様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 3 月 25 日現在

機関番号：13904
研究種目：基盤研究（B）
研究期間：2008～2010
課題番号：20300058
研究課題名（和文）
入力文の分割・翻訳・連結による同時通訳システム
研究課題名（英文） Simultaneous interpreting system based on segmentation, translation and connection of spoken sentences
研究代表者
稲垣 康善（INAGAKI YASUYOSHI）
豊橋技術科学大学・大学院工学研究科・副学長
研究者番号：10023079

研究成果の概要(和文)：同時通訳機能を備えた音声翻訳システムの新しい翻訳方式として、入力文を（1）単文相当に分割し、（2）各々を翻訳し、（3）翻訳結果をうまくつなげながら訳す、という分割・翻訳・連結による機械翻訳方式の開発を推進した。本研究組織が有する自然言語処理と音声言語処理に関する基礎研究の成果を活用し、音声の入力途中の段階で「うまく分割する」技術と翻訳された結果を「うまく連結できる」技術に焦点をあてて実施した。

研究成果の概要（英文）：A method for simultaneous speech translation has been developed. The method is executed based on segmentation of spoken sentences, translation of segmented chunks, and connection of translated chunks. The results of fundamental research on incremental speech and language processing have been utilized. The points are the technologies for finely segmenting the spoken sentences and naturally connecting the target fragments while listening the speech.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	7,000,000	2,100,000	9,100,000
2009 年度	4,200,000	1,260,000	5,460,000
2010 年度	3,500,000	1,050,000	4,550,000
年度			
年度			
総　計	14,700,000	4,410,000	19,110,000

研究分野：情報工学
科研費の分科・細目：情報学・知能情報学
キーワード：同時通訳、機械翻訳、自然言語処理、音声言語、構文解析、意味解析、コーパス、文分割

1．研究開始当初の背景

多言語間コミュニケーション環境を整えることは、「世界に開かれた社会」を実現するための重点課題の一つであり、音声翻訳はそのような環境における中核的技術となる。音声翻訳に関する研究は、音声・言語処理技術の進展を背景に、近年、活発に進められてきた。しかし、これまでに実現されている音声翻訳技術の多くは、対話文を単位とした逐次通訳を採用している。逐次的に翻訳する方式では、コミュニケーションの効率が大幅に低下することがわかっており、同時通訳の機能を備えたシステムの実現が大いに望まれる状況にある。

本研究の代表者ならびに分担者は、以前より同時通訳システムの研究開発に取り組ん

でおり、構文トランスファ方式に基づく英日同時通訳の様々な要素技術を開発するとともに、旅行対話をドメインとした実験システムを構築するに至っている。しかし、このシステムには、

- 翻訳単位として「句」を採用しており、翻訳単位の粒度が極めて細かく、現実性に欠ける。
- 変換の方式が英日方向に限定されており、対象言語に関する汎用性を備えていない。

などの問題があった。

２．研究の目的

本研究では、同時通訳機能を備えた音声翻訳システムの新しい翻訳方式として、入力文を、
（１）単文相当に分割し、
（２）各々を翻訳し、
（３）翻訳結果をうまくつなげながら訳す、
という分割・翻訳・連結による同時翻訳方式の開発を目的に推進した。この方式は、「プロの通訳者は、話者の発話を聞き、ある程度の情報を取得した段階でそれらをまとめ訳出している」という、通訳データの分析により得られた知見に基づいて設計されたものである。

３．研究の方法

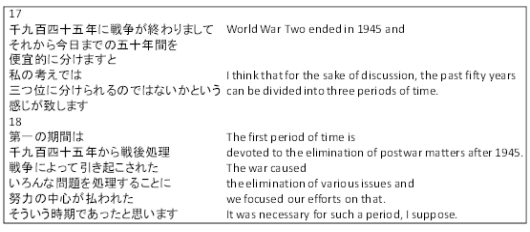
本方式を実現する上でのポイントは、音声の入力途中の段階で「うまく分割すること」と翻訳された結果を「うまく連結できること」にある。このため本研究では、研究代表者ならびに分担者らがこれまで推進してきた自然言語処理及び音声言語処理に関する基礎的研究の成果を活用し、入力文の「分割」と「連結」の技術に焦点をあてて研究を実施した。具体的には、以下の項目を推進した。

① 言語資源の開発と利用：同時通訳システムの開発での利用を目的としたコーパスを、既存の音声言語データに同時通訳を考慮した対訳文を付与することにより構築する。音声の入力途中の段階で意味まとまりを認識し、適切なタイミングで訳出する機能を実現するために、訳出タイミングの検出可能性を検証する。
② 講演文の分割：講演音声による入力を、同時翻訳のための処理単位に分割する手法を開発する。講演音声の書き起こしテキストへの漸進的改行挿入のアプローチを採用し、入力音声に追従して、意味的なまとまりを考慮しつつ、文より短い単位に分割する機能を実現する。
③ 対話文の分割：対話音声に対する同時翻訳のための単位を、対訳データを用いて定める方法を導入する。同時翻訳システムが入力に追従した訳出を実現するために、翻訳単位が、文に比べて十分に細かく、かつ、入力と同時進行的に検出可能であることを検証する。
④ 同時翻訳のための構造解析：入力文を単語の出現順序に従って同時進行的に解析する機構を実現するために漸進的構文解析と呼ばれる技術を開発する．接合操作と呼ばれる構文木上の操作を漸進的構文解析に取り入れることにより、解析精度が向上することを実験的に検証する．
⑤ 翻訳単位間の関係抽出：分割された翻訳単位に対応して生成された訳文が、自然に繋がるように訳文を連結し整えるための要素技術として、翻訳単位間の意味関係の推定手法を開発する。推定した意味関係に基づき、自然に連結するために挿入されるべき接続詞を選択する機構を実現する。

４．研究成果

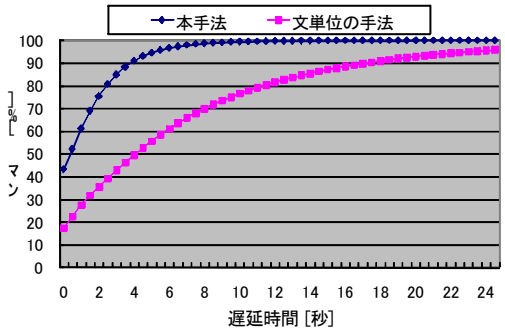
本研究を推進した結果、以下に示す成果が得られた。

① 同時通訳研究のための言語資源の開発：講演データを用いて同時通訳研究のための対訳コーパスを作成した。データとして、名古屋大学同時通訳データベースを使用した(1,935 文､60,829 形態素)。講演文を、同時翻訳のための処理単位（以下、チャンク）に分割する作業を人手で実施した。プロの同時通訳者における訳出遅延の様相を考慮し、チャンクの長さを 4.3 秒以内とし、その制限のもとで意味的にまとまる単位に分割した。その結果、8,644 チャンク（1 文あたり平均 4.47 チャンク)に分割された。また、各チャンクに対して、対訳を付与することによりコーパスを構築した（図参照)。対訳作成は、通訳業務に精通するプロの翻訳者により実施した。



17	
千九百四十五年に戦争が終わりまして	World War Two ended in 1945 and
それから今日までの五十年間を	
便宜的に分けますと	
私の考えでは	I think that for the sake of discussion, the past fifty years
三つ位に分けられるのではないかという	can be divided into three periods of time.
感じが致します	
18	
第一の期間は	The first period of time is
千九百四十五年から戦後処理	devoted to the elimination of postwar matters after 1945.
戦争によって引き起こされた	The war caused
いろんな問題を処理することに	the elimination of various issues and
努力の中心が払われた	we focused our efforts on that.
そういう時期であったと思います	It was necessary for such a period, I suppose.

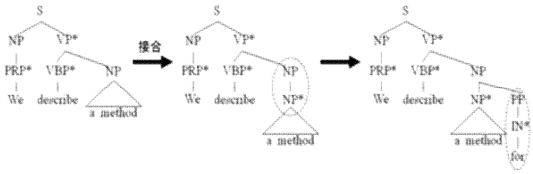
② 講演の同時通訳における入力文の分割：講演音声入力を、同時翻訳のための処理単位として利用可能な単位に分割する方法を開発した。本改行挿入手法は、入力音声に追従して、意味的なまとまりを考慮しつつ、文より短い単位に分割できる。そのため、本手法により分割された単位は、同時翻訳に適した単位として利用することができる。



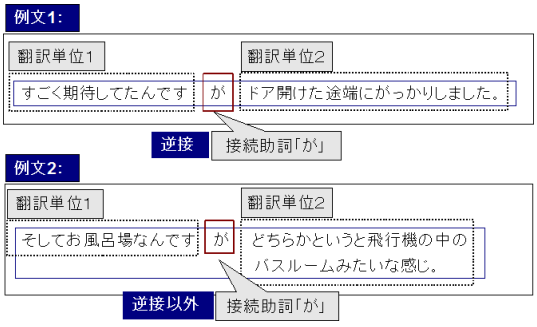
本手法では、講演全体の文節列を入力とし、節境界が検出されるごとに、それまでに入力された文節列の各文節境界に対して、改行位置を同定する。日本語講演データを用いて改行挿入実験を行った結果、文単位の従来手法と比較して、改行位置の再現率・適合率をそれほど低下させることなく、短い遅延時間での字幕テキスト表示を実現しており、本手法の有効性を確認した（図参照）。

③　対話の同時通訳における入力文の分割：同時翻訳単位への検出可能性を検証した。そのために、対訳コーパスを用いて訳出単位データを作成した。コーパスとして、名古屋大学同時通訳データベースに収録された旅行対話データの日本語話者発話データを利用した。訳出単位の検出では、形態素境界が訳出単位の境界になる確率を算出し、その確率が閾値以上の場合に訳出単位の境界であると判定する。確率計算のモデルとして、最大エントロピー法に基づくモデルを用いた。日本語対話文を訳出単位に分割する実験を実施した。実験では、本研究で同時翻訳単位に分割した対話データを用いた。素性には、前後の形態素の出現形と品詞（品詞、活用形、活用型）と発話単位境界の有無を選択した。精度にして 80.8%、再現率にして 74.3%を達成しており、訳出の検出可能性を確認した。

④　同時的な翻訳のための入力文の解析：左再帰構造に起因する局所的曖昧性の問題を解決するために，漸進的構文解析に接合操作を導入する．接合操作により，左再帰構造が必要になった段階で，それを後から部分構文木に追加することにより，局所的曖昧性の問題を回避できる．提案手法では，部分構文木だけ生成すればよい．左再帰構造が必要になった段階で，接合操作により挿入できるからである．解析が進行し，単語を読んだ時点で，図に示すように，まず接合操作により補助木を挿入し，その後に，構造を付加する．この例が示すように，従来の手法と異なり，あらゆる深さの左再帰構造を想定して部分構文木を生成する必要がない，すなわち，左再帰構造による局所的曖昧性の問題を回避できる．



⑤　連結に基づく訳文生成のための翻訳単位間の関係抽出：本研究では接続助詞「が」に着目してそのあいまい性を解消し、翻訳単位間の意味関係の推定を行った。具体的には、図のように、入力された文の接続助詞「が」の前後の翻訳単位が「逆接」であるか「逆接以外」の意味関係であるかを推定する。推定には機械学習手法を用いるが、機械学習のための十分な量の学習データを人手で作成するには膨大なコストが必要となる。そこで、学習データを自動的に作成し、それを使用して機械学習を行った。学習データの自動生成は、「しかし」のような「逆接」の意味をもつ接続詞を文頭にもつ文とその前の文に出現する単語列を正例、「そして」や「それでは」のような接続詞を文頭にもつ文とその前の文に出現する単語列を負例として、学習データを自動生成した。読売新聞 308,388 記事を使用し、学習データとして 118,724 個を自動的に生成した。そして、SVM で分類器を生成した。



５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１３件）

① Yoshihide Kato, Shigeki Matsubara, Correcting Syntactic Annotation Errors Using a Synchronous Tree Substitution Grammar, IEICE

Transactions on Information and Systems, 査読有, Vol.E93-D, No.9, pp. 2660-2663 (2010).

② Masaki Murata, Tomohiro Ohno, Shigeki Matsubara, Construction of Linefeed Insertion Rules for Lecture Transcript and Their Evaluation, International Journal of Knowledge and Web Intelligence, 査読有, Vol.1, No.3/4, pp.227-242 (2010).

③ 坂地 泰紀, 野中 尋史, 酒井 浩之, 増山 繁, Cross-Bootstrapping 特許文書からの課題・効果表現対の自動抽出手法, 電子情報通信学会論文誌, 査読有, Vol.J93-D, No.6, pp.742-755 (2010).

④ 鈴木 佑輔, 横田 隼, 酒井 浩之, 増山繁, Web 掲示板からの質問・回答対応の自動抽出, 査読有,人工知能学会論文誌, Vol.25, No.1, pp.168-173 (2010).

⑤ Hiroyuki Sakai, Shigeru Masuyama, Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies, IEICE Transactions on Information and Systems, 査読有, Vol.E92-D, No.12, pp. 2341-2350 (2009).

⑥ Yoshihide Kato, Shigeki Matsubara, Incremental Parsing with Adjoining Operation, IEICE Transactions on Information and Systems, 査読有, Vol.E92-D, No.12, pp.2306-2312 (2009).

⑦ 笠 浩一朗, 松原 茂樹, 同時通訳者の話速に影響を及ぼす要因の定量的分析, 通訳翻訳研究(日本通訳翻訳学会誌), 査読有, No.9, pp.21-32 (2009).

⑧ 酒井 浩之, 野中 尋史, 増山 繁, 特許明細書からの技術課題情報の抽出, 人工知能学会論文誌, 査読有, Vol.24, No.6, pp.531-540 (2009).

⑨ 村田 匡輝, 大野 誠寛, 松原 茂樹, 読みやすい字幕生成のための講演テキストへの改行挿入, 電子情報通信学会論文誌, 査読有, Vol.J92-D, No.9, pp.1621-1631 (2009).

⑩ 笠 浩一朗, 松原 茂樹, 稲垣 康善, 英日同時翻訳のための依存構造に基づく訳文生成手法, 電子情報通信学会論文誌, 査読有, Vol.J92-D, No.6, pp.921-933 (2009).

⑪ 大野 誠寛, 松原 茂樹, 柏岡 秀紀, 稲垣 康善, 節の始境界検出に基づく独話文の係り受け解析, 情報処理学会論文誌, 査読有, Vol.50, No.2, pp.553-562 (2009).

⑫ 松原 茂樹, 同時通訳の工学と科学-次世代自動通訳技術の実現に向けて-, 情報処理, 査読有, Vol.49, No.6, pp.617-623 (2008).

⑬ Hiroyuki Sakai, Shigeru Masuyama, Cause Information Extraction from Financial Articles Concerning Business Performance, IEICE Transactions on Information and Systems, 査読有, Vol.E91-D, No.4, pp.959-968 (2008).

〔学会発表〕(計３４件)

① Masaki Murata, Tomohiro Ohno, Shigeki Matsubara, Automatic Comma Insertion for Japanese Text Generation, The 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), Oct. (2010), 米国.

② Yoshihide Kato, Shigeki Matsubara, Correcting Errors in a Treebank Based on Synchronous Tree Substitution Grammar, The 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Jul. (2010), スウェーデン.

③ Koichiro Ryu, Shigeki Matsubara, Yasuyoshi Inagaki, Translation Unit for Simultaneous Japanese-English Spoken Dialogue Translation, The 2nd International Symposium on Intelligent Decision Technologies (IDT 2010), Jul. (2010), 米国.

④ Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, Shigeki Matsubara, Collection of Usage Information for Language Resources from Academic Articles, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), May (2010), マルタ共和国.

⑤ Masaki Murata, Tomohiro Ohno, Shigeki Matsubara, Yasuyoshi Inagaki, Construction of Chunk-Aligned Bilingual Lecture Corpus for Simultaneous Machine Translation, The 7th International Conference on Language Resources and Evaluation (LREC 2010), May (2010), マルタ共和国.

⑥ 酒井 浩之, 松原 茂樹, 増山 繁, 稲垣 康善, 文中の接続助詞「が」に着目した翻訳単位間の意味関係の推定, 言語処理学会第 16 回年次大会, Mar. (2010), 東大（東京都）.

⑦ Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, Shigeki Matsubara, Utilization of Usage Information for Language Resource Searches, The 2nd

International Conference on Global Interoperability for Language Resources (ICGL 2010), Jan. (2010), 香港.

⑧ 村田 匡輝, 大野 誠寛, 松原 茂樹, 稲垣 康善, 講演の同時翻訳のための対訳データの作成と分析, 第 8 回情報科学技術フォーラム, Sep. (2009), 東北工大（宮城県）.

⑨ Yoshihide Kato, Shigeki Matsubara, Incremental Parsing with Monotonic Adjoining Operation, The Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009), and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP 2009), Aug. (2009), シンガポール.

⑩ Tomohiro Ohno, Masaki Murata, Shigeki Matsubara, Linefeed Insertion into Japanese Spoken Monologue for Captioning, The Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009), and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (IJCNLP 2009), Aug. (2009), シンガポール.

⑪ Masaki Murata, Tomohiro Ohno, Shigeki Matsubara, Automatic Linefeed Insertion for Improving Readability of Lecture Transcript, The 2nd International Symposium on Intelligent Interactive Multimedia Systems and Services (IIMSS 2009), Jul. (2009), イタリア.

⑫ Hiroyuki Sakai, Shigeru Masuyama, Polarity Assignment to Causal Information Extracted from Financial Articles Concerning Business Performance of Companies, The AI-2008 28th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Dec. (2008), 英国.

⑬ Tomohiro Ohno, Shigeki Matsubara, Hideki Kashioka, Yasuyoshi Inagaki, Dependency Parsing of Japanese Spoken Monologue Based on Clause-Starts Detection, The 9th Conference in the Annual Series of Interspeech Events (Interspeech 2008), Sep. (2008), オーストラリア.

⑭ Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, Hitoshi Isahara, Construction of an Infrastructure for Providing Users with Suitable Language Resources, The 22nd International Conference on Computational Linguistics (COLING 2008), pp.119-122, Aug. (2008), 英国.

⑮ Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, Hitoshi Isahara, Construction of a Metadata Database for Efficient Development and Use of Language Resources, The 6th International Conference on Language Resources and Evaluation (LREC 2008), May (2008), モロッコ.

⑯ Takahiro Ono, Hitomi Tohyama, Shigeki Matsubara, Construction and Analysis of Word-level Time-aligned Simultaneous Interpretation Corpus, The 6th International Conference on Language Resources and Evaluation (LREC 2008), May (2008), モロッコ.

⑰ Shunsuke Kozawa, Hitomi Tohyama, Kiyotaka Uchimoto, Shigeki Matsubara, Automatic Acquisition of Usage Information for Language Resources, The 6th International Conference on Language Resources and Evaluation (LREC 2008), May (2008).

〔図書〕（計１件）

① Shigeki Matsubara, Hitomi Tohyama, Nobuo Kawaguchi, Kazuya Takeda, c/o The Americas Group, An evaluation database for in-car speech recognition and its common evaluation framework, in Computer Processing of Asian Spoken Languages, 2010,372 ページ(200-203)

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
ホームページ等
http://slp.itc.nagoya-u.ac.jp/~matubara/kaken/si/

６．研究組織
(1)研究代表者
稲垣 康善（INAGAKI YASUYOSHI）
豊橋技術科学大学・大学院工学研究科・副学長
研究者番号：10023079

(2)研究分担者
松原 茂樹 (MATSUBARA SHIGEKI)
名古屋大学・大学院情報科学研究科・准教授
研究者番号：20303589
加藤 芳秀
名古屋大学・情報基盤センター・研究員
研究者番号：20362220
笠 浩一朗
名古屋大学・大学院国際開発研究科・助教
研究者番号：40397451
大野誠寛
名古屋大学・大学院国際開発研究科・助教
研究者番号：20402472
酒井浩之
豊橋技術科学大学・大学院工学研究科・助教
研究者番号：70402659
(3)連携研究者