様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月１日現在

機関番号：３４３１６
研究種目：基盤研究(C)
研究期間：2007～2010
課題番号：19500133
研究課題名（和文）　日英混在型入力による動的文脈照合英作文システムに関する研究

研究課題名（英文）　Development of English-Writing Support Systems

研究代表者
馬　青（MA QING）
龍谷大学・理工学部・教授
研究者番号：30358882

研究成果の概要（和文）：
英作文において、部分的に適切な英語表現が思い浮かばないとき、本来言い表したい日本語表現（単語またはフレーズ）をそのまま入力するだけで、すなわち、日英混在の入力文から、適切な英語表現を生成してくれる英作文支援システムを開発した。単語レベルでの支援においては最適な文脈による訳語選択手法と大規模で高品質な英語コーパスと超大規模な Web データの統合利用手法を提案した。フレーズレベルでの支援においては日本語フレーズを構成する各単語の訳語候補の組み合わせによる英語フレーズの生成手法と、大規模で高精度な日英対訳表現抽出手法とそれにより抽出した日英対訳表現を利用した用例ベースに基づく英作文支援手法を開発した。

研究成果の概要（英文）：
English-writing support systems that enable non-native speakers to produce nearly perfect English sentences for mixed English-Japanese sentences, in which expressions without know translations are simply written in Japanese, have been developed. For word-level support, the methods for equivalent selection using optimal contexts and for integrated use of the high-quality English corpora and the huge amounts of Web data have been proposed. For phrase-level support, the method for English phrase generation using the combinations of the candidate translations of the single Japanese words have been proposed. Furthermore, the methods for extracting broad-scale, high precision parallel translation expressions and the approach based on translation patterns using these parallel translation expressions have been proposed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	800,000	240,000	1,040,000
2008 年度	900,000	270,000	1,170,000
2009 年度	800,000	240,000	1,040,000
2010 年度	900,000	270,000	1,170,000
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：英作文支援、日英混在文、高品質英語コーパス、Web データ、統合利用、日英対訳コーパス、日英対訳表現、用例ベース

１．研究開始当初の背景

高度情報化・グローバル化社会の到来に伴い英語によるコミュニケーションが益々重要になってきている今日、あらゆる分野・

様々な場面で英作文する必要性が高まってきている。英語を母語としない、初心者からプロの翻訳者までさまざまなレベルにある人々にとっては、丸ごと翻訳という、精度が低いため「使えない」和英翻訳システムよりも、機能が部分的にせよ、「確実に使える」英作文支援システム（ツール）があったほうが助かることは英作文を経験した人なら誰でもうなずくことである。

２．研究の目的

提案研究はこのような「確実に使える」英作文支援システムの開発を目的としている。提案システムをより具体的に述べると以下になる：ユーザが英作文する過程において、部分的に適切な英語表現（単語、フレーズ、または節など）が思い浮かばないとき、本来言い表したい日本語表現をそのまま入力するだけで、すなわち、日英混在の入力文から、適切な英語表現を選択（生成）してくれるようなシステムである。

３．研究の方法

（１）単語レベルでの英作文支援システムの開発　①英語コーパスとWebを対象とした柔軟で高速な検索システムの作成　②訳語選択に用いる最適な文脈の構成方法の考案　③ステップ①で作成した検索システムを利用した最適な文脈による訳語選択手法の提案と評価　④提案手法のシステム実装

（２）英作文支援のフレーズレベルへの拡張　①英語フレーズを訳語候補の組み合わせから生成する手法の開発と評価　②大規模で高精度な日英対訳表現の抽出手法の提案と評価　③上記ステップで開発した手法で抽出した対訳表現を利用した用例に基づくフレーズレベルでの支援手法の提案と評価　④提案手法のシステム実装

４．研究成果

（１）単語レベルでの英作文支援システムを開発した。

開発したシステムではまず、「The no-nuke 運動 is as active as ever before」のような日英混在入力文に対し、日本語部分「運動」を辞書引きし、動詞などの場合は原形を求めるなど形態素解析を行い、訳語候補(motion, exercise, sport, campaign, movement など)を取得する。そして、個々の訳語候補に対し、訳語候補の前後の文脈情報を用いて検索クエリを構成し、高品質コーパス・Web 検索を行い、検索ヒット数を取得する。検索ヒット数が最も多い訳語候補(今の例の場合は movement)をシステムの回答として出力する。

上記システムの開発において、まず、大規模で高品質な英語コーパスを収集するとともに、その高品質英語コーパスを対象とした、単語列だけでなく品詞やワイルドカードなどを用いた柔軟な検索を高速に行えるシステムを開発した。また、超大規模なWebデータも利用可能にするために Google API を利用したWebデータの検索プログラムを開発した。次に、訳語選択にもっとも重要な検索クエリの構成方法（文脈可変手法やルールベース手法など）を考案し、最適な文脈による訳語選択の手法を提案した。さらに、高品質な英語コーパスと大規模 Web データの統合利用手法を提案した。最後に、上記研究成果に基づき、前述のような日本語単語が混ざっている日英混在文を入力とした、日英対訳辞書と検索エンジンから構成される単語レベルでの英作文支援システムを開発した。開発したシステムをプロの日英翻訳者により作成された評価データを用いて評価した。その結果、英単語支援（訳語選択）に 80%という高い正解率が得られ、単語レベルでの英作文支援の有効性が確認できた。

（２）英作文支援をフレーズレベルへ拡張した。

フレーズレベルでは“宗教評論 has not been established as a discipline in this country.”のような日英混在文が入力となる。入力された混在文中の日本語を特定し，形態素解析ツール茶筌を用いて分割する。分割した単語を辞書引きしその訳語候補を取得する。次に訳語候補の組み合わせで検索クエリを構成し、クエリの高品質英語コーパスまたは Web でのヒット回数を調べる。ヒット数の多いものを英訳として出力する。

フレーズレベルでの英作文支援への拡張にあたって、日本語フレーズを構成する個々の単語の訳語候補の組み合わせから最適な英語フレーズを生成する手法を提案し、名詞句と動詞句を支援できるようにした。名詞句は、各訳語候補の全通りの組み合わせ」「単語の間に“of”を入れた全通りの組み合わせ」「前後の単語を逆にしてその間に“of”を入れた全通りの組み合わせ」の３通りの組み合わせである。動詞句は，「各訳語候補の全通りの組み合わせ」「前後の単語を逆にした全通りの組み合わせ」の２通りの組み合わせである。

フレーズレベルへ拡張したシステムの評価結果、名詞句の英作文支援が 71%、動詞句のそれが 67%という正解率が得られ、その有効性が確認できた。しかし、英語フレーズは訳語候補の組み合わせのみから生成されるため、支援できるフレーズの範囲が大きく限定されてしまう。また、名詞句などを不定詞句に英訳するような、異なるパターンへの英訳も支援できない。さらに、単語の訳語候補の数が多いためその組み合

わせの数が膨大となり処理時間がかかってしまう。このような問題を解決するために日英対訳パターンに基づくアプローチを導入し、そのパターン辞書を作成するにあたり必要となる大規模な日英対訳表現を大規模な日英対訳コーパスから抽出することを次に述べるように試みた。

（3）大規模で高精度な日英対訳表現の抽出手法を提案した。

日英対訳表現はたとえば「J: 暗殺されたラビン首相, E: Prime minister Rabin who was assassinated」や「J: 不均衡を縮小するため, E: to reduce imbalance」のようなものである。日英対訳表現を日英の対訳文から構成される日英対訳コーパスから抽出する研究はこれまで多数存在する。しかしそのいずれも抽出対象となるコーパスがきわめて小規模であり、計算機マニュアルや科学雑誌といったドメイン限定のものであった。また、抽出できた対訳表現も極めて小規模であり、抽出精度・再現率も非常に低かった。すなわち、先行研究で提案した抽出手法はいずれも実用レベルにほど遠かった。それらに対し本研究はNICTコーパス、JENNADコーパス、さらにロイター日英記事対応付けコーパスなどさまざまな分野の表現を有する計 28 万文の大規模なコーパスを収集し、そられから実用指向で大規模で高精度な対訳表現の抽出を図った。本研究は、まず、先行研究の提案手法をベースライン手法とし、対訳表現の抽出を行った。その結果、約 19 万対の対訳表現を 0.09 の精度で得られた。次に対訳表現抽出の改良を行い大幅な抽出数の向上と精度の向上に取り組んだ。まず、より多くの対訳表現が抽出できるように日英それぞれからの単語列の抽出方法に改良を加えた。その結果、抽出された対訳表現が約 80 万対に激増し、精度も 0.18 まで向上した。次に、単語列の先頭に「いる」,「こと」,「みたい」などの文法的にありえない不適切な単語がつく表現を、日英それぞれの単語列作成時に取り除けるように人手で作成した規則を導入した。その結果、抽出された対訳表現が約 48 万対に低下したが精度が 0.35 まで大幅に上昇し、結果的に再現率が 0.52 から 0.61 まで向上した。さらに，抽出した対訳表現をできるだけ正しいものに絞り込めるように対訳辞書情報を適用した。その結果、計 28 万文対の日英対訳コーパスに対し、対訳表現約 12 万 5 千個を精度 0.96 で抽出することができた。その結果から、改良手法が従来手法に比べ抽出数と精度の両方において飛躍的に向上し、実用レベルに達したことが分かった。

（4）用例に基づくフレーズレベルでの英作文支援システムを構築した。

構築した用例ベースに基づくフレーズレベルでの支援システムでは、（3）で抽出した対訳表現を加工して日英対訳パターン辞書を作成し、作成した日英対訳パターン辞書の中から入力された混在文の日本語表現と最も類似した対訳表現を取り出し、その英語表現を回答として出力する。より具体的には、入力はこれまで同様、「The President tackle 労働組合の構造改革」のような日英混在文である。このような入力から日本語表現を取り出し、日英対訳パターン辞書と照合を行う。照合を効率的に行うために、日英対訳パターン辞書にある対訳表現は「N の N」や「N を V ため」といったパターンで分類されている。そのため、入力された日本語表現もまず、「N の N」のようにパターン化される。次に、その入力と対訳パターン辞書間の、パターン照合と日本語表現照合の二段階での照合処理が行われる。ただし、同義語や類義語なども取り扱えるように、照合処理における類似度計算には日本語シソーラスを用いている。また、例えば「国民の支持を勝ち取るため」という日本語表現の入力に対し、仮に「勝ち取る」という日本語を含む対訳表現が対訳パターン辞書に存在していないとすれば、システムは「to V the people's support (V: 勝ち取る：carry, conquer, gain, score, walk, win 訳語候補 6 個)」のような、「勝ち取る」の英訳の代わりにその品詞とすべての訳語候補を出力する。本システムへの初歩的な評価によりシステムの有用性と有効性が確認されている。

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 Q. Ma, S. Sakagami, and M. Murata: Extraction of Parallel Translation Expressions for English-Writing Support Systems, *ICIC Express Letters, Part B: Applications*, 査読有, Vol. 2, No. 1, 2011, pp. 113-118

②M. Murata, K. Uchimoto, M. Uchiyama, Q. Ma, and et al.: Using the Maximum Entropy Method for Natural Language Processing: Category Estimation, Feature Extraction, and Error Correction, Cognitive Computation, 査読有, Vol. 2, No. 4, 2010, pp. 272-279

③K. Kotani and T. Yoshimi: Second Language Writing Classification System Based on Word Alignment Distribution, Themes in Science and Technology Education: Special Issue on ICT in Language Learning, 査読有, Vol. 3, No. 1&2, 2010, pp. 223-238

④M. Murata, T. Shirado, K. Torisawa, M. Iwatate, K. Ichii, Q. Ma and T. Kanamaru:

Extraction and Visualization of Numerical and Named Entity Information from a Very Large Number of Documents Using Natural Language Processing, International Journal of Innovative Computing, Information and Control, 査読有, Vol. 6, No. 3(A), 2010, pp.1549-1568

⑤吉見毅彦, 小谷克則, 九津見毅, 佐田いち子, 井佐原均:単語アライメントを用いた英日機械翻訳文の流暢さの自動評価, 自然言語処理, 査読有, Vol.17, No.1, 2010, pp.7-28

⑥吉見毅彦, 小谷克則, 九津見毅, 佐田いち子:逐語訳に着目した日本語学習者作文の自動評価, 教育システム情報学会誌, 査読有, Vol.26, No.2, 2009, pp.191-196

⑦神崎享子, 馬青, 山本英子, 白土保, 井佐原均:コーパスからの形容詞概念階層の構築と評価-実データによる形容詞オントロジーの構築にむけて-, 自然言語処理, 査読有, Vol. 15, No. 4, 2008, pp. 59-88

⑧吉見毅彦, 小谷克則, 九津見毅, 佐田いち子, 井佐原均:テキストの言語的特徴と英語学習者の英文読解能力に基づく英文読解時間予測モデル, 教育システム情報学会誌, 査読有, 2008, Vol.25, No.3, pp.272-281

⑨吉見毅彦, 小谷克則, 九津見毅, 佐田いち子:単語対応付けに基づく日本語学習者による作文の自動識別,情報処理学会論文誌（テクニカルノート）, 査読有, Vol.49, No.12, 2008, pp.4039--4043

⑩村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 塚脇幸代, 井佐原均:大規模記事群からの数値情報に関わるテキストマイニング・可視化, システム制御情報学会論文誌, 査読有, Vol. 20, No. 12, 2007, pp. 482-484

⑪M. Murata, Q. Ma, K. Uchimoto, Toshiyuki Kanamaru, and Hitoshi Isahara: Japanese-to-English Translations of Tense, Aspect, and Modality Using Machine-learning Methods and Comparison with Machine-translation Systems on Market, Language Resources and Evaluation, 査読有, Volume 40, 2006 (Published online: 19 July 2007), pp. 233-242

⑫吉見毅彦, 小谷克則, 九津見毅, 佐田いち子, 井佐原均:英語学習者の英文読解能力推定のための読解時間予測法,情報科学技術レターズ, 査読有, Vol.6, 2007, pp.461--464

⑬小谷克則, 吉見毅彦, 九津見毅, 佐田いち子, 井佐原均:英語学習者の英文読解時間に統語的要因が及ぼす影響,情報科学技術レターズ, 査読有, Vol.6, 2007, pp.457--460

〔学会発表〕（計 28 件）

①坂上信也, 馬青, 村田真樹:辞書情報と規則を用いた大規模な日英対訳表現の抽出, 言語処理学会第17回年次大会、pp. 983-986、2011 年 3 月 10 日、豊橋技術科学大学

②K. Kotani and T. Yoshimi: Classification of Language Learners' Sentences Into Native-Like or Non-Native-Like Sentences Using Learner Sentences and Machine Translation Sentences as Learning Data, Proceedings of International Conference of Education, Research and Innovation, pp.2672-2677, Nov. 15-17.2010, Madrid

③坂上信也、馬青、村田真樹:英作文支援のための大規模な日英対訳表現の抽出、言語処理学会第16回年次大会、pp. 660-663, 2010 年 3 月 10 日、東京大学本郷キャンパス

④K. Kotani, T. Yoshimi, and H. Isahara.: An Evaluation of Reading Assistance Tools in Foreign Language Reading: Syntactic Parsing and Machine Translation, Proceedings of International Conference of Education, Research and Innovation, pp.5677--5684, Nov. 17.2009. Madrid

⑤A. Nishikawa, R. Nishimura, Y. Watanabe, M. Murata., and Y. Okada: Dominant Variant Dictionaries for Supporting Variant Selection, Proc. of IADIS AC 2009, pp.265-269, Nov.20 .2009, Rome

⑥Q. Ma, R. Mori, M. Murata: Development of English-Writing Supporting Systems, 11th Conference of the Pacific Association for Computational Linguistics (Pacling2009), pp. 171-176, Sep. 1-4, 2009, Hokkaido University, Sapporo

⑦K. Kotani, T. Yoshimi, and H. Isahara, H.: Automatic Evaluation of Foreign Language Reading Proficiency based on Reading Time and Linguistic Features, Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics, pp.35-40, Sep.1. 2009, Hokkaido University, Sapporo

⑧K. Kotani, T. Yoshimi, T. Kutsumi, I. Sata, and H. Isahara: Predicting Foreign Language Learners' Reading Proficiency based on Reading Time and Text Complexity", Proceedings of International Technology, Education and Development Conference (INTED), pp.3040-3049, Mar. 2009, Valencia, Spain

⑨H. Hidaka, Y. Watanabe, and Y. Okada: Learning Support for English Composition by Asking Back Questions, Proceedings of CSEDU 2009, pp.367-370, Mar. 2009, LISBON, Portugal

⑩K. Kotani, T. Yoshimi, T. Kutsumi, and I. Sata: An Automatic Evaluation of Writing in Japanese as a Second Language Using a

Word-alignment-based Classifier, Proceedings of the International Conference on Asian Language Processing (IALP08), pp.210-216, Nov. 2008, Chiang Mai, Thailand

⑪M. Murata, M. Iwatate, K. Ichii, and Q. Ma, T. Shirado, T. Kanamaru, T. Torisawa: Extraction and Visualization of Numerical and Named Entity Information from a Large Number of Documents, IEEE NLPKE-08, pp.122-139, Oct. 2008, Beijing

⑫Q. Ma, K. Nakao, M. Murata, and H. Isahara: Selection of Japanese-English Equivalents by Integrating High-quality Corpora and Huge Amounts of Web Data, The sixth international conference on Language Resources and Evaluation (LREC2008), May.28. 2008, Marrakech, Morocco

⑬Y. Zhang, Z. Wang, K. Uchimoto, Q. Ma, and H. Isahara: Word Alignment Annotation in a Japanese-Chinese Parallel Corpus, The sixth international conference on Language Resources and Evaluation (LREC2008), May, 2008, Marrakech, Morocco

⑭M. Murata, S. T. Kanamaru, Q. Ma, Hitoshi Isahara: Non-Factoid Japanese Question Answering through Passage Retrieval that Is Weighted Based on Types of Answers, The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), pp.727-732, Jan.9. 2008, Hyderabad, India

⑮Y. Zhang, Q. Ma, and H. Isahara: Building Japanese-Chinese Translation Dictionary Based on EDR Japanese-English Bilingual Dictionary. In The 11th Machine Translation Summit Proceedings, pp. 551-557, Sep.12. 2007, Copenhagen, Denmark

６．研究組織

(1)研究代表者

馬　青（MA QING）
龍谷大学・理工学部・教授
研究者番号：30358882

(2)研究分担者

吉見 毅彦（YOSHIMI TAKEHIKO）
龍谷大学・理工学部・准教授
研究者番号：50368031
渡辺 靖彦（WATANABE YASUHIKO）
龍谷大学・理工学部・講師
研究者番号：10288665

(3)連携研究者