E-07

言語横断医療情報提供—翻訳機能の改善

Crosslingual Medical Information Provision — Improvement of Translation Function

磯崎　秀樹*
Hideki Isozaki

1　はじめに

病気になったときに、自分の受ける治療を選ぶ上で、参考になりそうな、信頼のおける日本語の医療情報は少ない。マスコミやインターネットなどで話題になる医療情報は、単なる宣伝であったり、十分な根拠がないものが多い。外国の状況を知ろうとしても、専門用語だらけの英文を読みこなせる人は、ごくわずかである。

この現状を打破するため、我々は、信頼できる海外の最先端医療情報を誰でも簡単に把握できるようにする『医療情報アクセスシステム』の研究に着手した [IHS+09a]。

このシステムは、PubMed のデータベースの言語横断検索をベースとして、情報検索・情報抽出・データマイニング・統計的機械翻訳・文書要約などの、様々な言語処理を統合したもので、その最初のバージョンの概要については、関西支部大会 [HIS+09] と昨年のユニバーサルコミュニケーション研究会 [IHS+09b] で報告した。

我々が構築しようとしているのは、単なる機械翻訳システムではなく、世界の動向が一目でわかるような、言語横断データマイニングシステムである。

しかし、このシステムを構築していて最大の問題となったのは、統計的機械翻訳 (Statistical Machine Translation, SMT) の性能である。SMT は、これまで主流だった、ルールベース型機械翻訳 (Rule-based Machine Translation, RBMT) に比べて、ソフトウエアの構築や保守のコストが少なくてすむ、と期待されており、すでにヨーロッパ言語間では、実用レベルに達している。

我々が用いている SMT システムは、階層的句に基づき [WTI00]、ラージマージンで学習する [WSTI07] などの機能を持つシステムで、NTCIR-7 特許翻訳タスクなどで好成績を収めている [FUYU08]。医療の専門用語の翻訳については、京大で作成された『ライフサイエンス辞書 (LSD)』を利用している。しかし、英語と日本語の語順はあまりに違うため、意味不明な翻訳になってしまうことが少なくない。

英語と日本語の語順の克服は、RBMT ではあたりまえのことであるが、SMT ではまだ十分に対応できていない。このことは、インターネットの無料翻訳サービスの品質を比較すれば一目瞭然である。RBMT は多くの文を直訳風に訳せるが、SMT は直訳ですらない、でたらめな訳を出力する。以下は、インターネットで利用可能な、ある有名な SMT システムの翻訳結果である。

- 入力：ボブはジョンを殺した。
  出力：John killed Bob.
- 入力：メアリとボブはジョンの家に行った。
  出力：Bob went to the house of John and Mary.

こんなに簡単な文でも、このシステムは正しく訳せない。SMT がでたらめな訳を出す傾向は、このシステムに限らず、我々の翻訳器も同様の傾向を持つ。

SMT システムは、様々な語順を調べて、もっともスコアのよい訳を探し出す。ヨーロッパ言語間であれば、語句はあまり大きく移動せず、探索空間を絞ることができるが、英日翻訳では語句の移動距離に制限をつけることができないので、探索空間が膨大になってしまい、適切な訳を見つけることができない。素朴に考えても、1 文中に 30 個単語があれば、その順列は $30! = 2.65 \times 10^{32}$ 通りもあり、膨大な計算時間が必要になる。

そこで我々のグループでは、この 1 年、英日翻訳の品質改良に取り組んだ。その成果は以下の 3 点である。

- Head Finalization: 英文中の単語を日本語の語順に並び替えるための簡単な手法。[ISTD10]
- Divide & Translate: 英文を分割して翻訳した結果を統合する手法。[SDT+10]
- マルチタスク学習を用いることで、大量の素性を使いながら過学習が起こらないようにする手法。[DST+10]

以下では、このうちの Head Finalization について簡単に紹介する。詳細は上記論文を参照されたい。

RBMT では、様々なルールを人間がプログラミングすることによって、適切な語順になるようにしている。SMT でも、いくつかのルールによって、あらかじめ Subjet-Object-Verb (SOV) の語順に並べ替えておく、という手法が、昨年 Xu ら [XKRO09] や Hong ら [HLR09] により提案されている。しかし、これらの手法は複雑である。たとえば、Xu らの手法では、品詞・係り受けラベル・ルール重み・並べ替えの向きからなる 20 個のルールが

*日本電信電話株式会社、NTT

使われる。しかも、単語の移動範囲は、接続詞や句読点を超えないように制限されているが、ルール適用のアルゴリズムが明記されていないので、どう実装すればよいのかわかりにくい。

2　ヘッド・ファイナル化

我々は、これまでに提案されている手法よりずっと簡単な、**ヘッド・ファイナル化 (Head Finalization)** という方法を考案した。この手法では、「日本語はヘッド・ファイナルである」というよく知られた事実を使う。つまり、日本語の語順に並び替える、ということは、ヘッド・ファイナルの語順に並び替える、つまり、各語句の中で、その語句のヘッドを最後に移動することであるとみなす。

ここでヘッドとは、ある語句の中で、その語句の構文的な性質を決定している部分語句のことである。たとえば、“the green table” という語句のヘッドは table であり、“on the green table” のヘッドは on であり、“a cup on the green table” のヘッドは cup である。つまり、構文的な観点から中心的な部分のことである。

ヘッドはしばしば「主辞」と訳されるが、どの国語辞典を見ても、「主辞とは論理学で主語のこと」「主語を見よ」などと書かれていて、「論理学では主語のことだが、言語学では主語ではない」というのはとても紛らわしいので、本稿では「主辞」と訳さない。

英文をヘッド・ファイナルの語順に並び替えるには、ヘッドを自動的に検出できればよい。ヘッドは以下のようないくつかの方法で検出できる。

- 英語の係り受け解析器を利用する。ヘッドは係られる単語として出力される。
- Penn Treebank 形式の構文解析器の出力に Collins [Col99] のヘッド・ルールを適用する。
- HPSG パーザを利用する。

我々は、東大辻井研で作成された、**Enju** (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.ja.html) という HPSG パーザを利用する。あとは、各語句の中で、Head を最後に持っていくだけである。

これは簡単に実装することができる。たとえば、「John can hit a ball with a bat.」という英文を Enju で解析して得られる構文木は、図 1 の上の図に示すものである。ここで ⋆ は、ヘッドのある方の枝である。たとえば、c0 にとっては c3 がヘッドであり、c3 にとっては c4 がヘッドである。c0 は後ろにヘッドがあるので問題ないが、c3 は前にヘッドがあるので、ヘッド・ファイナルになっていない。そこで、c4 と c5 をひっくりかえす。同様に、c7 と c8 などもひっくりかえす必要がある。これらの書き換えを行なうと、下の図の構文木が得られる。この「John a bat with a ball hit can.」は「ジョンはバットでボールを打つことができる」と逐語訳できる。このように、ヘッド・ファイナル化は、簡単に実装でき、多くの場合に日本語に近い語順が得られる。

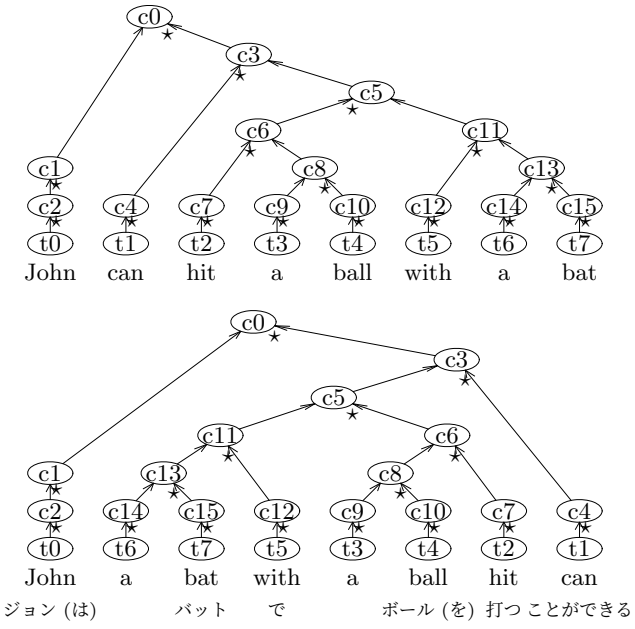

図 1: ヘッド・ファイナル化

一方、Xu らや Hong らは、スタンフォード・パーザ [KM03] と呼ばれる構文解析器の出力する係り受け情報を利用している。これは何の問題もなさそうに見えるが、実は大きな問題がある。スタンフォード・パーザの出力する係り受け情報は意味的なものであり、得られるのは、**意味的ヘッド (semantic head)**[dMM08] であって、上記で説明した**構文的ヘッド (syntactic head)** ではない。

たとえば

John can hit a ball.

という文全体の構文的ヘッドは can であるが、意味的ヘッドは hit である。したがって、ヘッド・ファイナルの語順に並び替えると、以下のようになる。

- 構文的ヘッド使用: John a ball hit can.
- 意味的ヘッド使用: John a ball can hit.

わずか 2 語の入れ替わりであるが、can hit は自然な日本語に逐語訳できない。助動詞の他に、because や whether などでも両者は食い違う。

では、Enju 以外のパーザでも、構文的ヘッドさえ用いれば、同じように日本語の順番が得られるのだろうか？

Charniak & Johnson のパーザ [CJ05] や構文的な単語係り受け解析器 [SICC09] などを試した結果、Enju 以外のパーザによる結果で Head を最後に移動しただけでは、逐語訳できる語順が得られにくいことがわかった。その原因を調べると、Enju の出力する構文木が**二分木**になっていることが、ポイントであることがわかった。

構文木が二分木の場合、各ノードの子ノードは、ヘッド・ファイナルにしても隣り合っている。しかし、ひとつのノードに 3 つ以上の子ノードがあると、英文では隣り合っている子ノードが、ヘッド・ファイナルにすることで切り離されてしまう。たとえば、

This toy is popular in Japan.

を Enju で解析してヘッド・ファイナルにすると、This toy Japan in popular is となり、「このおもちゃは日本で人気がある」と逐語訳できる。ところが、Charniak & Johnson のパーザで解析すると、is, popular, in Japan の 3 つが、ひとつの動詞句ノードの子になる。このうち is が Head なので、is を最後に持っていくと、This toy popular Japan in is となるが、これは逐語訳できない。これは一体であるべき popular と is が離れてしまったせいである。

3 実験と結果

前述のヘッド・ファイナル化を用いて英語を日本語の語順に近づけてから翻訳を行う実験を行なった。大量の訓練データがあること、多くの翻訳システムの性能がすでにわかっていて、提案手法の効果が明確になること、という観点から、ここでは、NTCIR-7 の特許翻訳タスク [FUYU08] で用いられたデータを使う。このタスクでは、180 万文の対訳データがトレーニングデータとして提供されていて、12 システムが参加し、その性能がわかっている。ただし、オーガナイザが用意したベースラインシステムの性能が、自動評価で最も良かった。

SMT システムとしては、この分野で標準的に用いられている GIZA++ (http://fjoch.com/GIZA++.html) と Moses (http://www.statmt.org/moses) を利用した。

提案手法と NTCIR-7 のベースラインシステムを比較したのが表 1 である。BLEU は機械翻訳の分野でデファクト・スタンダードとなっている自動評価法であり、数字が大きいほど良い。WER は Word Error Rate の略で、小さいほど良い。TER は Translation Edit (あるいは Error) Rate の略で、翻訳者が機械翻訳の結果を修正するときの手間を表すもので、やはり小さいほど良く、句のような長い単語列の移動のペナルティが WER よりも小さい。この結果によると、distortion limit が 6 のときの提案手法の性能がベストである。distortion limit は、語句の移動できる範囲を表す。

表 1: ヘッド・ファイナル化による翻訳品質向上:その 1 (括弧内は distortion limit)

	BLEU	WER	TER
提案手法 (0)	30.79	0.663	0.554
提案手法 (3)	30.97	0.665	0.554
提案手法 (6)	**31.21**	**0.660**	**0.549**
提案手法 (9)	31.11	0.661	**0.549**
提案手法 (12)	30.98	0.662	0.551
ベースライン (∞)	*30.58*	*0.755*	*0.592*

表 2: ヘッド・ファイナル化による翻訳品質向上:その 2

	ROUGE-L	IMPACT
提案手法 (6)	**0.480**	**0.369**
ベースライン (∞)	*0.403*	*0.339*

江原ら [EE09] は、日英翻訳の評価法として BLEU は良くなく、ROUGE-L や IMPACT という自動評価法 (いずれも大きいほどよい) が良いことを明らかにしている。英日翻訳での検証はまだないが、これらの指標を調べたのが表 2 であり、やはり提案手法の方がよい。

4 まとめ

言語横断医療情報提供では、統計的機械翻訳 (SMT) を用いるが、従来の SMT では、英語と日本語の語順があまりに違いすぎるために、原文に忠実な訳を得ることが難しかった。本稿では、この問題を解決するために、我々の行なった 3 つの改良のうち、ヘッド・ファイナル化という方法について説明した。この方法は、英語を日本語に近い語順に並びかえることが簡単にできる。その結果、英日翻訳の性能が向上した。

参考文献

[CJ05] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. of ACL*, pp. 173–180, 2005.

[Col99] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, Univ. of Pennsylvania, 1999.

[dMM08] Marie-Catherine de Marneffe and Christopher D. Manning. The Stanford typed dependencies representation. In *Proc. of the COLING-2008 Workshop on Cross-framework and Cross-domain Parser Evaluation*, pp. 1–8, 2008.

[DST+10] Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, Hideki Isozaki, and Masaaki Nagata. N-best reranking by multitask learning. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 375–383, 2010.

[EE09] 江原暉将, 越前谷博, 下畑さより, 藤井敦, 内山将夫, 山本幹雄, 宇津呂武仁, 神門典子. 機械翻訳精度の各種自動評価の比較, pp. 272–275. 日本特許情報機構, 2009.

[FUYU08] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Working Notes of the NTCIR Workshop Meeting*, pp. 389–400, 2008.

[HIS+09] 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. 自然言語処理による医療情報の読解支援. 情報処理学会関西支部大会, 2009.

[HLR09] Gumwon Hong, Seung-Wook Lee, and Hae-Chang Rim. Bridging morpho-syntactic gap between source and target sentences for English-Korean statistical machine translation. In *Proc. of ACL-IJCNLP*, pp. 233–236, 2009.

[IHS+09a] Hideki Isozaki, Tsutomu Hirao, Katsuhito Sudoh, Jun Suzuki, Akinori Fujino, Hajime Tsukada, and Masaaki Nagata. A patient support system based on crosslingual IR and semi-supervised learning. In *Proc. of SIGIR-2009 Workshop on Information Access in a Multilingual World*, pp. 59–61, 2009.

[IHS+09b] 磯崎秀樹, 平尾努, 須藤克仁, 鈴木潤, 塚田元, 藤野昭典, 永田昌明. 言語横断医療情報提供. 情報処理学会関西支部第 2 回ユニバーサルコミュニケーション研究会, 2009.

[ISTD10] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head Finalization: A simple reordering rule for SOV languages. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 244–251, 2010.

[KM03] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proc. of ACL*, pp. 423–430, 2003.

[SDT+10] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. Divide and translate: Improving long distance reordering in statistical machine translation. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 418–427, 2010.

[SICC09] Jun Suzuki, Hideki Isozaki, Xavier Carreras, and Michael Collins. An empirical study on semi-supervised structured conditional models for dependency parsing. In *Proc. of EMNLP*, pp. 551–560, 2009.

[WSTI07] Taro Watanabe, Jun Suzuki, Hajima Tsukada, and Hideki Isozaki. Online large-margin training for statistical machine translation. In *Proc. of EMNLP-CoNLL*, pp. 764–773, 2007.

[WTI00] Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. Left-to-right target generation for hierarchical phrase-based translation. In *Proc. of COLING-ACL*, pp. 119–126, 200.

[XKRO09] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. Using a dependency parser to improve SMT for Subject-Object-Verb languages. In *Proc. of NAACL-HLT*, pp. 245–253, 2009.