特　集　「アクティブマイニング」

利用者からの要求を考慮したテキストデータからの知識抽出

Goal-Oriented Knowledge Extraction from Scientific Text

新保　仁 Masashi Shimbo
奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology.
shimbo@is.naist.jp

松本　裕治 Yuji Matsumoto
（同　上）
matsu@is.naist.jp

山田　寛康 Hiroyasu Yamada
北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology.
h-yamada@jaist.ac.jp

Keywords: statistical natural language processing, active mining, text mining, Medline.

1. はじめに

インターネットによる情報革命は，過去には考えられないほどの情報の氾濫をもたらしたが，体系化されず未整理の情報の多くが言語による文書（テキスト）データとして存在している．人間がある目的で必要とする情報は，ほかの情報の中に埋没しており，これを適切な時期に適切な形で取り出すことは容易ではない．言語によって記述された情報の中から利用者の欲する情報を取り出し，適切な表現で提示することができれば，利用者は自分の目的に無関係な情報に煩わされることなく，情報の有効利用を実現することができる．中でも特に，医学分野の文書データからの情報抽出は，evidence-based medicine [Sackett 00]の観点からも有用性が高い．

我々は，このような現状認識に基づき，医学・生物学分野文書に含まれるさまざまな知識を，各利用者の要求に応じて抽出すること（文書データからのユーザ指向アクティブマイニング）を目指し，そのための要素技術の開発を行った．図1は，本プロジェクトの俯瞰図である．図に示すとおり，機械学習に基づく自然言語解析結果にテキストマイニング技術を組み合わせ，文書構造解析や，知識抽出といったタスクに応用する．本プロジェクトは，

（1）自然言語処理の基礎技術
（2）科学技術文書処理に特化した応用技術

の開発の2種類に大別でき，これらはさらに下の具体的なサブテーマに細分できる．

（1a）専門用語を含むシステムにとっての未知語の識別と処理技術の開発
（1b）科学技術論文中の自然言語文の自動解析精度の向上
（2a）論文アブストラクトの文書構造解析に基づく各文の文書内での役割解析
（2b）科学技術論文からの知識抽出のための手がかり語発見支援

2章以降は，これらについて概観する．なお，何れのサブテーマにおいても，高度に専門的な内容の文書を対象として想定していることから

- 一般の自然言語処理システムにとって未知である専門用語が頻出し，
- 訓練リソース作成の負荷が極めて高い，

といった対象文書の特徴を念頭に研究を行った．

2. 自然言語処理基礎技術

2・1　未知語を含む英文文書内の単語の品詞推定

一般の自然言語処理システムにとって，辞書は欠かせない構成要素であるが，すべての単語や用語が辞書に記述されている，というのは非現実的な仮定である．ことに，医学・生物学分野など，高度な専門性が要求される分野の文書には，一般辞書には含まれない語が多く出現し，品詞などの文法情報の特定に支障をきたす．分野別の専門用語辞書を用いることで問題はある程度軽減されるが，専門性の高い分野には次々に新しい用語が出現するため，専門用語が完全に網羅されていると考えるのはなお早計である．そこで，前後の文脈，あるいは単語の綴り（特に接頭，接尾表現）を手がかりとして，未知の単語の品詞を決定し，それによって，専門用語と考えられる名詞句の同定を柔軟に行うことを試みた．

具体的には，Support Vector Machine（SVM）を用

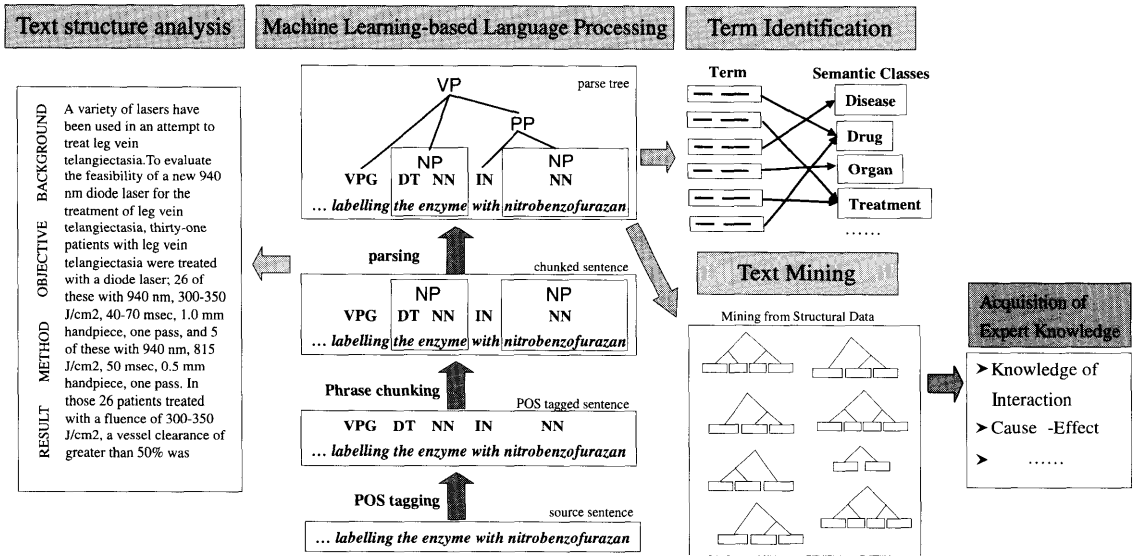


図1 文書データからのユーザ指向アクティブマイニング:プロジェクト概観

いた未知語の品詞推定，および文全体への品詞タグ付けを行う手法を開発した[Nakagawa 01]．しかし，問題が非常に巨大であるため，SVM を単純に適用しただけでは，実用的な計算時間で動作させることが困難であることがわかった．そのため，第一の処理として，従来型の N-gram に基づく統計的品詞付与システムによる学習を行い，そのシステムが誤りを生じる箇所を SVM によって学習するという 2 段階の構成に基づく処理手法（“修正学習モデル”）を提案した[Nakagawa 02, 中川 03]．これにより，英文文書に対して約 97 %という高い解析精度を維持しながら，実用的な解析時間で未知語を含む専門文書中の単語の品詞推定が可能となった．

2・2 日本語文書における未知語同定処理

日本語の医学・生物学文書処理を想定すると，英語の場合と同様，（システムにとって未知語である）専門用語の出現箇所を同定することが重要であることに変わりはない．ただし，日本語では単語間に明確な区切りがないため，処理内容は大きく異なったものとなる．

我々は，未知語の出現によって解析誤りが生じる場所を正確に同定するため，文の形態素解析結果の単語列を文字単位に分解して，形態素結果から得られる種々の属性を各文字に付与し，文字単位で未知語のチャンキングを SVM を用いて行う方法を提案した[Asahara 04a]．さらに同じ手法が，文中の未知語だけでなく，種々の固有表現の同定についても適用可能であることを示し，実験により従来のどの手法よりも高精度の結果を得た[浅原 04b]．

2・3 文中の基本句の自動抽出

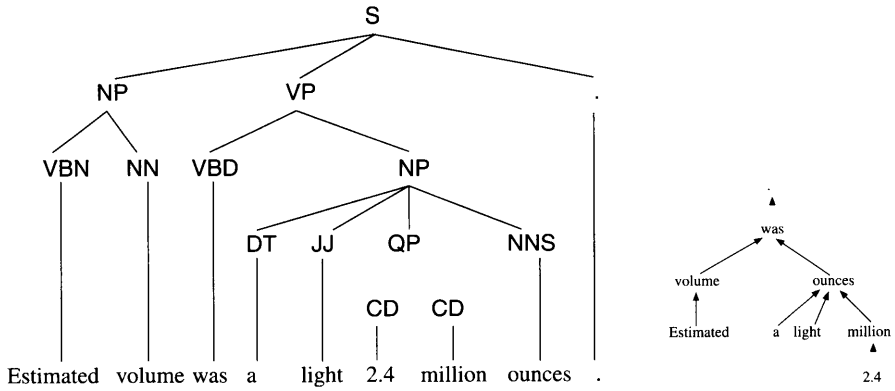
品詞付与が行われた文書に対して，そこに現れる名詞句や動詞句などの基本句を精度良く同定し，それらの間の文法的な係り受け関係を解析することは，その後の言語処理の性能に影響を与える．特に名詞句と動詞句の間の係り受けのような基本的な関係の同定は，専門用語の抽出にとって必須の処理である．本研究の一環として，英語の基本句の同定を SVM を用いた学習システムの混合モデルにより行い，高い解析精度が達成できることを示した[Kudo 01, 工藤 02]．

2・4 係り受け解析

言語の表現形式やスタイルは分野や応用によって異なることが多いため，統計的構文解析をさまざまな分野のテキストに適用する場合，適用分野ごとにある程度の構文解析済み訓練データが必要となる．

英語を対象とする統計的自然言語処理システムの場合，訓練用統語情報つきコーパスとして Penn Treebank [Marcus 94]が広く用いられているため*1，こういったシステムを特定分野文書に適合させるためには，訓練データとして，その分野の文書に Penn Treebank 準拠のタグを付与する必要がある．しかし，Penn Treebank には品詞タグ以外に粗く分類しても 26 種類のラベル（句）からなる複雑な句構造解析情報がタグ付けされている．したがって，同様の構文解析済み訓練データを作成（タグ付け）する作業者には，単に母国語話者として英語を理解できるだけではなく，句構造文法に関する専門的な知識が要求される．また適用分野が，新聞記事のような一般的な分野ではなく，医学生物学分野のような専門性の高い分野では，文法に関する知識に加えて，さらにその分野に関する専門知識を必要とする可能性が高く，タグ付けを行うことができる人の数はさらに限定される．しかも，複雑な句構造のタグ付けは，複数のタグ付け作業者間での一貫性を維持管理することが非常に困難であるが，

*1 Penn Treebank コーパスは新聞記事を中心に収録されている．



(a) 句構造木　　(b) 依存構造木

図2　文 "Estimated volume was a light 2.4 millon ounces." に対する（a）句構造木，および（b）依存構造木

一貫性を欠くタグ付けデータは，学習の際のノイズとなり，解析精度を低下させる主な原因となる．このような状況を考えると，さまざまな分野のテキストに対して，Penn Treebank のような複雑な句構造をタグ付けし，訓練データを準備することは現実的でない．

一方，単語間の係り受けよりなる依存構造は，一般の母国語話者にとって句構造木よりもはるかに直観的に理解しやすい．係り受け依存構造は，単語間の修飾関係で文の構造を表現した簡潔なものであり，タグ付け作業者に求められる用件および負担は大きく軽減される．さらに係り受け依存構造の簡潔さゆえ，タグ付け者間で揺れが小さく，専門分野においても一貫性の保たれた高品質の訓練データが作成可能になると期待できる．例として，同一の英文の句構造木と依存構造木を図2に示す．

以上の考察に基づき，単語間依存関係が付与された学習データを想定し，直接依存構造を学習する高精度な依存構造解析器を構築を試みた[山田 04]．我々が提案する依存構造解析手法は，各状況に対して3種類の手続き（Left，Right，およびShift）を繰り返し適用し，決定的に依存木を構築していく*2．

図3，図4，および図5にそれぞれの手続きの適用例を示す．図において破線で囲まれた2ノードが現在注目している解析位置を表す．また図の上方が手続き適用前を表し，下方が手続きを適用した結果を表す．3種の手続きの動作は以下のとおりである．

Right　隣接した2ノード間で，左ノードが右ノードに係るという依存関係を構築する（図3）．

Left　隣接した2ノード間で，右ノードが左ノードに係るという依存関係を構築する（図4）．

Shift　解析位置を一つ右へ*3移動する．この時点では依存木を構築せず，次のノード間の解析に移る（図5）．

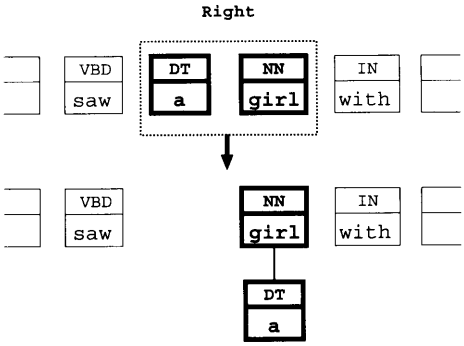


図3　手続き 'Right'

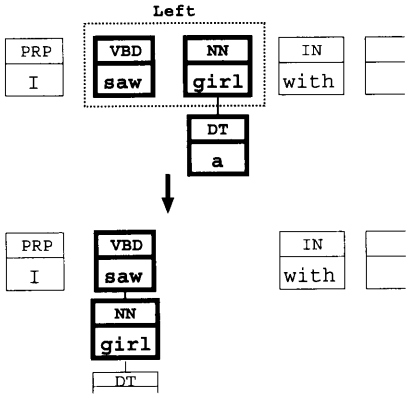


図4　手続き 'Left'

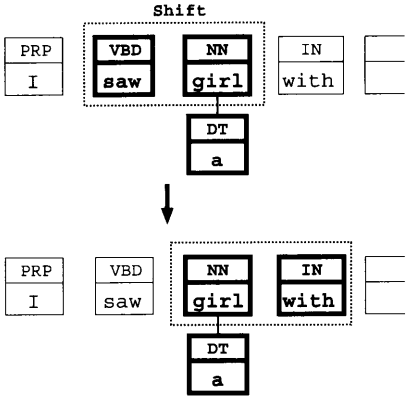


図5　手続き 'Shift'

*2 手続きをより細分化したモデル（4 ～ 6種類の動作モデル）に関する実験も行っている．詳細は文献[山田 04]を参照のこと．

*3 右向きの解析と左向き解析の何れの解析も可能であるが，ここでは右向き解析を仮定している．

解析の各段階で三つの手続きのうち何れを適用するのが適切かの判断は，SVMによって行う．すなわち，それぞれの手続きの適用可否を判定するSVMを訓練データから構築し，one-versus-the-rest法を用いてSVMの判定結果を組み合わせて一つの手続きを選択する．

以上の基本手続きを用いて，与えられた文の文頭から係り受け解析を行う具体的な手順は以下のとおりである．

(1) 文頭から順に，各ノード対に対して，one-versus-the-rest法によって選択したした手続き（Left, Right, Shift）を順次適用する．

(2) ステップ（1）で選択された手続きがすべてShiftであるか，残り1ノードになれば終了．さもなくば，再び文頭に戻ってステップ（1）を繰り返す．

Penn Treebankを主辞規則を用いて単語間係り受けによる依存木に変換し，これの一部を用いてSVMを訓練し，残りのデータによる評価を行った．その結果，係り先精度で約91％という結果を得た．

2・5 文中の基本句の自動抽出および係り受け解析の高性能化

前節までに提案した手法は，高次元の多項式カーネルを用いたSVMを用いることにより，従来法を上回る解析精度を達成することができたが，大規模な文書データに対しては，訓練時だけでなく解析時にも非常に長い実行時間を必要とするという問題があった．例えば，従来のルールベースの固有表現抽出システムに比べて，SVMとカーネル法を用いたシステムは，精度は高いとはいうものの，実行時間が数百倍遅いという報告がある．これを克服するために，学習過程で得られたサポートベクタ事例にデータマイニングの手法を適用することで，それらに共通に現れる属性情報の有効な共起情報を抽出し，選択された共起属性を単独の属性とみなすことによって線形カーネルの利用を可能にする手法を提案した[Kudo 03, 工藤 04]．結果として実行速度を数十倍から数百倍に向上させることができ，現実的な大規模文書データに対して高性能の言語解析を適用するめどが立った．

3. 科学技術文書からの知識抽出に向けて

3・1 テキストからの専門用語の抽出と分類

テキストから知識抽出を行うためには，単に個々の用語がどの品詞に属するか，という判別だけでなく，それらのより詳細な意味クラスまでをも判定する必要が生じる．我々は，医学・生物学関係の論文アブストラクトを提供しているMedlineデータベース[Med 03]を題材とし，テキスト中に現れる名詞句を対象にして，その意味クラスを推定する実験を行った．ある程度教師なしの手法を実践することを目指し，コーパスの精度は落ちるものの，次のように人手による作業をなるべく少なくしつつ，大規模なデータに対して，特定の意味クラスに属する語を推定する実験を行った．

既存の専門用語辞書の意味分類とそれに含まれる用語を学習データとし，そこに現れない未知の用語の意味クラスを自動的に推定を試みた．既存の専門用語シソーラス（MeSH [MeS 03]）の上位5種類の意味クラス（病気，治療法，薬品など）を対象に，未知の名詞句に対し，それがどのクラスに属する用語であるか，自動推定可能かどうかを確認することを目的とした．MeSHで定義されている上記意味クラスの用語のうち，曖昧性のない用語を対象にして（すなわち，MeSH上で複数意味クラスに登録されているものは除外し），Medlineアブストラクトでの出現データを収集し，その一部を訓練データ，残りを評価データとして実験を行った．用語の内部情報（単語および単語の接頭・接尾文字列）と前後の文脈情報によってどの程度用語の意味クラス推定が可能かを確認した．この際，文脈情報としては，固定長の隣接単語を見るのではなく，係り受け依存構造木*4を活用し，判定すべき単語と係り受け関係にある単語を特徴量として用いた．これによって，該当語と遠距離にあるがなお推定に有用な単語を手がかりとして用いることが可能となる．例として，“the epidemiology of human monocytotropic ehrlichiosis”という名詞句に含まれる，単語ehrlichiosisの意味クラスを判定する状況を考える．Epidemiologyは，ehrlichiosisが病名であることを強く示唆する用語であるが，隣接3単語以内を見ただけではその情報を取り出すことはできない．一般に隣接何単語を見れば必要な情報が得られるか，は状況に依存し一概にはいえない．また，観察する語数を大きくすることは，データスパースネスという新たな問題を引き起こす．これに対して，名詞句を係り受け解析することによって，ehrlichiosisが前置詞ofを介してepdiemiologyに係ることがわかるが，このように係り受け木構造上の近さを活用して素性として用いる単語を選択することができる．

表1は，文脈素性の表現方法を変えた場合の，SVMによる用語の意味クラス推定性能を示している．性能はF値（精度と再現率の調和平均）によって評価した．‘係り受け’欄は，判定したい単語に直接係る単語，および（前置詞を介して）該当単語が係る単語を素性として用いた場合の性能である．意味クラスChemicals and Drugs

表1 文脈情報の違いによる単語意味クラスの判定性能（F値）

		周辺N単語			
意味クラス	係り受け	$N=2$	$N=3$	$N=4$	$N=5$
Anatomy	**0.592**	0.514	0.514	0.510	0.498
Organisms	**0.462**	0.385	0.409	0.419	0.417
Diseases	**0.645**	0.516	0.546	0.552	0.544
Chemicals and Drugs	0.581	0.549	0.576	**0.584**	0.581
Techniques and Equipments	**0.498**	0.423	0.454	0.469	0.455

*4 2・4節の実験で用いたのと同様の方法で，句構造木から変換して作成した．

を除き，固定長隣接単語よりも，係り受けを用いた場合の性能が上回っていることがわかる．

さらに，単語の内部情報および係り受け木から抽出した文脈の双方の情報を用いてSVMによる用語の意味クラス推定を行ったところ，約70～88％の精度が得られた[Shimbo 02]．特に，病名か否かの判定に限れば，精度約90％，再現率約80％で判定が可能であった．

3・2 論文アブストラクトの文章構造解析

論文アブストラクトから利用者が所望の情報を検索する際に，アブストラクト内の各文が，そのアブストラクトの構成上，どのような役割をもっているかがわかれば，情報の絞り込みに役立つと考えられる．利用者が積極的に，文の役割と具体的な単語を指定して論文検索を行うことも有用である．そのような観点から我々は，論文アブストラクト中の文の役割を五つ（「背景」，「目的」，「実験方法」，「実験結果」，「結論」）に大分し，これら五つのクラスにアブストラクト中の文を分類する実験を行った．

このような検索システムの実現に向けて最大の問題は，文への役割タグ付与作業にかかるコストである．文の役割を絞り込んだ検索を実現するためには，データベース中のすべての文について，どのような役割をもっているか，というラベルが付与されている必要がある．Medlineの全アブストラクトの各文に役割ラベルを人手で付与することは，そのデータ量を考えると現実的ではなく，このラベルを自動的に付与する方法を考えなければならない．機械学習手法の適用が考えられるが，この場合にも訓練データ（役割ラベルが付与された文）は必要であり，その構築コストもなお無視できない．訓練データ構築コストを低減するために，我々は，Medlineに収録されたアブストラクト文のうち，“Background”，“Objective”，“Method”，“Conclusion”など，文書の構造を示すラベルが付けられている形式をもつ，いわゆる“構造化アブストラクト”[Ad Hoc Working Group for Critical Appraisal of Medical Literature 87]を訓練データとして活用することを考えた．そのために，構造化アブストラクトの役割ラベルとの対応について調査し，実験を通じて，このタスクに必要な属性情報について検討した．さらに，自動分類された文の役割を利用して，論文アブストラクトの検索を行うプロトタイプシステムを作成した（図6）[Shimbo 03]．

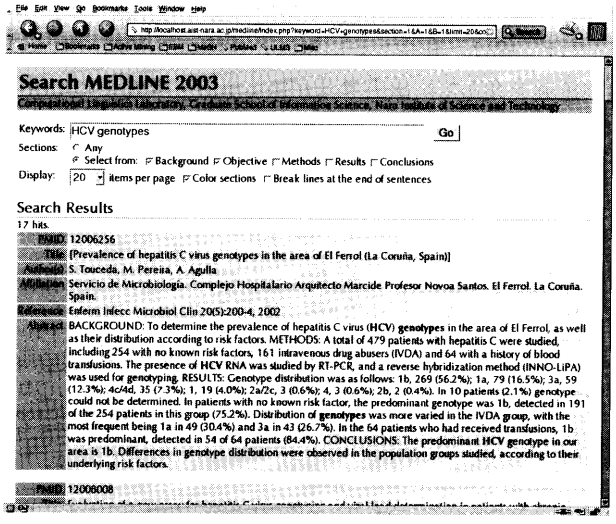


図6 検索システムプロトタイプのスクリーンショット

3・3 知識抽出のための手がかり表現発見支援

大規模なテキストコーパスの入手が容易になるにつれ，大量のテキストに埋もれている有用な情報をすばやく手に入れたい，という要求が高まっている．

Medline収録のアブストラクトから知識を抽出する，という研究は過去にも存在するが，これらの先行研究に共通するのは，抽出対象の知識がアブストラクト中でどのように記述されているか，があらかじめ（人手によって）決定されている点にある．この手がかりとなる記述パターンは，獲得したい知識の種類・分野に大きく依存するため，知識抽出を行うためにはまず，有効な‘手がかりパターン’を発見しなければならない．手がかりパターンの発見には人手の介在が多かれ少なかれ必要となるが，（人手による選別対象となる）候補パターンの数をいかに効率良く絞り込むか，という問題は先行研究の何れにおいても扱われていない．

我々は，そのために自然言語処理技術やテキストマイニング技術の適用を試みた．これまでに開発した係り受け解析システムによって，文の主語，主動詞，目的語などの主要素を高い精度で抽出することが可能なため，因果関係を表すことが可能な動詞を主として対象にすることにより，規則の自動獲得の可能性を探求した．具体的には，肝炎の検査項目に関する言い回しをマイニングし，出現頻度の高い言語表現を列挙する[Shimbo 04]．肝炎の検査項目を表す語句を入力として取り，係り受け解析器を用いて解析した文の係り受け木構造を利用して，

- （検査項目を含む名詞句）（動詞表現）（名詞句）
- （名詞句）（動詞表現）（検査項目を含む名詞句）

という表現を抽出し，名詞句，動詞表現，それぞれの頻度について調査した．この際，係り受け構造を活用し，不要な修飾語を取り除くことで，些細な表現の揺れを吸収した非連続表現を抽出することが可能である．例として，図7に示す，文“A stepwise increase in serum ADA level was observed with increasing severity of liver cirrhosis.”の係り受け依存木に対する解析の概略を述べる．まず，あらかじめ与えられた検査項目名である‘ADA’から，動詞（この木では‘was’）に出会うまで木の根に向かって上昇し，名詞句‘increase in ADA level’を得る．次に動詞‘was’を出発点に，上昇してきたほうとは反対側（右側）の枝を下降するが，まず品詞情報を手がかりに，動詞表現‘was observed with’のまとめあげを行う．さらに‘severity of liver cirrhosis’,

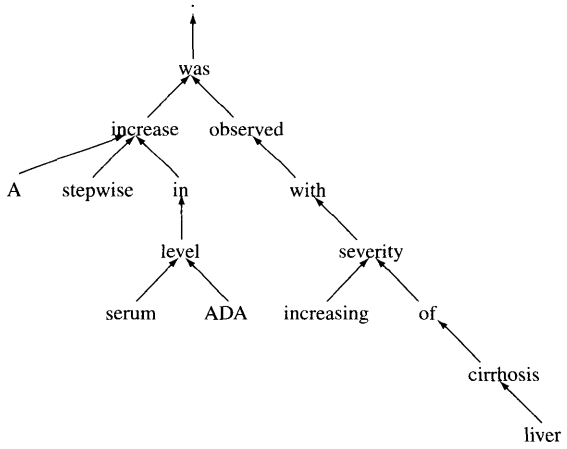


図7 "A stepwise increase in serum ADA level was observed with increasing severity of liver cirrhosis." に対する係り受け木

'increasing severity' といった名詞句が得られる.

このようにして得られた名詞句，動詞表現を頻度順に列挙し取捨選択することで，効率的に有効な手がかり表現を抽出することが可能となる．Medline で 'hepatitis'（肝炎）という単語を含むアブストラクトを対象に本手法を適用したところ，特に，動詞表現の効率的な収集に有効であった.

4. むすび

医学・生物学分野文書からの知識抽出処理のためには，(i) 専門用語が頻出し，(ii) 訓練リソース作成の負荷が極めて高い，といった一般文書の自然言語処理とは異なった問題に対処する必要がある．我々は，これらの問題の解消を念頭に，主として Medline に収録の論文アブストラクトを対象として，未知語（専門用語）とその意味クラス推定，基本句同定，単語間係り受け解析などの言語処理を行うための基本システムを作成した.

今回開発した未知語処理を伴う文書の品詞推定，および基本句へのまとめ上げプログラムは，現在発表されているシステムの中では最も高い精度を示しており，現時点では十分優れたものであるといえる．しかし，それぞれのシステムは，現在入手可能な解析付きコーパスである Penn Treebank という主に新聞記事よりなるデータから学習したものであり，今回解析対象とした医学生物学分野の論文アブストラクトとは内容が大きく異なる．今後，医学分野の解析付きコーパスを蓄積することによって，より精度の高い解析を行える可能性がある．また，今回高精度化した英語の単語および句の間の係り受け解析の結果を利用して専門用語推定の実験を再試行することを考えている.

謝辞

本稿で報告した一連の研究は，科学研究費補助金特定領域研究（B）"情報洪水時代におけるアクティブマイニングの実現" の一環として行ったものである．研究に協力いただいた，浅原正幸，工藤 拓，玉森彩弥香，中川哲治，山崎貴宏の各氏に深く感謝する.

◇ 参考文献 ◇

[Ad Hoc Working Group for Critical Appraisal of Medical Literature 87] Ad Hoc Working Group for Critical Appraisal of Medical Literature: A proposal for more informative abstracts of clinical articles, *Annals of Internal Medicine*, Vol. 106, No. 4, pp. 598-604 (1987)

[Asahara 04a] Asahara, M. and Matsumoto, Y.: Japanese unknown word identification by character-based chunking, *Proc. 20th Int. Conf. on Computational Linguistics*, pp. 459-465, Geneva (2004)

[浅原 04b] 浅原正幸, 松本裕治：日本語固有表現抽出におけるわかち書き問題の解決, 情報処理学会論文誌, Vol. 45, No. 5, pp. 1442-1450 (2004)

[Kudo 01] Kudo, T. and Matsumoto, Y.: Chunking with Support Vector Machines, *Proc. 2nd Meeting of North American Chapter of Association for Computational Linguistics* (*NAACL*), pp. 192-199 (2001)

[工藤 02] 工藤 拓, 松本裕治： Support Vector Machine を用いた chunk 同定, 自然言語処理, Vol. 9, No. 5, pp. 3-22 (2002)

[Kudo 03] Kudo, T. and Matsumoto, Y.: Fast methods for kernel-based text analysis, *Proc. 41st Annual Meeting of the Association for Computational Linguistics* (*ACL 2003*), pp. 24-31 (2003)

[工藤 04] 工藤 拓, 松本裕治：カーネル法を用いた言語解析における高速化手法, 情報処理学会論文誌, Vol. 45, No. 9, pp. 2177-2185 (2004)

[Marcus 94] Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B.: The Penn Treebank: Annotating predicate argument structure, *Proc. Human Language Technology Workshop*, pp. 114-119 (1994)

[Med 03] Medline, U. S. National Library of Medicine. `http://www.nlm.nih.gov/databases/databases_medline.html` (2003)

[MeS 03] MeSH: Medical Subject Headings, U. S. National Library of Medicine. `http://www.ncbi.nlm.nih.gov/mesh/` (2003)

[Nakagawa 01] Nakagawa, T., Kudo, T. and Matsumoto, Y.: Unknown word guessing and part-of-speech tagging using Support Vector Machines, *Proc. 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS2001*), pp. 325-331 (2001)

[Nakagawa 02] Nakagawa, T., Kudo, T. and Matsumoto, Y.: Revision learning and its application to part-of-speech tagging, *Proc. 40th Annual Meeting of Association for Computational Linguistics* (*ACL-02*), pp. 497-504, Philadelphia, PA, USA (2002)

[中川 03] 中川哲治, 工藤 拓, 松本裕治： Support Vector Machine を用いた形態素解析と修正学習法の提案, 情報処理学会論文誌, Vol. 44, No. 5, pp. 1354-1367 (2003)

[Sackett 00] Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W. and Haynes, R. B., eds.: Evidence-Based Medicine: *How to Practice and Teach EBM*, Churchill Livingstone, second edition (2000)

[Shimbo 02] Shimbo, M., Yamada, H. and Matsumoto, Y.: Using syntactic dependency information for classification of technical terms, *Proc. 7th Pacific Rim Knowledge Acquisition Workshop* (*PKAW 2002*), pp. 131-143, Tokyo, Japan (2002)

[Shimbo 03] Shimbo, M., Yamasaki, T. and Matsumoto, Y.: Using sectioning information for text retrieval: a case study with the Medline abstracts, *Proc. ISMIS Workshop on Active Mining* (*AM-2003*), pp. 32-41, Maebashi, Japan (2003)

[Shimbo 04] Shimbo, M., Tamamori, S. and Matsumoto, Y.: Finding cue expressions for knowledge extraction from scien-

tific text: early results, *Proc. 9th Pacific Rim Knowledge Acquisition Workshop* (*PKAW 2004*), Auckland, New Zealand (2004)

[山田 04] 山田寛康, 松本裕治：Support Vector Machine を用いた決定性上昇型依存構造解析, 情報処理学会論文誌, Vol. 45, No. 10, pp. 2416-2427 (2004)

2005年1月9日 受理

著者紹介

新保 仁（正会員）

1992年京都大学工学部電気工学第二学科卒業. 1994年同大学院工学研究科修士課程電気工学第二専攻修了. 1997年同大学院工学研究科博士後期課程情報工学専攻指導認定退学. 2001年より奈良先端科学技術大学院大学情報科学科助手. 博士（工学）.

松本 裕治（正会員）

1977年京都大学工学部情報工学科卒業. 1979年同大学院工学研究科修士課程情報工学専攻修了. 同年電子技術総合研究所入所. 1984～85年英国インペリアルカレッジ客員研究員. 1985～87年（財）新世代コンピュータ技術開発機構に出向. 京都大学助教授を経て, 1993年より奈良先端科学技術大学院大学教授, 現在に至る. 工学博士. 専門は自然言語処理. 情報処理学会, 日本ソフトウェア科学会, 言語処理学会, 認知科学会, AAAI, ACL, ACM 各会員.

山田 寛康

1997年山梨大学工学部電子情報工学科卒業. 1999年同大学院工学研究科博士前期課程修了. 2002年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了. 同年北陸先端科学技術大学院大学助手. 現在に至る. 博士（工学）.