文書の論理構造を利用したウェブベース翻訳支援システム

高橋　亜希子[†]　　是津　耕司[††,†††]　　小山　聡[††]　　田中　克己[††]

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
††† 独立行政法人 通信総合研究所　〒 184 － 8795 東京都小金井市貫井北町 4 － 2 － 1
E-mail: {takahasi,zettsu,oyama,ktanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　英文を翻訳していると訳が不確かな語と遭遇する．このような場面で我々は辞書を引き，例文を頼りに，辞書に載っているどの訳が現在読んでいる話題に合っているのか判断しようとする．しかし，多義語の場合，しばしば訳語を選択し兼ねることがある．訳の対象としている文書には論理構造があり，ウェブ上に存在する文章にも構造がある．このような文書の論理構造を利用して文脈や話題に沿った訳を提示することを考える．本論文では訳の対象としている文書の論理構造から Web に存在する文書を検索し，これらをもとに訳語の適切さを提示する手法を提案する．
キーワード　情報検索, 情報抽出, 多言語, 辞書, 翻訳

A Web-based Translation Support System with Document Structure Analysis

Akiko TAKAHASHI[†], Koji ZETTSU[††,†††], Satoshi OYAMA[††], and Katsumi TANAKA[††]

† School of Informatics, Kyoto University　Yosida-Honmachi, Sakyo-ku, Kyoto,606–8501 Japan
†† Graduate School of Informatics, Kyoto University　Yosida-Honmachi, Sakyo-ku, Kyoto,606–8501 Japan
††† Communications Research Laboratory 4-2-1 Nukui-Kitamachi, Koganei, Tokyo 184-8795 Japan
E-mail: {takahasi,zettsu,oyama,ktanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　We often encounter words of doubtful meaning when we read and translate documents written in English. Then, we consult with a dictionary to select the expression which perfectly matches the sentence concerned. In case of polysemic words, however, we find it difficult to select the optimum translation. The document to be translated and the electronic documents on the Web have their own logical structure. We wonder if such logical structure can be applied to effectively present the translation reflecting the contexts and topics of original documents. In this paper, we propose a system which enables presentation of appropriateness of translation through the process of searching the documents on the Web for the target expression by using the logical structure of the document to be translated.
Key words　information retrieval, information extraction, multilingual, dictionary, translation

1. はじめに

国際化が進み，外国語で書いた文書を読んだり翻訳したりする機会が増えた．これに伴い，読んでいる文書に対して正確な訳をすぐに得なければいけなくなった．このような状況下で，機械翻訳システムが注目されるが，現行の機械翻訳システムには以下のような問題点がある:

- 単語間の距離的のみから訳語を判断しているため，訳語の選択に迷っている語が所属する文脈を考慮して訳されていない．
- 最新の用例に対応していない．翻訳の際に用いられる辞書が更新されない限り，最新の訳で翻訳をすることができない．

機械翻訳システムには上記のような問題があり，支持されているとは言い難い．

通常，我々は単語の訳語の選択に迷うと，辞書を用いて訳語の候補を得る．しかし，辞書を用いる方法では，多くの例文の中から適切な例文を自ら選ばなければならず，その上，適切な訳語が見つからないことさえある．このような場面で，今日，我々はウェブ上の文書の表現を参考にして訳語の選択の手がか

りを得ることができる．しかし，我々にとって訳語の選択に役立つ情報を得ることが目的であって，それを探すプロセスは煩雑である．多くの場合，ウェブから翻訳対象語の訳語が適切に用いられている文書を探そうとすると，膨大な文書に目を通さなければならず，時間を要する．そこで，本研究では訳語の選択を支援するシステムを提案する．

訳語の選択を支援するシステムでは，以下のニつのステップが重要である．

（1） 翻訳対象語と関係のある語を取得する．

（2） ウェブを探索し，訳語候補をランキングする．

翻訳対象語と関係のある語を取得するには，翻訳対象語と共起度の高い共起語を取得する方法が考えられる．本研究では，文書の論理構造を反映させて対象語との共起度が高い共起語の抽出を試みる．また，ウェブを探索し，訳語候補をランキングするには，翻訳対象語と共起度の高い共起語の訳語を組み合わせて質問を生成し，検索結果を用いてランキングを行えばよい．

本研究では，はじめに文書の論理構造を反映させることで共起語の翻訳対象語に対する立場がどのように明確化されるか述べ，その取得方法について述べる．次に，共起度の高い共起語の訳語と翻訳対象語の訳語候補の組み合わせを質問とし，翻訳対象語の訳語が含まれる文書をウェブ上から検索し，これを根拠に訳語候補を妥当な順にランキングする方法について述べる．最後に，上記を実現する訳語選択支援システムを提案する．

2. 基本的事項

2.1 目　　的

翻訳対象語の訳語選定の際に，辞書を使っても機械翻訳を使っても適切な訳語候補を得られないケースがある．具体的には以下のようなケースである．

- 辞書に適切な訳語自体が載っていない場合
- 辞書には訳語が載っているが多義語で判断を迷う場合

このような場面でウェブを用いて訳語を発見することができる．具体的な手順は以下の通りである．

（1） 翻訳対象語の訳語候補の見当をつける．

（2） 翻訳対象語の関連語を，文書の論理構造と単語間の距離を反映させて翻訳対象文書中から見つける．

（3） 訳語候補と関連語の訳語を組み合わせて質問を生成し，ウェブを検索する．

（4） 検索の結果をもとにどの訳語候補が妥当か判断する．

本研究で扱う訳語選択支援とは，翻訳対象語の使われ方（背景）を提示し，単語の選択を容易にすることである．上記のステップを支援することで，ウェブ上で翻訳対象語がどのような文書で使われているのか，ユーザに提示することができる．本システムでは，上記の手順のステップ（2）から（4）を支援することを目的とする．

2.2 文書の論理構造

文書は通常複数の段落が入れ子状になって成り立っている．図 1 に示すように，注目している段落に対して関係のある段落を親段落，兄弟段落，子段落に分類することができる．翻訳対象語の所属する対象段落と各段落の関係は以下に示す通りである．

- 親段落

対象段落の話題を抽象化し，大まかに表すものである．

- 兄弟段落

対象段落の話題と並列に扱われる話題を表す．

- 子段落

対象段落の話題の一部を具体化し，詳細に表すものである．

2.3 文書の論理構造を反映した共起語

前節で述べたような文書の論理構造を前提として，翻訳対象語に対して文書の論理構造を反映して得られる共起語は以下の通りである．

- 親段落から得られた共起語：

翻訳対象語の所属する段落の一般的な話題を表す．

- 兄弟段落から得られた共起語：

翻訳対象語の所属する段落と並列に扱われる話題を表す．

- 子段落から得られた共起語：

翻訳対象語の所属する段落の具体的な話題を表す．

2.4 訳語選択支援システムの概要

前節までの内容を踏まえて，本研究では訳語選択を支援するシステムを提案する．訳語選択支援システムの動作の概要は以下の通りである．

（1） 翻訳対象文書と翻訳対象語を翻訳支援システムに渡す．

（2） 翻訳対象文書中の対象語と共起度の高い語（以下，共起キーワード）を抽出する．

（3） 翻訳対象語の訳語候補，共起キーワードの訳語を辞書またはユーザーから得る．

（4） 翻訳対象語の訳語候補，共起キーワードの訳語を組み合わせて質問を生成し，ウェブに対して問い合わせを行う．

（5） 妥当な順に語の組み合わせをユーザに提示する．

訳語選択の過程で，特に共起キーワードを翻訳対象文書中から抽出するステップ，翻訳対象語の訳語候補と共起キーワードの訳語を用いて質問を生成し，ウェブ上を検索するステップに注目する．本研究では，翻訳対象語の訳語を選択するのに有効な共起情報について考察し，これらを用いた質問生成・訳語候

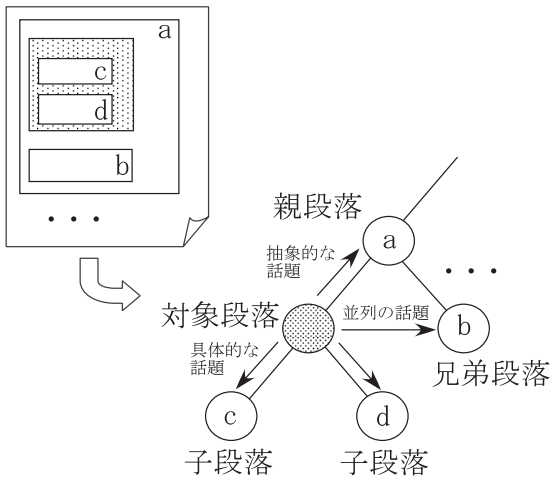

図 1 文 書 構 造

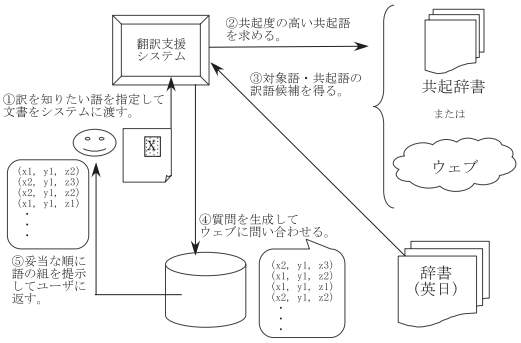

図 2　翻訳支援システムの概要

補のランキングについて検証する．最後に実際に訳語選択支援システムのプロトタイプを試作する．

3. 共起の種類

従来の共起情報を用いた手法では，文脈に応じて適切な訳語を選択するために，周辺の近接する単語との共起関係を用いていた．例えば，”domain”という単語を訳す場合，3語以内に”knowledge”という単語があれば，この単語は”領域”（領域知識）と訳される．一方，より複雑な訳語選択の方法としては，言語構造に基づいて，単語間の係り受け関係などから適切な訳語候補を選択することが行われてきた．

しかし，いずれも翻訳対象語の距離的周辺の共起情報のみを用いる手法で，文脈に沿った訳を提示するには不十分だと考えられる．本研究では文書の論理構造を反映した共起情報の利用について考察する．

3.1 文書の論理構造を反映した共起情報

本研究では，翻訳対象語の近傍のみからではなく，翻訳対象語の話題に対する立場が明確にわかっている親段落・兄弟段落・子段落からも共起キーワードを抽出し，利用することを考える．より広い話題を表す親段落，並列な話題を表す兄弟段落，より詳細な話題を表す子段落，それぞれでの共起キーワードを利用する必要があると考え，単語間の距離のみではなく，文書の論理構造も反映させた共起キーワードの利用を検討する．

3.2 共起の分類

ここで改めて翻訳対象語に対する語の共起の種類について述べる．はじめに，文書の論理構造を反映しているか否かで分類することができる．次に，文書の論理構造を反映させる場合は翻訳対象語の所属する段落と共起語の所属する段落の関係によって分類することができる．分類は以下の通りである．

- 文書の論理構造を反映しない場合：

翻訳対象語の周辺 n 文字での共起．近傍共起と呼ぶ．

- 文書の論理構造を反映する場合：

– 親共起

翻訳対象語の所属する段落の親段落での共起．

– 兄弟共起

翻訳対象語の所属する段落の兄弟段落での共起．（これは距離的周辺での共起を拡張したものである．）

– 子共起

翻訳対象語の所属する段落の子段落での共起．

以降，兄弟共起に関しては，翻訳対象語の周辺 n 語での近傍共起を拡張したものとし，これと合わせて考えることにする．文書の論理構造を反映した共起としては，親共起・子共起のみを考える．

4. 訳語選択アルゴリズム

訳語選択アルゴリズムとは，ユーザがある Web 文書から訳語を知りたい英単語を選択すると，文書内でその英単語の構造的周辺（親段落，子段落）および距離的周辺（翻訳対象語の周辺 n 語）に出現する周辺キーワードとの共起関係に基づき，その英単語の日本語訳候補の中から適切な訳語を選択するアルゴリズムである

今，翻訳対象となる英単語を翻訳対象語と呼び，k で表す．翻訳対象語 k の訳語候補の選択は以下の手順によって行われる：

（1） 翻訳対象語 k が含まれる文書（英語）から，k と共起度の高いキーワードの集合 $C(k)$ を抽出する．

（2） 翻訳対象語 k とその共起キーワード集合 $C(k)$ を合わせた翻訳キーワード集合 $E = \{e_i | 1 \leqq i \leqq n, 2 \leqq l \leqq n, e_1 = k, e_l \in C(k)\}$ に対して，英日辞書を使って各キーワード e_i の日本語訳語候補集合 $J(e_i)$ を取得し，E の各キーワードの訳語候補を組み合わせた訳語組合せ集合 $Q(E) = \{(j_1, j_2, \ldots, j_n) | 1 \leqq i \leqq n, j_i \in J(e_i)\}$ を生成する．

（3） 訳語組合せ集合 $Q(E)$ の各訳語組合せ $q = (j_1, j_2, \ldots, j_n)$ に対し，q に含まれる全ての訳語を含む Web ページ（日本語）を検索し，そのページ数 $N(q)$ に基づいて $Q(E)$ 中の訳語組合せをランキングする．

（4） 訳語組合せ集合 $Q(E)$ の中で高位にランキングされた訳語組合せ q から，翻訳対象語 k に対応する訳語 (j_1) を，この文書における k の適切な訳語として抽出する．

翻訳対象語 k に対する共起キーワード集合 $C(k)$ は，以下のように定義される：

$$C(k) = \{w | w \in D(k), coocc_x(w,k) > \theta\} (x = p, c, n)$$

ここで，$D(k)$ は翻訳対象語 k を含む文書内の全キーワード集合を表し，$coocc_x(w,k) (x = p, c, n)$ は翻訳対象語 k に対するキーワード w の段落別共起度 (p は親段落，c は子段落，n は k の近傍 n 語）を表す．即ち，$coocc_p(w,k)$ はキーワード w が翻訳対象語 k の親段落に出現する頻度，$coocc_c(w,k)$ はキーワード w が翻訳対象語 k の子段落に出現する頻度，$coocc_n(w,k)$ はキーワード w が翻訳対象語 k の n 語彙内の近傍に出現する頻度を表す．θ は，共起度閾値である．各々の共起度の取得方法，さらに共起度を用いた共起キーワードの抽出方法については次章で詳細を述べる

上記のアルゴリズムでは，訳語組合せ q の妥当性を q に含まれる訳語を含む Web ページ（日本語）数で判定したが，より厳密に妥当性を評価するには，q に含まれる訳語の Web ページ内での段落別共起関係を調べるように基本アルゴリズムを拡張することが考えられる．即ち，翻訳キーワード集合 $E = (e_1, e_2, \ldots, e_n)$ において，例えば e_2 が e_1（翻訳対象語 k に対応）の親段落に出現していれば，対応する訳語組合せ

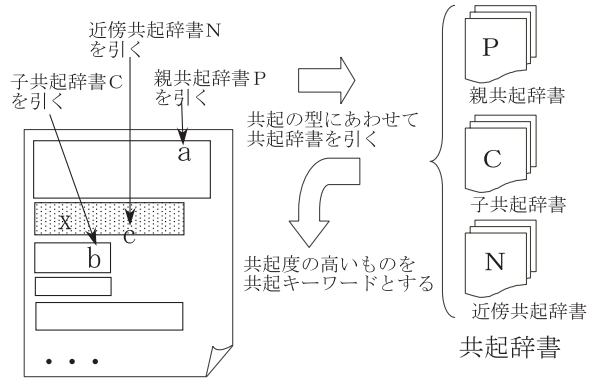

図 3　共起辞書を利用して共起度を求める方法

$q = (j_1, j_2, \ldots, j_n)$ において j_2 が j_1 の親段落に出現している Web ページの数を評価する．

5. 共起キーワードの抽出

共起キーワードを抽出する際には，ウェブに存在する文書全てから，文書の論理構造を反映させて抽出するのが理想的である．しかしながら，ウェブ上の文書は厳密に構造化されていないため，本研究では，翻訳対象語の共起キーワードを抽出する方法として以下のような近似的な方法を提案する．

それぞれの方法とその長所・短所は以下のような点が挙げられる．

- 方法１：共起辞書を利用して共起キーワードを求める方法

特定の分野の文書群（ニュース，論文など）を対象に厳密な文書構造を定義して共起辞書を作成するし，共起キーワードを抽出する．限られた文書群から共起辞書を作成しているため，ウェブの”最新の用例が存在する”という長所を活かしきれておらず，従来の辞書と同様，掲載されている語に制限されてしまうという問題が残る．

- 方法２：ウェブを利用して共起度を求める方法

常に最新のウェブ上の用例を用いて，共起度を求め，共起キーワードを抽出しているため，ウェブの成長性を活かしている．しかしながら，ウェブ上の全文書を対象としているため，厳密な文書構造を反映させることができない．また，実際利用する際の問題として，実行速度の問題が挙げられる．

本章ではこれらのふたつの方法についてそれぞれの詳細を述べる．

5.1 方法１：分野に特化した厳密な文書構造を反映して共起キーワードを抽出する方法

この方法では予め，一定の構造を持ったウェブ上の文書群から，共起の種類別に共起辞書を作成する．これら作成した共起の種類別の共起辞書を用いて，翻訳対象文書から共起キーワードを取得する．

この方法の概略は図 3 に示す通りである．

5.1.1 共 起 辞 書

翻訳対象語の共起キーワードを知るために，文書構造を反映した構造共起辞書と語間の距離を反映した近傍共起辞書を作成する．それぞれの共起辞書については以下の通りである．

- 構造共起辞書

– 親共起辞書

親段落に共起する語とその共起度を記載した辞書である．例えば，語 A について親段落に 100 語，語 B が共起しているならば，語 A について語 B の親段落での共起度は 100 であるといったことが書かれている..

– 子共起辞書

子段落に共起する語とその共起度を記載した辞書である．記載内容は親共起辞書と同様である．

- 近傍共起辞書

近傍 n 語に共起する語とその共起度を記載した辞書である．子段落に共起する語とその共起度を記載した辞書である．記載内容は親共起辞書と同様である．

これらは信憑性のために構造を持つ多数の文書からなる文書群抽出されなければならない．構造を持つ多数の文書群としてウェブ上の文書群が考えられる．

5.1.2 共起辞書の作成

共起辞書の作成対象

共起辞書の作成対象のウェブ上の文書群として，新聞記事や論文など一定の基準で構造化された文書群を考える．本研究では www2000 paper proceedings(http://www9.org/w9cdrom/index.html) に挙げられている論文 54 件を利用して共起辞書を作成した．

論文の論理構造

本節で共起辞書を作成する際には前述のように論文を用いる．論文の論理構造として，題名とアブストラクトを最上位の親とし，次に章がその子，節がその孫として扱うものとする．

文書の論理構造を反映した共起辞書・近傍共起辞書の作成

翻訳対象語に対して www 2000 paper proceedings に挙げられている論文群から共起辞書を作成する．

- 構造を反映した共起辞書

親共起辞書と子共起辞書の二種類の辞書を作成する．文書群中の単語全てについて，親段落の共起語を抽出し，親共起辞書に書き込み，同時に出現回数を数え，これを共起度として書き込む．同様のことをを子段落についても行う．作成された共起語の辞書は以下のようなものである．[共起語]:[共起度] という形式で表されている．

```
---------
親
[enables]:[1]
[attacks]:[3]
 ...
子
[require]:[19]
[inside]:[2]
 ...
---------
```

- 近傍共起辞書

共起語の種類	共起語
親共起語	pages(137), information(135), site(85)
子共起語	user(226), information(156)
近傍共起語	information(6), scope(2), component(2)

表 1　翻訳対象語”domain”に対して共起度の高い共起語

翻訳対象語の近傍 n(ここでは $n = 20$ とする) 語以内に出現する共起語とその共起度を記載した近傍共起辞書を作成する．作成された共起辞書は以下のようなものである．[共起語]:[共起度] という形式で表されている．

```
---------
近傍共起語
[resnick]:[1]
[awarded]:[1]
[compared]:[5]
…
---------
```

次節ではこのようにして作成した共起辞書を利用して共起キーワードを抽出する方法について述べる．

5.1.3　共起キーワードの抽出

翻訳対象文書中から翻訳対象語の共起キーワードを抽出する方法は以下の通りである．

- 構造共起辞書の場合：

翻訳対象語の親段落にある共起キーワードを抽出する場合，親段落の語すべてについて親共起辞書を用いて共起度を調べ，共起度の高いものを共起キーワードとする．子共起辞書も親共起辞書と同様に利用できる．

- 近傍共起辞書の場合：

翻訳対象語の近傍 n 語を取得し，近傍共起辞書を用いてそれぞれの共起度を調べ，共起度の高いものを共起キーワードとする．

5.1.4　具　体　例

翻訳対象文書を http://www9.org/w9cdrom/263/263.html，翻訳対象語を”domain”として抽出された共起キーワードを表 1 に示す．

5.2　方法２：ウェブページに共通の緩い文書構造を反映して共起キーワードを抽出する方法

ウェブには以下のような性質がある：

- ウェブは時々刻々成長している
- 多くの人が書いた莫大な数の文書が存在する

このような性質を利用することで，既存の辞書に頼って翻訳を行う場合と異なり，最新の語の用法を翻訳に取り入れることができる．本節ではウェブ上の全文書を利用して共起キーワードを翻訳対象文書中から抽出する方法について述べる．

この方法の概略は図 4 に示す通りである．

5.2.1　ウェブページに共通の緩い文書構造

すべてのウェブページに共通する文書構造として以下のような文書構造を考えることができる．

- title 内の語は text 内に存在する語の親共起語である．
- text 内の語は title 内に存在する語の子共起語である．

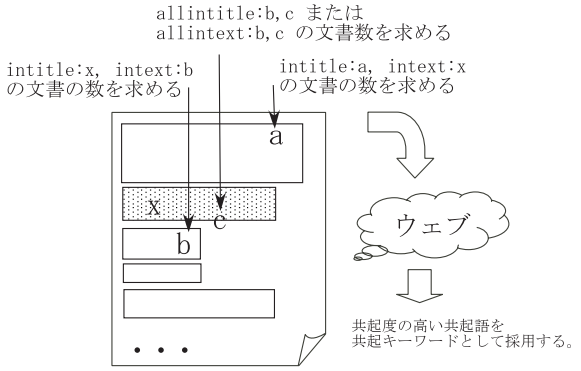

図 4　共起辞書を利用して共起度を求める方法

このような文書構造を利用する手段として，検索エンジン Google が提供する intitle, intext 検索という機能がある．それぞれの機能については以下の通りである．

- intitile 検索：質問 X で intitle:x として指定された文字列 x がタイトルに含まれる文書を検索する．
- intext 検索：質問 X で intext:x として指定された文字列 x がテキストに含まれる文書を検索する．

例えば，”macintosh”がタイトルに，”computer”がテキストに含まれる文書を検索する場合，Google に対して”intitle:macintosh intext:computer”というように質問を生成する．

次に，このような緩い文書構造を反映させて共起キーワードを抽出する方法について述べ，共起キーワードを抽出する具体例を述べる．

5.2.2　共起キーワードの抽出

翻訳対象語を k とし，翻訳対象文書中の共起キーワードの抽出方法について述べる（但し，k は翻訳対象文書中でテキスト内にあるものとする）．

- 親共起語と k の共起度の求め方：

（1）　翻訳対象文書中のタイトルに含まれる語を k の親共起語とし，親共起語の集合を $P = \{p_1, p_2, \ldots\}$ とする．

（2）　共起度は，k がテキストに (intext)，親共起語 $p(\in P)$ がタイトルに含まれる文書の数で表すものとする．全ての親共起語 $p(\in P)$ について，k との共起度を求める．

- 近傍共起語と k の共起度の求め方：

（1）　翻訳対象文書中の k の近傍 m 語を k の近傍共起語とし，近傍共起語の集合を $N = \{n_1, n_2, \ldots\}$ とする．

（2）　共起度は，k と近傍共起語 $n_i(\in N)$ がテキストまたはタイトルに同時に含まれる文書の数で表すものとする．全ての近傍共起語 $n_i(\in N)$ について，k との共起度を求める．

このようにして，共起語の共起度を求め，共起度の高い共起語を共起キーワードとする．

5.2.3　具　体　例

具体例として，文書 (http://www.women.com/style/clothes/photo/0,13829,535086_535310,00.html) 中の”macintosh”という語について考える．この文書内で語”macintosh”は布地の加工の一種（特にコート）という意味で用いられている（コンピュータのマッキントッシュ，りんごの品種のマッキントッシュ

という意味合いではない）．親共起語（title），周辺 20 語は以下の通りである．

- 親段落 (title)：7, Classic, Pieces, Every, Gal, Should, Own, Trench, Coat
- 周辺 (20 語)：Pricey, Atsuro, Tayama, from, Boudoir, in, Nolita, NYC, 244, Mulberry, St, 212.965.9925, 578.00, $Moderate$, $Brooks$, $Brothers$, $Three-Button$, $Coat$,168

これらについて，前述の方法で共起度を求め，ストップワードを除き，共起キーワードを抽出すると，以下のような語が得られた:

- 親段落 (title)：classic(5950)，pieces(485)，coat(167), gal(76), trench(10)
- 周辺 (20 語)：
  - allintitle: mulberry(34)
  - allintext: brothers(55800), brooks(32000), moderate(32000), coat(22900), NYC(27200), mulberry(10900)

6. 共起語の訳語を用いた質問生成・ランキング

前章では翻訳対象文書中から翻訳対象語の共起キーワードを抽出する方法について述べた．特に，方法 2 のウェブ文書に共通する文書の論理構造を利用する方法では，定義した文書の論理構造の浅さから，子段落の共起キーワードを得るのは困難である．また，方法 1 の分野に特化した厳密な文書構造を利用する方法でも，文書に子段落自体がない場合が多い．そこで，本章では近傍の共起キーワードと親段落の共起キーワードを質問生成に利用することにする．但し本章で挙げる例では前章の方法 1 で得られた共起キーワードを用いるものとする．翻訳対象語”domain”の翻訳対象文書中での妥当な訳語は”ドメイン”である．

本章では，はじめに，質問生成のパターンを提示し，その質問を用いた検索結果の利用方法について述べる．続いて提示した質問のパターンの具体例を示し，生成された質問での検索結果を示す．

6.1 質問生成・ランキング

本研究で検証した共起キーワードの組み合わせのパターンは以下の通りである．ここで，翻訳対象語に対して，親段落の共起キーワードの集合を $P = \{p_1 \cdots p_x\}$，近傍の共起キーワードの集合を $N = \{n_1, \cdots, n_y\}$ とする．また，訳語は，英単語が A 場合，$J(A)$ と表すものとする．

- 訳語候補と親段落の共起キーワードの訳語すべての組み合わせ

$Q = J(k) \wedge (J(p_1) \wedge \cdots \wedge J(p_x))$

- 訳語候補と近傍の共起キーワードの訳語の一部の組み合わせ

$Q = J(k) \wedge (J(n_1) \vee \cdots \vee J(n_y))$

- 訳語候補と親段落の共起キーワードの訳語すべてと近傍の共起キーワードの訳語の一部の組み合わせ

$Q = J(k) \wedge (J(p_1) \wedge \cdots \wedge J(p_x)) \wedge (J(n_1) \vee \cdots \vee J(n_y))$

翻訳対象語の訳語候補	検索ページ数
専門	508000
分野	362000
ドメイン	179000
領域	126000

表 2　親段落の共起キーワードすべてと訳語候補の組み合わせ

翻訳対象語の訳語候補	検索ページ数
専門	7280
分野	6840
領域	6660
ドメイン	2780

表 3　近傍の共起キーワード ”スコープ（scope）” と訳語候補の組み合わせ

翻訳対象語の訳語候補	検索ページ数
専門	14200
分野	25100
領域	22500
ドメイン	14100

表 4　近傍の共起キーワード ”コンポーネント（component）” と訳語候補の組み合わせ

我々は，ウェブを利用して検索を行う際に，知りたい語が所属する分野名や話題を用いて大まかな絞込みを行う．また，更に絞込みを行いたい場合は，その語に関連する具体的な語を質問に加えて検索を行う．このような行動をもとに生成した上記のような質問のパターンについて検証を行う．

また，ランキングの際には，質問を表す語の組み合わせがウェブ上の文書中に多く含まれるものほど，妥当な組み合わせとし，組み合わせに含まれる訳語候補を妥当な訳語候補として扱う．本研究ではページ数を指標としてランキングを行う．

6.2 親段落の共起キーワードを利用する場合

翻訳対象語にすべての親段落の共起語を加えて検索を行った結果を 2 に示す．親段落のすべての共起キーワードと各訳語候補を組み合わせると，(専門，ページ，情報，サイト）という語の組み合わせを含むウェブページが最も多いことがわかる．続いて，(分野，ページ，情報，サイト），(ドメイン，ページ，情報，サイト），(領域，ページ，情報，サイト）の順に支持されている語の組み合わせである．

しかしながら，親段落の共起キーワードと訳語候補の組み合わせによる問い合わせでは，全体的にページ数が多く絞込みが完全に行われていないと考えられる．親段落の共起キーワードは翻訳対象語が所属する話題を表していると考えることができ，話題の大まかな指定に利用できる可能性がある．

6.3 近傍の共起キーワードを利用する場合

翻訳対象語に近傍の共起キーワードを加えて質問を生成し，ウェブに問い合わせを行った結果を表 3，表 4，表 5 に示す．親段落の共起キーワードを用いた場合と同様，各質問の結果から妥当な語の組み合わせがわかり，語の組み合わせから各訳語候

翻訳対象語の訳語候補	検索ページ数
領域	826
ドメイン	619
分野	486
専門	384

表 5　近傍の共起キーワードすべてと訳語候補の組み合わせ

翻訳対象語の訳語候補	検索ページ数
専門	2070
分野	1350
領域	1340
ドメイン	1230

表 6　訳語候補とすべての親段落の共起キーワードと近傍の共起キーワード ”スコープ（scope）” の組み合わせ

翻訳対象語の訳語候補	検索ページ数
領域	6380
分野	6160
ドメイン	6010
専門	4930

表 7　訳語候補とすべての親段落の共起キーワードと近傍の共起キーワード ”コンポーネント（component）” の組み合わせ

翻訳対象語の訳語候補	検索ページ数
ドメイン	334
領域	333
専門	167
分野	165

表 8　訳語候補と親段落・近傍のすべての共起キーワードの組み合わせ

補の用いられる背景を読み取ることができる．

近傍の共起キーワードを用いた結果では，親段落の共起キーワードを用いた場合に比べ，検索の結果得られるページ数が極端に少ない．これは狭い話題の範囲から共起語を取得したため，話題を限定してしまっているからであると考えられる．しかしながら，近傍の共起キーワードは話題を絞り込む際に役に立つ可能性がある．

6.4　親段落の共起キーワードと近傍の共起キーワード両方を利用する場合

翻訳対象語と親段落の共起キーワードすべてに近傍の共起キーワードをひとつずつ加え，問い合わせを行った結果を表 6，表 7，表 8 に示す．親段落の共起キーワードを用いた場合と同様，各質問の結果から妥当な語の組み合わせがわかり，語の組み合わせから各訳語候補の用いられる背景を読み取ることができる．

親段落・近傍両方の共起キーワードを用いると，ひとつの種類の共起キーワードのみを用いた場合に比べ，キーワードの数が多いことからも翻訳対象語の背景がより明確になると考えられる．また，近傍の共起キーワードすべてを加えた場合では，求めている訳語候補”ドメイン”のランクが最も高くなっている．このことからも，ひとつの種類の共起キーワードを用いるよりも，親段落・近傍の共起キーワードを組み合わせて用いる方が良いと考えられる．

6.5　質問生成とランキングに関する考察

本研究では，多くの人が使用している語の組み合わせは妥当な用例であるとして，検索結果のページ数を訳語の適切さの指標としている．しかしながら，検索の結果得られるページ数は当然のことながら質問の生成の仕方で異なる．実際には以下のような問題が考えられる．

- 絞り込み過ぎると合計ページ数が減ってしまい，結果の信頼度が薄れる．
- ランキングの際にどこまで絞込みを行った結果を最良のものとして採用するか．

共起キーワードの組み合わせのパターンでは，どの方法が最も優れているとは言い難い．種類の異なる共起キーワードを併用することによって，より翻訳対象語の背景を知ることができる点では，単一の種類の共起キーワードを用いるよりも組み合わせて用いる方が良いと考えられる．

7.　翻訳支援システムの実装

これまでに述べた方法に基づいて実装した訳語選択支援システムの実行画面を図 5 に示す．訳語支援システムは，翻訳対象文書に対して別ウィンドウで現れる．ユーザが翻訳対象語を指定すると，共起キーワードを提示し，指定された共起キーワードをもとに検索をし，妥当な訳語を提示する．

翻訳支援システムのプロトタイプのシステムは以下のようにして処理を行う．

（1） ユーザから与えられた翻訳対象文書と翻訳対象語をもとに，親段落の共起キーワード，近傍の共起キーワードを発見する．

（2） 共起度キーワードをユーザに提示する．

（3） ユーザが指定した共起キーワードの組み合わせで質問を生成する．

（4） 検索結果から妥当な訳語の組み合わせを提示する．

8.　関 連 研 究

単語の用例は，あるキーワードがどの様な文脈でどの様な単語と一緒に使われれるかということを，そのキーワードを含む周辺テキストを使って表す．KWIC [1] は，単語の用例提示の方法としてよく知られており，キーワードを中心に複数の周辺テキストをリスト表示し，キーワードの様々な文脈を一覧することが出来る．KWIC on Google [2] は，キーワードの Web 上での用例を抽出し KWIC 形式で提示する．英次郎 [3] は，こうした単語の用例提示を翻訳に応用し，ある英単語の翻訳結果を，その単語の周辺テキスト（英語）および訳語の周辺テキスト（日本語）とともに提示する．この様に，翻訳結果を英日両方の用例とともに提示することで，英単語がどの様な文脈で用いられているときにどの様に日本語に翻訳されるのかを対比させながら適切な訳語を選択することができる．これらは，単語の周辺テキストを利用している点では本研究と関連があるが，本研究では文書構造を用いている点，および単なる用例提示だ

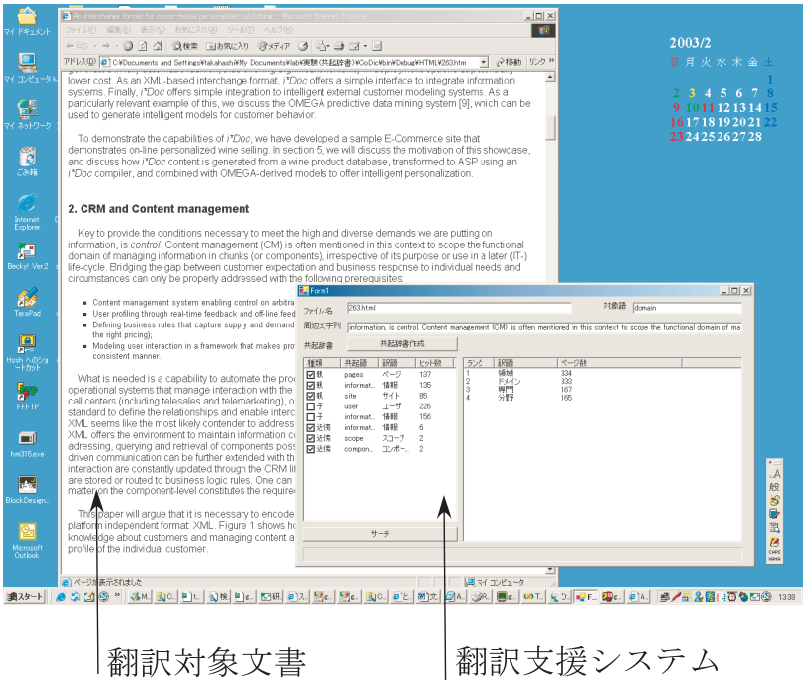

図 5　翻訳支援システムの実行画面

けではなく訳語候補の選択を行っている点が異なる．

また，是津ら [4] は，Web から画像などのマルチメディア・オブジェクトの周辺コンテンツを Web 文書構造やハイパーリンクに基づいて抽出し，マルチメディア・オブジェクトの用例の生成を行っている．周辺コンテンツの抽出に文書構造を用いている点は本研究と関連しているが，本研究は単に周辺を抽出するだけでなく，周辺コンテンツ（キーワード）との共起関係を調べ翻訳に活用している点が異なる．

9.　今後の課題

今後の課題として，質問の生成方法と質問の結果を利用したランキングについて特に検討が必要である．検討すべき事項は以下の通りである．

- 共起キーワードの組み合わせ方法

本研究では限られた共起キーワードの組み合わせで実験を行った．しかし，どの質問が最良という答えは得られず，プロトタイプシステムでは共起キーワードの組み合わせをユーザに任せる方式を採用した．今後，一般に有効な共起語の組み合わせ方法や，翻訳対象語により組み合わせを変化させる方法などの検討が必要である．

- 訳語候補の絞込み

訳語候補を絞り込む際にどこまでページ数を絞り込めばよいのか，という問題が残されている．”それぞれの候補に対するページ数に有意な差があり，かつ信頼に足るページ数”でなければならないが，これをどういう指標をもって実現するか検討しなければならない．

10.　おわりに

文脈を考慮した機械的な翻訳は未だ困難な状況にある．このような状況下で本研究では以下の方法を提案した．

- 文書の論理構造を反映した共起キーワード抽出

文書構造に注目し，従来困難だった文脈を考慮した訳に近づくことができた．

- ウェブ上の文書を用いた訳語候補の選択

ウェブを文書の宝庫と見ることで，従来得ることのできなかった膨大な文書をもとに訳語選択の支援を行える．

本研究では，上記のようなアプローチで，翻訳作業中の訳語選択を支援するシステムを想定し，文脈を考慮するための指標として文書の論理構造を反映して共起キーワードを抽出・利用する方法を提案した．また，翻訳作業中のユーザにとって副次的な作業であるにも関わらず，時間を取られてしまっていたウェブ検索を容易にするという点でも効果があったと考えられる．

謝辞

本研究は，一部平成 14 年度科研費特定領域研究 (2) 「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」（課題番号：14019048，代表：田中克己）による．ここに記し謝意を表します．また，本研究は，一部２１世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」による．ここに記し謝意を表します．

文　　献

[1] H. P. Luhn. Keyword in context index for technical literature (kwic index). *American Documentation*, No. 11, pp. 288–295, 1960.
[2] *KWIC on Google, http://163.136.182.112/xyz01/.*
[3] 道端秀樹．英次郎．アルク，2002.
[4] 是津耕司，角谷和俊，田中克己．Multimedia corpus: マルチメディアの用例のデータベース化．情報処理学会研究報告 2002-DBS-128, pp. 367–374, 2002.