2004年度修士論文

検索エンジンを使った
英作文支援システムの構築

提出日： 2005年2月2日

指導： 山名 早人 助教授

早稲田大学大学院理工学研究科
情報・ネットワーク専攻

学籍番号：3603U027-5

大鹿 広憲

概 要

近年、英語の必要性はますます高まってきており、英作文をする機会が増えてきた。それに伴い、多くの Web 和英辞典や翻訳システムが開発されてきている。しかし、和英辞典だけでは文の構造に関しての情報が得られず、十分な英作文ができない。また機械翻訳は、原文の構文と字句を反映させしまう直訳と呼ばれる訳文に近くなってしまう問題点がある。さらに、対訳を行うときに一つの名詞や動詞に対して複数の英単語が存在したり、前置詞も数多く存在するので、どれを使ったらいいか迷うことが多い。

以上の問題点を解決する方法として、検索エンジンを使った翻訳の方法がある。Web ページは、人手で作成されたものが多い。従って、検索エンジンを用例データベースにすることによって、多量の Web ページを用例として参照できる。また、作成した英語の文章の文型に対し、汎用性の高い文型を用例と共に検索結果件数で調べることができるという利点もある。

しかし、このような作業においては、それぞれのフレーズを検索エンジンで検索し、検索結果を見て比較するという手間がかかってしまったり、ワイルドカードを使用して検索を行ったときにそれぞれの検索結果を見ていくのが大変であるという問題がある。

そこで、本稿では以上の問題を解決するために検索エンジンを使った英作文の作業を自動化した翻訳サポートシステムを構築した。本システムを構築することにより、各フレーズにおいて検索式を入力して調べるという手間が省くことができ、英作文作業の支援をすることができると考えられる。検索エンジンは GoogleAPI を用いた。

実験として、任意の英日対訳集から選んだ日本語文と英語文を正解データとし、日本語文に対し翻訳ソフトで英訳を行い、ユーザが本システムを使って修正を施した。評価の結果、本手法の有効性を示すことに成功した。

目 次

第1章　はじめに

近年、企業の海外事業展開が活発になってきたことや、英語教育の推進から英語に触れる機会が増加してきている。また、インターネットの普及によって、アルファベット順に並んでいる辞書を引いて単語を調べるより、単語を入力してその単語に関する情報が瞬時に出てくる Web 上のシステムの方が便利なことから、「Web 上で調べる」機会が多くなってきた。

本章では、英作文の作業の支援に着目し、研究の背景と目的、本論文の構成について述べる。

1.1　研究の背景

「Web 上で調べる」といった行為が多く見られるようになった今日、多くの Web 上の和英辞典や翻訳ソフトが多く開発されてきた。しかし、和英辞典だけで英作文を行うには以下の問題点が挙げられる。

- 意味情報と簡単な使い方の情報しかないので、単語の使い方についての情報が十分ではない。
- 一つの日本語に対する英訳は複数あるので、どのように使い分ければいいのか分からない。

英語において、前置詞の使い方は様々あり、文型のパターンも数え切れないほど存在する。英語に熟達したノンネイティブな人でない限り、和英辞典だけで英作文を行うことは難しいと考えられる。

また、翻訳ソフトを使った機械翻訳による英作文においても、状況を判断せず、原文の構文と字句をそのまま反映させてしまう直訳と呼ばれる訳文に近くなってしまうという問

題が挙げられる。

現在、日英に関する翻訳の関連研究として、コーパス作成によるデータベースを作成し、様々な訳に対応する方法 [5][6][7][8][9]、機械翻訳が英訳しやすいように日本語を適切な表現に言い換える方法 [11][12] 、英作文の作業を支援する手法 [13][14][15] が提案されている。しかし、英語には様々なパターンがあり、網羅性に欠けるといった欠点があったり、法則の自動化を実現するには困難であるために、実際に Web で提供されているものが少ないのが現状である。

1.2 研究の目的

以上の問題を解決する方法として、検索エンジンを使った検討がある。例えば、「軌道に向けて打ち上げられた」という日本語文を英語にするとき、「〜に向けて」の部分はどのような前置詞を使ったらいいか迷うことがある。「〜に向けて」の英訳に当たる前置詞の部分をワイルドカードに置き換えて、「launched * the orbit」でフレーズ検索を行うとどの前置詞が使われているかを調べることができる。フレーズ検索とは、語句の並びをそのまま検索する機能である。

また、「医療施設」を英語にする場合、「medical facility」「medical institution」の 2 通りが考えられる。どちらが一般的に使われているかを調べるときに、以上の熟語を検索エンジン Google[1] でフレーズ検索を行う。フレーズ検索を行うと、「medical facility」の方が検索結果件数が多いことがわかる。従って、「医療施設」の訳は「medical facility」として使われるのが多いことがわかる。

従って、普段「情報検索」として利用する検索エンジンを、「表現検索」という形で利用することによって、英作文の検討を行うことができる。しかし、以上のテクニックは検索エンジンの検索テクニックを知っているものでないと活用することができず、またフレーズ検索を行ったときに検索結果を比較することや用例を参照するときに効率が悪いという問題点が生じる。

以上のことから、本論文では検索エンジンを使った英作文の検討の作業を自動化した、英作文支援システムの構築について述べる。本システムを構築することにより、各フレー

ズにおいてクエリを入力して調べるという手間が省くことができ、スムーズな英作文の検討が行えると考えられる。また、あまり英語を話せないネイティブな人達が英作文を行うときに本システムを活用することによって、英作文の作業を支援することができると考えられる。

1.3 本論文の構成

本論文は、本章を含めて 6 章から構成される。以下、第 2 章では日英翻訳に関する関連研究を述べる。第 3 章では検索エンジンを使った英作文の検討について述べる。第 4 章では、本システムの構成として第 3 章で述べたテクニックを自動化するための技術について述べる。第 5 章で本システムの評価及び実験を行い考察を行う。第 6 章で本論文のまとめとして今後の課題を述べる。

第2章　日英翻訳に関する関連研究

機械翻訳とは、コンピュータプログラムによって機械的に翻訳を行なうことであり、定められた法則に基づいて、データベースを使って翻訳を行う。

本章では、日英に関する翻訳の関連研究について述べる。日英翻訳に関する関連研究は、主に 3 種類の分野に大別できる。

- コーパス作成によるデータベース
- 機械翻訳のための日本語の検討
- 英作文支援

以下、それぞれについて述べ、本研究の位置づけについて述べる。

2.1　コーパス作成によるデータベース構築

日本語は複雑で、特定の単語や文について幾通りもの翻訳・解釈の仕方があるために、完全な翻訳が難しい。従って近年では、幾通りもの意味解析に対応した対訳コーパスの作成の研究が多くなされている。本節では、コーパス作成の関連研究について述べる。

2.1.1　日英対訳パターンの自動抽出

鳥取大学の道祖尾らは、日英対訳パターンの候補を自動的に抽出する方法を提案している [5]。対訳コーパスから、N-gram 統計処理方法によって日本語表現と英語表現を抽出する。N-gram 統計処理方法とは、複数の文から共通の文字列を自動的に発見し、抽出する方法である。

N-gram 統計処理方法の例を図 2.1 に示す。文 (1) は「ABCDE」、文 (2) は「FBCG」という文字列である。文 (1) と文 (2) の共通な文字列は「BC」である。N-gram 統計処理方法では、共通の連続文字列である「BC」を抽出する。

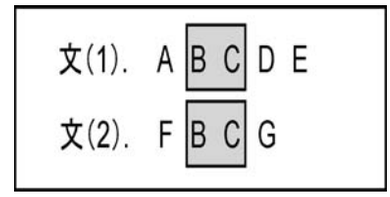

図 2.1: N-gram 統計処理方法

そして、同じ対訳文から抽出されている日本語表現と英語表現を探すことで、連続する単語から成る日本語表現と英語表現の日英対訳パターンの候補を抽出する。

対訳コーパス 36,500 文に対して、日英の対訳パターンの候補 803 個抽出することに成功している。しかし、英語と日本語文のパターンは無数に存在するので、更なる対訳パターンを作成する必要があることを今後の課題に挙げている。

2.1.2 情報検索システムを利用した日英対訳語推定

豊橋技術科学大の鈴木らは、情報検索システムを利用した対訳語抽出モデルを提案している [6]。

モデルの処理手順を図 2.2 に示す。情報検索システム IR は、日本語検索語 $query_s$ とコーパス C_s を入力とし、C_s から $query_s$ に関連する日本語文書集合 D_s を出力する。次いで、言語横断情報検索システム $CLIR$ は、日本語検索語 $query_s$、コーパス C_s、対訳辞書 $Dict_{st}$ を入力とし、$query_s$ に関連する英語文書集合 D_t を出力する。そして、訳語抽出システム TE では、文書集合 D とコーパス C を入力することによって、それぞれの統計量を用い

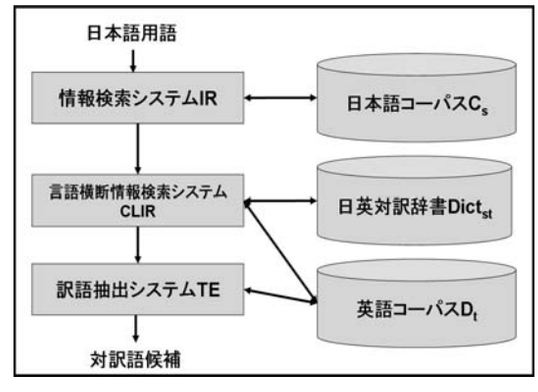

図 2.2: モデルの概要 ([6] より引用)

て、D に含まれる語に対してスコアをつけ、対訳語候補を抽出する。

提案されたモデルで実験を行った結果、正解率は 49 %と精度に不十分さが残ったが、語の相関をとるシステムより効果があったことを実証している。

2.1.3 パラレルコーパスからの対訳発見

京都大の荒牧らは、文同士の対応がとられた日本語と英語の対訳文を入力とし、句レベルで対訳対を発見するシステムについて述べている [7]。システムの構成は以下のとおりになっている。

1. 日英両言語の文を構文解析し、句を単位とした依存関係を得る。
2. 辞書引きによって、日英両言語の語の対応を調べ、句レベルの対訳対を発見する。
3. 対応がつかず残った句については、依存関係等の統語情報や全体の整合性から、対訳対を発見する。

対訳対の発見の例を図 2.3 に示す。図 2.3 では、2 つの対訳対「America ／アメリカ」、「role ／役割」は既に辞書引きで発見することができる。「play」と「果たす」は辞書引きでは対応することができず残ってしまうので、「play ／果たす」を対訳対として発見する。

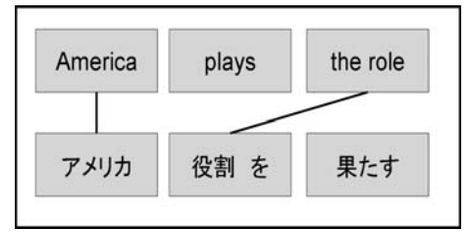

図 2.3: 句レベルの対応 ([7] より引用)

システムが発見する対訳対の精度を調べるために、科学技術庁と経済企画庁の白書と学研辞書の用例の句を人手で対応づけた評価セットを用いて評価を行ったところ、80%の精度で対訳対を発見することに成功している。しかし、翻訳システムに搭載しての評価は行っておらず、今後の課題となっている。

2.1.4 用例翻訳と統計翻訳の混合

用例翻訳は、対訳コーパスを一種のデータベースとして見なし、入力文と似た用例を用例ベースから検索する。統計翻訳は単語翻訳と語順調整を組み合わせて翻訳を行う。

ATRの今村らは、統計翻訳のモデルを利用して最適訳選択を行う、構文トランスファ方式の用例翻訳器を提案している [8]。

システムの構成を図 2.4 に示す。用例ベースの構文変換では変換規則を参照しながら構文解析やマッピングを行い、目的言語の木構造を作成する。例えば、「バスは 11 時に停まる」という文は、原言語文法で構文解析を行うと、「X_{NP}は Y_{VP}ます」となる。次いで、目的言語文法で構文解析を行うと、「X_{NP} $will$ Y_{VP}」となるので、用例として「((バス, 停まる)..)」がデータベース化される。次に統計的表層生成で、木構造から生成される単語列のうち最適な列を探索する。「最適」な組み合わせは、言語モデルと翻訳モデルから決定する。言語モデルには n-gram、翻訳モデルには語彙モデルだけを用いる。

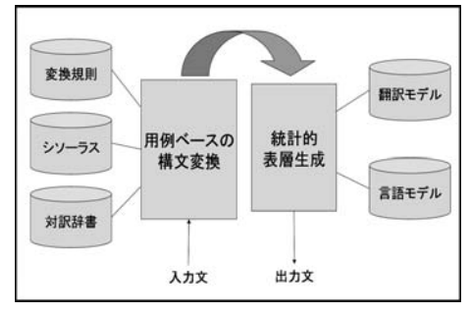

図 2.4: システムの構成([8] より引用)

旅行会話に頻出する表現を集めた 510 文のテストセットに対して英会話ネイティブ話者 1 名による主観評価を行ったところ、「理解可能訳」と判断したものが 70.4%に達していたことを示している。

2.1.5 結合価文法による訳語選択能力の評価

鳥取大の金出地らは、動詞と名詞の訳語選択における、結合価文法の効果を定量的に検証している [9]。結合価パターンは、用言を中心に意味的に必須の格要素 (名詞＋助詞) を記述したものである。表 2.1 にパターンを例示する。意味的関係の記述により用言と名詞の間に意味的な制約が生まれ、日本語解析で発生する意味上の多義が解消されると考えられる。

表 2.1: 結合価パターンの例 ([9] より引用)

見出し語	日本語文型	英語文型
送る	N1(人) が N2(休暇) を送る	N1 spend N2
送る	N1(人) が N2(生活) を送る	N1 live N2

まず対象とする文の用言により構文意味辞書を検索し、対応しうるパターンを大まかに選び、格要素の意味属性などが最も多く一致するパターンを一つ選択する。日英パターン対で登録されているため、パターンが決定することにより動詞の訳語が決定する。次に日本語パターンの格要素に対応する名詞の意味属性から名詞の訳語を選択する。

IPAL 辞書 [10] に登録されている基本動詞 861 語と基本名詞 1,081 語に対し評価を行ったところ、平均で 90%以上の精度を出すことに成功している。

2.2　機械翻訳のための日本語検討

日本語には同じ言葉でも文の構造によって複数の意味を持つ。文の構造を判断しなくては、正確な訳が行えない。従って、今日では英語に翻訳する際に日本語を翻訳しやすい表現に直す方法や、特定の品詞における英訳の規則化などの日本語の検討に関する手法が提案されている。

しかし、日本語の検討の手法は、手動で行ったものであったり、自動化はされていない。さらに、助詞の種類は多く存在するので、全てを網羅するのは難しいと考えられる。本節では、機械翻訳のための言い換えの関連研究について述べる。

2.2.1　機械翻訳のための助詞の言い換え

機械翻訳は、文のパターンによって複数の意味を持つ語句を使い分けるのは難しい。従って、日本語を機械翻訳が翻訳しやすい表現に言い換える手法が提案されている。

京都大の松吉らは、機能表現「なら」をどのように訳すかを調査し、「なら」を適切に言い換えることによって翻訳品質が向上したことを示している [11]。同じ「なら」でも、「名詞＋なら」と「形容詞＋なら」では表現の方法が違う。「名詞＋なら」において、名詞の部分が話題を表している場合は、「なら」を「は」に置き換えることができる。

2.2.2　数量表現の翻訳方法

鳥取大の延原らは、従来の翻訳規則、精細度、数詞の桁数、助数詞、名詞の意味属性などに着目した接頭・接尾辞の翻訳方法を提案している [12]。

「程度」は、「砂糖を 10kg 程度」と「予算は 1500 円程度」とで訳し方が異なってくる。前者は「only」とするのが良いと考えられるが、後者は複数の候補が存在する。従って、後者は訳語を生成せず、複数の解を正解とする規則と定義する。また、「以上」においては「度数を表す助数詞」は「over」、「avobe」を使い、桁数の多い数値の場合は「more than」と定義している。

以上のように、文献 [12] では、「程度」、「以上」などの使用頻度の高い 19 種類の接頭・

接尾辞について翻訳規則を提案している。

新聞記事、機能試験文集の数量表現に対して実験を行ったところ、頻度が最も高い翻訳を正解例とする翻訳方法と比べて、平均で精度が 20%向上していることを示している。

2.3 英作文支援に関する関連研究

日英の機械翻訳においては、あらかじめ「A の訳は B」と 1 通りに限定してしまう傾向がある。従って近年では、データベースを利用して英作文作業を支援する研究も行われている。本節では、英作文支援に関する関連研究について述べる。

2.3.1 英文アブストラクト作成支援ツール

リコーの成田は、文間のつながりを重視した、パラグラフ単位での英文作成支援ツールを試作している [13]。

アブストラクト作成支援ツールでは、アブストラクト全体の文章構成に着目してお手本となる用例を検索するための支援機能と、アブストラクトを構成する特定の役割を持った文の用例文を検索するための支援機能を実装している。また、文単位での英作文において、統語的あるいは語彙的な側面からの支援を行うための機能も実装している。言語資源として以下のようなものを利用している。

- 言語情報つき英日対訳論文アブストラクトコーパス
- 文構造パターンデータベース
- コロケーションデータベース
- エラーサンプルデータベース
- テンプレートファイル

英日対訳論文アブストラクトコーパスは、マルチメディア、画像処理、自然言語処理の 3 分野の用例を収録している。

以上のシステムをユーザに評価してもらったところ、以下のような課題が残った。

- Web-Based のツールが望まれている。
- 用例の収録分野を大幅に増やす必要がある。
- 和英辞書引き機能を取り込む必要がある。

また、アブストラクト作成支援ツールは UNIX 上でしか動作しないことから、Windows での実装も課題に挙げられていた。

2.3.2 TransAid

電気通信大学の Takakura らは、機械翻訳の質が低いことを指摘し、文書作成支援システム TransAid を提案している [14]。

処理の流れを図 2.5 に示す。日本語の文章と、市販の機械翻訳システムによるその翻訳例を入力とする。次いで、翻訳システムの出力を訂正したり、洗練したりして、特定の目的に合った英語にするため、有用な英語の例文はインターネット・コーパスから抽出したデータベースを使用する。

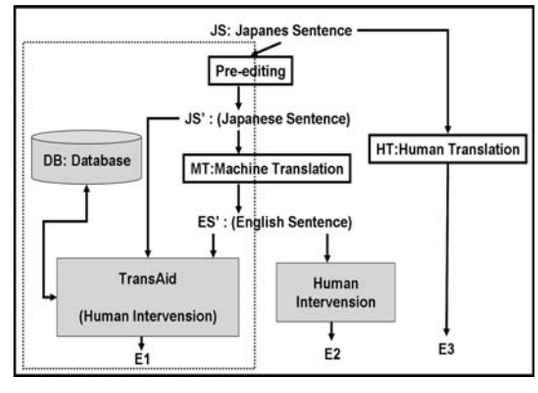

図 2.5: モデルの概要 ([6] より引用)

文献 [14] で取り上げられている例を以下に示す。「第 5 回自然言語処理会議を 2000 年 7 月 20 − 23 日に東京で開く」という文について検討を行うとする。日本語文を S_1 をすると、S_1 から $W =$ (自然言語処理, 会議, 開く) を得る。日英辞書を開くことにより、W 中の各語に対して以下のようなリスト M を得る。

- 自然言語処理：”natural language processing”
- 会議：conference, meeting, table, council, congress, convention, consultation
- 開く：hold open ”set up” ”yield to” throw

次に以上の語義の組み合わせを含む文がデータベース中に存在するかを調べると、「natural language processing」を含む例文が 75 文、「natural language processing」と「meeting」を含む例文が 1 文、「natural language processing」と「conference」を含む例文が 5 文、「natural language processing」と「meeting」、そして「meeting」を含む例文が 1 文、「natural language processing」と「conference」、そして「hold」を含む例文が 4 文検索されることがわかる。以上から、「自然言語処理の会議を開く」という訳は、「natural language processing」と「conference」、そして「hold」を使うのが良いということをシステムが提示する。

評価方法として、5 人の大学院生に TransAid を使って書かせた英文書を 3 人のネイティブスピーカーが評価する方式を用いている。評価の基準として、機械翻訳システムの出力と比較して、文構造の質の改善度と、意味の捉えやすさの改善度を 100 〜-100%で表した。結果として、全体平均で 70%の改善度があったことを示している。

インターネット・コーパスに検索エンジンを用いているところは本研究と似ているが、学会に関するページのみを集め、動詞名詞に限定した訂正を行っているため、前置詞を使った熟語やその他の品詞の分野に対して、汎用性がないのが欠点である。

2.3.3 WebLEAP

鹿児島大の Yamanoue らは、Web の知識を使った文書作成支援システム (WebLEAP) を構築している [15]。WebLEAP は入力された文や表現に含まれる単語の列の、WWW 上の出現頻度をグラフィカルに表示するものである。

WWW 上の出現頻度を調べるには Google を使っている。フレーズ検索を行うことによって、単語の出現頻度を調べている。文章中の各部の出現頻度が色分けされてグラフィカルで表示され、グラフィカルなバーをクリックすることによって、キーワード前後の文脈を一緒に表示する索引方式である KWIC(KeyWord In Context) によって、用例を参照することができる (図 2.6)。

WebLEAP を使った検討の例を示す。前置詞の検討について、”please use this by your own risk”と入れて解析すると「by your own risk」の件数が少ないことがわかる。そこで、

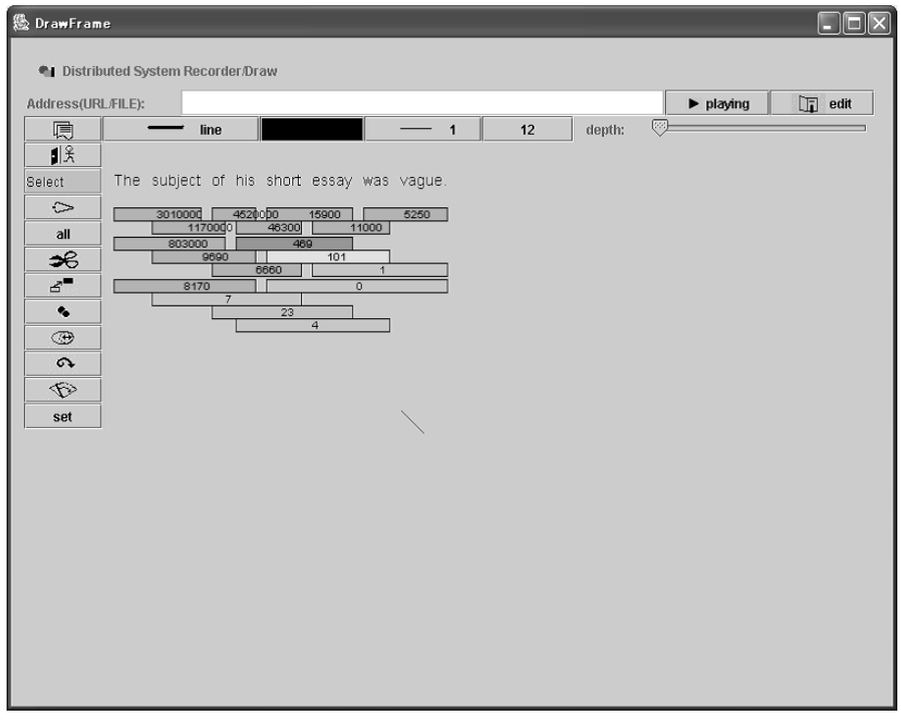

図 2.6: WebLEAP の結果表示

「by your own risk」をクリックすると KWICが表示される。KWIC を参照すると、日本やメキシコなど英語のネイティブスピーカーが少ない地域の URL が多いことがわかる。次に、your own risk をクリックして KWIC を表示して例文を見ると、「at」 が、「your own risk」の前に付く場合が多いことがわかる。従って、「your own risk」の前は「at」が良いということになる。

検索エンジンのフレーズ検索を用いているところは本研究と非常によく似ているが、この WebLEAP はフレーズ検索をした結果を提示しているだけに過ぎず、最適なフレーズを提示する機能がないので、修正に関してはユーザ自身が用例を参照しながら思いつく必要があるという点が欠点である。

2.4 関連研究のまとめ

本節では、関連研究のまとめを表 2.2 に示す。

第 2 章で紹介した手法は、翻訳の法則を自動化を実現するには難しいものであったり、実際に Web で提供されているものが少ないのが現状である。また、英作文のコーパスを構築する際にも、英語には様々なパターンがあり、網羅性に欠けるといった欠点がある。

また、文献 [16] では、機械翻訳において以下のような問題点を挙げている。

- 多義性の理解
- 構文解析に関する問題点

例えば、名詞の「問題」の訳は、「question」「problem」「issue」などがあり、共起する名詞の種類や文の構造によって訳し方が変わってくる。人間用の辞書であれば、意味を列挙しておくだけで人間が判断することができるが、システムが多義性を判断するのは難しいと考えられる。また、人間でもネイティブスピーカーではない人であったり、英語の理解が乏しい人だと使い分けが判断できない場合がある。

また、構文解析に関しては、「疑問文」や「否定」「比較」といった語順に関するルールが決まっているものに対しては機械翻訳で対応できるが、「挿入」「省略」「倒置」などの語順のルールの対応が悪い。また、「〜する人」や「〜するもの」などの形容詞節においては、関係代名詞を使った構文にしてしまう傾向もあり、関係代名詞を使って修飾されない語句にも関係代名詞をつけてしまう傾向がある。機械翻訳において、辞書データを拡充する方が翻訳精度を向上できるので、構文解析の性能が悪いということが考えられる。

本研究では、Web 上での翻訳のサービスを実現することを目的に、検索エンジンを用いて、訳語候補を自動的に提示できる英作文支援システムを構築した。

表 2.2: 関連研究のまとめ

分野	提案者	概要	問題点
コーパス作成	道祖尾 [5] (2003)	対訳コーパスから、N-gram 統計処理方法によって日本語表現と英語表現を抽出する。	自動抽出したコーパス数が少ない。
	鈴木 [6] (2002)	情報検索システムを利用してコーパスを作成し、日英対訳語推定	評価対象を名詞に限定している。精度が不十分。
	荒牧 [7] (2001)	文同士の対応がとられた日本語と英語の対訳文を入力とし、句レベルで対訳対を発見する。	翻訳システムに搭載しての評価は行っていない。
	今村 [8] (2004)	用例翻訳と統計翻訳のモデルを利用して最適訳選択を行う翻訳	動詞と名詞の関係だけに着目しているため、網羅性に欠ける。
	金出地 [9] (2003)	動詞と名詞の訳語選択における、結合価文法の効果の検証	
日本語検討	松吉 [11] (2004)	機能表現「なら」をどのように訳すかを調査し、「なら」を適切に言い換える。	自動化を行っておらず、Web 上で提供されていない。
	延原 [12] (2001)	数詞の桁数、助数詞、名詞の意味属性などに着目した接頭・接尾辞の翻訳方法の提案	
英作文支援	成田 [13] (2001)	文間のつながりを重視した、パラグラフ単位での英文作成支援ツールの試作	分野を限定している。UNIX 上でしか動作しない。
	Takakura[14] (2002)	インターネット・コーパスを使った英作文支援システム	分野が限定されており、特定の品詞しか評価していない。
	Yamanoue[15] (2004)	入力された文や表現に含まれる単語の列の WWW 上の出現頻度をグラフィカルに表示する。	フレーズ検索をした結果を提示しているだけで、修正を提示する機能がない。

第3章　検索エンジンを使った英作文の検討

英訳した文章に対して、検索エンジンのフレーズ検索を行うことによってその文章がどのくらい使われているかの汎用性を調べることができる。また、作成した英訳の気になる部分をワイルドカードに置き換えることによって、ワイルドカードにした部分にどのような単語が使われているかを調べることができ、的確な英訳が見つかることがある。

本章では、検索エンジンを使った英作文の検討について述べる。まず英作文の作業に検索エンジンを使うことの利点をのべ、検索エンジンの代表的な存在である Google[1] を紹介する。以下、文献 [3] を参考に、検索エンジンを使って英作文の検討する検索テクニックについて、例を挙げながら説明する。

3.1　英作文に検索エンジンを使うことの利点

英作文に検索エンジンを使うことの利点は以下のとおりである。

1. 多量の Web ページの参照できる。
2. 全ての品詞について検討ができる。
3. Web 上のサービスとして構築しやすい。

Web ページは、人手で作成されたものが多い。従って、検索エンジンを用例データベースにすることによって、多量の Web ページを用例として参照できるほか、汎用性の高い文型を用例と共に検索結果件数で比較ができるという利点がある。また、多量の Web ページを参照するので、どの品詞においても汎用性を調べることができ、網羅性が高いという利点もある。更に、Google は API として公開されているので、API を使うことにより、Web 上でのサービスを実現することができる [4]。

3.2 検索エンジン Google[1] の紹介

Google は、1998 年に現 CEO の Larry Page と社長の Sergey Brin が米スタンフォード大学大学院在籍中に開発した検索エンジンによるサービスであり、検索が速く、求めている情報が見つかりやすいことから世界一のシェアを誇っている検索エンジンである [4]。

図 3.1: Google のトップページ

Google のトップページを図 3.1 に示す。Google のトップページは、1 つのテキストボックスと 2 つのボタンだけから構成されており、テキストボックスにキーワードを入力することによってそのキーワードを含んだページを検索結果として表示する。

一見簡素に見えるシステムだが、膨大な情報が蓄積されている。そこで、Google には多くの検索構文があり、効果的に使うことによってより的確な情報が見つかることがある。

Google で複数のキーワードを検索したいときには、「スペース」で区切って検索を行う。例えば、「情報」と「通信」というキーワードを含んだページを検索したい場合は、「情報 通信」という形でキーワードを入力する。すると、Google は「情報」と「通信」の両方のキーワードを含んだページを検索する。以上の検索の方法を AND 検索という。

また、「情報 OR 通信」という形でキーワードを入力すると、「情報」と「通信」のどちらかを含んだページを検索する。以上の検索の方法を OR 検索という。

3.3　フレーズ検索

パソコンで英文を入力する時、英語の単語間の区切りには「スペース」を用いる。従って、入力した英文をそのまま Google にクエリとして入力すると、全ての単語の AND 検索になってしまい、語順を保ったままの検索ができない。

以上の問題点を解決する方法として、「フレーズ検索」がある。語順を保ったまま検索を行いたい文章に対して半角の二重引用符 (ダブルコーテーション) で囲んで検索を行うと、その単語の並びのものだけを検索することができる。

以下にフレーズ検索を用いた英作文の検討について述べる。

3.3.1　フレーズ検索の特徴

例えば、「副詞の重要性」という文章の英訳の候補として、「importance of welfare」が挙げられるが、これをそのまま検索する (図 3.2 左) と、「welfare」と「importance」を含んだページの検索になってしまい、訳語をチェックすることができない。さらに Google では、頻繁に使われる言葉や文字 (「I」,「a」,「the」,「of」) はストップ語句として扱われるので、これらの単語を検索キーワードとして入力しすると自動的に除かれてしまう。

そこで、「”importance of welfare”」と半角の二重引用符で囲んで検索を行う (図 3.2 右)。すると、「”importance of welfare”」の並びのものだけを検索することができるので、検索結果件数などの情報から英訳した文型の汎用性を調べることができる。フレーズ検索を行った結果、検索結果件数が 0 件であると、その表現は使われていないということが分かる。

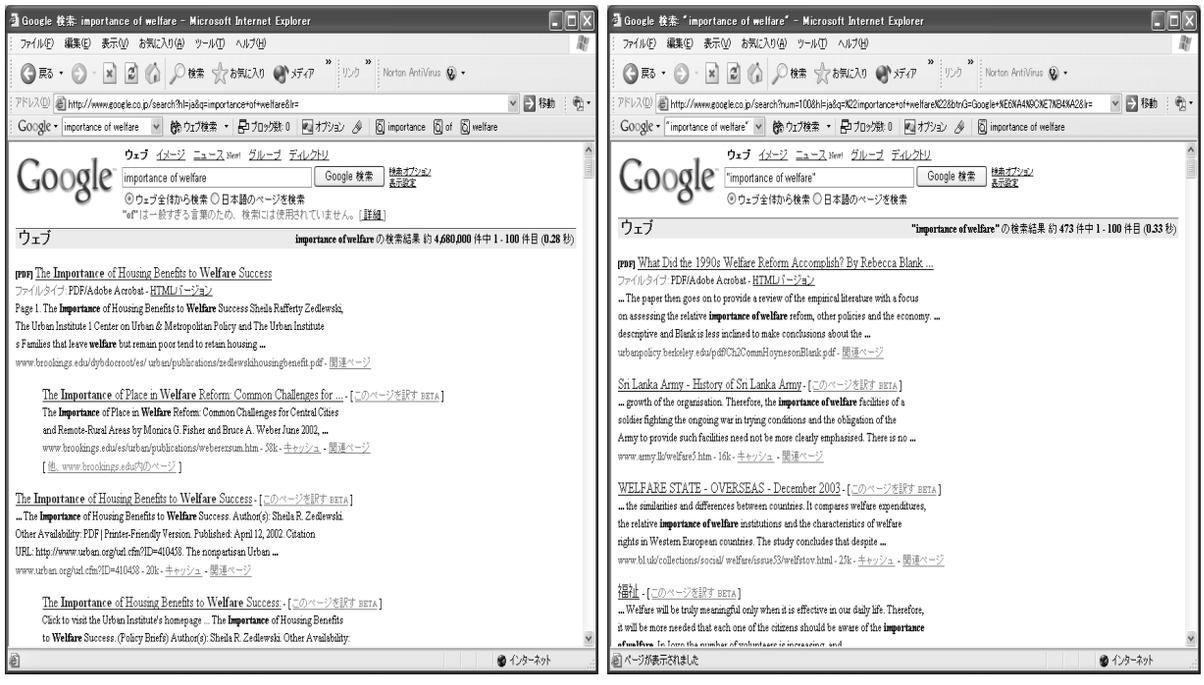

図 3.2: フレーズ検索の有無の違い

3.3.2 フレーズ検索を用いた冠詞の検討

フレーズ検索を活用することによって、冠詞の有無をチェックすることができる。例えば、「幾何学の概念」という日本語を英作すると、「concept of geometry」となるが、ofの後ろの「geometry」には冠詞が必要なのだろうかと迷うことがある。

そこで、「”concept of geometry”」と「”concept of the geometry”」でフレーズ検索を行って検索結果数を比較してみると、表 3.1 のようになる。

表 3.1: 冠詞の検討

検索文字列	検索結果件数
”concept of geometry”	493
”concept of the geometry”	60

表 3.1 より、「”concept of geometry”」の方が検索結果数が多いことから、「of」の後ろの「geometry」には冠詞を付けない方が適切であると考えられる。

3.4　ワイルド カードを使った検討

英訳を行うときに、文法的な構造は分かっているものの、以下のような疑問点が出てくることがある。

- 前置詞を挿入するのはわかっているが、どの前置詞を使ったらいいのか分からない。
- ある単語とある単語の間にはどんな単語が入るべきなのだろうか。

以上の疑問点を解決する方法としてワイルド カードを使った検討がある。「wild」という単語は「野性的な」という形容詞で、「wild card」は、人間が決めた規則に従わずに自由に使えるというカードという語源であり、検索エンジンで「*」で示すことによって任意の 1 単語を表すことができる。

例えば、「〜において」という日本語に対する前置詞は「in」「at」など複数考えられる。どんな前置詞を使ったらいいのか分からない場合、文型における前置詞の位置をワイルド カードに置き換えてフレーズ検索を行うことにより、どんな前置詞が使われているかを調べることができる。調べた結果、複数の前置詞を使った文型のパターンが出てくる。そして、それぞれをフレーズ検索を行うことによって、英作した文型においてどの前置詞がよく使われているかを知ることができる。

また、文型が「SVC」の形を取っている場合、動詞 V の部分をワイルド カードにして検索することによって、主語 S と形容詞 C との関係を表すべき動詞 V を見つけることもできる。

以上から、単語と単語の間にワイルド カードを指定することによって、どんな単語が使われているかを調べることができる。本節では、ワイルド カードを使った英作文の検討について述べる。

3.4.1 ワイルド カード を使った前置詞の検討

例として、以下のような文章の英訳の検討を行うとする。

その選手は汗でびしょ濡れだった。

まず、この文章を Excite[2] で英作文してみると、以下のようになる。

The player was dripping wet in the sweat.

以上の文章において、「汗で濡れる」という英訳に着目する。3.3.2 と同様、「sweat」に係る冠詞の有無についてフレーズ検索を行う。検索結果件数の比較を表 3.2 に示す。表 3.2 から「sweat」には冠詞を付加しない方が適当であることが分かる。

表 3.2: sweat の冠詞の検討

検索文字列	検索結果件数
”wet in sweat”	655
”wet in the sweat”	7

次に、「汗で濡れる」という英訳に対して前置詞の使い方が気になったとする。そこで、前置詞の部分を

”wet * sweat”

と in の部分をワイルド カードに置き換えて検索する (図 3.3)。検索結果として表示された用例の中から、

with, from

を使っているものがあった。Excite[2] で訳したときの前置詞「in」を含め、「汗で濡れる」の英訳として、

1. **wet in sweat**
2. **wet with sweat**
3. **wet from sweat**

以上の 3 つを英訳の候補とする。

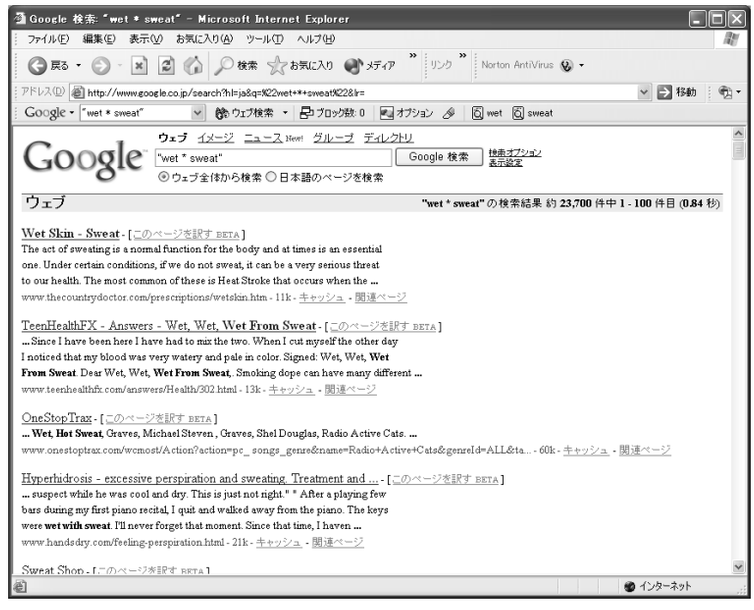

図 3.3: 「”wet * sweat”」で検索した結果

次に、以上 3 つの英訳の候補に対し、それぞれフレーズ検索を行う。結果は表 3.3 のようになった。

表 3.3 から、「with」を使った場合が一番ヒット件数が多いことがわかり、この文型がよく使われているということがわかる。

表 3.3: 前置詞の検討

検索文字列	検索結果件数
”wet with sweat”	3,280
”wet from sweat”	854
”wet in sweat”	274

3.4.2 ワイルド カードの複数指定による検討

3.4.1 では、調べたい部分をワイルド カードに置き換える方法について述べたが、動詞を調べたい場合、ワイルド カードの単数指定では、前置詞を含んだ熟語の検索ができないという欠点がある。そこで、ワイルド カードの個数を増やすことによって、新たな熟語や語句が見つけることができる。

例として、以下のような英訳を行うとする。

空気は窒素と酸素から成り立つ

まず、この文章を Excite で訳すと、以下のようになる。

Air consists of nitrogen and oxygen.

「成り立つ」の部分の英語訳は「consists of」であることがわかる。この部分を表 3.4 のようにワイルド カードを複数個利用して置き換えて、フレーズ検索してみる。

表 3.4 のようにそれぞれ検索してみると、検索結果から「consist of」の他に「is made of」という熟語も使われていることがわかる。以上の 2 つの熟語についての汎用性を調べるためにフレーズ検索をしてみると、

表 3.5 よりいずれも用例が存在しヒット件数にも大差がないことから、「成り立つ」の訳として、「consist of」「is made of」のどちらを使っても良いということがわかる。従っ

表 3.4: ワイルドカード複数指定

検索のタイプ	検索文字列	検索結果件数
ワイルドカード 1 個指定	”Air * nitrogen and oxygen”	180
ワイルドカード 2 個指定	”Air * * nitrogen and oxygen”	193
ワイルドカード 3 個指定	”Air * * * nitrogen and oxygen”	251

表 3.5: フレーズ指定

検索文字列	検索結果数
”Air consists of nitrogen and oxygen”	6
”Air is made of nitrogen and oxygen”	2

て、ワイルドカードを複数指定することによって、「成り立つ」という熟語を見つけることができる。

しかし、ワイルドカードは任意の文字を検索するので、訳語の指定はできない。そのため、違った英訳の動詞が検索されてしまうという欠点がある。

3.5　和英辞書を使った多義語の検討

英作文の作業において、一つの日本語に対する英訳は複数存在するので、用いられる文によって訳し方が変わってくるので、どのように使い分ければいいか分からない場合がある。以上の問題点を解決するために、各々の単語を使った語句をフレーズ検索することによって、汎用性を調べることができる。本節では和英辞典を使った多義語の検討について述べる。

例えば、以下のような日本語文において英訳を検討する。

その候補者は選挙の結果に失望した。

同じように、Exciteを使って英訳してみると、以下のようになる。

The candidate was disappointed at the result of the election.

「選挙の結果」という部分の英訳が気になったとする。「結果」という単語は複数あるが、この場合における「結果」はどの単語を使ったらいいか迷うことがある。

和英辞典で「結果」を調べると、「result」の他に「outcome」、「conclusion」がある。従って、「選挙の結果」の訳は

1. result of the election
2. outcome of the election
3. conclusion of the election

が候補として挙げられる。そこで、これらの語句をそれぞれフレーズ指定して検索結果件数を調べると、表3.6のようになる。

表3.6から、「outcome of the election」として使用する方が検索結果数が多いことがわかる。従って、「選挙の結果」の訳は「outcome of the election」として使うのが適切であると考えられる。

表 3.6: 「選挙の結果」の訳語

検索文字列	検索結果件数
”result of the election”	11,300
”outcome of the election”	24,300
”conclusion of the election”	1,100

3.6 ドメインの参照

検索結果には、そのページのリンクと共にURLが記載される。そのURLのドメインを調べることによって、どのような分野または国で使われている語句なのかを調べることができる。

例えば、「激しい論争」という語句で「激しい」という部分の単語について検討するとき、「violent controversy」, 「fierce controversy」, 「furious controversy」, 「vehement controversy」の候補が挙げられた。

「vehement controversy」が一番検索結果件数が少なかった。そこで、「vehement controversy」の検索結果のURLを調べてみると、

`home.wanadoo.nl/piet.fontaine/volumes/vol13.htm`

`www.nvvs.nl/medisch/admiraal_ch1.htm`

`www.qantara.de/webcom/show_ article.php/_c-327/_nr-8/_p-1/i.html`

などがあり、「nl」(オランダ)、「de」(ドイツ) などのドメインを含んだページが出てきた。以上のことから、「vehement controversy」は、非英語圏で使われていることが多い語句だということが分かる。

また、国別ドメインだけでなく、「edu」などの分野ドメインなどの比較によって、語句の使い分けを検討することができる。

3.7　検索エンジンからの用例の参照

Google の検索結果を図 3.4 に示す。

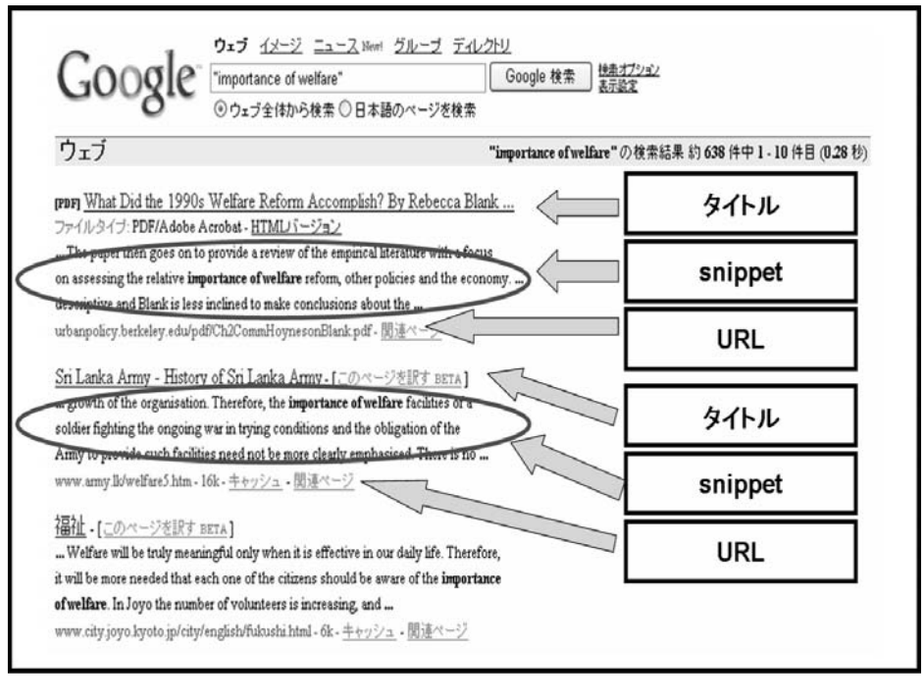

図 3.4: Google の検索結果

検索結果には、ページタイトルと URL の他に検索キーワードの周辺の文章を含んだ snippet が表示される。図 3.4 の例で挙げると、

```
... The paper then goes on to provide a review of the empirical literature with a focus
on assessing the relative importance of welfare reform, other policies and the economy. ...
descriptive and Blank is less inclined to make conclusions about the ...
```

の部分に相当する。検索キーワードを含む 1 文は表示されるので、snippet を参照することによって、検討したい語句がどのように使われているかを調べることができる。また、周辺の文章やタイトルを参照することによって、どのような分野のページなのかを参照することができる。

第4章　検索エンジンを使った英作文支援システムの構築

第3章で、検索エンジンを使った英作文の検討の方法について述べたが、以下のような問題点が存在する。

1. フレーズ検索を行う際に「”」(ダブルコーテーション) を付加するのが面倒である。
2. ワイルドカードを用いる際、品詞の特定ができないと、関係のないものが検索結果に紛れてしまう。
3. それぞれの英語をフレーズ検索して検索結果ウィンドウを切り替えながら、検索ヒット件数を比較するのは効率が悪い。
4. ドメインの参照や用例の参照を行う際に、画面をスクロールしながらの参照になるので見にくい。

問題点1を解決するためには、入力された文に対してあらかじめフレーズ検索を行うように設定することによって解決できる。問題点2については、品詞分解を行うことで解決することができる。問題点3及び問題点4については、それぞれの結果を整理して提示する形にすれば、便利なシステムとして提供できると考えられる。

以上の問題点を解決するために、本論文ではGoogleを使った英作文支援システムを構築した。ユーザには英作文の作業において、気になる部分の英文を入力してもらい、本システムで解析を行う。第3章で述べた検索エンジンを使った英作文の検討の作業を自動化することによって、スムーズに英作文の作業が行えると考えられる。

本システムの構成を図 4.1 に示す。

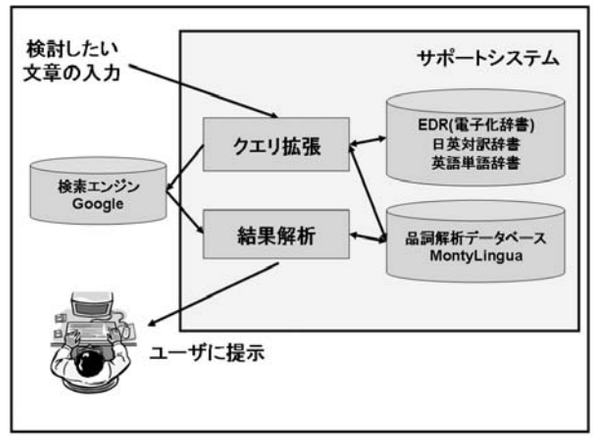

図 4.1: 英作文支援システムの構成

本システムを構成するツールとして以下のものを利用した。

1. GoogleAPI[3] によって Google にアクセス
2. 辞書データベースとして、EDR[17] を使用
3. 品詞の特定には、Eric Brill の MontyLingua[18] を使用

GoogleAPI は、Google のデータベースにアクセスできるインターフェースで、2002 年春にサービスが開始された。Google の膨大な量のデータベースを利用して、検索結果を好きなように利用することができる。プログラムから Google にアクセスできるので、あらかじめフレーズ検索を指定したり、ワイルド カードに置き換えて検索するといった作業が可能である。従って、GoogleAPI を使うことによって検索エンジンを使った英作文の検討の自動化が行えると考えられる。

辞書データベースには通信総合研究所の EDR を使用した。和英辞典を使った検討では、EDR の日英対訳辞書を使用した。また、単語の品詞の情報として英語単語辞書を使用した。

品詞の特定には Eric Brill の MontyLingua[18] を使用した。入力した単語において品詞分解を行うもので、品詞の特定を行うことによって必要な情報だけを検索することができる。

以下、Google を使った翻訳サポートシステムの処理の流れについて述べる。

4.1 ワイルドカードを使った検討の自動化

3.4.1 では、Google を使った前置詞の検討について述べた。この一連の作業を自動化する方法について、再度 3.4.1 の例を挙げて説明する。

まず、ユーザはテキストボックスに気になる部分の英語文 (wet in sweat) を入力してもらう。ここで、文章全体を入力してしまうと、文章全体のフレーズ検索になってしまい、定型文でない限り検索結果が 0 件になる可能性が高い。従って、文章の一部分について入力をしてもらう。

そして、その中で検討したい部分においてドラッグで囲んでもらい、「送信」ボタンを押す。3.4.1 の例の場合、前置詞を検討したいので「in」をドラッグで囲むことになる (図 4.2)。

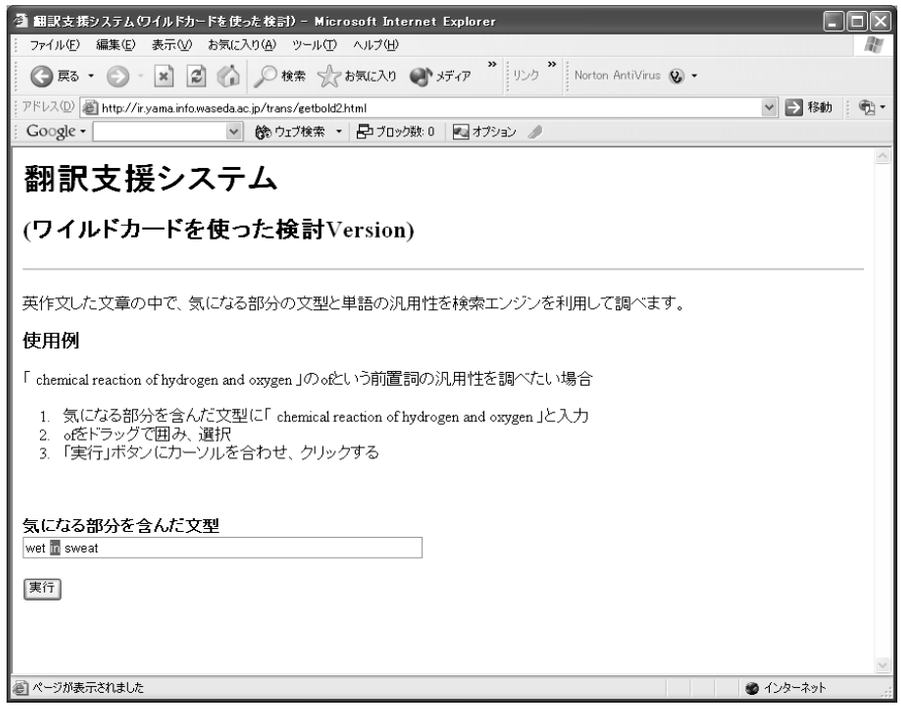

図 4.2: 「ワイルドカードの検討」のシステム画面

「送信」ボタンが押されたときのシステムの概要を図 4.3 に示す。

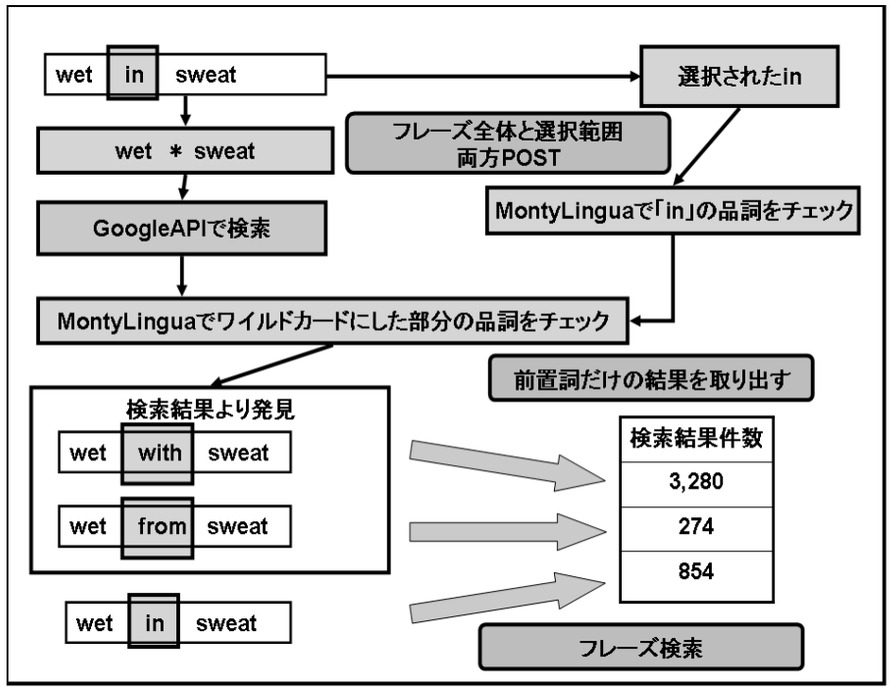

図 4.3: ワイルドカードを使った検討の自動化

システムは、入力された英文と選択した部分の両方を POST し、選択した部分「in」をワイルドカードに置き換えて、Google でフレーズ検索を行う。検索結果の snippet 部分の中で、< b > 〜 < /b > タグで囲まれているところがクエリの部分なので、ワイルドカードに相当した部分の品詞を判定する。3.4.1 の例の場合、「in」は MontyLingua から前置詞と判断する。そして、検索結果からワイルドカードに対応する部分に現われる単語を抽出する。MontyLingua により、前置詞のものだけの用例を取り出す。この結果、「with」、「from」が使われているのがわかる。

システムは、クエリを「wet in sweat」と「wet with sweat」「wet from sweat」として GoogleAPI を用いて並列処理で検索する。GoogleAPI による検索では一度に 10 件の検索結果しか得ることができないため、1 件目〜 10 件目、11 件目〜 20 件目とそれぞれ並行して GoogleAPI にクエリを送り並列で検索を行う。その後、検索結果件数を比較する表にまとめ、ユーザにどちらの前置詞がよく使われているかの提示を行う。

3.4.1 の例で前置詞を検討した場合のシステムの結果画面を図 4.4 に示す。図 4.4 のように提示された情報をもとに、ユーザは最適な前置詞を判断することができる。

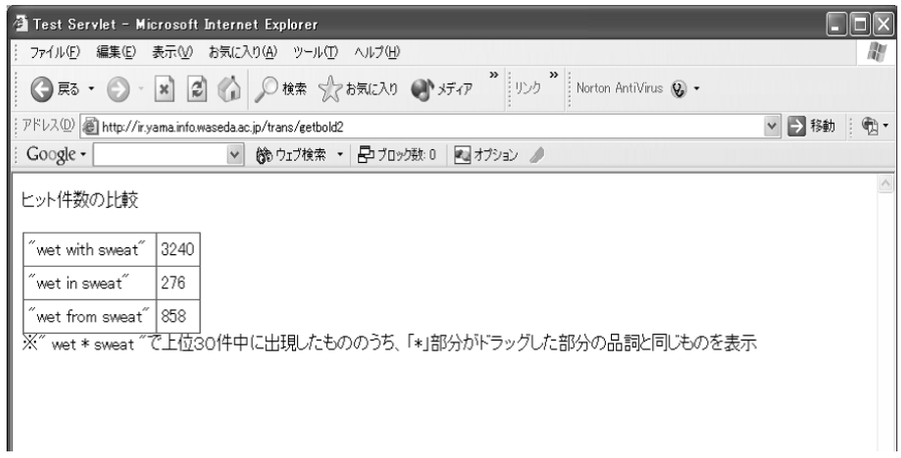

図 4.4: ワイルドカードを使った検討を行ったシステムの結果画面

4.2 多義語の検討の自動化

多義語の検討において、ユーザに入力してもらう情報は以下のとおりである。

- 検討したい英語の語句 (3.5 の例の場合、「選挙の結果」に対応する英訳)
- 調べたい英語の部分の日本語訳 (3.5 の例の場合、「結果」)

「結果」という日本語を入力してもらうのは、和英辞典で類語を探すためである。英単語をもとに類語を探すと、もととなる英単語が複数の異なる意味を持つ場合、的確な英単語表示をリストアップできない。そこで、最適な単語を探す場合においては、日本語訳を入力してもらうようにした。また、英語で「選挙の結果」の部分まで入力してもらうのは文型のパターンを読みとるためで、この場合における「結果」の使い方について調べるためである。

本稿では、辞書データベースとして EDR の日英対訳辞書を用いた。収録されているレコード数は、364,430 個で、対訳情報は約 19 万語である。EDR のレコードの構成は、単語見出しとその対訳情報となており、更に日本語の見出しにおいての概要説明の情報も格

納されている。ユーザは、概要説明の検索も行うことができ、訳語の完全一致だけでなく、概要説明からの訳語候補の選択を行うことができる。

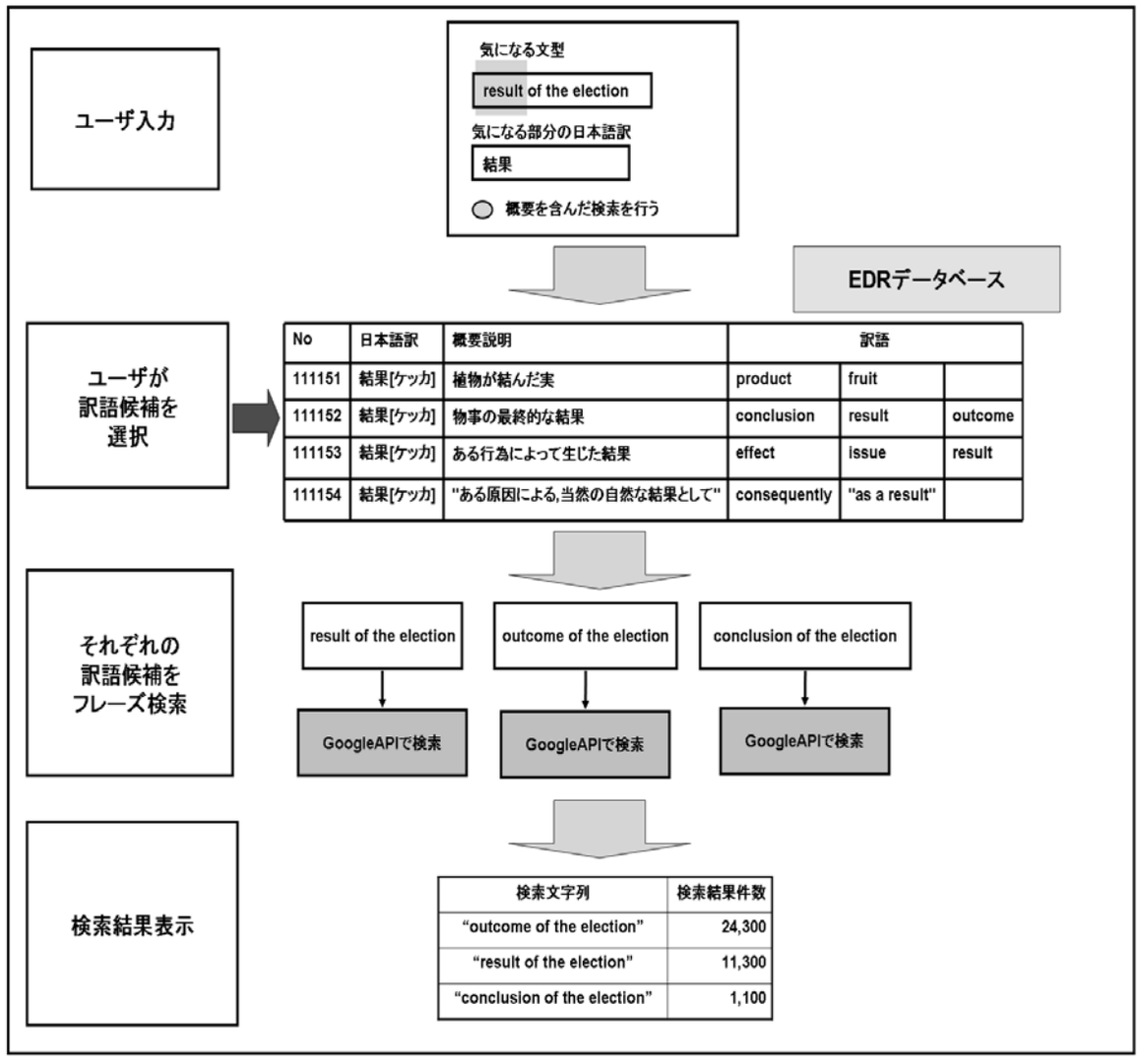

図 4.5: 多義語の検討の自動化

「結果」に対する訳語を検索したときのシステムの処理の流れを図 4.5 に示す。同じ「結果」でも「物事の最終的な結果」や「ある行為によって生じた結果」など複数の意味を持つ「結果」が存在するので、ユーザは状況に合わせて訳語候補を選択することができる。システムは選択された訳語候補をそれぞれ GoogleAPI で検索を行い、検索結果件数の比較を表示する。

4.3 活用形の対応

フレーズ検索はそのままの形で検索されるので、動詞において現在形で指定されると、現在形のものだけを検索する。動詞は現在形だけでなく、時制に応じて様々な活用変化がなされている。従って、多くの用例を参照するために、動詞の検討を行う際には活用形の変化にも対応させる必要がある。また、品詞特定で名詞と判断したときに、冠詞の検討が自動的に行えるような機能をオプションとして追加した。

さらに、名詞においては状況に応じて、単数形と複数形どちらが一般的に使われるのか問われるときがある。従って、本システムでは名詞の複数形と単数形を使ったフレーズ検索を自動的に行えるような機能をオプションとして追加した。

本節では、活用形の対応におけるシステムの処理の流れについて説明する。

4.3.1 動詞の活用形の対応

本システムでは、入力された文章を MontyLingua で品詞を特定し、動詞が存在した場合は、以下のように活用形を変化させて OR 検索でフレーズ検索を行うようにした。活用形を変化させてフレーズ検索を行うことによって、より幅広い用例の検索を行うことができる。

EDR の英語単語辞書を使用し、動詞の活用形を調べる。調べた活用形の種類を「(現在形 OR 進行形)」のような形で指定することによって、検索結果件数に活用形を変化させたものを含むことができる。

例えば、3.4.1 で例を挙げた例文の中で、「dripping wet」という表現は一般的に使われているかのどうかのフレーズ検索を行うとする。するとシステムは、「dripping」の品詞を解析し、EDR の英語単語辞書で構成した品詞解析データベースから原形「drip」、3 人称単数「drips」、過去分詞形「dripped」を抽出し、それぞれ OR 検索で囲った形で検索を行う (図 4.6)。

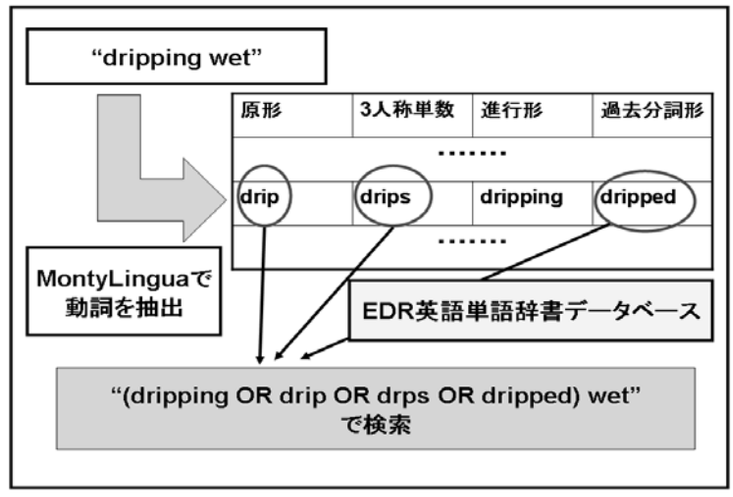

図 4.6: 動詞の活用形の対応

4.3.2 名詞の複数形の対応

名詞は単数形だけでなく複数形も使われている。従って、単数形でフレーズ検索を行うと単数形の用例しか検索できない。以上の問題を解決するために、フレーズ検索の際、**(複数形 OR 単数形)** と指定することによって、単数形と複数形を検索対象に含めることができる。

本論文では、オプション機能として「名詞の活用形の対応」を実装した。検討したい語句の中に名詞が含まれていた際、自動的に (複数形 OR 単数形) に置き換える機能である。複数形への変換のために EDR の英語単語辞書を利用し、名詞の活用形のパターンをデータベース化した。表 4.1 に名詞の活用形のパターンの一例を示す。

また、辞書に掲載されている名詞に対して、どのパターンを適用させるかのデータベースを作成し、データベースを元に名詞を複数形に変換する。名詞の活用形のタイプのデータベースの一例を表 4.2 に示す。例えば、「bus」という単語の活用形は表 4.2 から ECN2 だということがわかり、表 4.1 から ECN2 のパターンは、複数形に es が付くことが分かる。従って、「bus」の複数形は「buses」となり、「bus OR buses」で検索を行うようにする。

一方、複数形から単数形を調べる際は、MontyLingua で品詞解析を行った際に、単数形のデータを抽出し、活用形のデータベースから複数形を再度抽出する。

表 4.1: 名詞の活用形パターンのデータベースの一例

line_id	type_name	single	plural
1	ECN1		s
2	ECN2		es
3	ECN3	y	ies
4	ECN4	fe	ves
5	ECN5	f	ves
6	ECN6		(e)s

表 4.2: 名詞の活用形のデータベースの一例

line_id	standard	unchanged_part	change_type
19908	bus	bus	ECN2
19909	busboy	busboy	ECN1
19910	busby	busb	ECN3
19911	bush	bush	ECN2

4.3.3 冠詞の検討の自動化

図 3.3.2 で冠詞の有無の検討について述べた。本システムでは、MontyLingua の品詞解析で名詞と判定したものに対し、オプション機能で「the」をつけた場合と、「the」をつけない場合のフレーズ検索を自動的に行うようにした。ユーザによっては既に冠詞の検討においては理解している者も存在するかもしれないので、本システムでは、検討したいときに検討が行えるような形として「オプション機能」を実装する。

4.4 ドメインの参照

3.6で検索エンジンを使ったドメインの参照を用いて説明した。しかし、1件1件がURLが分かれており、画面をスクロールしながら、ドメインを調べていくというのは時間のかかる作業である。

本システムでは、GoogleAPIを用いて検索結果のURL部分だけを収集し、ドメイン別の統計を出すことにより、見やすい結果を表示する機能を実装した。

4.5 用例の参照

3.7で、検索エンジンの用例の参照について説明した。しかし、URLと同様に1件1件用例が分かれており、画面をスクロールしながら見ていくという形になってしまう。

そこで、本システムではGoogleAPIを使って、用例の部分のsnippetだけに着目し、検索結果上位100件のsnippetを一度に参照できる機能を実装した。検索結果からsnippetだけについて整理を行い、キーワードのみを含む文を抽出することによって、検討したい語句がどのように使われているかを一見で参照することができる。

1画面に複数の検討したい語句の用例を表示するのは困難なので、用例を参照したい語句について表示する形にした。

4.6 品詞分解による構文解析の検討

機械翻訳の構文解析に関する問題点で、関係代名詞を使って翻訳する傾向が多いことや、「挿入」「省略」「倒置」に関するルールには弱いことが挙げられていた。

従って、本節では、MontyLingua によって品詞分解を行った結果を解析し、関係代名詞を使った構文と副詞の挿入の位置に関する検討を自動的に行う機能をオプションとして実装した。

以下に、システムの流れについて説明する。

4.6.1 関係代名詞を使った構文の検討

機械翻訳において、特に精度が低いのが関係代名詞を使った文の翻訳である。「〜な人」や「〜なもの」などの名詞節を機械翻訳で英訳すると、that や which を使った関係代名詞を使った文にしてしまうことが多い。しかし、関係代名詞を使った英訳に対しフレーズ検索を行うと、件数が少なく汎用性が低いことが分かる。従って、以上のような名詞節を英訳する場合、関係代名詞を省略し名詞を修飾する動詞の「〜 ing」や「〜 ed」を使って、名詞に係る英訳の方法がある。

例えば、「Web 上に存在するデータ」を英訳するとき、Excite の翻訳では、「data that exists on the Web」となる。「data that exists on the Web」の語句をフレーズ検索すると、0 件になる。従って、一般では使われない表現だということが分かる。そこで、「倒置」などを行った様々な文型に置き換えてフレーズ検索を行ってみた。表 4.3 に結果を示す。

表 4.3: 文型の汎用性

検索語句	検索結果件数
”data that exists on the Web”	0
”data existing on the Web”	6
”existing data on the web”	88
”data on the web”	37,900

表 4.3 から that を使った修飾よりも、「ing」を使った名詞の修飾、あるいは「ing」を前に置いた名詞の修飾の表現の方が使われていることが分かる。更に「exist」は省略可能な動詞なので、「data on the Web」としても可能な表現であることが分かる。

以上の検討を本システムで自動的に行えるように、MontyLingua で品詞解析を行って「SVO」の関係を抽出し、それぞれのパターンについてフレーズ検索を行う機能を実装した。

構文解析機能の概要を図 4.7 に示す。

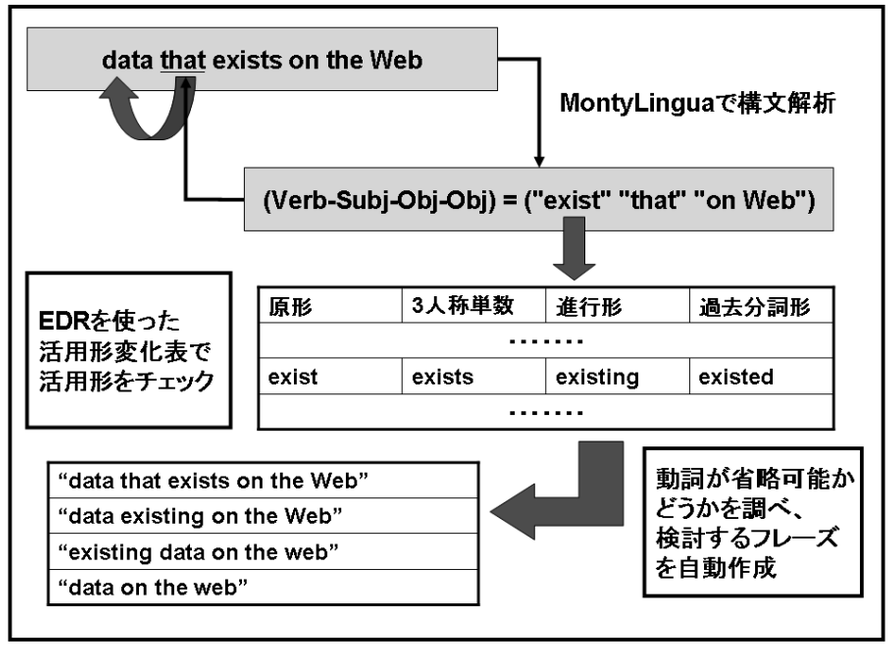

図 4.7: 構文解析機能の概要

例えば、「data that exists on the Web」に対しての「SVO」の関係を抽出すると、以下のようになる。

```
(Verb-Subj-Obj-Obj) = ("exist" "that" "on Web")
```

本来なら主語に当たる部分は「data」が抽出されるべきだが、MontyLingua は関係代名詞の「that」を抽出してしまう。S に「that」や「which」が抽出された場合、一つ前の単語に着目することにする。「that」が「あの、その」の指示語として使われている場合、

検討したい語句として入力するときには「that」が一番最初に来るので、指示語と関係代名詞の判断は前に名詞が存在しているかどうかで判断する。

次いで、EDR を使って、V の活用形を調べる。V に当たる部分が「exist」や「get」の場合、省略可能動詞と判断し、省略した形も検討する。MontyLingua では、C に当たる部分は O として抽出する。例えば、「data which get interested」のような「get + C」は「〜になる」という文法なので、「data interested」とすることも可能である。本システムでは、「exist」「get」を省略可能な動詞としてデータベース化を行った。

4.6.2 副詞による修飾の位置の検討

英訳する際、副詞の挿入位置に疑問を感じることがある。例えば、「影響を強く受ける」という文を英訳する際、翻訳ソフトで英訳すると「The infuluence was received strongly」となる。副詞が受身形の後ろで修飾されているが、副詞を受身形の前に持ってきて、修飾することもできる。従って、どちらの置き方がよく使われているかを検討することができる。実際に Google でフレーズ検索をすると表 4.4 のようになる。

表 4.4: 副詞の位置の検討

検索文字列	検索結果件数
(be OR is OR was OR were OR been OR being) strongly received	145
(be OR is OR was OR were OR been OR being) received strongly	74

以上の検討を本システムで自動的に行う機能を実装とした。処理の流れは、以下のとおりになる。まず、MontyLingua の品詞解析で、「be 動詞＋動詞の受身形＋副詞」と解析する。品詞解析は以下のとおりに行われる。

```
was/VBD/be received/VBN/receive strongly/RB/strongly
```

原形の形も解析されるので、以上から「be 動詞＋動詞の受身形＋副詞」と判断することができる。従って、副詞の位置を置き換え、それぞれにおいてフレーズ検索を自動的に行うようにした。

4.7 ワイルドカードの応用的使用

ワイルドカードは、「*」一つで任意の単語1語に相当するので、1語のものしか検索できない。従って、複数指定することによって熟語が検索できたり、また、ワイルドカードの置き方によっては新たな文型の単語の使い方が発見できることがある。

3.4.2において、ワイルドカードの複数指定による検討について述べた。3.4.2のように前置詞を含んだ動詞句や熟語を検索するには最適の機能である。しかし、「多義語の検討」と違い、日本語訳を指定して検索することはできないので、単語と単語の間に入りそうなものを探す、いわゆる共起を調べる程度に過ぎない。

本システムでは、オプション機能として、ワイルドカードを使った検討を行う際に、複数調べることを自動的に行うかどうかの選択をユーザが選ばせる機能を実装した。検討したい語句について、ワイルドカードの部分を「A * B」と置き換えるのに加え、「A * * B」や「A * * * B」と自動的にワイルドカードの個数を増やしてフレーズ検索を行う。また、AがS、BがVの場合にはよく使われる副詞や助動詞も発見することができる。

更に、MontyLinguaで品詞特定を行うことによって、検討したい品詞だけを抽出機能もオプションとして選択できるようにした。以上の機能を実装することにより、ワイルドカードの複数指定による検討を効率良く行うことができる。

4.8 英作文支援システムの機能のまとめ

本節では、第 4 章で紹介した機能の一覧を表 4.5 に示す。英作文の検討の際に生じた問題点と問題点を解決するための本システムの機能を対応させると表 4.5 のようになる。

表 4.5: 本システムの機能一覧

従来の提案手法における問題点及び英作文の作業における問題点	問題点を解決するための本システムの機能	機能の概要
従来の英日に関する翻訳の提案手法では、検討する品詞が限定されている。	ワイルドカードを使った検討	気になる部分をワイルドカードで指定することによってどのような単語が使われているかを調べる。全ての品詞に対応できる。
用いられる文によって単語の英訳が変わってしまうので、使い分けが分からない。	多義語の検討	気になる部分を同じ意味を持つ別の単語に置き換えてフレーズ検索を行うことにより汎用性を調べる。 前置詞を除く全ての品詞に対応できる。
文の状況に応じて、冠詞をつけたらいいのかわからない。また、フレーズ検索では 1 種類の活用形しか検討できない。	活用形の対応 (名詞の冠詞の検討及び複数形の対応)	それぞれのパターンにおいて、フレーズ検索を行う。また、OR 検索を行うことによって、活用形の対応を自動的に行う。
英作した文型が英語圏で使われているかどうかを調べたい。	ドメインの参照	検索結果 URL のドメインを抽出して統計を表示することにより、英語圏で使われている表現かどうかを提示する。
従来の英日に関する翻訳の提案手法では、コーパスの分野が限定されてしまっている。	用例の参照	検索結果の snippet を抽出することにより、検討したい語句がどのように使われているかを見やすく提示することにより、用例を参照できる。
従来の英日に関する翻訳の提案手法では、構文解析による翻訳の精度が良くない。	品詞分解による構文解析の検討	それぞれの文型のパターンでフレーズ検索を行い、検索結果件数を表にまとめて提示
検討したい単語がどのように使われているかを調べたい。2 つの語句の関係を調べたい。	ワイルドカード複数指定による熟語発見及び共起による語順の検討	検討したい語句についての新しい使い方を発見できる。

第5章　システムの評価

本章では構築したシステムの有用性を調査するための評価を行う。今日、英作文は様々な分野で行われている。従って、検索エンジンを用いた英作文の検討の評価を行うことにより、幅広い分野の対応が行えると考えられる。

本論文では、英訳が頻繁に行われる以下の分野において英作文の評価を行った。

- 一般の英作文 (日常生活で用いられる)
- 新聞記事の英文化 (時事問題として用いられる)
- 学術論文の英文化 (教育分野で用いられる)

本章では、以上の分野でランダムに抽出した日本語訳に対する英作文の作業において、システムの評価と考察を行う。評価は、第 4 章で説明した以下の機能に対して行うものとする。

- 4.1:ワイルドカードを使った検討
- 4.2:和英辞書を用いた多義語の検討
- 4.3.3:フレーズ検索による冠詞の検討
- 4.6:構文解析の検討
- 4.7:ワイルドカードの複数指定による英訳の発見

以下、評価方法として評価基準、評価対象データについて述べる。次いで、評価結果について述べ、考察を行う。

5.1 評価方法

本節では、評価対象データ、評価する検討方法について述べ、評価基準については実例を用いながら説明する。

5.1.1 評価対象データ

本項では、評価の対象となるデータについて述べる。

一般の英作文の評価

あらかじめ日本語文と英語文の対訳情報のある正解データを用意し、日本語文に対し翻訳ソフトで英訳を行い、本システムを使って修正を施す。修正を施した場所において、正解データである英語文と使用している単語が一致しているならば「正解」とする。検討する語句は、主に名詞、前置詞、動詞、および形容詞について検討する。なお、熟語については特殊構文なので、それぞれの単語においてはチェックしない。翻訳ソフトは、「The 翻訳プロフェッショナル V10」(製品版) を用いた。

正解データには、任意の英文集 [19] を用いた。翻訳ソフトを使う際の日本語に関しては、一字一字正しく入力するものとする。

新聞記事の英作文の評価

近年、日本語の新聞が英語化されていることが多い。本論文では、「DAIRY YOMIURI」[20] の社説の英訳を本システムを使って適切な訳にすることができるかどうかの検討を行った。

「読売新聞」に掲載されている「社説」の部分をまず翻訳ソフトで、英語文に翻訳する。翻訳した英語文を本システムを使って修正を施す。修正を行ったところに対し、「DAIRY YOMIURI」の社説の英訳で使用している単語と一致しているならば、「正解」とする。翻訳ソフトは、5.1.1 と同様に「The 翻訳プロフェッショナル V10」(製品版) を用いた。

論文の英作文の評価

学術論文の海外発表を行う際、日本語で作成した論文を英文化することが多い。本論文では以上のことに着目し、日本語で作成された学術論文に対し、5.1.1 と同様に評価を行う。日本語で作成された学術論文の各文において翻訳ソフトで翻訳する。次に、本システムを使って修正を施す。修正した英語文が人手で作成された英語論文と一致しているかどうかを調べる。

評価データは、研究室内で書かれた論文を使用し、人手で書かれた英語論文をネガティブスピーカーによって添削されたものを正解データとする。翻訳ソフトは、5.1.1 と同様に「The 翻訳プロフェッショナル V10」(製品版) を用いた。

5.1.2 評価を行う検討項目

評価を行う検討項目を表 5.1 にまとめた。精度に関しては、表 5.1 の 7 項目に対して算出するものとする。

表 5.1: 評価を行う検討項目

評価の対象となる検討項目	検討内容
ワイルド カードの検討	前置詞の検討に関しては、4.1 のような方法でワイルド カードを使って評価を行う。その他の品詞に関しては、「ワイルド カード 複数指定の検討」の方法で、評価を行う。
動詞における多義語の検討	動詞において 4.2 の和英辞書を用いた多義語の検討の方法で、評価を行う。
名詞における多義語の検討	名詞において 4.2 の和英辞書を用いた多義語の検討の方法で、評価を行う。
形容詞における多義語の検討	形容詞において 4.2 の和英辞書を用いた多義語の検討の方法で、評価を行う。
冠詞の検討	気になった冠詞の有無について 4.3.3 のような方法で検討し、評価を行う。
構文解析の検討	気になった構文について、4.6 の構文解析の検討の方法で、評価を行う。
ワイルド カード 複数指定の検討	気になった動詞の使い方や SVO の関係について、4.7 のワイルド カードを複数置いた検討の方法で、価を行う。検討を行う品詞は限定しない。

5.1.3　評価基準

本項では、評価基準について実例を用いながら説明する。

検索結果件数が一番多いのフレーズが他のものより極端に多い場合

次の文において英訳を検討する。

賃料が手頃ならどんなマンションでもいいです。

「The 翻訳プロフェッショナル」で訳すと以下のようになる。

If a rent is handy, what kind of apartment house is sufficient.

「手頃な」の英訳に着目する。「賃料」に対する「手頃」なので、「手頃」において本システムの 4.2 の機能を使って検討すると、表 5.2 のような結果が得られる。

表 5.2: 「賃料」に対する「手頃」の多義語の検討

検索文字列	検索結果件数
”rent is handy”	3
”rent is suitable”	40
”rent is reasonable”	925

表 5.2 から「reasonable」を使ったものが圧倒的に検索結果件数が多い。以上の場合、「rent is handy」はおかしな表現と判断する。また表 4.3.3 の機能から「a」より「the」の方が適切だと判断する。従って、以下のように書き換える。

If the rent is reasonable, what kind of apartment house is sufficient.

正解データの文献 [19] に掲載されている英文は以下のとおりである。

Any apartment will do as long as the rent is reasonable.

文は倒置になっているものの、「家賃が手頃な」という部分においては、修正した通りになっている。従って、「形容詞」の「多義語の検討」および「冠詞の検討」において成功したことになる。

検索結果件数が一番多いフレーズが他のものと大差がない場合

次の文において英訳を検討する。

個人は地域社会の基本的な構成要素である。

「The 翻訳プロフェッショナル」で訳すと以下のようになる。

An individual is a fundamental component of a community.

「構成要素」の英訳に当たる「fundamental component」について検討する。「要素」において、本システムの 4.2 の機能を使って検討すると、表 5.3 のような結果が得られる。

表 5.3: 「構成要素」の多義語の検討

検索文字列	検索結果件数
”fundamental element”	25,400
”fundamental factor”	6,780
”fundamental component”	23,900
”fundaental constituent”	1,100

表 5.2 から、「fundamental component」と「fundamental element」は同じくらいの件数であることがわかる。以上から「fundamental element」と書き換えても良いということがわかり、以下のような 2 パターンの書き方があることを示す。

An individual is a fundamental [element/component] of a community.

正解データの文献 [19] に掲載されている英文は以下のとおりである。

An individual is a fundamental element of a community.

全く同じ文になり、本システムを使って検討することができたことになる。以上の場合、「名詞」において「多義語の検討」が成功したことになる。

書き換えを判断する検索結果件数は、一番検索結果件数が多いものとするが、2 番目に多いものとの差が 2 倍以下の場合は、どちらの表現でもおかしくないと判断する。

評価基準のまとめ

評価基準を表 5.4 にまとめた。

表 5.4: 評価基準

評価結果	実験結果
正解	検索結果件数が 1 位で 2 位との差は倍以上でその単語に直した結果、正解で使われている単語と一致していた場合
正解	検索結果件数が 1 位であるが、2 位との差は倍以下で両方使えると判断した結果、正解で使われている単語が検索結果件数が 1 位だった場合
正解	検索結果件数が 2 位であるが、1 位との差は倍以下で両方使えると判断した結果、正解で使われている単語が検索結果件数が 2 位だった場合
不正解	フレーズ検索を行った検索結果件数が 0 件だったが、適切な単語に直すことができなかった場合
不正解	検索結果件数が 1 位で 2 位との差は倍以上でその単語に直した結果、正解で使われている単語と違った場合
評価の対象にしない	機械翻訳の単語が検索結果件数 1 位だった場合 (修正する必要がなかった)

5.2 評価結果

以上の評価データと評価基準を元に、3 分野の評価対象データから無作為に抽出した 96 文に対して、本システムの評価を行った。3 分野の評価対象のデータを無作為に抽出することによって、総合的に有用性のあるシステムであることを評価できると考えられる。評価結果は表 5.5 のようになった。

表 5.5: 評価結果まとめ

	検討数	正解数	精度
ワイルド カード の検討	7	6	85.71%
多義語の検討 (動詞)	16	10	62.50%
多義語の検討 (名詞)	39	27	69.23%
冠詞の検討	7	6	85.71%
多義語の検討 (形容詞)	30	20	66.67%
構文解析の検討	11	9	81.82%
ワイルド カード 複数指定	9	7	77.78%

5.3 考察

本節では、評価に対する考察を行う。検索エンジンを使った検討で修正が上手くいった場合と、修正が上手くいかなかった場合においてそれぞれ考察を行う。

5.3.1 修正が上手くいった場合

「〜な問題」や「〜なもの」などの名詞節、冠詞の検討およびワイルド カード を使った前置詞に関する検討は精度が高かった。構文が一律に定まっている場合や、単純な「SVO」、「SVC」構文における検討は汎用性が調べやすかったと考えられる。従って、任意の単語集からの英訳の検討は全体的に精度が高かった。

5.3.2　修正が上手くいかなかった場合

修正がうまく行かなかった原因について以下のような原因が挙げられる。

- 意訳による日本語の変化
- 構文解析による検討の限界
- ユーザの英語に対する知識

以下、それぞれについて説明を行う。

意訳による日本語の変化

新聞記事や海外論文では、意訳が多かったため、十分な精度が出なかった品詞があったり、前置詞の検討を行っても正解では前置詞が使われていないことがあった。

翻訳ソフトは、前後に共起する単語を把握しながら、基本的に日本語を忠実に翻訳する。検索エンジンを使った検討も構文解析の場合を除き、翻訳ソフトで英訳したものに基づいて検討を行うので、意訳が行われると精度が下がってしまう現象が起きた。

和英辞書を使った検討では、見出し語概念を用いることにより、意訳の対応を行った。しかし、本来は「怪しい」と訳される「strange」という単語を、文脈上から「未知」という意味で翻訳しているなど、意訳の対応が十分でなかった点があった。従って、見出し語概念の機能を拡張する機能が必要だと考えられる。

また、意訳として省略を行った構文があった。「get」や「exist」など、どの構文においても省略しやすいものであれば、検討が行えるが、文章として成り立っている日本語を省略する「予測が難しい省略」の傾向があったため、的はずれな単語に修正してしまったケースもあった。

どのような意訳が行われるかは正解データを作成した人物次第になってしまう。検索エンジンは英語について検討を行っているので、検索エンジンを使っての意訳の推測は難しいと考えられる。従って、意訳に対応するためには、英語だけではなく日本語の検討を行う必要があると考えられる。

構文解析による検討の限界

新聞記事や海外論文は、日本語文が長いために複雑な構文になっていることが多かった。単語の検討はできたが、構文解析による検討は十分に行えなかった。従って、構文解析の検討やワイルドカード複数指定の検討に関しては、検討数が少なかった。例えば、「不十分だ」という日本語文を英訳する際、翻訳ソフトで翻訳を行うと、「as inadequate as」と英訳しているのに対し、正解では「hardly at an advantage」と訳してあった。肯定文を否定文で表現する場合において、文の構造が変わってしまうので、検索エンジンで検討するには難しいことが分かった。

しかし、翻訳ソフトと正解データをフレーズ検索で比較すると、汎用性を調べることができる。以上の場合、検索エンジンとコーパスを連携させて検討できる機能を追加できれば、否定文に関する文の構造の変化に対応できると考えられる。

また、日本語で 2 文で書かれている文章を、正解データで 12 文に統合して英訳を行っているケースもあった。2 文を 1 文にすることによって、文の構造は大きく変わってくる。機械翻訳では、句読点に対し忠実に翻訳を行っているので、2 文を 1 文に統合するといった翻訳が行えないのが現状である。検索エンジンを使った検討でも、フレーズ検索で「汎用性の低い文型」は検出できるが、修正を提示するところまでは導けないことが多かった。

ユーザの英語に対する知識

検索エンジンを使った英作文の検討は、1 文そのものではなく、気になった部分に対して検討を行う方法である。従って、ユーザによって「気になる文型」は異なってくる。従って、評価を行うユーザが異なると、精度も変わってくると考えられる。

和英辞典を使った検討においては、気になった部分のみの日本語訳の入力であるので、文脈を考慮することができない。また、気になった部分の日本語訳の入力の仕方によって検索される語句は変わってくるので、正解データの単語が見つかる割合も異なってくると考えられる。上記に挙げた今回の実験では、1 人に対して行ったので、今後は多くの人に対して実験を行うべきだと考えられる。

第6章　おわりに

近年、日常生活で英語を扱う機会が増えてきており、教育分野でも小学校に英語教育が導入されるなど、英語の重要性は高まってきている。

本論文では、日本語から英語に直す英作文の作業を支援するために、検索エンジンを使った英訳の検討の方法を提案し、検討の作業を自動化する英作文支援システムの構築を行うことによって、有用性があることを明らかにした。

しかし、日本語のパターンは無数に存在し、数多くの英訳の方法があるので、検索エンジンを使った検討で以上のパターンを網羅するためには、本論文で紹介した以外にも様々な検討方法を発見する必要性がある。そして、今回評価を行った英文以外にも多くの英訳に対し検討し、評価を行う必要がある。

また、本論文で構築したシステムは、Web 上のデータを用例として参照することによって、どのように修正すればいいかを提示するシステムである。従って、ある程度の英語知識を持った人にとっては使いやすいシステムとなったが、英訳を行う際、どのように英訳したらいいか分からない人にとっては使いにくいシステムであることが分かった。

今後は、英作文支援の範囲を広げ、英語の知識を問わず誰でも使いやすい英作文支援システムを構築する必要があると考えられる。

謝辞

本研究を行うにあたり、数々の指導を頂いた山名早人助教授、「翻訳に役立つ Google 活用テクニック」の著者の安藤進先生、辞書データベースを提供してくださった独立行政法人通信総合研究所、研究室の同輩の本田大、蛭田智則、海外論文の対訳集を提供して頂いた斉田直幸、そして共同研究者の佐藤学に厚く御礼申し上げます。

研究業績

- 大鹿広憲, 佐藤学, 安藤進, 山名 早人:”検索エンジンを使った翻訳サポートシステムの構築”, 電子情報通信学会技術研究報告 Vol.104,No107,pp.237-242 (2004.7)
- 大鹿広憲, 佐藤学, 安藤進, 山名 早人:”Google を活用した英作文支援システムの構築”, 電子情報通信学会第 15 回データ工学ワークショップ (DEWS2005)

参考文献

[1] Google

http://www.google.com

[2] EXCITE 翻訳

http://www.excite.co.jp/world/

[3] 安藤進著, ”翻訳に役立つGoogle活用テクニック”, 丸善, ISBN4-621-07294-3 (2003.10)

[4] 山名早人監訳, 田中裕子訳,Tara Calishain & Rael Dornfest 著：”Google Hacks”, オライリー・ジャパン,ISBN4-87311-136-6 (2003.8)

[5] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟:”日英対訳パターンの自動抽出に向けて”, 情処研報,NL-153-15,pp.113-124 (2003)

[6] 鈴木健二, 梅村恭司:”情報検索システムを利用した日英対訳推定”, 情処研報, NL-151-1,pp.1-6 (2002)

[7] 荒牧英治, 黒橋禎夫, 佐藤理史, 渡辺日出雄:”用例ベース翻訳のためのパラレルコーパスからの対訳対発見”, 情処研報,NL-144-4,pp.23-30 (2001)

[8] 今村賢治, 大熊英男, 渡辺太郎, 隅田英一郎:”統計翻訳指標を導入した構文トランスファに基づく用例翻訳”, 情処研報, NL-162-11, pp. 71-77 (2004)

[9] 金出地真人, 徳久雅人, 村上仁一, 池原悟:”結合価文法による動詞と名詞の訳語選択能力の評価”, 情処研報,NL-153-16,pp.119-124 (2003)

[10] 情報処理振興事業協会:”計算機用日本語基本動詞辞書 IPAL”

http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

[11] 松吉俊, 佐藤理史, 宇津呂武仁:”機能表現「なら」の機械翻訳のための言い換え”, 情処研報,NL-159-28,pp.201-208 (2004)

[12] 延原由高, 池原悟, 村上仁一:”接頭・接尾辞を含む数量表現の翻訳方法”, 情処研報,NL-142-11,pp.75-82 (2001)

[13] 成田真澄:”英文アブストラクト作成支援ツールのユーザ評価”, ”COE 形成基礎研究費研究成果報告 (5)”, 神田外国大学 (2001)

[14] Sawa Takakura, Takeshi Ito and Teiji Furugori:”TransAid: a writer's aid system for translating Japanese into English”, WA2E3, The Proceeding of IEEE SMC 2002 Vol.6, Oct.(2002)

[15] Takashi Yamanoue, Toshiro Minami, Ian Ruxton and Wataru Sakurai: ”Learning Usage of English KWICly with WebLEAP/DSR”, Proceedings of the 2nd International Conference on Information Technology and Applications (ICITA-2004), 14-6, Harbin, China January. 8-11 (2004)

[16] 国際情報化協力センター:”平成 13 年度 多言語情報セキュリティ関連情報収集分析システム開発事業 -業務報告書-”, pp.70-78 (2003)

[17] 通信総合研究所,EDR 電子化辞書仕様説明書 (2003)

[18] MontyLingua Project
http://web.media.mit.edu/ hugo/montylingua/

[19] 鈴木陽一著,”現代英語の重要単語・熟語 2400”,ICP, ISBN4-900790-00-1 (1995)

[20] DAILY YOMIURI
http://info.yomiuri.co.jp/company/shimen/03.htm

付 録A　GoogleAPI

本論文では、検索データベースとしてGoogleAPI[4]を利用した。APIとはApplication Program Interfaceの略で、あるプラットフォーム(OSやミドルウェア)向けのソフトウェアを開発する際に使用できる命令や関数の集合のことを言う。付録では、GoogleAPIで使用できる機能や関数について説明する[4]。

A.1　GoogleAPIの概要

GoogleAPIを利用するには、Googleのデータベースにアクセスするための認証であるデベロッパーズキーが必要である。また、1つのデベロッパーズキーで1日1000回のままでGoogleへのデータベースのアクセスが可能である。

Google Web APIデベロッパーズキットのJAR(Java Archive)ファイルを使うことで、javaを用いたGoogleAPIのプログラミングが可能である。

A.2　検索要求オブジェクト

本節では、検索要求のオブジェクトについて説明する。また、GoogleAPIにおいて、指定できる特別構文について説明する。

A.2.1　クエリーの要素

GoogleAPIにおけるクエリーの要素の構成は以下のようになっている。

1. key(キー)
2. query(クエリ)

3. start(スタート)
4. maxResult(最大検索結果数)
5. filter(フィルタ)
6. restrict(制限)
7. safeSearch
8. lr
9. ie
10. oe

以下にそれぞれの要素について説明をする。

key(キー)

GoogleAPI デベロッパーズキーを置く場所である。キーがないとクエリは実行されない。

query(クエリ)

実際の検索質問の部分を配置する。プログラミングによって、あらかじめフレーズ検索を行うようにすることも可能である。

start(スタート)

オフセットとも言う。GoogleAPI は一度のアクセスで、10 件の検索結果しか返すことができないので、検索結果を返す場合、何番目の結果を返すのかを指定する。指定する整数は「Zero-based index 指標」であるので、1 からでなく 0 から数え始める。例えば、この値が 15 ならば、GoogleAPI は結果 15 から結果 24(16 番目の結果から 25 番目の結果) までを返す。

従って、30 件の検索結果を得たい場合、start 値が 0, 10, 20 と 3 回指定することによって、GoogleAPI は 3 度アクセスし、1 件目〜 10 件目、11 件目〜 20 件目、21 件目〜 30 件目の結果を得ることになる。

maxResult(最大検索結果数)

GoogleAPIから返される検索結果の件数を指定する。GoogleAPIは1回に最大10件までの結果を返すことができるので、設定可能な値は1から10までとなる。

filter(フィルタ)

内容が同じページが複数存在する場合(タイトルとページの抜粋の類似性から判断)、あるいは同じホストやサイトから3件以上の結果がある場合に、クエリフィルタを通して重複した結果を自動的に除外するかどうかという指定を行う。値はブール値(trueかfalse)で行い、trueが指定されている場合、1セットの結果には、同一ホストからの結果は最初の2件に限定される。

safeSearch

結果の中で、子供に見せたくない内容をフィルタするかどうかの指定をブール値(trueかfalse)で指定する。

lr

「language restrict」の略で、検索結果のWebページが書かれている言語を制限する。検索結果を英語だけに制限したい場合は、「`lang_en`」を値として指定する。複数の言語に制限したい場合、例えば、英語とドイツ語に限定したい場合は、「`lang_en|lang_de`」と指定する。また、「-」(負符号)を指定することによって、検索結果に含めない言語を指定することができる。

ie

「input encording」の略で、APIに渡すクエリーで使用される文字エンコードを指定する。現時点では「UTF-8」のみがサポートされている。今後、サポートされるエンコーディングは増えていくと考えられる。

oe

「output encording」の略で、ie と同様に「UTF-8」のみがサポートされている。

A.2.2 GoogleAPI における特別構文の扱い

Google では、クエリとして入力するときに、AND 検索や OR 検索、さらにフレーズ検索といった細かい検索オプションを指定できるが、GoogleAPI でも同様に使うことができる。さらに GoogleAPI では、きめ細かい検索を可能にするために以下の特別構文が使用できる。

site:

サイトかトップトップレベルドメインのいずれかに検索を制限することができる。例えば、`site:edu` と指定することによって、edu ドメインのページのみを検索することができる。

daterange:

指定された日あるいは期間にデータベースに登録されたページに検索を制限する。`daterange:` の指定には、カレンダーのグレゴリオ暦ではなく、ユリウス暦が使われる。ユリウス暦は紀元前 4713 年 1 月 1 日世界時正午を基点とした通日 (紀元前 4713 年 1 月 1 日から数えて何日目か) を表すもので、2005 年 2 月 2 日の場合は、2,453,404 日である。

intitle:

検索対象を Web ページのタイトルだけに限定して検索を行う。「`intitle:tennis`」と設定すると、ページのタイトルに「tennis」を含むものだけが検索される。

inurl:

検索対象を Web ページの URL だけに限定して検索を行う。「`inurl:help`」と設定すると、URL に help を含むものだけが検索されるので、ヘルプ関連のページが見つかりやすい。

allintext:

検索対象を Web ページの本文のテキストのみに限定して検索を行う。リンク、URL、タイトルは検索の対象外となる。

link:および allinlinks:

指定された URL にリンクしているページを結果として返す。「`link:www.google.com`」と指定すると、Google にリンクするページのリストが返される。URL を複数指定したい場合は、「`allinlinks:`」で指定する。

filetype:

「`filetype:`」は、URL のサフィックス (URL を右から見て最初に「.」が現われるまで) を対象に検索する。検索対象がファイルであれば、ファイルの拡張子を対象に検索する。「`filetype:pdf`」と指定すると、pdf のファイルだけを検索することができる。

info:

「`info:`」は、指定された URL に関する様々な情報へのリンクを返す。以下の情報が結果に含まれる。

- 指定されたページの Google のキャッシュへのリンク
- そのページへリンクしているページのリスト
- そのページと関連したページ

- URL を含むページの情報

related:

指定したページと関連したページを検索するのに用いる。「`related:www.oreilly.com`」と指定すると、Perl.com や Amazon.com のページを返す。

cache:

Google がデータベースに登録した時点でのページのコピーを表示する。頻繁に変更されるページや、既にページが存在しなくなった場合に検索するときに便利である。

A.3 検索結果オブジェクト

GoogleAPI からの検索結果は、検索結果のサマリーデータと個々の検索結果である Web ページのデータから構成される。本節では、検索結果のサマリーデータの構成と個々の検索結果データについて述べる。

A.3.1 検索結果のサマリーデータ

検索結果のサマリーデータは、クエリ自体とそれに関する情報および検索結果の個数で以下のものから構成されている。

$\langle documentFiltering \rangle$ **(ドキュメントフィルタリング)**

非常に類似した内容の検索結果、あるいは同じ Web サイト上の検索結果がフィルタリングされたかどうかが、ブール値で示される。

$\langle directroyCategories \rangle$ **(ディレクトリカテゴリ)**

クエリに関連するディレクトリのカテゴリのリストを返す。

⟨$EstimatedTotalResultsCount$⟩ **(検索結果件数)**

検索結果件数を数値で返す。

A.3.2 個々の検索結果データ

URL、ページタイトル、ページの抜粋など検索結果の「中身」は「⟨$resultElements$⟩」リストの中に返される。結果は以下の要素から構成されている。

⟨$summary$⟩

Google ディレクトリとマッチすれば、そのサマリーが返される。

⟨URL⟩

検索結果の URL を返す。

⟨$snippet$⟩

簡潔なページの内容の抜粋で、キーワード周辺の文章を返す。クエリ用語は太字 (HTML の $< b >$ 〜 $< /b >$ タグ) で強調される。

⟨$title$⟩

HTML のページタイトルを返す。

⟨$cacheSize$⟩

Google に Web ページがキャッシュされていれば、そのキャッシュされたバージョンのページの大きさで、KB(K バイト) を返す。

付 録B　評価データ

本章では、評価データに用いた日本語文と英語文の一覧を示す。

表 B.1: 評価データ一覧 (任意の単語集 DUO[19])

日本語文	正解英語文
個人は地域社会の基本的な構成要素である。	The individual is the fundamental element of a community.
賃料が手頃ならどんなマンションでもいいです。	Any apartment will do as long as the rent is reasonable.
脳の構造は複雑だ。	The structure of (the) brain is complicated.
現状では倒産は避けられない。	Under the circumstances, bankruptcy is inevitable.
諺にある通り、「目的が手段を正当化する」だ。	As the proverb goes, "The end justifies the means."
まずは公式を暗記しなさい。	First of all,learn the formula by heart.
観客達は彼女の優雅な演技に感動した。	The spectators were moved by her graceful performance.
もう寝る時間です。ラジオを消しなさい。	It is (high) time (that) you went to bed.Turn off the radio.
主婦達が日々の集まりきった仕事に不満を言うのももっともだ。	Housewives may well complain about their daily routine.
彼の小論文は主題が曖昧だった。	The theme of his essay was obscure.
その運河は大西洋と太平洋とをつないでいる。	The canal connects the Atlantic with the Pacific.
人工衛星が軌道に向けて打ち上げられた。	The artificial satellite was launched into orbit.
福祉の重要性はいくら強調してもしすぎることはない。	Once cannot emphasize too much the importance of welfare.
その付き合い人はお世辞がうまい。	The attendant is good at flattery.
暴力犯罪は郊外にも広がった。	Violent crime spread into the suburbs.
遅かれ早かれ捕虜達は釈放されるだろう。	Sonner or later,the hostages will be released.
近ごろでは、結婚の動機は必ずしも純粋なものでない。	These days,the motive for marriage is not necessarily pure.
わんぱくな子は道に迷ってあたりを見回した。	The naughty boy got lost and looked around.
今日では、通勤者たちは交通渋滞を当たり前のことと思っている。	Nowdays,commuters take traffic jams for granted.
私の前の彼はポルトガル育ちでした。	My ex-boyfriend was brought up in Portugal.
その専門家は国際緊張は高まっていくと予測している。	The specialist predicts international tension will buld up.
言うまでもないことだが、その思想は時代遅れだ。	It goes without saying that the ideology is behind the times.
その政治家はその込み入った問題に何とか対処した。	The statesman barely coped with the intricate issue.
彼はその浅い溝を飛び越えた。	He leaped over the shallow ditch.
私達は損失をできるだけ正確に見積もった。	We estimated the losses as exactly as possible.

日本語文	正解英語文
彼の評論は簡潔で要点を押さえたものだった。	His comment was concise and to the point.
ここだけの話だけど、あの太った醜い魔女は減量中なんだ。	Between ourselves,the fat ugly witch is on a diet.
散歩しませんか。	How about going for a walk?
いいですね、喜んで。	Why not? I'd be glad to.
原爆は人類にとって重大な脅威だ。	The atomic bomb is a grave threat to mankind.
平等は憲法で保障されている。	Equality is guaranteed by the Constitution.
予算はかろうじて議会の承認を得た。	The budget was narrowly approved by Congress.
時代遅れのその政権は崩壊寸前だ。	The obsolete regime is about to collapse.
その病気の初期症状は高熱と喉の痛みです。	The initial symptoms of the disease are fever and sore throat.
アンケート用紙が無造作に配布された。	The questionnaires were distributed at random.
議長は彼のばかげた提案を一蹴した。	The chairman rejected his absurd proposal.
これらのデータはちっとも正確ではない。	These data are anything but accurate.
その広大な大陸には化石燃料が豊富にある。	Fossil fuels are abundant in the vast continent.
軍隊の規律は文字どおり厳しいものだ。	Military discipline is literally rigid.
株式投資は必ずしも利益を生むとは限らない。	Stock investments do not always yield profit.
赤字は徐々に減少している。	The deficit has been diminishing little by little.
その仮説は徹底的な実験に基づいている。	The hypothesis is based on the through experiments.
湿気の多い気候はその半島の特色です。	The humid climate is characteristic of peninsula.
彼のめいは年の割には魅力的で大人っぽい。	His niece is attractive and mature for her age.
彼が絶えず侮辱したので私の怒りをかった。	His constant insults aroused my anger.
聴衆は意味の深い講演に感妙を受けた。	The audience was impressed by his profound lecture.
消極的なその男はめったに自己表現をしない。	The passive man seldom,if ever, expresses himself.
この州には鉱物資源が豊富です。	The province is rich in mineral resources.

表 B.2: 評価データ一覧 (読売新聞社説 2003 年 3 月 25 日付 [20])

日本語文	正解英語文
日本アニメ界にとって歴史的な意味を持つ受賞である。	It is the award with a meaning historical for a Japanese animation community.
宮崎駿監督の「千と千尋の神隠し」が米アカデミー賞で、長編アニメ映画賞を受けた。	Director Hayao Miyazaki's "Spirited Away" won the animated features movie prize by U.S. Academy Awards.
日本映画が長編作品でアカデミー賞を獲（と）るのは初めてだ。	This is the first time that a Japanese movie obtains Academy Awards with a long work.
既に、ベルリン国際映画祭のグランプリ「金熊賞」や、米国内でも、国際アニメ映画協会のアニー賞を始め、数々のアニメ賞を獲得している。	Many animation prizes including the Annie prize of an international animated film association are already won also Grand Prix "Golden Bear" of the Berlin International Film Festival, and in the U.S.
快挙を 心から喜びたい。	Inspiring feat I want to be glad from the bottom of my heart.
アカデミー賞は 5,000 人以上のハリウッド関係者の投票によって決まる。	Academy Awards is decided by vote of the 5,000 or more Hollywood persons concerned.
芸術性を重視する他の映画賞と異なり、商業性や、アメリカでの興行成績も重要な判断材料となる。	Unlike other movie prizes which think art as important, a commerce sex and the box-office record in the United States also serve as an important judgment material.
「千と千尋」はアメリカの上映期間が短く、事前の PR と不十分だった。	"1000 and 1000 fathoms" had the short U.S. show period, and it was as inadequate as prior PR.
日本人の心の深層を掘り下げた作品が、アメリカでどれだけ理解されるかも懸念された。	We were [which is understood in the United States] anxious about the work which investigated the depths of the Japanese alignment.
そうしたハンデを乗り越え、ディズニーなどの米製作会社による他作品を抑えて受賞したのは、いかにこの作品への評価が高かったかを示している。	Such a handicap was overcome, it called at U.S. production companies, such as Disney, and also the work was pressed down and it was awarded.
いかにこの作品への評価が高かったかを示している。	It is shown how the evaluation to this work was high.
日本のアニメ制作は、1950 年代後半から本格化し、テレビの放映で盛んになった。	Animation work of Japan got into stride from the second half of the 1950s, and prospered by televising of television.
アメリカの影響を強く受けた。	The influence of [U.S.] was received strongly.
今回の受賞は、日本アニメが既に独自の表現領域を確立している証しである。	This award is a proof which has established the expression area where Japanese animation is already original.
アカデミー賞受賞は「オスカー効果」と呼ばれる、経済効果ももたらす。	Academy Awards for an award also brings about the economic effect called the "Oscar effect."
日本のアニメのアメリカ市場への進出にも、弾みのつくことが期待される。	It is expected that the advance into the U.S. market of animation of Japan will also gain momentum.
だが、心せねばならないことがある。	But, an alignment may have to be carried out.
日本のアニメを支える基礎の部分が危うくなっていることだ。	It is that the part of the foundation supporting animation of Japan is dangerous.
日本で、劇場用アニメは年間 30 本以上、テレビアニメは 3,000 本以上も、制作されている。	In Japan, the animation for theaters is made or more in 30 per year, and 3,000 or more television animation is made.

日本語文	正解英語文
その基盤を支えるアニメ制作者の人材不足が、原画、演出、キャラクター・デザインなどに携わる職種で目立つ。	The run short of of talented people of the animation producer supporting the base is conspicuous with the occupational description engaged in an original picture, production, a character design, etc.
若い制作技術者の定着率が悪く、十分な経験を積めないままでいる。	A young work technical expert's fixing rate is bad, and keep sufficient experience not stacked.
契約社員やアルバイト社員が多い。	There are many contract employees and nonregular workers.
業界独特の雇用環境が影響している。	The employment environment peculiar to the industry has influenced.
原画制作などを、人件費の安い韓国や中国に発注し、「空洞化」が進んだことの原因のひとつだ。	It is one of the causes of having ordered original picture work etc. from the Rok and China where personnel expenses are cheap, and "emasculation" having progressed.
下請けとして経験を積んだ韓国や中国の制作者が、自主作品で高い評価を得ていることも脅威だ。	It is also a threat that the producer of the Rok which gained experience as a subcontract, or China has got high evaluation with the independence work.
デジタル化をどう生かしていくかも課題である。	Another challenge facing this nation 's animated film-making is how to take advantage of digital technology in movie production.
複雑な商習慣が国際的な流通を阻んでいる。	In addition, complicated commercial practices in this country hinder efforts to distribute domestic animated films in the international market.
日本のアニメが国際競争に生き残るには業界の体質改善が避けて通れない。	Given all this, it is essential to reform the filmmaking industry in this country as a whole so it will be able to flourish in the global market.
「千と千尋」の受賞を機に、日本の知的財産であるアニメの振興をさらに図りたい。	We hope that the honor accorded " Spirited Away " serves an important step in the advancement of Japan 's animated filmmaking as part of its intellectual property.
課題は山積している。	There are numerous tasks to be tackled in accomplishing that goal.

表 B.3: 評価データ一覧 (SPIRE2004 に投稿した海外論文)

日本語文	正解英語文
ＷｅｂページとＷｅｂコミュニティが扱う話題との関連性によって定義される「距離量」を利用したＷｅｂページ探索を行い、従来手法と比較して、精度と網羅性を向上させたＷｅｂコミュニティ抽出手法ＰｌｕｓＤＢＧを提案し、これについて評価を行う。	The proposed scheme adopts the distance defined by the relevance between a Web page and a Web community, and extracts the community with higher precision and recall.
また、その結果として、従来手法と同等の精度を維持したまま、従来手法と比べて約３．２倍のＷｅｂコミュニティのメンバーを発見することが可能であるとわかった。	Moreover, we implement and evaluate the proposed scheme. As a result, we have confirmed that the proposed scheme is able to extract the member of Web community about 3.2 times as large as the member of the conventional scheme, keeping equivalent precision compared with the conventional scheme.
近年の急速なWWWの普及に伴い、WWW上に存在するデータ数も爆発的に増加してきている。	The rapid growth of WWW is explosively increasing the data on WWW.
実際に検索エンジンなどがインデクシングしているＷｅｂページ数に注目すると、Ｇｏｏｇｌｅが２００４年４月の段階で実際に保有しているＷｅｂページ数が４２．８億ページを超えており、非常に大規模なデータベースを構成していることがわかる。	Indeed, Google has over 4 billion Web pages in April, 2004, which shows search engines have constructed huge-scale database.
また、WWWには様々な目的や意図を持った無数のユーザーが存在し、WWWは、部分的にはこれらのユーザーの意図に従って構築される。	WWW tends to be built according to the intent of users with diverse and often conflicting purposes at extremely individual level.
しかし、WWWは、全体としては計画性なしに構成されており、	Moreover, WWW is built without planning at global level.
WWW は情報の質や幅、扱われる話題など、場所によって様々である。	Thus, the Web data are various at quality, scale of information, topic, and format.
このような大規模かつ複雑なWWWの中から、価値のあるデータを抽出するＷｅｂマイニング技術が、現在では必須のものとなっている。	Therefore, Web mining techniques extracting valuable data from such large-scale and complex WWW are indispensable.
Ｗｅｂマイニングの問題の一つとして、Ｗｅｂコミュニティ抽出がある。	Web community extraction is one of the Web mining problems.
Ｗｅｂコミュニティとは「特定のトピックに対して興味を持つ、又は特定のトピックを扱うＷｅｂページの集合」と定義される。	A Web community is defined as a set of related Web pages interested in the same topic.
Ｗｅｂコミュニティ抽出では、WWWからこのようなＷｅｂコミュニティを抽出することにより、WWW中のＷｅｂページを意味的な集まりへと分類、整理する。	We call it “ Web community extraction ” to cluster Web pages into semantic sets.
従来、このようなＷｅｂページをトピックごとに分類する作業は、Ｙａｈｏｏ！などのようなＷｅｂディレクトリサービスの場において人手で行われてきた。	Conventionally, the Web directory service such as Yahoo! clusters Web pages into well-known topics by humans.
しかし、このような人手での分類作業には限界があり、膨大で、頻繁に更新されているＷｅｂページのすべてを適切に分類することは不可能に近い。	However, it is impossible to cluster all Web pages properly by humans because they are updated frequently and the number of them is huge.

日本語文	正解英語文
そこで、大規模なＷｅｂデータ中のＷｅｂページを自動的に分類・整理する手法として、Ｗｅｂコミュニティ抽出手法が注目されている	Consequently, Web community should be extracted automatically from large-scale Web data.
Ｗｅｂコミュニティの自動抽出における目的には２つの異なる視点がある。	The automatic extraction of Web community has two different purposes in comparison with Yahoo! ?like clustering..
(1) 既知のトピックに関する価値のあるページ、有用なページの集合を抽出する。	(1) Extracting the Web communities related to some known topics.
(2) WWW上で、未知の話題を持ったＷｅｂコミュニティを、可能な限り全て抽出する。	(2) Extracting all Web communities as many as possible without specifying any seed topics.
1. は、Ｋｌｅｉｎｂｅｒｇによる ＨＩＴＳアルゴリズムに代表され、ユーザーに対し、ユーザーが興味を持っているトピック中の代表的なページを提供する。	(1) is represented by HITS proposed by Kleinberg, that extracts the pages related to the topic in which a user is interested.
2. はＫｕｍａｒらによるｔｒａｗｌｉｎｇによって代表され、Ｗｅｂ中に潜在的に存在すＷｅｂコミュニティを、ユーザーに対し提供する。	(2) is represented by Trawling proposed by Kumar et al., that extracts implicit Web communities in WWW without specifying any seed topics.
しかし、これらの Web コミュニティ抽出手法は、抽出されるコミュニティにおいて、精度と網羅性の両方を満足させることを目的としてはいない。	However, these conventional schemes are not able to satisfy both precision and recall.
そのため、コミュニティ中にトピックとは関係のない Web ページが含まれていたり、逆に、重要なページのいくつかがコミュニティに含まれていない、といった場合がある。	This results in including the unrelated pages into the extracted Web community or excluding the related pages from the extracted Web community..
従って、ユーザーにとって有用で、理解可能な形の Web コミュニティを抽出するためには、精度と網羅性の双方を満足させる Web コミュニティ抽出が必要となる。	Thus, the Web community extraction scheme satisfying both precision and recall is indispensable in order to extract full advantages of Web communities.
そこで本稿では、話題が既知でない Web コミュニティを、可能な限りの精度と網羅性を持たせて抽出することを目的とする。	In this paper, we propose a new Web community extraction scheme improving both precision and recall, called PlusDBG, which extracts Web communities without specifying any topics.