コンパラブルコーパスによる単語共起頻度を用いたクロス言語検索

5Y－9

奥村 明俊　石川 開　佐藤 研治
NEC C&C メディア研究所

1　はじめに

インターネットの普及と共に、ある言語で別の言語のテキストを検索するクロス言語情報検索(CLIR)のニーズが高まっている。CLIRでは、検索要求(キュエリ)を翻訳して情報検索(IR)の手法を適用することが一般的であり、キュエリの翻訳のために、対訳辞書、シソーラス、コーパス、機械翻訳システムなどを用いる[1]。キュエリの翻訳においては、訳語選択に曖昧性が存在するために、単言語でのIRと比べて精度が落ちる。とりわけ、日本語と英語間のCLIRを行なう場合、ヨーロッパ言語間の検索と比べて、翻訳されたタームの意味が大きく異なることが多いので、キュエリの翻訳精度が検索精度に大きく影響する。CLIRにおいては、機械翻訳における訳語選択と異なり、必ずしもキュエリタームを一つに決定する必要はない。IRにおいてキュエリの拡張が有効であるように、類義語や同義語などを含めた適切な訳語の集合に変換することが有効である。当社では、機械翻訳における訳語選択のために、日英のコンパラブルコーパスを活用するDMAX訳語選択法を開発し、ユーザインタラクション機構と共に、CLIRに適用する機構を提案してきた[2; 3; 4]。本稿では、DMAXで効果があったコンパラブルコーパスにおける共起頻度情報を多次元化によって一般化し、類似ベクトル検索として訳語集合を求めるGDMAX訳語選択法を提案する。GDMAX法は、入力言語コーパスにおけるキュエリターム間の共起頻度を求め、共起頻度を成分とするベクトルによって入力キュエリを表現する。コンパラブルなコーパスにおいて訳語候補のターム間の共起頻度を求め、訳語候補のキュエリもベクトル化する。入力キュエリベクトルと類似した訳語キュエリベクトルから訳語集合を得て、翻訳キュエリとする。GDMAX法を用いて日本語キュエリから英語キュエリを作成し、TREC[5]のデータによって英語テキストを検索する実験を行なったので報告する。

2　翻訳キュエリタームの選択

CLIRのキュエリの翻訳における課題を簡単化すると、表1に示すように、検索の入力となるキュエリの各タームに対する訳語の候補から、検索に最適な訳語集合を見つけることである。

表1: キュエリタームの訳語

入力キュエリターム	翻訳キュエリターム候補
j_1	e_{11} e_{12} … e_{1p}
j_2	e_{21} e_{22} … e_{2q}
j_3	e_{31} e_{32} … e_{3r}
…	…
j_n	e_{n1} e_{n2} … e_{nm}

キュエリタームを翻訳する方法は、主に、パラレルコーパスやコンパラブルコーパスを用いて翻訳するコーパスベース手法、辞書や機械翻訳を用いて翻訳する辞書ベース手法、両者をハイブリッドに組み合わせる手法がある[1]。コーパスベースの課題としては、パラレルコーパスは十分大量に集めるのが容易ではなく適用可能なドメインが限られること、コンパラブルコーパスによる翻訳では関連性の低いタームも翻訳キュエリに含まれやすいことがある。一方、辞書ベースの手法はドメインによって訳語選択を切替えることが容易ではない。GDMAX法は、辞書とコンパラブルコーパスを用いるハイブリッドな構成によってそれぞれの欠点を補完する。まず、対訳辞書によって入力キュエリの訳語候補を抽出し、次に、コンパラブルコーパスを用いて訳語候補から翻訳キュエリタームの集合を選択する。

3　GDMAX訳語選択方法

GDMAX法は、DMAX法の実験結果に基づき、キュエリをキュエリタームの共起頻度を成分とするベクトルとして表現した場合、コンパラブルなコーパスの空間において、入力キュエリと翻訳キュエリは類似のベクトルとなると仮定している。例えば、日本語キュエリターム j_1, j_2, j_3 のそれぞれ2つのタームの日本語コーパスでの共起頻度を成分とするベクトルは、図1の日本語コーパス空間における三角形に置き換えることができる。GDMAX法は、j_1, j_2, j_3 の三角形と相似形となる三角形 e_{1i}, e_{2j}, e_{3k} を英語コーパス空間において見つけ、ある閾値以上の類似度をもつ英訳語を翻訳キュエリタームとする。

一般に、n個のタームからなる日本語キュエリにおいて、各ターム間の共起頻度を成分とするベクトルの列 $\mathbf{F}_{jap}$ は以下のように表現できる。

$$\mathbf{F}_{jap} = (\mathbf{f}_1, \mathbf{f}_2, .., \mathbf{f}_n)$$

GDMAX Query Translation Model for Cross-Language Information Retrieval
Akitoshi Okumura, Kai Ishikawa, Kenji Satoh
NEC C&C Media Research Laboratory

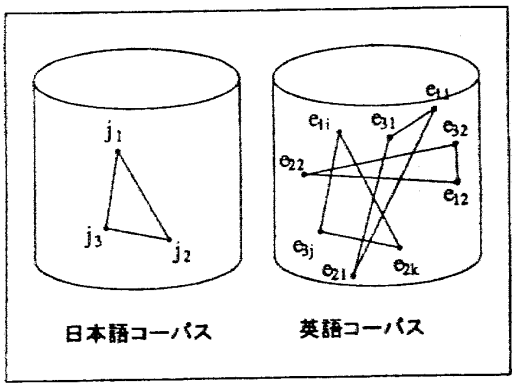

図 1: コンパラブルコーパスでの類似性

$\mathbf{f}_p$ は、任意の p 個のタームが共起する頻度を成分とする ${}_nC_p$ 次元のベクトルである。例えば、任意の 2 つのタームの共起頻度ベクトル $\mathbf{f}_2$ は、以下のようになる。$f(j_i, j_j)$ は j_i と j_j の共起頻度を正規化した値である。

$$\mathbf{f}_2 = (f(j_1, j_2), f(j_1, j_3), \ldots f(j_{n-1}, j_n))$$

同様に、英語翻訳キュエリは以下のように表現できる。

$$\mathbf{F}_{eng} = (\mathbf{e}_1, \mathbf{e}_2, .., \mathbf{e}_n)$$

表 1のような訳語候補があった場合、$p * q * r * .. * m$ 個の $\mathbf{F}_{eng}$ の候補が存在する。キュエリタームをこのようにモデル化した場合、データスパースネスの問題から 3 個以上のタームの共起頻度は小さくなり、また、ターム単独の出現頻度は、日英の単語の多義性の分布の違いから類似性が少ないと考えられる。そこで、実際に、コンパラブルコーパスにおいて、$\mathbf{F}_{jap}$ と類似の $\mathbf{F}_{eng}$ を探索する時には、2 個のタームの共起頻度を成分とするベクトル $\mathbf{f}_2$ に着目する。日英コンパラブルコーパスによって 2 個のタームの共起頻度を抽出し正規化したベクトルの内積をとることにより、類似度のランキングを行なう。

4　実験

実験のために、TREC6 の英語キュエリから日本語キュエリを 17 セット用意した。この日本語キュエリを入力キュエリとして TREC6 英語テキストデータに対して検索実験を行なう。キュエリは、100 タームほどの経済・社会的なものであり、データはウォールストリートジャーナルや AP 通信記事など約 30 万記事である。共起頻度データは、日英約 50 万件の記事から抽出した。実験で用いた英語キュエリは以下の 4 種類である。それぞれのキュエリのタームに対しては、リレバンスデータから重みを与えた[6]。

1) 等価英語キュエリ (E_{hum}):
日本語キュエリと等価な人手によるキュエリ
2) 可能英語キュエリ (E_{all}):
訳語候補すべてを含むキュエリ
3) 代表英語キュエリ (E_{rep}):
対訳辞書の代表訳語によるキュエリ
4) GDMAX キュエリ (E_{gdx}):
GDMAX による選択された訳語集合のキュエリ

ランキング上位 100 件に関する平均精度 (average precision) は、以下の通りである。

表 2: 実験結果

翻訳方法	E_{hum}	E_{all}	E_{rep}	E_{gdx}
平均平均精度	64.36	34.77	45.72	61.05

5　おわりに

CLIR におけるキュエリ翻訳手法として GDMAX 訳語選択方法を提案し、TREC データを用いて実験を行った。実験の結果、人手によるキュエリと比べてほぼ同等の精度を確認した。今後は、1 タームの出現頻度と 3 つ以上のタームの共起頻度やリレバンスによる重み付けとの統合を行なって GDMAX 法の改良を図り、インターネット上の実際の検索サービスの中で効果を検証していく。

謝辞　DMAX 法、DMAX 法の CLIR への適用および GDMAX アルゴリズムの検討に関して土井伸一氏と山端潔氏より貴重な助言を頂き、感謝致します。

参考文献

[1] David Hull, "Using Structured Queries for Disambiguation in Cross-Language Information Retrieval," Technical Report of 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.

[2] Kiyoshi Yamabana, et all, " A Language Conversion Front-End for Cross-Language Information Retrieval," Workshop on Cross-Linguistic Information Retrieval, SIGIR '96, 1996

[3] Shin-ichi Doi, et all, " Translation ambiguity resolution based on text corpora of source a nd target languages,"In Proc. of the 15th Conference on Computational Linguistics (COLING '92), vol.2, pp 525-531,1992

[4] 土井伸一,"文内共起頻度データを利用した段落内共起に基づく訳語選択," 情報処理学会第 48 回全国大会, 7Q-02,1994.

[5] Donna Harman, editor "The 4th Text Retrieval Conference (TREC-4), NIST SP 500-236, 1996.

[6] 佐藤研治, 他 "単語共起によるクエリー展開を用いた大規模テキスト検索," 情報処理学会第 52 回全国大会, 5P-04,1996.