様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １８ 日現在

機関番号：１３８０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５０１１８５

研究課題名（和文）コーパスを活用した英語技術文書の作成を支援するＷｅｂアプリケーションの開発

研究課題名（英文）Development of a Corpus-Based Web Application to Support Writing Technical Documents in English

研究代表者
宮崎　佳典（MIYAZAKI, Yoshinori）

静岡大学・情報学研究科・准教授

研究者番号：００３０８７０１

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：技術系の人間に必要不可欠な，英語による技術文書作成を支援するアプリケーションを開発した．ユーザが英文を入力すると，技術論文を集めたコーパスの中から類似した文を抽出し例文として提示し，必要に応じてその例文の構造を簡略化して表示する．また，ユーザの学習履歴から各自の語学レベルを推測し，提供する例文をアダプティブに変化させるパーソナライゼーションの仕組みを持たせる．以上のような実用性の高い英文書作成支援システムをWeb上に公開することで，理系学部の学生対象の英語教育に資することも可能となる．

研究成果の概要（英文）：A Web application which supports non-native speakers of English to write technical documents in English was developed. This application presents English texts extracted from corpora, with high similarities with the original English texts input by users, and also with a provision of a series of generalized form of the original sentence extracted by the application, if necessary. Additionally, the function of personalization is employed which guesses language levels of individual users based on their study logs. Allowing the open access of such a Web application on the web will enable students of science majors to have a better learning environment of technical English composition.

研究分野：教育工学

キーワード：e-Learning　英語技術文書　コーパス　Webアプリケーション　簡略化

１．研究開始当初の背景

（１）日本は技術立国でありながら，現状ではその技術を表現する英語力が不足し，世界に充分発信できていない．従って技術者の英語による情報発信をバックアップする体制作りは急務である．

（２）専門的な技術文書に類するものは英語で書かれる場合が多い．技術文書には日常では使用されない単語や表現が頻繁に現れることから，特に EFL(English as a Foreign Language)環境下にある大学生などの英語学習者にとってアカデミックな英文の文書作成は簡単な作業ではない[1]．このような背景から，英語学習者向けに技術文書の書き方を指南する本も多く出版されている．しかしながら，これらの本は技術文書に多く見られる基本的な単語や文法の紹介に終始しており，それだけでは英語で技術文書を書くことはまず不可能である．また，一般的な英文書作成支援ツールは今までにもいくつか開発されているが，技術に特化した英文書作成支援ツールにフォーカスをあてたものは見当たらない．さらに一般英文書用ツールでも，その多くは日本語を入力するシステムであり（[2]など），日本人にとって利用しやすいシステムである反面，機械翻訳の不正確さが原因で不適切・無関係な例文が提示されることも少なくない．それはまたユーザを日本語の母語話者とする者に限定してしまうことにもなる．

＜引用文献＞

[1] Evans,S. & Green,C., Why EAP is necessary: A survey of Hong Kong tertiary students, Journal of English for Academic Purposes, 6(1), pp.3-17, (2007).
[2] 高倉佐和，古郡廷治，TransAid-英文書作成支援システム-，電子情報通信学会技術研究報告, 102(199), pp.7-14, (2002).

２．研究の目的

（１）「１．研究開始当初の背景」に既述したような先行研究の状況に対して，本研究では英語学習者の実際の学習プロセスに可能な限り符合し，学習者のニーズにあった学習効率の高い技術文書作成支援システムを目指している．その意味で，本システムは先行研究の英文書作成支援ツールとはその立場を異にしている．

（２）英語学習者の英文作成のプロセスに即して本アプリケーションの意義を考察する．学習者は英文技術文書作成の際，【Point 1】当該分野のオーソライズされた技術文書から自分が書きたいものに似た英文を探し，それを手本にしようとする．しかし，書きかけの英文中の単語では類似した英文を探すことは容易ではない．そこで，【Point 2】随時類似した単語や表現に置き換えて検索することが必要になる．

（３）さらに学習効率を考えれば，コーパスから類似英文の検索時に【Point 3】技術文書に特有の，当該分野毎の特徴表現やコロケーション（語の組み合わせ）に重み付けすることで他の一般的な表現とは差別化することが必要である．さらに実際に検索される例文は構造が複雑で難解なことが多く，【Point 4】文構造が明確化されるような簡略化機能（文内の部分的な構造を省いたり記号化することで簡略に見せる機能）を付加することによってそれが実現する．

（４）以上の全項目を実装することで実用的な英文書作成の支援が可能となると考える．この理念に基づき，本研究グループは例文提示型支援システム構築のための，初期段階のプロトタイプ作成を終えている．しかし現段階では上記【Point 1】～【Point 4】の動作を実現するには至っておらず，理論に基づく仕様策定および実用に耐えられるレベルの実装が今後必要となる．一般公開を想定した実用性の高い Web アプリケーションを構築するには，詳細な開発工程に裏付けされた動作確認やデバッギング作業，ならびに開発後の複数の実験を通じた有用性の実証が不可欠であり，本申請はこれを目的とするものである．また，本システムは入力には完結した英文のみならず，フレーズ単位や単語入力だけでもそれに応じた適切な出力を返す特性を持ち，ユーザに負担を感じさせないシステムとすることが特長に挙げられる．加えて出力表示用インタフェースには対応語が特定しやすいよう，目的に応じた多段階のハイライト機能を装備する予定である．

３．研究の方法

（１）「２．研究の目的」において，研究内容及びフレームワークについて詳述を行った．一方で，実用化に際し，この英文書作成支援システムには解決しなければならない問題点も多い．それは計算速度と精度の問題に大別される．

（２）本システムの計算速度に関しては，特に入力文に類似した文を大量の用例（コーパス）から検索する際に行われる大規模なベクトル計算時間に依るところが大きい．その計算の抑制のために，すでにデータ構造に工夫を重ねている．さらに本研究では，近年情報科学で身近になった並列化の導入で大幅短縮化を実現できると見込んでいる（提唱しているアルゴリズムは並列処理に親和性が非常に高く，並列化の際はさらに効果的に並列処理されるようコーディング最適化を行う）．また，入力に対して参考になる例文提示の精

度に関しては，多面的アプローチが必要と考えている．コーパスを拡充すれば当然，入力に対するユーザが満足する適切な例文が含まれる可能性は高くなる．これには技術ジャーナルから版権を購入することで質・量ともに充実化を図る．入力文内の単語を適宜置き換えるための同義語グルーピングの精度向上も重要なポイントである．特に技術用語の同義語群はコーパスに依存し，複数のシソーラスなど参照しながらの生成は大量の計算を要し，動的に対応するため新規導入するサーバで即時処理を行う予定である．(【Point 2】)．分野別の特徴表現やコロケーションを取得することにも重点を置き，文中の n-gram（連続する語 n 組）を採用することで，単語間の修飾関係を擬似的に扱うものとする．(【Point 3】)．最後にかつ最重要課題はユーザのパーソナライゼーションである．どの文を類似度が高い・参考になるとみなすかはユーザの研究分野や英文の嗜好性に依存し，また提示すべき例文の簡略化のレベルも同様であろう．本テーマは精度やユーザの満足度に大きな影響を及ぼすことは間違いなく，ユーザの各種学習履歴情報（閲覧したコーパス文，実際に推敲に使用された表現，入力英文，過去に選んだ汎化レベルなど）を戦略的に保存・利用していき，必要に応じて因子分析法などの統計的手法を駆使する．特に簡略化のレベルは現時点においてユーザが指定する形式になっており，システム使用の都度設定する手間を考えると，パーソナライゼーションの必要性は高い（【Point 4】)．

（3）以上のように，本研究で開発する英文書作成支援アプリケーションには実用性のみならず，簡略化機能，パーソナライゼーションに代表される学術的な意義も有する．対象を技術文書作成とし，固有の特徴表現や n-gram を用いて類似度を計算する点も本研究の独創的な点であると考える．

４．研究成果

(1) 2012 年度-2013 年度前半は，調査活動なども含めて理論的側面を重視し，Web アプリケーション全体の枠組みを確定した．特に実用性に向けて処理の高速化，出力精度の向上，パーソナライゼーション，そしてユーザビリティを考慮したインタフェース開発は最重要項目で，各自の役割分担に沿って検討を行い，数回の全体会で摺り合わせを行った．年度末から次年度前半にかけてはシステムのプロトタイプを元に（非公開）Web アプリケーションとして完全実装し，サーバへの移行を行った．併せて，内部の情報系院生などによるパイロット実験を実施し，特にデータの流れを含むモジュール間の接続処理の動作確認作業を行った．

2012 年度-2013 年度前半に行ったこと：

Web アプリケーション全体の枠組みと理論構築に専念し，研究代表者ならびに研究分担者間でミーティングを行い，以下を討議した：
・システム処理の高速化（類似文の検索法の改善，処理の並列化など）
・出力の精度改善（同義語グループ再編成，コーパス拡充など）
・パーソナライゼーションへの詳細アプローチの決定
・学習者特性を考えたインタフェースの考案
・システム処理の高速化（類似文の検索法の改善，処理の並列化など）
→グリーディー算法を採用し，見込みのあるコーパス文飲みに対する計算処理の実行．

・出力の精度改善（同義語グループ再編成，コーパス拡充など）
→不適切なコーパス文に対する修正

・パーソナライゼーションへの詳細アプローチの決定
→文例の簡略化に対する個人嗜好に照らし合わせたパーソナライゼーションの提案

・学習者特性を考えたインタフェースの考案
→類似文検索に加え，フレーズ検索機能を追加したことに対する，GUI のフレキシブルな挙動を実現

また，データの流れを精査しながら，システムの実現可能性／可用性についての議論も重ねた．技術的には，本研究は以下の手順によるアプローチを基本として上記（「研究目的」）の【Point 1】～【Point 4】の実現を目指した．

□Step 1．［同義語グループによる置換］入力された文に対して形態素解析（文中の単語を同定し，品詞を付与）を適用し，文内のキーワードを抽出する．キーワードが同義語グループ内にある場合，その同義語グループ名で置き換える（上述の【Point 2】に対応)．同義語グループはシソーラスや技術用語集などから事前に生成しておく．

□Step 2．［n-gram，特徴表現の抽出］n-gram 要素ならびに特徴表現を抽出する．これらを用いることで，擬似的に文内の修飾関係が検索に考慮されることとなる（【Point 3】)．特に特徴表現については[4]の結果を活用する．

□Step 3．［類似文抽出］技術文献コーパスから入力文に類似した文を見つける方法として，基本的な検索モデルの一つであるベクトル空間モデルを採用する．具体的には，入力文と技術文献コーパス中の各文を Step 1,Step 2 から得られるキーワードや n-gram・特徴表現の使用状況から，それぞれベクトル化することで類似度が与えられる

（【Point 1】）.

□Step 4.［文例の簡略化］技術文献コーパス内の文は実際に論文等で使用された文なので，参考にするには複雑すぎる場合も少なくない．そこで，Step 3 で検索されたものについては，その文構造を簡略化するような処理を行う．最適な簡略化レベルは統計モデル選択の問題に帰着することができ，またその度合いをユーザ自身が制御することもできる（【Point 4】）.

(2) 2013 年度後半-2014 年度前半は,

2013 年度後半-2014 年度前半に行ったこと：

初年度に詳細に行った Web アプリケーションの全体の枠組みに沿って実装を進めた．現時点で動作しているアプリケーション（言語）は Apache，PHP，Perl,Javascript，HTML，そして DBMS として MySQL を組み合わた．またコーパスには，名古屋工業大学の研究グループによって編纂されたコーパスの使用許可をすでに得ており，総計で既に 30 万文を超えている Association for Computational Linguistics, Nature, Scientific American, Biology 系のテキストで構成されるものを利用した．次に，公開サーバとして本システムを実装し，研究者間において実装途中のアプリケーションが常に観察・アクセス可能な環境を整えた.

(3) 2014 年度後半は,

2014 年度後半に行ったこと：

実証実験を行った，加えてデータ解析を行い，システムの有用性を検証した．また，さらに先を見据えて本 Web アプリケーションのさらなる機能を追加すべく設計を開始した.
また，本システムの評価における比較対象として，Google などのサーチエンジンによる英作文支援がある．サーチエンジンによる英作文支援は，先行研究でも比較対象としてよく挙げられており，また現実的にもその精度，計算速度などを相対評価するのに適している．また，各種履歴を出すことで学習者が本 Web アプリケーションを用いてどのように英作文を遂行していくのか，途中に表出した間違いをどのように訂正していくのか，などに焦点を当てて観察した.

さらに，今後のアプリケーション機能拡充を視野に置き，類似の表現やコロケーションを抽出するための方策を考えた．今回の類似性計算には根底に n-gram を採用しているが，それは単語の組み合わせとして擬似的な類似性を考えているに過ぎず，学習者の考える句単位やチャンク単位での配慮が必要である（例えば“ take A into consideration ”の A に単語が来ても句やチャンクが来ても類似表現として認識させる）．また，例文提示の精度に関しては，多面的アプローチが必要と考えている．また，分野別の特徴表現やコロケーションを取得することで，コーパス内の各文章に“ 技術文章らしさ ”の指標を与えることで最終的な類似文ランキングのアルゴリズムを検討した.

最後に研究の結実として，以下の語学関係学会や e-Learning 関係の学会などで発表などを行った（外国語教育メディア学会中部支部第 84 回支部研究大会，Proceedings of The 7th International Multi-Conference on Society, Cybernetics and Informatics (IMSCI 2013), 5th Independent Learning Association Conference 2012 (ILAC2012), 日本教育工学会 第 28 回全国大会）.

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６件）

① 戸沢 信晴，宮崎 佳典，田中 省作，チャンク情報を考慮した例示型英文書作成支援ツール，統計数理研究所共同研究リポート 338，査読無，pp. 23-35 (2015).

② Y. Miyazaki, S. Tanaka, Y. Koyama, Development of a Corpus-Based Web Application to Support Writing Technical Documents in English, World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 査読有, Vol. 2014, No. 1, pp. 1371-1380（2014）.

③ 戸沢 信晴，宮崎 佳典，田中 省作，技術文献コーパスを用いた例文提示型英文書作成支援ツールの開発，電子情報通信学会技術研究報告 信学技報 114(82),査読無，69-72 (2014).

④ 田中 省作，宮崎 佳典，小山 由紀江，藤枝 美穂，分野依存性を考慮した用例提示型英文書作成支援ツールの開発，教育システム情報学会 研究報告，査読無，Vol.28，No.2，pp. 79-84 (2013).

⑤ Y. Koyama, S. Tanaka, Y. Miyazaki, M. Fujieda, Development of a Corpus-assisted Writing System for Research Papers by Science and Technology Students, ILAC Selections -- Autonomy in a Networked World, 査読有，65-67 (2013).

⑥ 山本 昇平，宮崎 佳典，技術文献コーパ

スを用いた英文書作成支援ツールの開発－類似文検索機能とパターン検索機能－，統計数理研究所共同研究リポート 295，査読無，pp. 71-95 (2013).

〔学会発表〕（計４件）

① 戸沢 信晴，宮崎 佳典，田中 省作，チャンク情報を考慮した例示型英文書作成支援ツール，外国語教育メディア学会中部支部第 84 回支部研究大会，査読無，2014 年 11 月 22 日，静岡大学浜松キャンパス（静岡県，浜松市）．

② Y. Miyazaki, S. Tanaka, Y. Koyama, A Tool Supporting Writing Technical Documents in English Using Corpora: Retrieving Functions by Cosine Similarity and Pattern Matching, Proceedings of The 7th International Multi-Conference on Society, Cybernetics and Informatics (IMSCI 2013), 査読有, 2013 年 7 月 10 日, オーランド（フロリダ州, 米国), pp. 129-134 (2013).

③ Y. Koyama, S. Tanaka, Y. Miyazaki, M. Fujieda, Development of Corpus-assisted Research Paper Writing System for Science and Technology Students, 5th Independent Learning Association Conference 2012 (ILAC2012), 査読有, 2012 年 8 月 31 日, ウェリントン（ニュージーランド), p. 33.

④ 小山 由紀江，宮崎 佳典，藤枝 美穂，田中 省作，科学技術コーパスに基づいた英語論文ライティング支援システムの構築とその評価，日本教育工学会 第 28 回全国大会，査読無, 2012 年 9 月 16 日，長崎大学文教キャンパス（長崎県，長崎市)，pp. 523-524 (2012).

〔その他〕
ホームページ等

□技術文書作成援用 Web アプリケーション
http://lmo.cs.inf.shizuoka.ac.jp/~tozawa/ewss/web/index.php

６．研究組織
(1)研究代表者
宮崎 佳典（MIYAZAKI, Yoshinori）
静岡大学・情報学研究科・准教授
研究者番号：00308701

(2)研究分担者
小山 由紀江（KOYAMA,Yukie）
名古屋工業大学・工学教育総合センター・教授
研究者番号：20293251

田中 省作（TANAKA,Shosaku）
立命館大学・文学部・教授
研究者番号：00325549