コンパラブルコーパスと対訳辞書による 日英クロス言語検索

奥村 明俊[†] 石川 開^{††} 佐藤 研治[†]

クロス言語検索手法 GDMAX は、日本語入力から英語ドキュメントの検索を可能にする。GDMAX は、対訳辞書によって入力キューリから翻訳キューリ候補を生成し、キューリからそれぞれの言語のコーパスにおけるキューリタームの共起頻度を成分とする共起頻度ベクトルを生成する。入力共起頻度ベクトルと翻訳共起頻度ベクトルとの距離によって、翻訳キューリ候補をランキングし、上位の英語キューリ集合を検索キューリとする。この手法によって、一つの対訳だけでなく適切な複数の訳語集合を英語キューリとして得ることができる。ウォールストリートジャーナルや AP 通信など 2 ギガの英語ドキュメントについて適合率と再現率で評価したところ、理想訳と比べて約 62% の精度を得て、対訳辞書のすべての訳語候補を用いる場合と比べて 12%、機械翻訳による訳語選択と比べて 6% 高い精度を得ることができた。

キーワード: クロス言語検索, コーパス, キューリ翻訳, 情報検索

Japanese-English Cross Language Information Retrieval based on Comparable Corpora and Bilingual Dictionary

AKITOSHI OKUMURA [†], KAI ISHIKAWA [†] and KENJI SATOH [†]

This paper proposes a method to translate query terms for cross-language information retrieval (CLIR). CLIR is generally performed by query translation and information retrieval (IR). CLIR is less precise than IR because of query term translation ambiguities, especially in Japanese and English CLIR. We developed Double MAXimize criteria based on comparable corpora (DMAX), which is an equivalent translation selection method for machine translation (MT), by using term co-occurrence frequency in comparable corpora. Though a term should be translated into one word for MT, a query term should be translated into several appropriate terms for CLIR. This paper describes a generalized query term selection model, the GDMAX for CLIR. In this model, a source query is represented in the vector form of the term co-occurrence frequency in source corpora. Translation queries are searched by vector similarity calculation between a source query and a target query represented by the co-occurrence frequency in comparable target corpora. GDMAX was evaluated by using TREC6 (Text Retrieval Conference) English data and 15 Japanese queries. GDMAX queries had approximately 62% accuracy of human queries, and 6% higher accuracy than machine translation queries and 12% higher accuracy than bilingual dictionary-based queries.

KeyWords: *Cross Language Information Retrieval, Comparable Corpora, Query Generation*

[†] 日本電気 (株) C&C メディア研究所, NEC Corp., C&C Media Research Laboratories

^{††} ATR 音声翻訳通信研究所, ATR Interpreting Telecommunications, NEC 在籍中の研究成果

1 はじめに

WWWの普及とともに多言語情報検索、とりわけ、クロス言語検索 (cross language information retrieval, CLIR) に対するニーズが高まっている。CLIRによって、例えば、日本語の検索要求 (クエリ) によって英語ドキュメントの検索が可能となる。CLIRは、クエリもしくは検索対象となるドキュメントの翻訳が必要となるので、IRよりも複雑な処理が必要となる (Hull 1997)。

CLIRの多くは、クエリを翻訳した後、情報検索を行なう。クエリの各タームには、訳語としての曖昧性が存在するため、CLIRの精度は単言語でのIRよりも低い。特に日英間では、機械翻訳の訳語選択と同様に、対訳の訳語候補が多いので困難である (Yamabana, Muraki, Doi, and Kamei 1996)。機械翻訳の訳語選択手法として、コンパラブルコーパスでの単語の文内共起頻度に基づいた Double MAXimize (DMAX) 法が提案されている (Yamabana et al. 1996; Doi and Muraki 1992, 1993; Muraki, Akamine, Satoh, and Ando 1994)。DMAX法は、ソース言語コーパスにおいて最大の共起頻度を持つ2つの単語に着目し、その2つの単語の訳語候補が複数ある場合、正しい訳語は、コンパラブルなコーパスにおいても最大の共起頻度を有するという事実に基づいた訳語選択手法である。機械翻訳においては、一つの単語は一意に訳されるべきであるが、CLIRにおいては、クエリのタームは適切な複数のタームに訳される方が精度良く検索できることもある。シソーラスや他のデータベースによって適切に展開されたクエリのタームは良い検索結果を導くことが報告されている (Harman 1996; Satoh, Akamine, and Okumura 1996)。

CLIRにおけるクエリタームの訳語選択の問題を解決するために、DMAX法を一般化したGDMAX法を提案する。GDMAX法では、コンパラブルコーパスを用いてクエリタームの共起頻度を成分とする共起頻度ベクトルを生成し、入力クエリと翻訳クエリの類似度をベクトルとして計算して類似性の高い翻訳クエリを選択する。

本報告では、まず、CLIRにおけるクエリの翻訳の課題について説明し、次に、GDMAX法によるクエリタームの翻訳・生成について説明する。GDMAX法に関して、TREC6 (Text Retrieval Conference) の50万件のドキュメントと15の日本語クエリを用いて実験したので報告する (Harman 1996)。

2 クエリタームの訳語選択

クエリタームの翻訳では、検索精度が向上するよう適切な訳語集合を得ることが課題である。一般には、表1に示すように、ソースクエリのターム j_i には、ターゲットクエリのタームとして対訳辞書から $e_{i1}, \dots, e_{ik}, \dots, e_{ir}$ の候補を得る。それぞれのソースクエリのタームに対して適切な訳語集合を得なければならない。

クエリタームの翻訳には、パラレルコーパスやコンパラブルコーパスを用いるコーパス

表 1 キュエリタームの選択

Source query terms	Target query terms
j_1	e_{11} e_{12} ... e_{1p}
j_2	e_{21} e_{22} ... e_{2q}
...	...
j_i	e_{i1} ... e_{ik} ... e_{ir}
...	...
j_n	e_{n1} e_{n2} ... e_{nm}

ベースの手法, 機械翻訳を含めて対訳辞書を用いる辞書ベースの手法, コーパスベースと辞書ベースを統合したハイブリッドな手法, の3種類がある (Hull 1997). パラレルコーパスは, 異なる言語の文書セットで, 対訳関係が保証されたものであって, パラレルコーパスによって, キュエリの選択を高い精度で行なうことができるが, 一般にはなかなか整備されていないので, 適用ドメインが限定される. コンパラブルコーパスは, 対訳コーパスほど異言語間で文書の対応が保証されていないが, 記述内容の分野的同一性が保証されたものであり, 同一の概念や表現がコーパス中に含まれる. コンパラブルコーパスは, 収集しやすいが, 関連性の低いタームもキュエリの中に含みがちである. また, 辞書ベースの手法では, ドメイン毎に適切な訳語を柔軟に得ることが困難である. こういった問題を解決するためにハイブリッドな手法が研究されている. GDMAX 法は, 対訳辞書とコンパラブルコーパスを用いるハイブリッドな手法で, 対訳辞書から得られた訳語候補の中から適切なものをコンパラブルコーパスを用いて抽出する.

3 GDMAX 訳語選択法

GDMAX 法は, DMAX 法と同様に, 対訳辞書およびコンパラブルコーパスにおけるタームの文内共起頻度を利用する. DMAX 法は, ソース言語コーパスにおいて最大の共起頻度を持つ2つの単語に着目し, その2つの単語の複数の訳語候補の中から, コンパラブルコーパスにおける共起頻度を用いて訳語を選択する. DMAX のアルゴリズムは, 以下の通りである (Doi and Muraki 1993).

- (1) コンパラブルコーパスにおけるタームの1文内の共起頻度をカウントする.

$j_i (i = 1 \cdots n)$: ソース言語のターム

$e_{ij} (j = 1 \cdots m_i)$: j_i に対する m_i 個のターゲット言語の訳語

$f(j_a, j_b)$: ソース言語でのターム j_a, j_b の共起頻度

$f(e_{ai}, e_{bj})$: ターゲット言語のターム e_{ai}, e_{bj} の共起頻度

- (2) $\max_{p,q} f(j_p, j_q)$ となる j_p, j_q を選択する.

この段階では, j_p と j_q のターゲット言語の訳語が決定されていない.

- (3) $\max_{r,s} f(e_{pr}, e_{qs})$ となる e_{pr}, e_{qs} を選択する.

- (4) j_p のターゲット言語の訳語を e_{pr} に決定する.

- (5) j_q のターゲット言語の訳語を e_{qs} に決定する。
 (6) すべての $j_a (a = 1 \dots n)$ のターゲット言語の訳語が決まるまで、2-5 のステップを繰り返す。

DMAX 法は、機械翻訳における訳語選択のために開発されたもので、最大の共起頻度をもつ訳語のペアに着目して、決定的に訳語のタームを探索していく。一方、GDMAX 法は、CLIR のために訳語集合を得るもので、ターゲット言語のタームをランキングするために、すべてのタームのペアの文書内共起頻度を考慮しながら探索する。共起頻度データがスパースな場合、ひとつの候補を選択できても、すべての候補をランキングすることは困難である。そこで、GDMAX 法では、文内共起頻度よりも文書内共起頻度を用いる。GDMAX 法では、ソース言語のキューリタームとターゲット言語のキューリタームからコンパラブルなそれぞれの言語コーパスを用いて、各タームの共起頻度を成分とするベクトル、共起頻度ベクトルを生成する。

例えば、 n タームからなる日本語キューリからは、共起頻度ベクトルの列である $\mathbf{F}_{jap}$ が生成される。

$$\mathbf{F}_{jap} = (\mathbf{f}_j^1, \mathbf{f}_j^2, \dots, \mathbf{f}_j^n)$$

$\mathbf{f}_j^p$ は、 nC_p 次元のベクトルで、同一文中の p 個のタームの共起頻度である。つまり、GDMAX 法は、キューリを表現するために、ベクトル空間法におけるタームの代わりにタームの共起頻度を成分として用いる。以下の式において、 $\mathbf{f}_j^2$ は、任意の 2 つのタームが 1 文書中で共起する頻度から構成される。ここで、 $f(j_i, j_j)$ は、日本語コーパスにおけるターム j_i と j_j の共起頻度を正規化した値である。

$$\mathbf{f}_j^2 = (f(j_1, j_2), f(j_1, j_3), \dots, f(j_{n-1}, j_n))$$

同様に、英語翻訳キューリの共起頻度ベクトル列 $\mathbf{F}_{eng}$ は、以下のように表現される。

$$\mathbf{F}_{eng} = (\mathbf{f}_e^1, \mathbf{f}_e^2, \dots, \mathbf{f}_e^n)$$

表 1 に示すような訳語候補が存在する場合、 $\mathbf{F}_{eng}$ として、 $p * q * r * \dots * m$ 通りの可能性が存在する。

$\mathbf{F}_{jap}$ と $\mathbf{F}_{eng}$ の類似性、 $\text{Sim}(\mathbf{F}_{jap}, \mathbf{F}_{eng})$ は、各成分の類似性、 $\text{Sim}(\mathbf{f}_j^1, \mathbf{f}_e^1), \text{Sim}(\mathbf{f}_j^2, \mathbf{f}_e^2), \dots, \text{Sim}(\mathbf{f}_j^n, \mathbf{f}_e^n)$ の関数と考えられる。ここでは、類似性 $\text{Sim}(\mathbf{f}_j^p, \mathbf{f}_e^p)$ を以下のようにコサインによって定義する。

$$\text{Sim}(\mathbf{f}_j^p, \mathbf{f}_e^p) = \frac{(\mathbf{f}_j^p, \mathbf{f}_e^p)}{|\mathbf{f}_j^p| |\mathbf{f}_e^p|}$$

実際には、データスパースネスの問題もあって、3 ターム以上の共起頻度は、2 タームの共起頻度と比べて無視できるほど小さくなることが多い。また、1 タームの出現頻度は、非常に大き

く、日英の単語がカバーする意味の違いを考えた場合、類似性における 1 タームの出現頻度の影響は抑制されるべきである。例えば、「米」というタームには、rice と USA などの意味があり、「米」の出現頻度が日本語コーパスにおいて多いからといって、訳語の一つである rice がコンパラブルな英語コーパスにおいて出現頻度が多いとは限らない。そこで、ここでは、2 タームの共起頻度のみ着目してモデルを簡略化する。つまり、 $\mathbf{F}_{jap}$ と $\mathbf{F}_{eng}$ は、 $\mathbf{f}_j^2$ と $\mathbf{f}_e^2$ を用いて類似性を照合する。類似性は、以下に示すように、 $\mathbf{F}_{jap}$ とその訳語候補である $\mathbf{F}_{eng}$ の内積を計算しコサインによって照合する。

$$\begin{aligned}\text{Sim}(\mathbf{F}_{jap}, \mathbf{F}_{eng}) &= \text{Sim}(\mathbf{f}_j^2, \mathbf{f}_e^2) \\ &= \frac{(\mathbf{f}_j^2, \mathbf{f}_e^2)}{|\mathbf{f}_j^2| |\mathbf{f}_e^2|}\end{aligned}$$

訳語候補は、ベクトルの距離の近いものから順にランキングされる。

例えば、日本語クエリが 3 つのターム、 j_1, j_2, j_3 で表現される場合、2 つのタームの共起頻度として、 $f(j_1, j_2)$, $f(j_1, j_3)$, $f(j_2, j_3)$ の 3 つの値がある。この日本語クエリは、図 1 の示すように日本語コーパス空間の三角形で表現できる。GDMAX 法は、コンパラブルな英語コーパス空間において、 j_1, j_2, j_3 と相似に近い三角形 e_{1i}, e_{2j}, e_{3k} を探索する。ある閾値を越えた類似性をもつ訳語集合が、検索で用いられる英語訳語クエリタームとなる。

図 1 コンパラブルコーパスにおける類似性

4 実験・評価

4.1 実験方法

TREC6 のドキュメントとトピックを用いて、GDMAX 法を実験・評価する (Harman 1996; Satoh et al. 1996). TREC6 では、図 2 に示すように、トピックと呼ばれる英語検索要求文が用

```
<num> Number: 012
<dom> Domain: Environment
<title> Topic: Water Pollution
<desc> Description:
Document is about the pollution of a body of water.
<narr> Narrative:
To be relevant, a document will discuss a body of
water being polluted or at risk of being polluted and
the source of pollution. The body of water must be
identified (e.g. Lake Erie, or the water supply of
Limestone, Maine).
```

図 2 TREC 英語トピック例

```
<num> Number: 012
<dom> Domain: 環境
<title> Topic: 水質汚染
<desc> Description:
水の濃度の汚染についてである。
<narr> Narrative:
汚染されていたり、汚染にさらされる危機のある
水の濃度と汚染源について論じている。水の濃度
は確認されなければならない。
(例えば、エリー湖やメイン州のリムストーンの
水の供給)。
```

図 3 TREC 日本語トピック例

意されている。それぞれのトピックには、関連があることが人手により確認されたりレバントドキュメントと呼ばれる正解データが準備されている。今回の実験では、図 3 に示すように、英語トピックと等価な日本語トピックを 15 作成した。つまり、15 の日本語トピックに関しては、英語ドキュメントのレバントドキュメントが存在しており、この 15 の日本語トピックを入力として実験を行なう。トピックの内容は、政治、経済、科学、社会的なもので、およそ、50 から 100 のタームのクエリで表現される。検索対象となる英語ドキュメントは、ウォールストリートジャーナルや AP 通信などから抽出された約 50 万件の記事である。

訳語集合による検索の効果を確認するために、英語トピックから生成したクエリによる検索、日本語クエリの各タームの可能な英語訳語をすべて用いた検索、日本語クエリの各タームの英語訳語を一つに絞った検索との比較を行なう。図 4 に示すように、以下の 4 種類の方法で生成された英語クエリに関して、実験・評価を行なう。英語のクエリタームの重みは、それ

図 4 実験環境

それぞれのトピックに対するリレバントドキュメントを含むトレーニングデータから与える。評価用のドキュメントには、トレーニングデータは含まない。重みの与え方としては、クエリタームの全ドキュメント中の出現頻度とリレバントドキュメント中の出現頻度に基づく対数尤度比を用いて行なう (Sato et al. 1996)。検索は、ベクトル空間モデルを用いており、重み付きのベクトルとして表現されたクエリとドキュメントの内積を計算することでランキングする (Harman 1996)。

・理想訳語クエリ (理想訳):

クエリタームは、人手によって作成された英語タームである。TREC の英語トピックから約 50 のストップワードを除外し、クエリのタームとなる英単語を抽出して生成した英語クエリで、理想訳として扱う。シソーラスや人手によるクエリの拡張は行なわない。

・可能訳語クエリ (可能訳):

クエリタームは、日英対訳辞書による訳語すべてである。日本語トピックは、形態素解析によって単語単位に分割され、約 50 のストップワードを除く自立語をタームとし、日英機械翻訳システムの約 5 万語の一般辞書 (対訳辞書) を用いて、それぞれの日本語タームにリンクされた英語訳語すべてを可能訳としてクエリタームとする。

・機械翻訳クエリ (機械翻訳):

クエリタームは、日英機械翻訳システムによって選択されたものである。日本語トピック

は、約5万語の一般辞書(対訳辞書)をもつ日英機械翻訳システムによって英語に翻訳される。英語に翻訳されたトピックは、理想訳語キューリ生成と同様の手順で、英語キューリに変換される。対訳辞書は、可能訳語キューリ生成に用いたものと同一のものである。訳語を一つに絞る方法としては、機械翻訳システム以外に、代表訳語によるもの、DMAX法などコーパスを用いる方法などが考えられるが、ドメインに影響されずに安定した訳出を行なうルールベースの機械翻訳システムを用いることとした。

・GDMAX キューリ (GDMAX):

キューリタームは、GDMAX法によって日英対訳辞書から選択されたものである。共起頻度データは、図5に示すように、コンパラブルコーパスと対訳辞書を用いて準備する。日本語トピックは、形態素解析によって単語に分割され、約50のストップワードを除く自立語をタームとして、すべての2つのタームの文書内共起頻度を日本語コーパスにおいてカウントする。日本語タームは、対訳辞書を用いて訳語となり得るすべての英語タームに変換する。対訳辞書は、可能訳と機械翻訳に用いたものと同一のものである。日本語コーパスとコンパラブルな英語コーパスにおいて、英語タームについても、すべての2つの組み合わせの文書内共起頻度をカウントする。例えば、図5に示すように、日本語ターム、「危機」と「水」には、英語タームとして、それぞれ、“crisis, precipice, critical situation, pinch, imminent danger, emergency”と“water, aqua, plasma”が存在し、それぞれの組み合わせに対する共起頻度がカウントされる。日本語キューリと英語キューリから共起頻度を成分とするベクトルを生成し、ベクトルの類似度計算によって、類似した英語キューリを求める。今回、ひとつのタームあたり平均3個程の訳語となるように、実験的にベクトルの類似性の閾値を0.85とした。この閾値を越える共起頻度ベクトルをもつすべての英語キューリを類似キューリとし、類似キューリのすべてのタームから構成されるキューリを作成して検索を行なう。

4.2 日英コンパラブルコーパス

共起頻度データは、50万記事ずつの日英コンパラブルコーパスから抽出する。コーパスは、ドメインとサイズに関して、以下の点を考慮して収集した。

・ドメイン:

実験対象となるドメインが、政治、経済、社会、科学の分野であることから、これらの分野に関する新聞、雑誌記事を収集する。コンパラブルコーパスを用いることの利点の一つは、その収集しやすさであることから、ジャンルや話題の一致などの特別な調整は行わずに収集を行なう。ただし、新聞記事の時間情報に着目して、日英ともに10年未満のものとし、時間的な共有を図る。実験をオープンテストとするために、検索対象となるドキュメント(TRECドキュメント)は、コーパスの中に含まない。

図 5 共起頻度データの作成フロー

・サイズ:

コンパラブルコーパス作成のために収集した記事の平均的サイズは、日本語約 500 文字、英語約 300 単語であり、共起頻度は記事毎にカウントする。コーパスのサイズが小さく共起頻度データがスパースになると、GDMAX の有効性が低下すると予想されるので、共起頻度データが 0 となるタームの対が、極力少なくなるまでデータを収集する。データサイズは、大きければ大きいほどデータスパースネスの問題は解消されるが実験コストが大きくなる。今回のトピックを用いて実験するために、9,815 対の日本語キューリタームの共起頻度データが必要であり、英語タームは、すべての訳語候補の組合せとして、166,111 対の共起頻度データが必要であった。日英 50 万件の記事を収集した時点で、頻度 0 の日本語ターム対は 96 対、英語ターム対は 2,503 対となった。ターム対総数に対する 0 件のターム対の割合が、日英ともに 1~2% 程度であり、共起頻度 0 のターム対がほとんど増加しなくなったので、データスパースネスの問題は少ないと判断し、かつ、現実的に実験可能なサイズであることから、50 万件ずつのコンパラブルコーパスを用いて実験を行なうこととした。

4.3 実験結果

4つの方法で生成されたキューリの各タームに先に述べたように対数尤度比によって重みをつけ、ベクトル空間モデルによってドキュメントとの類似性を計算する。キューリベクトルとドキュメントベクトルとの内積によってランキングされた 1000 件のドキュメントが、各キュー

り毎に出力される。適合率/再現率は、TREC の評価方法に基づいて 4 種類のクエリによる検索結果に対して計算した (Harman 1996)。TREC では、各クエリ毎にリレバントドキュメントが用意されており、このデータをもとに、interpolated precision-recall curve を描く。これは、リレバントドキュメントの 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% を含むように上位の検索結果をとった場合の精度をプロットして曲線を描くもので、それぞれ 4 つの方法に対して、15 件のクエリの平均をプロットした結果、図 6 に示すような適合率/再現率曲線 (precision-recall curve) を得ることができた。

図 6 適合率/再現率曲線

5 考察

GDMAX 法は、理想訳に比べて再現率の各ポイントの平均で約 62% の精度を得た。また、機械翻訳による手法と対訳辞書による手法と比べて、適合率/再現率ともに高い結果を得た。特に、実際にユーザがブラウズすることが多い再現率 10% までの精度に関して、GDMAX 法は、機械翻訳による手法に比べて、約 6%、対訳辞書による手法と比べて、約 12% 高い精度を得た。

理想訳のクエリよりも低かったのは、主に 2 つの理由による。第 1 は、対訳辞書が十分な語

彙を持っていなかったことである。理想訳語が対訳辞書のエントリとして含まれていない場合と、エントリとしては存在するが、入力訳語の対訳としてのパスがない場合がある。IRにおいては、専門用語や固有名詞が精度の向上に大きく寄与する。今回、5万語の一般的な語彙に関する対訳辞書を用いたので、必ずしも専門用語や固有名詞の語彙が十分ではなかった。そのため、3つの手法によるキューリの精度がすべて低くなっている。対訳の中に的確な訳語が含まれていなかったため、GDMAX法が正しい訳語を選択できなかった。第2は、タームの重み付けの問題である。タームの重みは、それぞれのトピックに対するリレバントドキュメントから計算されて付与されているが、GDMAX法によって得られた類似性の割合も考慮すべきである。類似性が低いタームであっても類似性の高いタームと同様の重みがつけられたことが、検索精度として大きな差を生じなかった原因と考えられる。今回の実験結果を基に、以下の課題に取り組む予定である。

・対訳辞書

キューリを翻訳するための適切な訳語を辞書の中に含むように対訳辞書を改良する。改良した対訳辞書を用いて、GDMAX法によるキューリ、機械翻訳によるキューリ、可能訳語すべてのキューリ、TRECの英語トピックから作ったキューリによる、4つの検索精度を比較する予定である。

また、対訳辞書を改良した後、訳語選択能力を評価する。予め入力訳語ごとに選択されるべき訳語(集合)を明らかにしておき、選択能力を適合率と再現率で評価する。選択されるべき訳語とは、検索精度をベストにするものであるべきだが、ここでは、検索能力と訳語選択能力を切り分けるために、理想訳語およびその同義語を選択すべき訳語として評価する。

・タームの重み調整と閾値の設定

今回の実験で、GDMAX法によるキューリのタームの重みは、ベクトルとしての類似性を考慮しなかった。基本的には、タームの重みに類似度を掛け合わせて調整を図る必要がある。その時、ワードネットのようなシソーラスを用いて重み調整をコントロールする(Miller 1990)。例えば、類似性の高いベクトルの訳語タームと類似性が低いベクトルの訳語タームが、シソーラスによって同義語関係があることが保証された場合、タームの重みを調整する必要はないと考えられる。タームの重みの調整方法は、実験的に最適なものを見つけていく。

また、今回の実験では、平均選択訳語数の観点から類似度の閾値を設定したが、閾値は、検索精度が最適となるように設定されるべきものである。実験的に最適な閾値を求めていく。類似ベクトルの選択方法としては、類似度の絶対度を閾値とするのではなく、類似度の差分を用いる方法も考えられる。例えば、ランキングされたベクトルの1位と2位の差分、2位と3位の差分、…というように差分を比較して、その差分がある一定の値を超えたところを境界として選択することができる。他の統計的な手法も含めて、ランキングされたベクトルからの選択手法について検討する。

・多次元的な類似性照合

今回のモデルでは、2 タームの共起頻度、 f_j^2 に着目して、 F_{jap} と F_{eng} の類似性を照合した。厳密には、 F_{eng} の他の次元の共起頻度ベクトルも F_{jap} と照合されるべきである。例えば、次元毎の重み w_p を各次元の類似度に掛け合わせて以下のような総合的な類似性が考えられる。

$$\text{Sim}(F_{jap}, F_{eng}) = \sum_{p=1}^n w_p \text{Sim}(f_j^p, f_e^p)$$

計算が複雑になるので探索アルゴリズムを工夫する必要があるが、少なくとも、主成分分析や他の統計的手法によって、重要な次元を認定する必要がある。

・コーパスへの依存性

コンパラブルコーパスは、パラレルコーパスに比べて収集しやすく、ドメインへの依存性が少ない。さらに、本方式では、訳語を複数選択するので、訳語を一つに選択する手法よりもドメイン依存の影響は小さいと考えられる。その代わり、不要な訳語がキューリに含まれる可能性が増えるが、選択された訳語集合の個々のタームにリレバンスデータから計算された重みをつけるので、不要な訳語の検索への悪影響を抑制することができる。実験によって、すべての訳語に重みをつけたもの（可能訳キューリ）と、一つに選択したものに重みをつけたもの（機械翻訳キューリ）の双方よりも精度が上回ったので、訳語集合としての選択と重み付けは効果があったと考えられる。

しかしながら、その精度は、作成された共起頻度データの質に依存するものであり、共起頻度データの質的評価方法の確立が必要である。一つの方法として、パラレルコーパスとの比較において、コンパラブルコーパスの質を評価することが考えられる。パラレルコーパスによる共起頻度データを用いた方が、より正確に類似度の高い共起頻度ベクトルを求めることができるので、訳語集合を選択するためには好ましい。コンパラブルコーパスによる共起頻度データと、パラレルコーパスによる共起頻度データとの差異が、一つの質的評価基準となりうると考える。コンパラブルコーパスのいずれかの言語に関してパラレルコーパスを作成して、パラレルコーパスにおける共起頻度データとコンパラブルコーパスにおける共起頻度データを比較することができれば、コンパラブルコーパスとパラレルコーパスのずれを分析できるが、大量データでの実現は困難である。そこで、いくつかのドメインをサンプリングして、小規模でもパラレルコーパスとコンパラブルコーパスを作成して評価する方法が考えられる。例えば、すでにパラレルコーパスが存在するものをいくつかの異なる分野で収集して、パラレルコーパスと同じ分野のコーパスを類似文検索の手法を用いてそれぞれの言語で独立に収集することによって、コンパラブルコーパスを作成する。また、実際に収集したコンパラブルコーパスの中には、パラレルコーパスとなっているものもある。対訳辞書と統計情報を用いてテキスト照合を行なうことにより、パラレルコーパス部分を抽出することも可能だと考えられる (Utsuro, Ikeda, Yamane, Matsumoto, and Nagao 1994)。パラレルコーパス部分とコンパラブルコーパス部分を分離して、共起頻度データを作成して比較することで、コンパラブルコーパスの質の評価が可能になる。

さらに、コーパスの量に関する評価として、コーパスの量と検索精度の関係を明らかにする必要がある。現状の50万記事以上を集めるのは、必ずしも容易ではないので、量を減少させた場合の検索精度の変化を測定する予定である。その際には、コーパスに含まれる記事のドメインの割合が同じくなるように変化させることが必要と考える。

今後、上述した手法でコーパスの質と量に関する分析を行ない、コンパラブルコーパスの収集方法と質的・量的評価手法を確立していく。

6 関連研究

CLIRのシステムは利用するリソースから以下のように分類される (Hull 1997)。

・コーパスベースシステム

コーパスベースシステムは、パラレルコーパスやコンパラブルコーパスをクエリの翻訳のために用いる。LSI (Latent Semantic Indexing) は、行列の次元を縮退させる手法で、パラレルコーパスから言語に依存せずに、タームとドキュメントを表現することができる (K.Landauer and L.Littman 1990; W.Oard. 1997)。LSIは、パラレルコーパスから抽出したタームとドキュメントの対応をターム/ドキュメント間頻度マトリクスに、特異値分析法 (singular value decomposition) を適用して、次元数を縮退し新たな空間を形成するベクトルを抽出する。LSIが、トレーニングコーパス以外のドメインでどの程度有効なのかは明らかではない。ETHは、疑似的なパラレルコーパスとシソーラスを用いて類似のクエリタームを拡張しながら、ドイツ語クエリをイタリア語クエリに変換した (Sheridan and paul Ballerini. 1996)。シソーラスを用いた類似性判定では、タームの分布状況によって、ドキュメントにまたがるタームを関連つける。拡張されたイタリア語クエリタームが、イタリア語ドキュメントと照合される。この手法は、平均適合率で単言語の検索の約半分の精度を得たが、ドメイン依存の問題点がある。

また、機械翻訳のためにコンパラブルコーパスから単語レベルの訳語知識を抽出する手法がいくつか提案されている (Fung 1995, 1997; Rappo 1995; Kaji and Aizono 1996; Tanaka and Iwasaki 1996)。訳語集合抽出に活用できる手法については比較・検討し、GDMAX法の改良のために参考としていく。

・辞書ベースシステム

辞書ベースシステムは、対訳辞書によってタームやフレーズを翻訳しすべての翻訳結果を結合してクエリを生成する。SPIRITは、ターム、複合語、イディオムの辞書を用いてクエリタームを翻訳しブーリアンモデルによって検索する (Radman 1994)。この辞書ベースのシステムは、単言語検索の75-80%の精度を得たが、機械翻訳システムでは、60-65%の精度であった。この性能は、ドメイン対応辞書を作ってクエリをマニュアル編集した結果によるものであり、機械翻訳においても、ドメイン辞書を用意して比較したものである。我々も、対訳辞書をドメインに適応させた後、GDMAX法に関して同様の比較実験を行なう予定である。辞書ベースシ

テムに関して、David Hull は、自身のシステムの実験の中でキューエリの約 20%は、キューエリ翻訳の訳語の曖昧性の問題により、不適切なものになっていることを報告している (Hull 1997). 対訳辞書の改良とともに、GDMAX 法に関しても同様の分析を行なう。

・ハイブリッドシステム

ハイブリッドシステムは、キューエリ翻訳のために、辞書、コーパス、ユーザインタラクションなどを組み合わせて用いるもので、GDMAX 法は、この範疇に属する。日英の言語対の報告ではないが、以下の評価結果と GDMAX 法の結果を比較検討する予定である。マサチューセッツ大学では、翻訳前と翻訳後のフィードバックを用いる手法を提案している (Ballesteros and Croft 1996). キューエリの拡張には、キューエリタームと高頻度で共起するタームを選択する自動フィードバック手法が用いられている。キューエリ拡張は、ソース言語で翻訳前に、ターゲット言語で翻訳後に行なう。翻訳前のフィードバックは、検索結果と関係のないコーパスを用いて、翻訳後のフィードバックは、検索結果のコーパスを用いて行なわれる。はじめは、単言語検索の 40-50 % の精度だったものが、キューエリ拡張により、60-65 % まで向上した (Ballesteros and Croft 1996). ニューメキシコ州立大学は、品詞タガーを使って、同じ品詞の訳語だけを選択するようにしている。さらに、各タームの訳語候補の中から、パラレルコーパス中のアラインメントされた文に着目して、最も対応性の高い訳語を選択する。CLIR の精度は、初期翻訳では 40-50 % の精度であったが、この 2 段階の訳語絞り込みにより 70-75 % の精度となった (Davis 1997). CNR は、コンパラブルコーパスを用いた自動キューエリ翻訳の手法を提案している (Peters and Picchi 1995). ソース言語のキューエリタームは、限られた範囲内で共起する単語集合というプロファイルで表現される。ターゲット言語のキューエリタームに関しても同様のプロファイルが作成される。ソース言語のプロファイルは、対訳辞書を用いて翻訳され、ターゲット言語のプロファイルと最も類似しているものが認定される。その中で上位のターゲット言語のタームが、翻訳キューエリとして用いられる。この手法は、ソース言語キューエリとターゲット言語キューエリをプロファイルという別の形式で表現し比較する点において、GDMAX 法と類似しているが、共起する単語というインスタンスで表現している点で異なる。GDMAX 法は、共起頻度という数値と共起の次元数も複数とれるので、より一般的な記述と考えられる。CNR については、具体的な評価結果が報告されていないが、同様の実験を日英に関しても行なって比較評価していきたい。

7 おわりに

本稿では、クロス言語情報検索のためのキューエリ翻訳手法として、GDMAX 法を提案した。GDMAX 法を TREC6 の 50 万件の英語ドキュメントと 15 の日本語キューエリを用いて実験評価したところ、理想訳に比べて再現率の各ポイントの平均で約 62%の精度を得た。また、適合率/再現率において、機械翻訳を用いる方法、対訳辞書を用いる方法よりも高い精度を得た。今後は、対訳辞書を整備しシソーラスと統合するとともに、GDMAX 法の一般化によって検索精度の向

上を図っていく予定である。

謝辞

本研究を進めるにあたって、DMAX 法に関して、C&C メディア研究所音声言語 TG の土井伸一氏より、また、関連研究並びに GDMAX 法の検討にあたって、C&C メディア研究所音声言語 TG の山端潔氏より、有意義なコメントを頂きました。また、東京工業大学の田中穂積教授と徳永健伸助教授からは、論文作成にあたり大変貴重なコメントを頂きました。

参考文献

- Ballesteros, L. and Croft, B. (1996). "Dictionary-based methods for crosslingual information retrieval." In *The 7th International DEXA Conference on Database and Expert Systems Applications*.
- Davis, M. (1997). "New experiments in cross-language text retrieval at NMSU's computing research lab." In *The 5th Text Retrieval Conference (TREC-5)*.
- Doi, S. and Muraki, K. (1992). "Translation ambiguity resolution based on text corpora of source and target languages." In *The 14th Conference on Computational Linguistics (COLING '92)*, Vol. 2.
- Doi, S. and Muraki, K. (1993). "Evaluation of DMAX Criteria for Selecting Equivalent Translation Based on Dual Corpora Statistics." In *Theoretical and Methodological Issues in Machine Translation (TMI 93)*.
- Fung, P. (1995). "Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus." In *The 3rd Workshop on Very Large Corpora*.
- Fung, P. (1997). "Finding Terminology Translation from Non-Parallel Corpora." In *5th Workshop on Very Large Corpora*.
- Harman, D. (Ed.) (1996). *The 4th Text Retrieval Conference (TREC-4) NIST SP 500-236*. NIST.
- Hull, D. (1997). "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval." In *Technical Report of 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*.
- Kaji, H. and Aizono, T. (1996). "Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information." In *The 16th Conference on Computational Linguistics (COLING '96)*.
- K.Landauer, T. and L.Littman, M. (1990). "Fully automatic cross-language document retrieval using latent semantic indexing." In *the 6th Conference of UW Centre for the New OED and Text Research*.

- Miller, G. (1990). "WordNet: An On-Line Lexical Database." *International Journal of Lexicography*, **3** (4). (Special Issue).
- Muraki, K., Akamine, S., Satoh, K., and Ando, S. (1994). "TWP: How to Assist English Production on Japanese Word Processor." In *The 15th Conference on Computational Linguistics (COLING '94)*.
- Peters, C. and Picchi, E. (1995). "Capturing the comparable: a system for querying comparable text corpora." In *Analisi Statistica dei Dati Testuali (JADT)*.
- Radman, K. (1994). *Vers l'Acces Multilingue en Langage Naturel aux Bases de Donnees Textuelles*. Ph.D. thesis, Universite de Paris-Sud, Centre d'Orsay.
- Rappo, R. (1995). "Identifying Word Translations in Non-Parallel Texts." In *the 33rd Annual Meeting of ACL*.
- Satoh, K., Akamine, S., and Okumura, A. (1996). "Improvements on Query Term Expansion and Ranking Formula." In *The 4th Text Retrieval Conference (TREC-4) NIST SP 500-236*.
- Sheridan, P. and Paul Ballerini, J. (1996). "Experiments in multilingual information retrieval using the SPIDER system." In *The 19th ACM/SIGIR Conference*.
- Tanaka, K. and Iwasaki, H. (1996). "Extraction of Lexical Translations from Non-Aligned Corpora." In *The 16th Conference on Computational Linguistics (COLING '96)*.
- Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., and Nagao, M. (1994). "Bilingual Text Matching using Bilingual Dictionary and Statistics." In *The 15th Conference on Computational Linguistics (COLING '94)*, pp. 1076-1082.
- W.Oard, D. (1997). "Alignment of Spanish and English TREC topic descriptions." In *The 5th Text Retrieval Conference (TREC-5)*.
- Yamabana, K., Muraki, K., Doi, S., and Kamei, S. (1996). "A Language Conversion Front-End for Cross-Language Information Retrieval." In *Workshop on Cross-Linguistic Information Retrieval SIGIR'96*.

略歴

奥村 明俊: 1984 年京都大学工学部精密工学科卒業。1986 年同大学院工学研究科修士課程修了。同年、日本電気株式会社入社。1992 年 10 月より 1 年半南カリフォルニア大学客員研究員として DARPA 機械翻訳プロジェクトに参加。現在、C&C メディア研究所、主任研究員。自然言語処理、自動通訳システムの研究に従事。情報処理学会、人工知能学会、ACL 各会員

石川 開: 1994 年東京大学理学部物理学科卒業。1996 年同大学大学院修士課程修了。同年、日本電気株式会社入社。1997 年 ATR 音声翻訳通信研究所出向。現在、第三研究室、研究員。音声翻訳、情報検索の研究に従事。情報処理学会会員

佐藤 研治: 1989 年京都大学工学部情報工学科卒業。1991 年同大学院工学研究科修士課程修了。同年、日本電気株式会社入社。現在、C&C メディア研究所、主任。自然言語処理、情報分類の研究に従事。情報処理学会、人工知能学会、各会員

(1998 年 2 月 4 日 受付)

(1998 年 4 月 7 日 再受付)

(1998 年 5 月 18 日 再々受付)

(1998 年 7 月 10 日 採録)