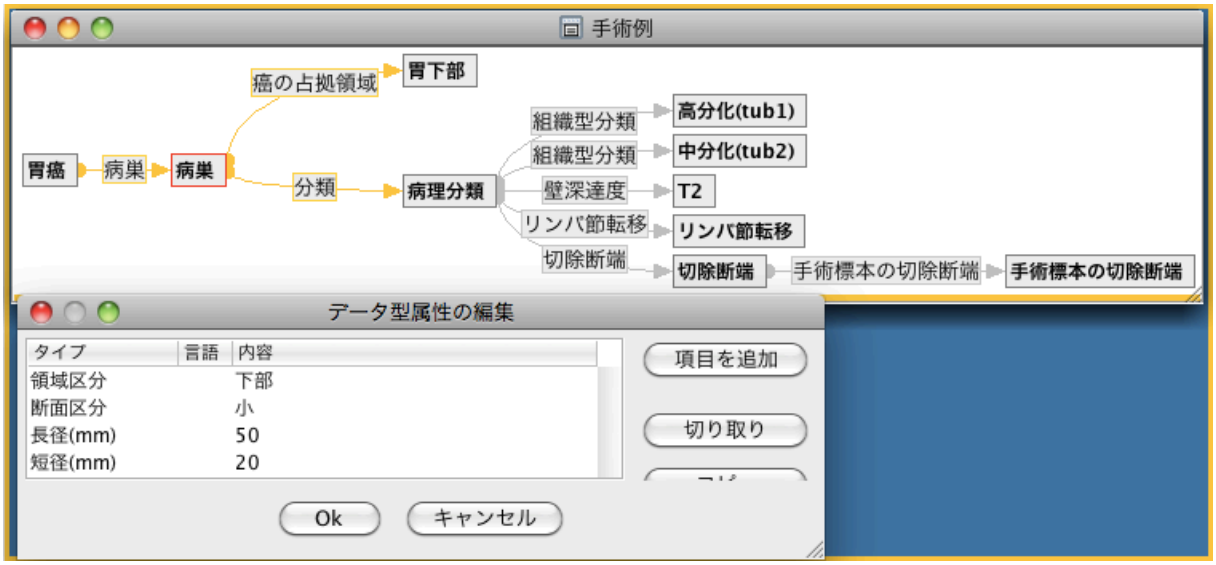

図 4: 「L,Less,Type 2, 50 × 20 mm, tub1 > tub2, pT2, int, INFb, ly1, v1, pN1(2/13), pPM0(40mm), pDM0(12mm)」の意味構造の例

pPM0(40mm), pDM0(12mm)」に対する意味構造の例を図 4 に示す.

4.5 データベースと知識マイニング

本システムは，持続的に大量のデータを収集し，多角的・科学的に分析して得られる知見を収集・活用する目的で，入力テキストのログ，構文木，意味構造，解析ログなどをデータベースへ保存する．そうすることで，意味構造付きのテキストが保存できる．構文解析および意味解析で失敗したテキストも同様に保存され，辞書，文法，オントロジーの改良・拡張のための材料となる．このように，正しい文章を入力することを支援するとともに，高度な知識の改善の材料を蓄えることを想定する．

5 まとめ

人々の日常の生活や業務を通じてデータを蓄積・分析しながら仮説を構築・反証・改良し，価値を創造し続けるという科学研究の実践を社会全体に広めるために，医療における病理診断報告書のテキスト入力を支援するシステムを提案した．従来のオントロジーを使ったテンプレート形式による入力支援と異なり，提案システムはテキスト入力時に構文解析器および意味解析器を動的に行いながら入力を支援することが大きな特徴である．また，入力支援を行うにあたり，胃癌取り扱い規約オントロジーと胃癌生検所見オントロジーの 2 種類の病理診断のためのオントロジーの作成を行っている．

今後の課題として，辞書，文法，オントロジーといった知識のさらなる拡張が必要である．また，本システムの入力支援の定性的・定量的な評価を行い，実際の医療現場への導入を目指す．さらに，文法やオントロジーをはじめとした知識を継続的に変更や改善をするためのメンテナンス方法についても検討が必要である．

参考文献

[1] 伊藤一成, 橋田浩一. 絵文字の作成と理解を促進するためのオントロジーマッピング. 電子情報通信学会技術研究報告. DE, データ工学, 第 106 巻, pp. 145–150, 2006.

[2] 産業技術総合研究所社会知能技術研究ラボ. セマンティックエディタ. http://i-content.org/semauth/intro/.

[3] 藤田伸輔, 今井健. Snomed-ct と icd-11 に見る医学・医療分野の ready to use ontology. 人工知能学会誌, Vol. 25, No. 4, pp. 501–508, 2010.

[4] 神崎正英. セマンティック・ウェブのための RDF/OWL 入門. 森北出版, 2005.

[5] 白井清昭, 植木正裕, 橋本泰一, 徳永健伸, 田中穂積. 自然言語解析のための MSLR パーザ・ツールキット. 自然言語処理, Vol. 7, No. 5, pp. 93–111, 2000.

[6] 長谷川雪憲, 松村泰志, 三原直樹, 川上洋一, 笹井浩介, 武田裕, 中村仁信. 胸部写真読影における医学的知識に基づいたレポーティング支援システム. 第 26 回医療情報学連合大会論文集, pp. 1117–1118, 2006.

[7] 日本胃癌学会（編）. 胃癌取扱い規約. 金原出版株式会社, 第 14 版, 2010.

[8] 川上洋一, 安永晋, 笹井浩介, 稲田紘, 木内貴弘, 黒田知宏, 坂本憲広, 竹村匡正, 田中博, 玉川裕夫, 仲野俊成, 朴勤植, 平松治彦, 松村泰志, 宮本正喜. 症例データベースから抽出した医学的知識のレポーティングシステムへの応用. 第 25 回医療情報学連合大会論文集, pp. 962–965, 2005.

[9] 大江和彦, 今井健. 臨床医学知識処理を目指した医療オントロジー開発. 人工知能学会誌, Vol. 25, No. 4, pp. 493–500, 2010.