C-11

自然言語処理による医療情報の読解支援

A Patient Support System based on Statistical Natural Language Processing

平尾 努* 磯崎 秀樹* 須藤 克仁* 鈴木 潤* 塚田 元* 藤野 昭典* 永田昌明*

Tsutomu Hirao, Hideki Isozaki, Katsuhito Sudoh, Jun Suzuki, Hajime Tsukada, Akinori Fujino, Masaaki Nagata

1 はじめに

近年，病に関する情報，すなわち，薬や治療法に関する情報を Web を利用して得ることが簡単になってきた (www.nytimes.com/2008/09/30/health/30online.html)．特に，英語を母国語とする人々であれば，以下のようなサイトから最先端の情報を簡単に得ることができる．

- 患者間のコミュニティサイト (たとえば，Patinets-LikeMe)
- WHO，NIH，NCI，FDA など公的機関のサイト
- PubMed(医学生物学分野の論文アブストラクトのレポジトリ)

しかし，多くの日本人は言語の違いもあり，このような英語で提供されてる情報にアクセスすることが困難であり，最先端の医療情報を手に入れることが難しい状況におかれている．

そこで，我々は英語を苦手とする患者 (あるいは，その身近な人々) が最先端の医療情報に手軽にアクセスできるようなシステムの実現を目指している．

2 研究位置づけ

英語で記述された最新の医療情報へ患者がアクセスすることを手助けする試みとしては，米国国立がんセンター (National Cancer Institute: NCI) の Physician Data Query (PDQ) を和訳した「がん情報サイト (cancerinfo.tri-kobe.org)，PubMed に掲載されている論文アブストラクトを和訳した「海外癌医療情報リファレンス (www.cancerit.jp/xoops/)」などがある．こうした情報は患者にとって非常に有益であるが，運営者側からは，翻訳などの様々な処理を人手でによって行っているため，多大なコストがかかるという問題点がある．

そこで，我々は，様々な自然言語処理技術を適用することによって先述した問題を解決し，英語で書かれた医療情報に患者がアクセスすることを手助けするシステムの実現を目指す．具体的な初期目標は，英語文献 (論文のアブストラクト) を対象とし，日本語で検索し，その結果も日本語で表示することとする．

なお，患者にとっては，公的機関のレポートや論文で発表される情報だけでなく，患者自らが発信するような情報もまた有用なケースが多々ある．しかし，このような情報に対しては，しばしば信憑性が問題とされる．そこで，本稿では，情報の信憑性が高い医学論文のアブストラクト (具体的には PubMed) を用いることとした．他の情報源に関しては今後，拡張することを検討したい．

また，生物医学ドメインを対象とした自然言語処理は，以前より盛んに行われており，近年ではワークショップにおいて shared task も開催されている (compbio.uschc.edu/BioNLP2009/index.html)．また，様々なツール類も NaCTeM (National Center for Text Mining in the UK: www.nactem.ac.uk) などが公開している．しかし，これらの研究は，医者あるいは研究者が利用することを想定しており，我々のように一般的な患者が利用することを想定したものではない．

3 システムの構成

現在のシステムは以下のモジュールにより構成される．

1. 辞書に基づく言語横断検索
2. 専門用語抽出
3. 修辞構造解析
4. 階層的な句構造に基づく統計的機械翻訳

*日本電信電話株式会社, Nippon Telegraph and Telephone Corporation

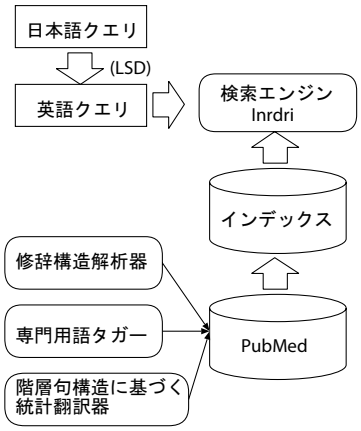

図 1: システム構成図

システム構成図を図 1 に示し，以下，それぞれのモジュールについて述べる．

3.1 言語横断検索

検索対象が英語文書であるため，検索要求として入力された日本語を英語へと変換する必要がある．この処理には，京都大学の金子周司氏らによって開発された医学分野の対訳辞書である Life Science Dictonary: LSD (`lsd.pharm.kyoto-u.ac.jp`) を用いた．さらに検索エンジンには Indri (`www.lemurpproject.org/indri`) を用いた．

入力された検索要求は LSD に基づき英単語に翻訳される．この時，ある日本語の表現に対し，複数の英語表現がある場合，それらすべてを検索エンジンへの入力とする．たとえば,「タミフル」に対しては,「Oseltamivir」,「Tamiflu」で文書を検索する．検索結果である文書は後に説明する階層的な句構造に基づく統計機械翻訳器により翻訳される．

なお，辞書に基づく検索語の翻訳は精度が高い反面，新語や辞書に登録されてないエントリに対応することができないという問題がある．これに関しては，後に説明する統計翻訳における対訳コーパスを拡充することで対処可能であると考える．

3.2 専門用語抽出

医学分野の専門用語抽出は，自然言語処理分野でも盛んに研究が行われているいわゆる固有表現抽出と同等である．これは，文中の単語に対し，以下に示すタグを付与する問題としてとらえることができる．

...	the	monoterpene	alchols	,	geraniol	...
O	O	B-subst	I-subst	O	B-subst	O

`O` は専門用語ではないことを表すタグ，`B-sbst` は専門用語 (この場合，物質名) の始まりを表すタグ (`B` は begin の略)，`I-sbst` は専門用語の途中を表すタグ (`I` は inside の略) である．上記の例では,「monoterpene alchols」という 2 単語,「geraniol」という 1 単語がそれぞれ物質名を表す．

このように単語にタグを付与することは一種の分類問題であり，自然言語処理の分野では多くの場合，教師あり学習が用いられることが多い．以前は，個々の単語に対し，独立にタグを予測する問題として，SVMs (Support Vector Machines) [Vapnik 95] がよく用いられていたが，近年では，タグを独立に予測するのではなく，タグ系列を予測する問題として CRFs (Conditional Random Fields) [Lafferty 01] が用いられることが多い．

ただし，このような教師あり学習の枠組みにおいて，精度を担保するためは，大量の訓練データが必要となる．さらに，訓練データに偏って学習される可能性もある．こうした問題を軽減するため，我々は，CRFs に半教師あり学習を取り入れた手法を採用した [Suzuki 08].

Penn BioIE コーパスの CYP450 ドメイン (`bioie.ldc.upen.edu/publication/latest-release/`) から 73,108 文を訓練，8.137 文をテスト，アンラベルドデータとして，PubMed から 10 万アブストラクトを用いて，物質名の抽出精度を評価した結果，教師あり CRFs を用いた場合の F 値は 0.897，半教師あり CRFs を用いた場合の F 値は 0.905 であった．

3.3 修辞構造解析

科学技術論文のアブストラクトは明確な修辞構造を持っている場合が多く，実際，PubMed に登録されているアブストラクトの一部には,「論文の目的 (`<OBJECTIVVE>`)」,「実験手法 (`<METHOD>`)」「実験手法 (`<METHOD>`)」「実験結果 (`<RESULTS>`)」「論文の結論 (`<CONCLUSION>`)」というタグが予め付与されているものもある．

このように文書の修辞構造が明らかであると，検索や読解支援として非常に有益である．たとえば，検索時には，検索語が出現する領域を指定することができる．ある病に関して有効な薬や治療法を知りたいのであれば，`<OBJECTIVE>`に病の名前を含み，かつ，`<CONCLUSION>` に「統計的有意である」と書かれている文書を検索してくれば，それらの多くは検索要求に適合するであろう．

表 1: 修辞構造解析の精度

	Hirohata et al.	提案手法
per sentence	0.943	0.951
per abstract	0.629	0.677

また，文書を読む際，実験の詳細について知りたいと思わないのであれば，<OBJECTIVE>，<CONCLUSION>領域を読むだけで良いなど，効率的に検索結果をブラウジングすることも可能である．

このような修辞構造解析は，文書中の各文に対し，上記タグのいずれかを付与する分類問題として捉えることができる．ただし，専門用語抽出と同様，個々のタグの間には依存関係があることに注意しなければならない．よって，専門用語抽出に用いた半教師あり CRFs をここでも採用することとした．

Hirohata ら [Hirohata 08] の実験データを用いた評価結果を表 1 に示す．訓練データは 1 万アブストラクト，テストは 1000 アブストラクトであり，提案手法のアンラベルドデータには専門用語抽出の実験と同じ 10 万アブストラクトを用いた．なお，Hirohatara らの手法は教師あり CRFs である．per sentence は文単位でのタグの正解率，per abstract はアブストラクト中全てのタグについて正解したアブストラクトの割合を示す．表より，双方の評価指標において，提案手法の有効性がわかる．特にアブストラクト中のすべてのタグを正解した割合では，既存手法を大きく上回った．

3.4 階層的な句構造に基づく統計的機械翻訳

論文アブストラクトの翻訳には，統計的機械翻訳 (Statistical Machine Translation: SMT) と呼ばれる機械翻訳手法を用いた．SMT の利点は，専門家による言語知識に頼らなくとも，学習用の対訳コーパスがあれば，翻訳器を構築できる点にある．

我々の SMT は，階層的な句に基づく手法を採用しており，対訳コーパスから統計量でスコア付けされた文法を自動獲得し，翻訳のモデルとすることを特徴としている [Watanabe 06]．入力された原言語の文に対して，もっとも適切な句の対応関係を与える目的言語側の文を求め，それを翻訳結果とする．図 2 に獲得した文法によって得られる日本語と英語の階層的な句の対応関係の例を示す．

この手法では，翻訳モデルの中に語の並び替えモデルが統合されているため，日本語と英語のように語順が大きく異なるような言語間の翻訳に適しているという特徴がある．

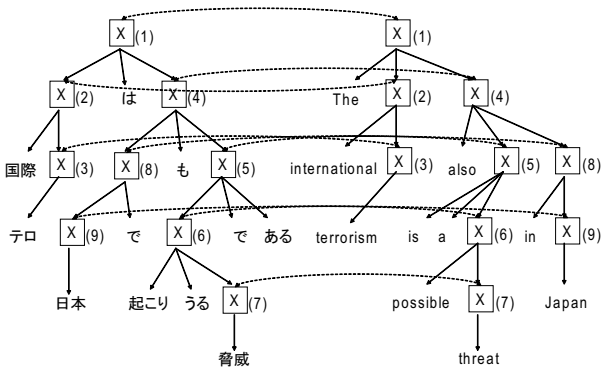

図 2: 階層的な句の対応関係

PubMed から抽出した 2000 アブストラクト (約 2 万文) に対し対訳を用意し，翻訳器を訓練した．約 4000 文を用いて評価したところ BLEU 値は約 0.3 であった．

図 3 にシステムの動作例を示す．検索後をフィールドを指定して入力すると該当する文書一覧が表示され，掲載誌，著者，英語タイトル，その翻訳一覧が表示される．英語タイトルをクリックすると，原文とその対訳を表示する．この時，専門用語抽出，修辞構造解析の結果もあわせて表示される．

4 まとめと今後の課題

本稿では，英語を得意としない人でも，英語で記述された医療情報にアクセスできるようにすることを目的としたシステムの現状について述べ，システムを構成する下記モジュールについて簡単に説明した．

1. 辞書に基づく言語横断検索
2. 専門用語抽出
3. 修辞構造解析
4. 階層的な句構造に基づく統計的機械翻訳

現状では，英語で記述された医学分野の論文アブストラクトデータベースである PubMed に対し，日本語で検索し，日本語で結果を読むことができるようになっている．また，修辞構造解析，専門用語抽出も行っている

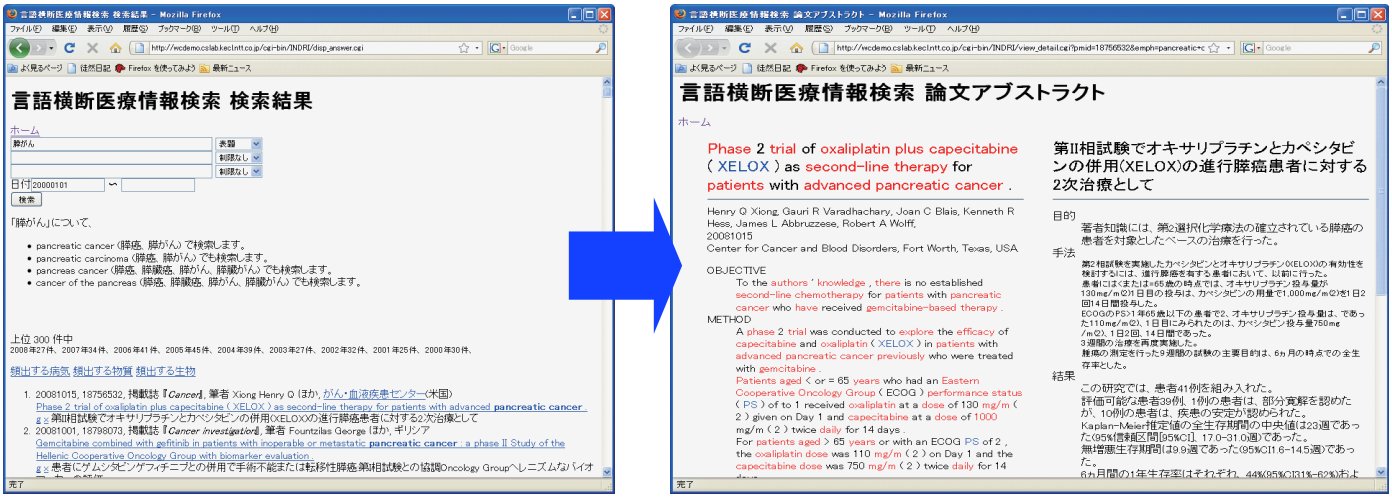

図 3: システムの動作例

ため，柔軟な検索が実現でき，ブラウジングも効率的に行える．

今後の課題としては，まず，翻訳の精度向上がある．現在の翻訳精度では，人間による翻訳にはまだまだ遠く及ばない．これは，特に，対訳コーパスのサイズが小さいことが原因であると考えるので，対訳コーパスの拡充することを考えたい．また，日本語の検索要求を英語に変換する際に，LSD を用いているが，対訳コーパスを拡充できると，LSD に登録されていない単語や句を単位とした検索ができるようになると考える．これらについても今後の課題としたい．

参考文献

[Hirohata 08] Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M.: Identifying Sections in Scientific Abstracts using Conditional Random Fields, in *Proceedings of IJCNLP* (2008)

[Lafferty 01] Lafferty, J., McCallum, A., and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in *Proc. of the 18th ICML*, pp. 282–289 (2001)

[Suzuki 08] Suzuki, J. and Isozaki, H.: Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data, in *Proceedings of ACL-08: HLT* (2008)

[Vapnik 95] Vapnik, V.: *The Nature of Statistical Learning Theory*, New York (1995)

[Watanabe 06] Watanabe, T., Tsukada, H., and Isozaki, H.: Left-to-Right Target Generation for Hierarchical Phrase-Based Translation, in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (2006)