厚生労働科学研究費補助金

感覚器障害研究事業

多機能高精度自動点訳エンジンの開発

平成２１年度　総括研究報告書

研究代表者　　石川　准

平成２２（２０１０）年　５月

目　次

厚生労働科学研究費補助金（感覚器障害研究事業）
総括研究報告書

多機能高精度自動点訳エンジンの開発

研究代表者　　石川　准　　静岡県立大学　国際関係学部　教授

研究要旨

本研究は、各種ドキュメントが有する構造情報、レイアウト情報、テキスト情報等を生かした高精度自動点訳、医学、法学等専門文献の高精度自動点訳、固有名詞、人名の高精度自動点訳、点訳分かち書き誤りの発見などの機能を有する自動点訳エンジンの開発を目的とする。

今年度までの成果は以下の通りである。

構造化点訳 XML(Structured Braille Translation Interface Format)の規格を策定し、Open XML を SBTIF に変換するモジュールを開発した。そのモジュールを既存の自動点訳システムのファイルインポート機能に組み込み、Open XML 文書から構造情報、各種レイアウト情報を抽出できるようにした。その結果、見出し付けやレイアウトなどで自動点訳の可読性が向上した。

医学等各分野の専門辞書を整備した。専門辞書が、対象分野の文書の点訳で効果を発揮することを確認した。一方で、専門辞書を一般辞書に統合すると、対象外の文書で変換精度の顕著な低下を招くことから、分野種別を自動判別するか、判別精度が悪い場合には、ユーザが明示的に指定できるようにする必要のあることを再確認した。

読みの誤り、分かち書きの誤り等の点訳誤りを項目別に自動的に算出するツールを開発し、市販、公開されている点訳ソフトウェアの自動点訳性能の客観評価を行った。また既存の自動点訳エンジンの形態素解析アルゴリズムを抽出、解析し、問題点を明確にした。

点訳規則のうち「分かち書き規則」には大量コーパスを用いる今日的形態素解析手法が有効であり、「切れ続き規則」には点訳コーパスを用いる誤り補正が有効であることを実験により確認した。

研究分担者
宮本　修（筑波技術大学 障害者高等教育研究支援センター　特任助教）

A. 研究目的

本研究は、近年の以下のような新しいニーズの高まりに対応した多機能高精度点訳エンジンの開発を目的とする。

1．各種ドキュメントが有する構造情報、レイアウト情報、テキスト情報等を生かした高精度自動点訳の実現
2．医学、法学等専門分野の文献の自動点訳の改良
3．固有名詞と一般名詞の判別及び人名の変換精度の向上
4．点訳分かち書き誤りの発見

B. 研究方法

上記目的を達成すべく以下のように研究を進める。

1．Open XML 等の構造情報を自動点訳に利用するための構造化点訳 XML の策定とその有効性評価の研究

構造化点訳 XML を策定し、Open XML やテキスト DAISY(DAISY XML, DAISY/NISO 2005)から構造化点訳 XML への変換を行うモジュールを開発する。

既存の自動点訳ソフトウェアに Open XML やテキスト DAISY からのインポート機能を実装し、構造情報、レイアウト情報を持ったワード文書を自動点訳した場合の点訳文書の可読性の向上を評価する。

2．専門分野文献の点訳精度向上、固有名詞と一般名詞判別性能の向上のための新しい形態素解析手法の研究

読みの誤り、分かち書きの誤り等の点訳誤りを項目別に自動的に算出できるツールを開発し、市販または無償公開されている点訳ソフトウェアの性能を客観的に評価する。

既存の自動点訳エンジンの形態素解析アルゴリズムの性能を評価する。

大量のコーパスデータを用い、自動学習機能を有する今日的な形態素解析手法を評価、分析し、それを応用して、確率論的方法論に基づいた新しい自動点訳エンジンを開発する。

専門分野ごとにコーパスを整備し、ドキュメント種別判別の精度を高め、各専門分野特有の読み方に対応し点訳精度を向上させる。

3．点訳分かち書き誤りの発見の研究

全国視覚障害者情報提供施設協会が運営する視覚障害者情報総合ネットワークサピエの点字図書データを用いて点字コーパスを作成する。

点字コーパスを用いた実用レベルの分かち書き誤り検出手法を開発する。

以上の成果に基づき多機能高精度自動点訳システムを開発する。

（倫理面への配慮）

研究の過程で知り得た個人情報の守秘義務を遵守した。

C. 研究結果

Open XML 等の文書から構造情報、各種レイアウト情報を抽出することで、見出し付けやレイアウトなどにおいて、自動点訳の可読性が向上した。

読みの誤り、分かち書きの誤り等の点

訳誤りを項目別に自動的に算出するツールを開発し、市販、公開されている点訳ソフトウェアの自動点訳性能の客観評価を行った。また、既存の自動点訳エンジンの形態素解析アルゴリズムを抽出、解析し、問題点を明確にした。

点訳規則のうち「分かち書き規則」には大量コーパスを用いる今日的形態素解析手法が有効であることを確認した。

一方「切れ続き規則」には点訳コーパスを用いる誤り補正が有効なことを確認した。

D. 考察

多機能高精度自動点訳エンジンの利用範囲は、自動点訳デスクトップパブリッシング、スクリーンリーダ、携帯端末、DAISY プレイヤ、携帯電話、電子読書プレイヤ、放送、ウェブサービス等と幅広い。多機能高精度自動点訳エンジンを実装した情報機器、情報サービスは、視覚障害者の就労支援、高等教育での情報保障支援等に資する。

E. 結論

Open XML や EPUB 等のいわゆるオープンな標準に基づき、かつアクセシビリティへの配慮が施されたドキュメントが至る所で作成され、だれもが利用できる社会の実現への努力が実を結びつつある。

本研究により、このようなドキュメントを高い精度で自動点訳できるようになる。

形態素解析、検索、自動翻訳等のために整備されている日本語コーパスおよびサピエの点字図書データを自動点訳に利用することで、自動点訳エンジンの開発と改良に要するコストを低減できる。

F. 健康危険情報

該当しない。

G. 研究発表

（口頭発表）

石川 准，宮本 修，多機能高精度自動点訳エンジンの開発，平成 21 年度感覚器障害研究・研究成果発表会，KKR ホテル東京，2010 年 2 月 10 日

H. 知的財産権の出願・登録状況

なし

厚生労働科学研究費補助金（感覚器障害研究事業）
総括研究報告書

多機能高精度自動点訳エンジンの開発

全文検索エンジンを使った点字のマス空け誤り検出

研究分担者　宮本　修　　筑波技術大学障害者高等教育研究支援センター　特任助教

研究要旨　日本語の点訳において、漢字かな混じり文をカナの分かち書き文に直すが、マス空けの困難さにより、機械ではできず、熟練した人間の作業が必要である。検索エンジンによる点字データベースとのマッチングにより、マス空けの誤りを検出する方法を開発する。

A. 研究目的

日本語の点訳については、「分かち書き」という、表意文字である漢字から表音文字であるカナに翻訳せざるをえない日本語特有の問題がある。「分かち書き」は計算機では 100%正確におこなえないし、人間がおこう場合であっても、技術を習得するまでには多くの場合、数年かかる。その「分かち書き」の誤りを検出することを目的とする。

「分かち書き」をしなければならない理由は以下のとおりである。

6 点式点字は、1 文字を 6 点のあるなしで表現するため、63 通りの文字を作ることができる。一般的な日本語の点訳においては、表意文字である漢字かな混じり文を、表音文字のカナだけに変換する作業が必要となる。カナだけの文は読みにくいので、適当な場所でマス空けを行うことによって読みやすくする。空白のあけ方は日本語点字の表記法にしたがって正確にあける必要がある。

機械による自動点訳には誤りがあるし、また、いかに熟練した点訳者といえど、誤ることはある。点訳後、修正することが必要である。誤りには、読みの誤りと、マス空けの誤りに分けられる。読みの誤りについては、点字の知識がなくても修正することができるが、マス空けの誤りについては、点字の知識が必要になる。

点字のマス空け規則には、「文節ごとに区切る」(分かち書き)という原則と、「長い複合語は途中で区切る」(切れ続き)というふたつの原則がある。文節ごとに区切る、という分かち書き原則については、日本語を自立語と付属語に分けることができれば、自動的に決まるので、比較的納得しやすいといえる。また、熟練した点訳者であれば、

間違うことはあまりない。

これに対して、長い複合語は途中で区切る、という切れ続き原則については、おおまかなルールはあるものの、熟練した点訳者でも判断に迷う場合が多々ある。例えば同じ「超」がつく語でも、「超特急」はマスあけをしないが、「超現実主義」は「超　現実　主義」とマスあけをする。また、「ミニバイク」はマスあけをしないが、「ミニスカート」は「ミニ　スカート」とマスあけをする。(表記辞典による。)

表記辞典が出版されており、それに従うものの、表記辞典に載っていない語については、そのつど判断するしかない。普通、点訳は1冊の本を複数人で点訳するため、判断した語はそのつど他の点訳者にも伝え、最低限1冊の本の中では統一しなくてはならない。鍼灸の分野についてはパソコン上で動く辞書も用意されているが、それ以外の分野の例えば化　学の専門書をもし点訳しようとすると、気の遠くなるような作業である。(そのため、点訳ボランティアには専門書は敬遠される傾向にある。)

B. 研究方法

点訳後のデータについて、全文検索エンジンによる、点字データベースとのマッチングをおこなう。これにより、点訳のマスあけが妥当かどうかを判断するための情報を、修正作業をおこなう点訳者に提示する。

全文検索エンジンの実装について、以下の検討をおこなった。

まずgoogle desktopなどの既成の検索エンジンを使うことを検討したが、ひらがなだけの文字列はうまく検索できなかった。これはgoogle desktopは、形態素解析を行ってインデックスを作っているが、ひらがなだけの文字列では形態素解析がうまく働かないためであると推測される。

今回、点訳ファイル専用の全文検索エンジンを作成した。検索エンジンの仕様の検討にあたっては、点字本10万タイトルが余裕をもって格納でき、そこから検索できることとした。この10万タイトルという数は点字図書館などの障害者情報提供施設の団体である全国視覚障害者情報提供施設協会が運用している「ないーぶネット」に2009年4月現在、登録されている点訳ファイルのタイトル数である。

データベースのインデックスにはマス空け単位の語をそのまま用いた。これはデータベースに登録された点訳ファイルは、点訳者によってすでにマス空けされたものであるからである。データベースへの登録は、登録文の1語から32バイトの情報ブロックを作成し、登録している。情報ブロックの内訳は、インデックスとなる語の前8文字を8バイト、語の後16文字を16バイト、文書IDと文書内での位置がそれぞれ4バイトで表し、合計32バイトとなっている。

C. 研究結果

点訳ファイルからデータベースを作成した。データベースは、約1500タイトルの点訳ファイルから、作成した。作成にはCore2Duo 2.6GHzのパソコンで104時間かかった。作成後のデータベースの大きさは約6ギガバイトであり、仮に10万タイトルを収めたとしても1テラバイトのハードディスクに収めることができると思われる。

形態素解析エンジンであるMeCabを用いて点訳エンジンを作成し、その点訳結果

を、作成したデータベースによって誤り修正した。この誤り修正後の結果を他の点訳エンジンと比較した。

ある新聞記事を点訳した結果は以下のとおりである。

表 1. 自動点訳の結果

点訳エンジン名	誤り数(個)
EXTRA 5.1	5 個
Ibuki-TenC	12 個
提案手法	6 個

D. 考察

新聞記事などでは EXTRA が最も良い結果であったが、他の分野でも検討する必要があると考えられる。

E. 結論

この結果は、提案手法が、代表的な市販ソフトの点訳エンジンにほぼ匹敵する性能を有していることを示していると考えられる。

KOKUYO セポ-50B