情報学基礎 51－8
自然言語処理 127－8
（1998. 9. 17）

多言語情報検索のための複合語翻訳

藤井敦, 石間衛, 石川徹也
図書館情報大学
{fujii,ishima,ishikawa}@ulis.ac.jp

現在我々は、英日/日英多言語情報検索システムを開発しており、本論文はその一つの基盤要素である複合語翻訳法を提案する。複合語は語基を組み合わせることで漸進的に生成されるため、網羅的に対訳を用意することが困難である。そこで本手法は、ユーザが検索質問として入力した複合語を語基の対訳辞書を用いて動的に翻訳し、検索に利用する。本論文では、語基対訳辞書の作成と語の共起に基づく統計的翻訳モデルについて説明し、英日多言語検索の実験を通して本翻訳法の有効性を示す。さらに、ユーザが検索結果を閲覧しながらシステムを学習させるための枠組を提案する。

Compound Word Translation for Cross-Langage Information Retrieval

Atsushi FUJII, Mamoru ISHIMA, Tetsuya ISHIKAWA
University of Library and Information Science
{fujii,ishima,ishikawa}@ulis.ac.jp

This paper proposes a method for translating compound words, which is one component of our prototype English-Japanese/Japanese-English cross-language information retrieval (CLIR) system. Since exhaustive enumeration of progressively created compound words is less than reasonable, our method dynamically translates English compound queries into Japanese, using an English-Japanese base word dictionary and Japanese collocational information. We demonstrate the effectivity of our translation method by way of experiments, in which our method improves on the performance of baseline CLIR systems. We also propose an interaction strategy to facilitate use feedback.

1 はじめに

多言語情報検索 (Cross-Language Information Retrieval: CLIR) の研究開発は、古くは 1960 年代の道路技術に関する文献検索システム [30] や 1970 年代の Salton の実験 [37, 38] にまで遡ることができる。近年は、コンピュータネットワークを通じて母国語以外の文書にアクセスできる機会が増えており、CLIR の研究は益々盛んになっている [1, 2, 3]。

では CLIR とは一体何だろうか? まずここで、その定義を明確にしておきたい。Hull と Grefenstette [20] は CLIR[1] の定義として以下の 5 つを挙げている。

(1) 英語以外の言語を用いた情報検索処理
(2) 検索質問 (query) と同じ言語の文書だけを多言語データベース (個々の文書は単一言語で書かれている) から検索する処理
(3) 多言語の検索質問を使って単一言語データベースから文書を検索する処理
(4) 多言語の検索質問を使って多言語データベースから文書を検索する処理
(5) 多言語文書 (一つの文書が複数の言語を含んでいる) を検索する処理

本論文では、その他多くの研究と同様に、定義 (3)(4) を限定的に用いる。すなわち CLIR とは「検索質問と異なる言語の文書を検索する処理」である (単一言語を対象とした通常の情報検索は information retrieval (IR) として、CLIR と区別する)。そこで、CLIR を実現するためには検索質問か検索対象文書のどちらかを翻訳するか、あるいは両方を中間言語に変換する必要がある。既存の CLIR 法の多くは、対訳辞書、コーパス、機械翻訳システムなどを用いて翻訳や変換を行う。しかし、訳語曖昧性や翻訳誤りのために、一般に CLIR の検索性能は IR には依然及ばない[2]。

現在我々は、英日/日英 CLIR システムの開発を進めている。本論文では、我々の CLIR システムの検索性能を向上させるための手段の一つとして、検索質問として入力された複合語の専門用語の翻訳に焦点を当てる。複合語は造語力が大きく、特に専門用語の複合語は学問の発展に伴って漸進的に作られるため、網羅的に対訳を用意することが困難である[3]。本論文では、複合語を構成する「語基」[4] の対訳辞書を用いて辞書未登録の複合語対訳を自動的に構成し、検索に利用するための手法を提案する。

2 節で従来の CLIR の研究について検討する。3 節で我々が開発中の CLIR システムについて概説し、4 節で複合語翻訳法について説明する。5 節で本翻訳法を用いた英日 CLIR の実験について説明し、6 節でユーザとのインタラクションを用いたシステムの学習機能を提案する。

[1] Hull と Grefenstette [20] は "multilingual information retrieval (MLIR)" という言葉を使っている。

[2] 近年のいくつかの実験では、CLIR の検索性能 (平均適合率) は IR の 60-70%程度であると報告されている。CLIR の評価法については、2.3 節で説明する。

[3] この問題は複合語解析の研究でも指摘されている。

[4] 本論文では、語基とは一つの形態素を指す。

2 先行研究

2.1 検索方式

従来の CLIR 法を図 1 に示すように大きく 3 つの方式に分類し、以下それぞれについて説明する[5]。

検索質問翻訳方式 検索質問を検索対象文書の言語に翻訳して検索を行う方式である。モジュラリティーに富んでおり、検索エンジンには既存の手法 (キーワードマッチングやベクトル空間法 [39] など) を用いることができる。翻訳のための知識資源には、機械可読の対訳辞書やコーパスなどが用いられる。

Hull と Grefenstette [20] は対訳辞書から得られる訳語候補を全て検索に利用している。Hull [19] は重み付き boolean を用いて訳語候補に重要度を付与している。Ballesteros と Croft [5] はフレーズ単位の翻訳と Local Context Analysis [49] を導入して、対訳辞書による手法の検索性能をさらに向上させている。Aone ら [4] は、人手で作成したローマ字とひらがな (カタカナ) の対応規則による翻字 (transliteration) と対訳辞書による翻訳を複合的に用いている。

Carbonell ら [8] は、文対応付き二言語コーパス (sentence-aligned bilingual corpus) から対訳を自動的に抽出する手法 [7] を用いている。彼らは、後で述べる GVSM、LSI などよりも文対応付きコーパスに基づく検索質問翻訳の方が良い検索性能であることを実験によって示している。Lee と Choi [28] は、英韓単語対応付きコーパスから抽出した統計情報を利用して検索質問の翻字を行っている。しかし、文/単語対応付きコーパスの作成は人手のコストが高いという問題がある。

対訳辞書には訳語の曖昧性があるため、全ての訳語候補を利用する手法 [20] では、検索結果に不必要な文書が含まれたり、検索時間効率を低下させる原因になる。そこで、コーパスを用いて訳語の曖昧性を解消する手法が提案されている[6]。本論文で提案す

[5] CLIR のサーベイとして他に文献 [21, 31, 32] などがある。

[6] 訳語曖昧性を解消する場合、(a) 訳語をユニークに決定する、(b) 尤度の高い複数の訳語を選択する二つの選択肢がある。

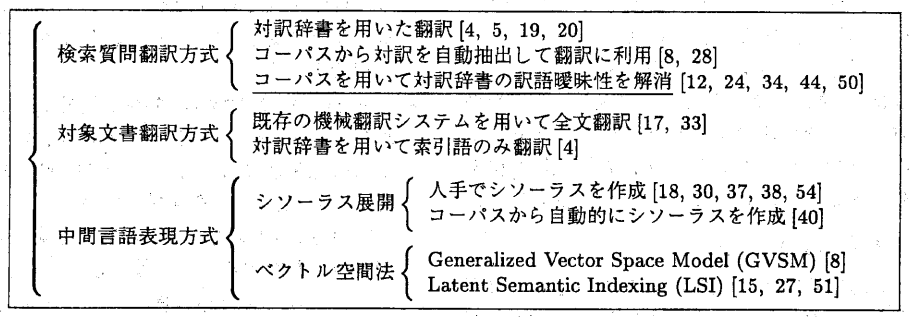

図 1: 従来の CLIR 検索方式の分類 (下線は本論文の提案手法が属するカテゴリを表す)

る翻訳法もこの範疇に属する。Kikui [24] と Suzuki と Hashimoto [44] は検索対象言語 (目的言語) における共起頻度を用いて訳語曖昧性を解消している[7]。Yamabana ら [50] と Okumura ら [34] は DMAX 法 [14] を用いて訳語曖昧性を解消している。DMAX 法は「comparable 二言語コーパス」[8]を用いて、原言語の単語の共起分布と最も類似するような訳語の組みを選択する翻訳方式である。Davis と Ogden [12] は、英仏 comparable コーパス[9]を用いて、以下の手順によって検索質問の翻訳を行っている。

(1) 原言語 (英語) の検索質問を用いて英語コーパスを検索する。

(2) 複数の翻訳候補 (仏語) を検索質問に用いて仏語コーパスを検索する。

(3) 英仏の検索結果が最も類似するような仏語訳を選択し、仏語の検索に利用する[10]。

対象文書翻訳方式 検索質問ではなく、検索の対象となる文書を翻訳する方式である。既存の機械翻訳システムを利用して全文を翻訳する手法 [17, 33] や索引語 (index) のみを翻訳する手法 [4][11]がある。Oard と Hackett [33] は、対象文書の全文翻訳方式は検索質問翻訳方式よりも検索性能が良いことを報告している。しかし一般に、機械翻訳システムによる全文翻訳はコストが高い。また、既存の膨大な索引データを CLIR 用に再構築する必要がある。

中間言語表現方式 この方式は、シソーラス展開による手法とベクトル空間法に細分類できる。

Salton [37, 38] や Sheridan ら [40, 42] が指摘するように、シソーラス展開による CLIR は、IR における「検索質問の拡張 (query expansion)」の一種である。すなわち、検索質問や検索対象文書を概念レベルに抽象化して言語の表層的な違いを吸収する。Salton [37, 38] は人手で作成した英独仏シソーラスを SMART システム [39] に実装し、CLIR の検索性能は IR とほぼ変わらないことを実験によって示している。International Road Research Documentation scheme (IRRD) [30] は、英独仏シソーラスを道路技術の文献検索に用いている。Gilarranz ら [18] は EuroWordNet [47] をシソーラスとして用いている。石川ら [54] は、図書分類に用いる「分類表」をシソーラスとして利用し、国ごとに異なる分類を人手で対応付けして日中書誌データ検索を実現している。以上の手法が人間の手作業に依存しているのに対して、Sheridan ら [40, 42] はシソーラス自動構築法 [35] を応用し、comparable コーパスから多言語シソーラスを自動的に作成して検索に利用している。

ベクトル空間法 (vector space model: VSM) による CLIR は、言語に依存しない軸によってベクトル空間を構成する点に特長がある。言語に依存しないベクトル空間法は IR の研究で提案されており、一般化 VSM (GVSM) [48] や Latent Semantic Indexing (LSI) [13] がある。Carbonell ら [8] は GVSM を、Dumais ら [15]、Landauer と Littman [27]、Young [51] は LSI を CLIR に応用している。いずれの手法も comparable コーパスが必要である。

2.2 検索結果の提示方法

CLIR の検索文書はユーザの母国語以外の言語で記述されており、ユーザが必ずしも内容を (素早く) 理解できるとは限らない。そこで、検索結果の提示方法は通常の IR 以上に重要な課題である。膨大な数の検索文書をそのまま提示するだけのナイーブな手

[7] 目的言語の共起頻度に基づく機械翻訳は、Dagan ら [11] によって提案されている。

[8] 同一の内容について複数の言語で記述したコーパスを指す。言い換えれば、記事単位の対応が付いた二言語コーパスである。

[9] Davis と Ogden [12] は "parallel corpus" と呼んでいる。

[10] この手法では、comparable コーパスは検索前の翻訳にのみ利用し、実際の検索対象となるデータベースとは別のものである点に注意する必要がある。

[11] 検索質問の翻訳と同じ手法 [4] によって索引語を翻訳する。

法では、ユーザの負担が大きくなり好ましくない。

Aone ら [4] は、検索文書中のキーワード (重要語)[12]のみを翻訳して提示している。Resnik [36] や鈴木ら [52] は、(優れた翻訳法でなくても) 重要語を翻訳した方が、しない場合よりもユーザの検索効率が向上することを報告している。文書要約技術が検索結果閲覧に及ぼす効果などは今後の研究課題である。

2.3 CLIR の評価方法

従来の CLIR 評価法は、通常の IR の評価とほぼ同じである。すなわち、あらかじめ用意した検索質問を用いて検索を行い、その結果に対して適合率 (precision) と再現率 (recall) で評価する。データには既存の IR 用テストコレクション (例えば TREC) を人手で翻訳して用いることが多い。Carbonell ら [8] は、人間による正解文書判定のコストを避けるために、単一言語検索の検索結果を正解とする自動評価法を提案している。これは「単一言語検索の性能をどれだけ保持できるか」という観点に基づく評価である。

2.2 節でも述べたように、CLIR では大量の文書よりも、ノイズを含まない少ない文書をユーザに提示することがより好ましい。そこで、CLIR の評価では再現率よりも適合率が重視されることもある [42]。そして、検索結果の提示方法やユーザインタラクション (relevance feedback [39] など) の効果を評価することも重要である。ユーザが正解文書を検索するまでの時間に基づく評価 [41, 52] はその一つである。

3 提案 CLIR システム概要

本研究で開発している英日/日英 CLIR システムの構成を図 2 に示す。本システムは「検索質問翻訳方式」(2.1 節参照) に基づいている[13]。システムは以下の 3 つの構成要素からなる。

- translator: ユーザ (user) の検索質問 (query) を語基対訳辞書 (bilingual dictionary) を用いて検索対象言語に翻訳し、翻訳候補 Q_i を出力する。検索に利用する翻訳候補数はユーザオプションとして与える。現在は、検索質問として複合語の専門用語のみを想定している。
- IR engine: Q_i を用いてデータベース (database) から文書 D_i を検索する。ここで、D_i は Q_i によって検索された文書集合を表す。今回の実験 (5 節参照) では単純なキーワードマッチングを用いている。「〜を買いたい/利用したい」のような検索質問を扱うための、より高度な検索機構を我々は提案しており [9]、今後本システムとの統合を行う。
- browser: 検索文書 D_i を翻訳候補 Q_i ごとに整理して「summary」をユーザに提示する。現在は TF·IDF 法 [39] による重要語抽出法を用いた summary 生成などを検討している。さらに、本論文の翻訳法を応用して、抽出した重要語をユーザ言語へ翻訳することも検討している。

本システムのもう一つの特長は、ユーザが検索結果を閲覧しながらシステム (translator) に feedback を返すことができる点にある (図 2 の破線部分)。これについては 6 節で説明する。

以下 4 節と 5 節では、translator 部分の実装方法及び評価実験について説明する。残り二つの構成要素の実装と評価は今後の研究課題である。

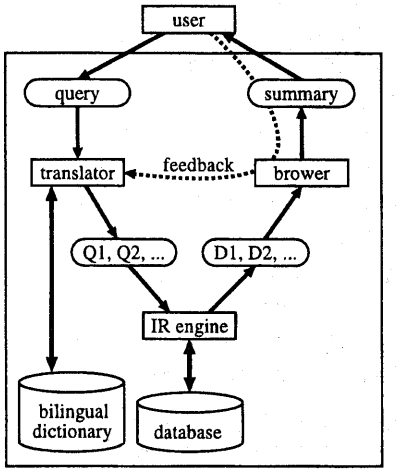

図 2: CLIR システム構成

4 複合語翻訳法

図 2 の translator は、対訳辞書を用いて入力された複合語を語基レベルで翻訳する。語基の語順は変更しない[14]。尚、本論文では英日翻訳についてのみ説明する (日英翻訳は今後対応する)。以下、複合語翻訳に必要な 3 つの要素技術についてそれぞれ説明する。

4.1 語基対訳辞書の作成

通常の対訳辞書は、専門的な用語を記述していなかったり、人間向けの説明文を含んでいるため、本

[12]従来の手法 [4, 52] は検索文書中の頻出語を重要語とする。

[13]「対象文書翻訳方式」「中間言語表現方式」との一般的な優劣については、本論文では議論しない。

[14]石川ら [54] によれば、既存の対訳辞書 [16] に登録されている複合見出し語の約 95%は原言語と目的言語で語基の語順が同じである。また、複合語のうち約 84%は 2 語基から構成されている。

システム用の語基対訳辞書には向いていない。そこで、EDR 日英専門用語対訳辞書 [55] (情報処理分野の専門用語 12 万語を収録) から語基対訳辞書を作成した。日本語には語基の区切りがないため、英語語基との対応付けが困難である。しかも、日本語分割の難しさは語基数の増加とともに顕著になる。そこで、2 語基から構成される英単語とその日本語対訳 59,533 対のみを抽出し、字種情報などを用いた簡単なヒューリスティックによって日本語を 2 語基に分割して (語順を保持したまま) 英語語基との対応付けを行った[15]。その結果、英語見出し数 6894、平均訳語数 2.1 の英日語基対訳辞書を作成した[16]。図 3 に日本語分割後の英日複合語対訳辞書の例を示す。図 4 は、図 3 から作成した英日語基対訳辞書である。作成した語基対訳辞書は「メモリ/メモリー」のような表記のゆれも含んでいる。

英語複合語	日本語複合語
CCD memory	CCD メモリー
IC memory	IC メモリ
associative learning	相関 学習
associative memory	連想 メモリ
associative record	結合 レコード
correlation function	相関 関数
error detection	誤り 検出
factor correlation	因子 相関
hybrid IC	ハイブリッド 集積回路

図 3: 英日複合語対訳辞書の例 (2 語基のみ)

英語語基	日本語語基
CCD	CCD
IC	IC, 集積回路
associative	相関, 連想, 結合
correlation	相関
detection	検出
error	誤り
factor	因子
function	関数
hybrid	ハイブリッド
learning	学習
memory	メモリ, メモリー
record	レコード

図 4: 英日語基対訳辞書の例

4.2 統計的翻訳モデル

翻訳手法には、品詞タグ付け (part-of-speech tagging) [10] や機械翻訳 [6] で用いられている統計的手法を用いた。まず英語検索質問 E と、その日本語翻訳候補 (の一つ) である J を以下のように定義する。

$$E = e_1, e_2, \ldots, e_n; \quad J = j_1, j_2, \ldots, j_n$$

ここで e_i と j_i は i 番目の語基を表す。翻訳のタスクは、$P(J|E)$ を最大化する J を選択することであり、ベイズ則によって式 (1) のように表現できる[17]。

$$\arg\max_J P(J|E) = \arg\max_J P(E|J) \cdot P(J) \quad (1)$$

さらに、$P(E|J)$ と $P(J)$ を式 (2) で近似する。

$$\begin{aligned} P(E|J) &\approx \prod_{i=1}^{n} P(e_i|j_i) \\ P(J) &\approx \prod_{i=1}^{n-1} P(j_{i+1}|j_i) \end{aligned} \quad (2)$$

4.3 確率値の推定

ここでは、式 (2) の各項の推定方法について説明する。$P(e_i|j_i)$ は、図 3 の日本語分割後の英日複合語対訳辞書を用いて推定する[18]。図 3 では、例えば「相関」は「associative」と 1 回、「correlation」と 2 回対応している。すなわち、式 (3) が成り立つ。

$$\begin{aligned} P(\text{associative} \mid \text{相関}) &= 1/3 \\ P(\text{correlation} \mid \text{相関}) &= 2/3 \end{aligned} \quad (3)$$

$P(j_{i+1}|j_i)$ は、図 3 で分割した日本語語基間の共起頻度を用いて推定する。また、EDR 日本語共起辞書 [55][19]から抽出した「名詞-名詞」「名詞-動詞」共起頻度も利用した。日本語の複合名詞には動詞性名詞とその格要素で構成されるものがあるため [26]、「名詞-動詞」共起の利用は妥当であると考える。例えば「誤り (を) 検出する」という共起から、図 3 の「誤り検出」を構成できる。EDR 共起辞書は共起語の間の関係子 (agent, object など) を定義している。今回は恣意的に「object」関係子を含む共起のみを利用した。さらに、データスパースネスを避けるために、日本語語基 (j_i) は分類語彙表 [53] の分類コード上位 5 桁を用いて意味クラスに抽象化する[20]。

5 複合語翻訳法の評価実験

4 節の翻訳法の有効性を評価するために、英日 CLIR の実験を以下の手順で行った。

[15] Tsuji と Kageura [46] は複合語対訳辞書から HMM を用いて語基対訳辞書を作成している。それに対して、我々の手法は計算コストが安いという利点がある。両手法のさらなる比較検討は今後の研究課題である。

[16] 同様の手法によって日英語基対訳辞書も作成可能である。

[17] 複数の翻訳候補を許容する場合は、$P(J|E)$ の値の大きい J から順に検索質問として利用する。

[18] 本来ならば、単語対応二言語コーパスの利用が理想である。しかし、現在の所このような言語資源は入手が困難である。

[19] 新聞記事から抽出した語の共起 1,140,000 エントリを収録している。

[20] 分類語彙表に未登録の語基は確率値の推定には使用しない。

(1) 英語の検索質問を日本語に翻訳する。EDR 日英専門用語対訳辞書の 2 語基複合語からランダムに抽出した 1000 語を検索質問に利用した。残りの 2 語基複合語は確率値の推定に利用した。

(2) 翻訳した検索質問を用いて日本語文書を検索する。今回は、既存の「goo 検索エンジン」[21]を用いてインターネット上の文書を検索した[22]。

(3) 検索文書と正解文書を比較し、適合率と再現率によって CLIR システムの検索性能を評価する。適合率と再現率は、式 (4) で計算する。

$$\begin{aligned} \text{適合率} &= \frac{\text{検索できた正解文書数}}{\text{検索文書数}} \\ \text{再現率} &= \frac{\text{検索できた正解文書数}}{\text{正解文書数}} \end{aligned} \quad (4)$$

正解文書とは英語検索質問の正しい日本語訳を含む文書であり、正しい日本語訳とは EDR 日英専門用語対訳辞書に定義されている対訳である。例えば図 3 では、「associative learning」の正しい訳は「相関学習」である。そこで、正解文書の判定は全自動で行うことができる。比較した手法を以下に示す。

(1) 全ての訳語候補を利用 (再現率は常に最大)

(2) ランダムに選択した k 個の訳語を利用

(3) 検索文書が多い訳語を利用 (目的言語における共起頻度を利用する手法 [24, 44] と類似する)

(4) 本手法 (共起頻度は対訳辞書からのみ抽出)

(5) 本手法 (名詞-名詞、名詞-動詞共起も利用)

手法 (4) と (5) に用いた共起情報 (のべ) はそれぞれ 8020 と 651,885 である[23]。検索質問あたりの平均訳語候補数は 12、平均検索文書数は 535 であった。各手法の適合率と再現率を表 1 に示す。手法 (3)-(5) では、上位 k 個の訳語候補を検索質問として用いる。表 1 の「optimal」欄は、正しい日本語訳の順位を k とした場合である。表 1 より、どの k の値についても、本手法 (4) は手法 (1)-(3) の適合率を向上させることを確認した。また、共起情報の追加 (手法 (5)) によって手法 (4) の適合率を高めることができた。このことから、comparable コーパスを用いることなく、単に目的言語 (日本語) のコーパスを追加することで今後さらなる検索性能の向上が期待できる。

6 ユーザとのインタラクション

CLIR システムの検索性能をさらに高めるためには、ユーザがシステムを学習できることが好ましい。

[21] http://www.goo.ne.jp/

[22] 本実験は 1998 年 7 月に行った。

[23] 分類語彙表に登録されている語基のみを使用した。

表 1: 各手法の適合率/再現率の比較

手法	適合率/再現率 (%)			
	$k=1$	$k=5$	$k=10$	optimal
(1)	39.0/100			
(2)	49.0/8.69	52.1/61.5	48.0/75.7	52.5/100
(3)	48.4/82.0	40.6/99.8	39.3/99.9	45.8/100
(4)	76.6/23.0	59.1/85.6	52.6/98.3	73.6/100
(5)	78.4/26.6	62.4/89.4	53.4/98.4	75.8/100

Yamabana ら [50] は、ワープロ仮名漢字変換のように、複数の検索質問翻訳候補からユーザが正しい訳語を選択する機能を設定している。しかし CLIR では、母国語以外の翻訳結果に対してユーザが適切な教示を行うことは困難な場合もある。

そこで我々は、翻訳結果ではなく、検索文書を用いたシステム学習法を提案する。ユーザは、検索文書中に現れる多くの単語や文脈を手掛かりにして、その文書が自分の欲しいものかどうかを (母国語以外の文書であっても) 判断できる場合が多いと考える。その上、システムは各検索文書がどの訳語候補によって検索されたのかを判断できる。その結果、ユーザは単に欲しい文書を選択するだけで、システムに間接的に正しい訳語を教示できる (図 2 の破線部分)。システムは、教示された正解訳語を (a) 複合語のまま対訳辞書に追加したり、(b) 確率値の推定 (4.3 節参照) に利用する。この手法は、広く捉えれば関連性フィードバック (relevance feedback) の一種と見なすこともできる。しかし従来の関連性フィードバック [8, 39, 40] とは異なり、本手法はユーザの教示をそれ以降 (将来) の検索に利用できる点に特長がある。

7 おわりに

本論文は、多言語情報検索 (CLIR) のための複合語翻訳法を提案し、(a) 語基対訳辞書の作成、(b) 翻訳手法、(c) 確率値の推定法について説明した。また英日 CLIR の実験を通して、本手法はその他の検索質問翻訳法よりも CLIR の検索性能を向上させることを示した。さらに、ユーザがインタラクティブにシステムを学習させる枠組を提案した。本論文で提案した複合語翻訳法は、既存の CLIR システムに「複合語翻訳モジュール」などの形で導入できる。また検索文書中の重要語の翻訳にも利用可能である。

今後の研究課題としては、まず browser 部分の実装がある。次に、英語シソーラス (WordNet [29] など) を用いた日英翻訳の実装がある。また、コーパス (新聞記事や論文など) から対訳を抽出する手法 [22, 25, 43] の利用を検討している。語基の抽象化に

用いた分類語彙表は登録語数が限られているので、専門用語をシソーラスに追加する作業 [45] も今後の課題である。最後に、本複合語翻訳法を CLIR 以外の多言語アプリケーション (Cross-Language Information Extraction [23] など) に応用することも検討している。

謝辞

小林義行氏 (日立中央研究所) には本研究に対して有益なコメントを頂きました。感謝致します。

参考文献

[1] *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996. `http://www.rxrc.xerox.com/research/mltt/DMHead/CLIR/SIGIR96CLIR.html`.

[2] *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996-1997.

[3] *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997. `http://www.clis.umd.edu/dlrg/filter/sss/papers/`.

[4] Chinatsu Aone, Nicholas Charocopos, and James Gorlinsky. An intelligent multilingual information browsing and retrieval system using information extraction. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 332–339, 1997.

[5] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 84–91, 1997.

[6] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[7] Ralf D. Brown. Automated dictionary extraction for "knowledge-free" example-based translation. In *Proceedings of the 7th Internatinal Conference on Theoretical and Methodological Issues in Machine Translation*, 1997.

[8] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artficial Intelligence*, pp. 708–714, 1997.

[9] Youlee Chun, Mamoru Ishima, Atsushi Fujii, and Tetsuya Ishikawa. A utility-based information retrieval system for user information usage -UBIR system-. In *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Languages*, 1998. (To appear).

[10] Kenneth W. Church and Robert L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, Vol. 19, No. 1, pp. 1–24, 1993.

[11] Ido Dagan, Alon Itai, and Ulrike Schwall. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130–137, 1991.

[12] Mark W. Davis and William C. Ogden. QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–98, 1997.

[13] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[14] Shinichi Doi and Kazunori Muraki. Translation ambiguity resolution based on text corpora of source and target languages. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 525–531, 1992.

[15] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[16] Gene Ferber. *English-Japanese, Japanese-English Dictionary of Computer and Data-Processing Terms*. MIT Press, 1989.

[17] Denis A. Gachot, Elke Lange, and Jin Yang. The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[18] Julio Gilarranz, Julio Gonzalo, and Felisa Verdejo. An approach to conceptual text retrieval using the EuroWordNet multilingual semantic database. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[19] David A. Hull. Using structured queries for disambiguation in cross-language information retrieval. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[20] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–57, 1996.

[21] Gareth J. F. Jones and David A. James. A critical review of state-of-the-art technologies for cross-language speech retrieval. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[22] Hiroyuki Kaji and Toshiko Aizono. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 23–28, 1996.

[23] Megumi Kameyama. Information extraction across linguistic barriers. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[24] Genichiro Kikui. Term-list translation using monolingual word co-occurrnce vectors. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, 1998.

[25] Judith Klavans and Evelyne Tzoukermann. Combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation*, Vol. 10, No. 3, pp. 185–218, 1995.

[26] Yoshiyuki Kobayashi, Takenobu Tokunaga, and Hozumi Tanaka. Analysis of syntactic structure of Japanese compound noun. In *Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium*, pp. 326–331, 1995.

[27] Thomas K. Landauer and Michael L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the 11th International Conference: Expert Systems and Their Applications*, pp. 77–85, 1991.

[28] Jae Sung Lee and Key-Sun Choi. A statistical method to generate various foreign word transliterations in multilingual information retrieval system. In *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages*, pp. 123–128, 1997.

[29] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller, and Randee Tengi. Five papers on WordNet. Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University, 1993.

[30] P. E. Mongar. International co-operation in abstracting services for road engineering. *The Information Scientist*, Vol. 3, pp. 51–62, 1969.

[31] Douglas W. Oard. Alternative approaches for cross-language text retrieval. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[32] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, 1996.

[33] Douglas W. Oard and Paul Hackett. Document translation for cross-language text retrieval at the University of Maryland. In *The 6th Text Retrieval Evaluation Conference (TREC-6)*, 1997.

[34] Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. Translingual information retrieval by a bilingual dictionary and comparable corpus. In *LREC workshop on translingual information management: current levels and future abilities*, 1998.

[35] Y. Qiu and H. Frei. Concept based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 160–169, 1993.

[36] Philip Resnik. Evaluating multilingual gisting of Web pages. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[37] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, Vol. 21, No. 3, pp. 187–194, 1970.

[38] Gerard Salton. Experiments in multi-lingual information retrieval. Technical Report TR 72-154, Computer Science Department, Cornell University, 1972.

[39] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[40] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58–65, 1996.

[41] Páraic Sheridan, Jean Paul Ballerini, and Peter Schäuble. Building a large multilingual test collection from comparable news documents. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[42] Páraic Sheridan, Martin Wechsler, and Peter Schäuble. Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, 1997.

[43] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, Vol. 22, No. 1, pp. 1–38, 1996.

[44] Masami Suzuki and Kazuo Hashimoto. Enhancing source text for WWW distribution – prototyping a cross-lingual information links server –. In *Proceedings of the International Workshop on Information Retrieval with Oriental Languages*, pp. 51–56, 1996.

[45] Takanobu Tokunaga, Atsushi Fujii, Makoto Iwayama, Naoyuki Sakurai, and Hozumi Tanaka. Extending a thesaurus by classifying words. In *Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pp. 16–21, 1997.

[46] Keita Tsuji and Kyo Kageura. An HMM-based method for segmenting Japanese terms and keywords based on domain-specific bilingual corpora. In *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium*, pp. 557–560, 1997.

[47] Piek Vossen, Pedro Diez-Orzas, and Wim Peters. Multilingual design of EuroWordNet. In *Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, pp. 1–8, 1997.

[48] S.K.M. Wong, W. Siarko, and P.C.N. Wong. Generalized vector space model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 18–25, 1985.

[49] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11, 1996.

[50] Kiyoshi Yamabana, Kazunori Muraki, Shinichi Doi, and Shin'ichiro Kamei. A language conversion front-end for cross-linguistic information retrieval. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[51] Paul G. Young. Cross-language information retrieval using latent semantic indexing. Technical Report CS-94-259, University of Tennessee, Knoxville, 1994.

[52] 鈴木雅美, 井ノ上直己, 橋本和夫. クロスリンガル情報検索結果の閲覧支援のための主要キーワード対訳表示の効果. 情報処理学会 自然言語処理研究会, Vol. 98, No. 63, pp. 99–106, 1998.

[53] 国立国語研究所（編）. 分類語彙表. 秀英出版, 1964.

[54] 石川徹也, 河手太士, 石間衛. Common indexing/retrieval language としての「分類表」資源の活用. 電子情報通信学会自然言語処理シンポジウム「実用的な自然言語処理に向けて」, 1997.

[55] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1995.