2008－NL－184 (9)
2008／3／27

自動点訳サーバ eBraille の医療文書点訳精度の向上に向けた IPADIC の最適化

菅野 亜紀[1] 三浦 研爾[1] 浅原 正幸[2] 池上 峰子[3] 前田 英一[4]
大島 敏子[3] 松本 裕治[2] 高岡 裕[1]

自動点訳サーバ「eBraille」の点訳精度は新聞記事などの通常文書では実用レベルに達した。しかし、病院で患者に渡す医療文書の点訳精度は通常文書よりも低く、疾患名などの医療に関連する語句を辞書に追加することが点訳精度向上に有効と考えられる。「eBraille」に使用している「茶筌」の IPADIC は、単語へ形態素生起コストを付与している。本研究では疾患名を含む 800 語の医療関連語を選び、四種類の方法で形態素生起コストを付与した IPADIC を作成した。これらの四種類の辞書間で点訳精度を比較し、医療文書の点訳精度向上への有効性を検討した。

Optimization of IPADIC (Dictionary for Morphological Analyzer) for Improvement of translating medical text by Japanese-into-Braille Translating Server

Aki Sugano[1], Kenji Miura[1], Masayuki Asahara[2], Mineko Ikegami[3], Eiichi Maeda[4], Toshiko Ohshima[3], Yuji Matsumoto[2] and Yutaka Takaoka[1]

The translation of newspaper articles by a Japanese-into-Braille Translating Server "eBraille 1.49" has been improved for practical use. However, the translation accuracy of medical texts is lower than that of newspaper article, so additional medical words or terms in its dictionary would be helpful. As the dictionary of ChaSen, IPADIC includes a cost for morpheme occurrence, we prepared four kinds of expanded IPADIC, which have the costs set differently, to examine and to find the optimal way for the accurate translation of medical texts.

1. はじめに

1824 年に、フランスの Louis Braille は視覚障害者用の記述文字として、点字を創出した。点字は、6 個の点の有無で構成された触読文字であり、我が国では日本語に翻案した点字を使用している。その特徴は、音を表す表音文字であること、分かち書きや前置点などの記述である[1]。現在我々は、視覚障害者への個別化医療対応の実現を目指し、患者向けの点字文書を簡便に点訳可能にする自動点訳システムを開発している。バリアフリー化に向けた社会的要請の結果、「視覚障害者等に対する服薬指導について」(1998 年 8 月 19 日 政医第 289 号、厚生省保健医療局) の通達や、2000 年の「平成 12 年度社会保険診療報酬改定等の概要」における視覚障害者に対する点字等を用いた薬剤情報提供料の加算可能化がなされた。このように、点字文書による視覚障害を有する患者のサポートは実施すべき課題の一つとなっている。

1997 年に、五十嵐 (東工大) と高岡 (東大医科研) は点字への機械翻訳を実現した自動点訳サーバ「eBraille 0.81」を、視覚障害者への簡便な文書アプローチ手段としてインターネット上で公開した[2]。この「eBraille0.81」は形態素解析器の「茶筌[3]」を利用し、cgi アプリケーションとしてクラ

1: 神戸大学大学院医学系研究科ゲノム医療実践講座
2: 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座
3: 神戸大学医学部附属病院看護部
4: 神戸大学医学部附属病院医療情報部
1: Laboratory for Applied Genome Science and Bioinformatics, Department of Applied Genome Medicine, Kobe University Graduate School of Medicine
2: Computational Linguistics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology
3: Nursing Department, Kobe University Hospital
4: Division of Medical Informatics, Kobe University Hospital

イアント・サーバ型システムとして構成されている。その後2007年に、高岡らは日本点字委員会による日本点字表記法改定（2001年）[1]に対応した、点訳エンジンKUICを新規に開発し、「eBraille 1.49[4]」として公開した。その結果、新聞記事などの点訳精度は実用レベルに達した。しかし、医療用文書の点訳精度向上に向けて更なる改良の必要性が判明した[4]。そこで今回、現行の「eBraille」が使用している「茶筌 2.3.3」が未知語と判定する医療に関連する語句を辞書であるIPADICへ追加した後、辞書の最適化による点訳精度の向上方法について検討した。

2. 方法

2.1 IPADICに追加する単語の選定

IPADICに追加する単語は、これまでの結果[5, 6]を基に疾患名と疾患名以外の医療に関連する語句を選択した。疾患名は、「ICD10対応電子カルテ用標準病名マスター V2.62」（財団法人医療情報システム開発センター）[7]および厚生労働省指定の特定疾患名を対象として、「茶筌2.3.3」で未知語と判定された語を選択した。次に、疾患名以外の医療に関連する語彙は、特定機能病院2007年4月の個人情報を除いた全看護記録を解析し、「茶筌 2.3.3」が未知語と判定した語の中から、看護記録で使用頻度の高い漢字語のみを選択した。以上のうち、誤字を除いた 800語を対象に、読み仮名・形態素生起コストを付与して辞書（ipadic 2.7.0）を拡張した。

2.2 追加する単語の形態素生起コスト

IPADICは、各単語に形態素生起コスト[9]を付与している。これは、解析済みのデータから学習した単語の出現確率を元に計算した数値であり、数値が小さいほど出現しやすい語であることを示している[8]。今回我々は800語の出現確率を計算する際、(1)通常文書コーパス(2,006 文)、(2)医療文書コーパス(1,003 文)、(3)通常文書と医療文書を合わせたコーパス(3,009 文)を対象に解析した。各々のコーパスの解析結果から得られた800語の出現回数から、形態素生起コストを計算した。なお、形態素生起コストは

$$-\alpha \log P(w_i \mid t_i)$$

と定義される[9]。さらに、形態素生起コストを全て 3,000 に設定した拡張辞書も作成し、比較に用いた。

ところで、辞書の形態素生起コスト値である3,000は、疾患名を数多く含む「ICD10対応電子カルテ用標準病名マスター V2.62」を用いて、以下の手順で決定した。

1.「ICD10対応電子カルテ用標準病名マスター V2.62」を「茶筌2.3.3」で形態素解析し、解析結果から名詞のみを抽出する。
2. 抽出した名詞のうちipadic 2.7.0に収録されている語（1,829 語）の、通常文書と医療文書、各々の文書での出現回数と形態素生起コストの分布を調べる。
3. 通常文書、医療文書、両者共に、分布の偏りが少ない形態素生起コスト値を決定し（2,500から3,500）、その中央値（3,000）を拡張辞書の形態素生起コストとする。

今回、通常文書には点字毎日の新聞記事（2003年から2005年）52,677文を、医療文書には特定機能病院の全診療科の看護記録（2007年4月分）112,917文を使用した。

2.3 評価用コーパス（通常文書・医療文書）の作成と評価実験

辞書拡張後の点訳プログラムの点訳精度を確認するために、点訳精度解析時に使用する評価用コーパスを準備した。まず、通常文書コーパスを点字毎日2002年4月から6月の記事からなるべく多様な内容を含むように選び、墨字文書に対応する点字文書を作成し、合計2,006文に拡大した。次に、通常文書と同様に、医療文書コーパス1,003文を作成した。
その内訳は、
「特定機能病院の全診療科の看護記録 2007年4月分（112,917 文）から無作為に選択した359 文」、「特定機能病院の患者向けの文書（各種同意書や検査票など）359 文」、「ICD10対応電子カルテ用標準病名マスター V2.62に含まれる疾患名（約2万語）から無作為に選択した疾患名を含む130 文」、そして「厚生労働省指定の特定疾患名を含む155 文」

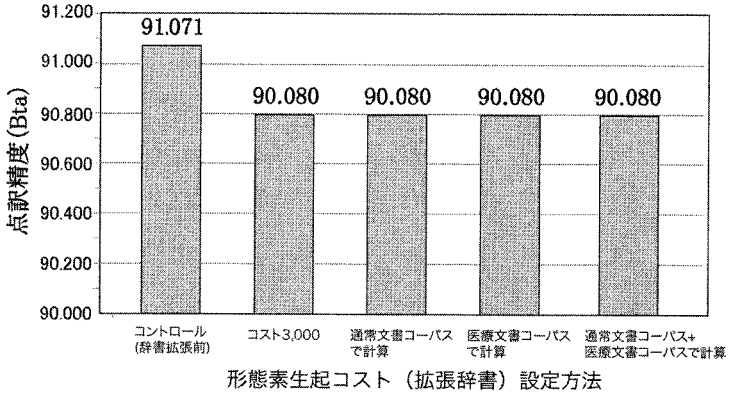

図 1　拡張辞書と点訳精度の解析結果
（通常文書コーパス）

である。
これらのうち、疾患名を含む文（合計 285 文）では、疾患名を「あなたは」と「です。」の間に配置した単純な文として作成した（例：「あなたはアミロイドーシスです。」）。

評価実験は、通常文書と医療文書の 2 つの評価用コーパスに対して、辞書を拡張する前後の「eBraille」の点訳精度の比較でおこなった。4 種類の拡張辞書の間での点訳精度を比較し、拡張辞書の点訳結果と形態素生起コストの数値の関係も精査した。最後に、辞書拡張後の「eBraille」と他の点訳ソフトの点訳精度を比較した。なお、点訳精度の評価指標は、分かち書きの精度の F-measure と読みの精度、正音率の積である Bta (Braille Translate Accuracy)を用いる既報の方法[10]で評価した。

3.　結果

3.1 辞書拡張後の点訳精度

通常文書コーパスを用いて点訳精度を辞書拡張前後で比較した結果、4 種類の拡張辞書全てのケースで拡張前に比較して点訳精度がわずかに低下した（図 1）。なお、4 種類の拡張辞書の間に点訳精度の差はなかった。

3.2 医療文書の点訳精度と生起コスト計算方法の関係

医療文書の点訳精度と拡張辞書の関係を検討した。その結果、医療文書コーパスのみを用いて形態素生起コストを計算した拡張辞書が最も高い点訳精度を示した（図 2）。

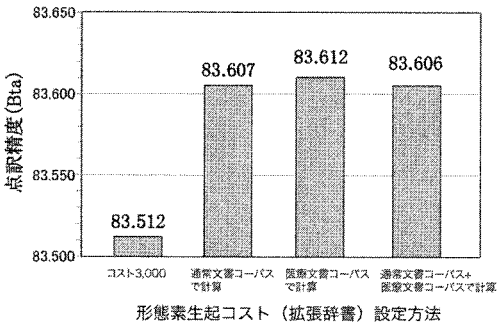

図 2　拡張辞書と点訳精度の解析結果
（医療文書コーパス）

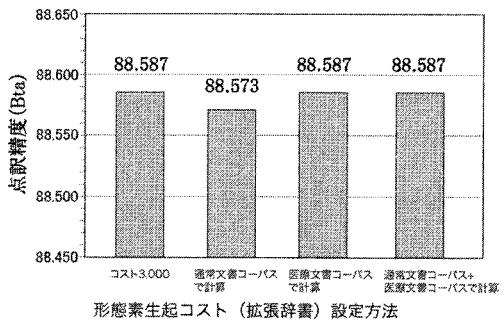

図 3　患者向け文書を用いた
点訳精度の解析結果

表１　辞書別の点訳結果の違い

墨字文	形態素生起コストの設定に用いたコーパス			正解点字文
	（辞書拡張なし）通常文書	（一律 3,000）医療文書	通常文書 + 医療文書	
細胞腫瘤と	サイボーシュ■コブト	サイボーシュリュート	サイボーシュリュート	サイボー■シュリュート
脳動脈瘤	ノードーミャク■コブ	ノードーミャクリュー	ノードーミャクリュー	ノー■ドーミャクリュー
顎下部腫瘤です。	アゴカブシュ■コブデス。	アゴカブシュリューデス。	アゴカブシュ■コブデス。	ガクカブ■シュリューデス。

また、医療文書コーパスを構成している４種類の文書の点訳精度を比較したところ、全てに共通して高い点訳精度を示した拡張辞書は存在しなかった。

3.3 形態素生起コストの違いに起因する点訳結果の精査

辞書の拡張で漢字の読みが改善された点訳結果の例を表１に示した。辞書拡張に「瘤（名詞 接尾一般)」が追加されたことにより、「細胞腫瘤（さいぼうしゅりゅう)」「脳動脈瘤（のうどうみゃくりゅう)」を４種類全ての拡張辞書で正しく読むことが可能となった。しかし、「顎下部腫瘤（がくかぶしゅりゅう)」の「瘤」は、医療文書コーパスで計算した形態素生起コストが付与された拡張辞書と、生起コストを 3,000 に設定した辞書の２種類で正しい読み「リュー」が付与され、他の拡張辞書では「コブ」と誤訳された。
そこで、各々の辞書の「瘤（名詞 接尾一般)」の形態素生起コストの数値を精査した。その結果、医療文書コーパスを用いて設定した生起コストは 3,272 で、通常文書と医療文書の両方のコーパスを用いて設定した生起コスト 3,371 よりも小さかった（表 2)。このことは、「瘤」の正しい読みを付与する生起コストの境界が、この２つの数値の間に存在することを示唆している。

3.4 他の自動点訳プログラムとの比較

「eBraille 1.49(20071126)」に辞書を追加し、その点訳精度を他の点訳ソフトの場合と比較した。この比較に用いた「eBraille」の辞書は、医療文書コーパスで計算した形態素生起コストを付与した辞書とした。
まず、医療文書コーパスの点訳精度を比較したところ、「eBraille 1.49(20071126)」の点訳精度は辞書拡張後に 83.606Bta となり、辞書拡張前から 8.856Bta 向上した。「お点ちゃん[11]」の点訳精度は、88.779Bta であった（図 4)。次に、医療文書コーパスの一部である患者向け文書の点訳精度を比較したところ、「eBraille」では 86.568Bta から 88.587Bta に向上していた（図 5)。

4. 考察

今回、IPADIC に 800 語の疾患名や医療関連語を追加したことで、医療文書コーパスの点訳精度が向上し、その有効性が示された。特に、医療文書コーパスのみで形態素生起コストを決定した拡張辞書で最も高い点訳精

表２　「瘤（名詞 接尾一般)」の形態素生起コスト

形態素生起コストの設定に用いたコーパス		
通常文書	医療文書	通常文書+医療文書
3999	3272	3371

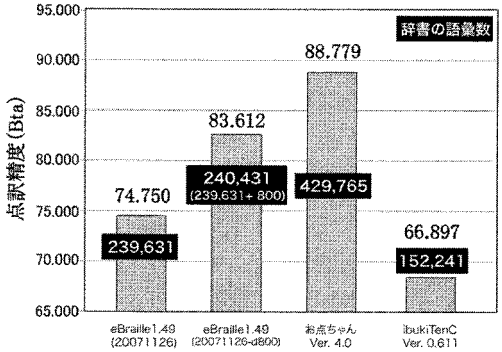

図4　eBrailleと他の点訳ソフトとの点訳精度の比較（医療文書コーパス）

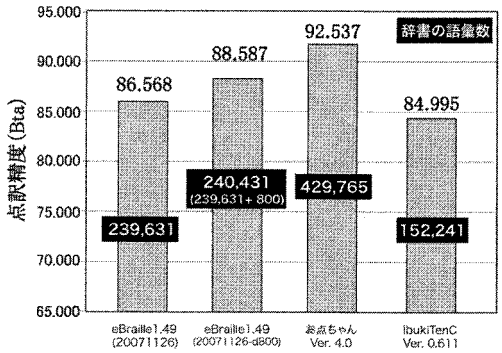

図5　eBrailleと他の点訳ソフトとの点訳精度の比較（患者向け文書）

度であった。

しかし、生起コストの算定方法による点訳精度の差は小さく、最適な辞書を作成する方法を最終決定するには至らなかった。その理由としては、辞書に追加した語数が僅か800語であったことに加えて、形態素生起コスト算定時に用いたコーパスが小さかったこと、さらにはコーパスが小さいために出現回数0の語の生起コストを3,999（これはコーパスでの出現回数が1回の時の生起コストに相当する）としたことが挙げられる。よって、今後はもっと大規模なコーパスを用意し、それを用いて生起コストを計算する必要がある。

次に、他の点訳ソフトとの比較から、たった800語の追加でも効果的に点訳精度を向上させることが可能であったので、他の辞書で採用している“あらかじめ分かち書きした語”を追加することなく点訳精度を向上可能といえる。特に、医療文書コーパスによる解析で最も点訳精度が高かった「お点ちゃんVer.4.0」では、[医学]にタグ付けされた語が7,960語も含まれており、このことが医療文書における点訳精度を規定していると考えられる（図4, 5）。よって今後「eBraille」のIPADICに医療関連語を追加していくことで、「お点ちゃん」と同程度かそれ以上の点訳精度に向上可能といえる。

さて、「eBraille」の場合、形態素解析結果に対し点字表記ルールを適用して点訳を実現しており、漢字などの読みの精度は辞書および形態素解析に、分かち書きの精度は点字表記ルールに大きく依存する。

今後の課題としては、(1) 辞書に追加する語句を検討する (2) 患者向けの文書コーパスの規模を拡大する、もしくは、疾患名や薬、検査・手術で用いられる医療関連の語句を幅広く含むコーパスを作成することが考えられる。今後は、これらに取り組む予定である。

謝辞

看護記録から抽出した未知語の読みについて御協力いただいた、神戸大学医学部附属病院の高橋副看護部長と藤原師長に深謝します。本研究は、科学研究費補助金萌芽研究「自動点訳プログラムを利用した視覚障害者向け点字文書提供システム構築の試み」（課題番号19659563、代表 大島敏子）による研究成果の一部である。

参考文献

[1] 日本点字委員会. 日本点字表記法 2001年版, 日本点字委員会 (2001).

[2] eBraille 0.81, 1997年版日本語自動点訳システム, https://www.ebraille.org/eBraille/.

[3] 松本裕治,北内啓,山下達雄,平野義隆,松田寛,高岡一馬,浅原正幸. 形態素解析システム『茶筌』version 2.3.3 使用説明書, 奈良先端科学技術大学院大学情報科学研究科松本研究室 http://chasen-legacy.sourceforge.jp/ (2003).

[4] eBraille 1.49. 2007年版日本語自動点訳シ

ステム, https://ebraille.med.kobe-u.ac.jp/eBraille2/.
[5] 菅野亜紀, 池上峰子, 松浦正子, 大田美香, 前田英一, 大島敏子, 松本裕治, 高岡裕. 医療文書の自動点訳システムの開発, 医療情報学(Suppl.) 27, pp.1214-1216(2007).
[6] 菅野亜紀, 大田美香, 三浦研爾, 松浦正子, 高橋京子, 池上峰子, 前田英一, 松本裕治, 大島敏子, 高岡裕. 自動点訳サーバ eBraille の開発, 信学技報 Vol.107, No.368 (WIT2007-32~54), pp.93-98(2007).
[7] MEDIS 標準マスター,病名マスター(ICD10 対応電子カルテ用標準病名マスター), 財団法人医療情報システム開発センター, http://www.medis.or.jp/4_hyojyun/medis-master/index.html
[8] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル, 奈良先端科学技術大学院大学情報科学研究科松本研究室(2003).
[9] 山下達雄, 規則と確率モデルの統合による形態素解析, 奈良先端科学技術大学院大学修士論文,NAIST-IS-MT9551119, (1997).
[10] 高岡裕, 菅野亜紀, 五十嵐大和, 三浦研爾, 大田美香, 小田剛, 松浦正子, 松本裕治, 大島敏子. 日本語自動点訳サーバ eBraille の改良と評価, 医療情報学(Suppl.) 27, pp.1210-1213(2007).
[11] 勝沼貞幸. お点ちゃん, http://www17.plala.or.jp/otenchan/.