コンパラブルコーパスによるキュエリタームの拡張とクロス言語検索 *

2P－5

中澤 聡　奥村 明俊　佐藤 研治 †　落合 尚良 ‡

NEC C&C メディア研究所 §　NEC 情報システムズ オープン技術システム事業部 ¶

1　はじめに

近年 WWW などの普及にともない、ある言語 (ソース言語) で書かれた検索要求 (キュエリ) から、別の言語 (ターゲット言語) で書かれた文書を検索するクロス言語検索 (CLIR) もニーズが高まってきている。CLIR では、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る。この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が用いられる [1]。

表 1 に、CLIR のために、与えられたキュエリから検索キーとなるキュエリタームを抽出し、それぞれに対して対訳辞書を引いて作成された、各ソースキュエリターム j_i の訳語候補の表を示す。

表 1: キュエリタームとその訳語候補

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}

このようにキュエリタームの翻訳には訳語選択の曖昧性が存在するため、単言語での IR に比べて精度が落ちる。よって、この訳語選択にどのような手法を用いるかは、最終的な検索精度に影響する大きな問題である。

CLIR を目的とした訳語選択法の 1 つに、コンパラブルコーパスにおけるソース言語とターゲット言語それぞれの単語共起頻度を利用した GDMAX 訳語選択法がある [2]。この GDMAX 法には、

- 出力として、複数の訳語選択の順序づけが可能。
- 比較的入手が容易な対訳辞書とコンパラブルコーパスをリソースとして用いるので、適用分野を広げやすい。

といった特長があるが、一方、キュエリから抽出したタームの数が少ないと精度が悪い、可能な訳語候補の組合せ数に従って必要な計算量が増大する、といった問題もある。

本稿では、この GDMAX 法を基に、キュエリタームエクスパンションと段階的な訳語選択を行う CLIR 手法を提案する。

2　キュエリタームエクスパンション

本提案手法では、まずソース言語においてキュエリタームエクスパンションを行い、ついで元のキュエリタームと拡張されたタームの訳語選択を実行するという手順を取る。ソース言語上でキュエリタームエクスパンションを行うことにより、キュエリから抽出されたターム数が少ない場合、訳語選択の手掛かりとする共起頻度データが不十分になるという GDMAX 法の問題を解決できるからである。本手法におけるキュエリタームエクスパンションのアルゴリズムは以下のとおり。

(1) キュエリに形態素解析などの処理を行い、キュエリタームを抽出。このときストップワードをキュエリタームから除外。

(2) 各キュエリターム j_i に対して

$$\frac{f(j_i, j_a)}{f(j_i) \cdot f(j_a)} \geq TETH1 \qquad (1)$$

を満たすターム j_a を全て求めて、ターム j_i の拡張候補タームリストとする。ただし $f(j_i, j_a)$ はターム j_i と j_a の共起頻度、$f(j_i)$ と $f(j_a)$ はそれぞれターム j_i、ターム j_a 単独での出現頻度、$TETH1$ は任意の閾値とする。

---

*Query Term Expansion and Cross Language Information Retrieval based on Comparable Corpora

†Satoshi Nakazawa, Akitoshi Okumura, and Kenji Satoh

‡Takayoshi Ochiai

§NEC C&C Media Research Laboratories

¶NEC Information Systems, Ltd. Open Technology Systems Division

(3) 元のキュエリタームを $j_1 \cdots j_n$ としたとき、全ての拡張候補タームリストに含まれているタームに対して

$$\sum_{i=1}^{n} \frac{f(j_i, j_a)}{f(j_i) \cdot f(j_a)} \geq TETH2 \qquad (2)$$

となるようなタームを最終的に拡張するタームとして求める。$TETH2$ は任意の閾値。

このようにタームのエクスパンションは2段階に分けて、行われる。第1段階は追加するタームの候補を絞り込むためのフィルタであり、第2段階によって、元の検索キュエリに含まれている複数のキュエリタームにある程度まんべんなく共起するタームが選ばれる。

3 段階的 GDMAX

本手法では、訳語選択に基本的に GDMAX 法を用いる。ただし、GDMAX 法では訳語候補の可能な組合せ数に従って計算量が増大してしまう。そこで、共起頻度を用いて、タームのグループ分けを行い、各グループごとに段階的に GDMAX 訳語選択法を施すことで、訳語候補の組合せ爆発を防ぐ。この処理手順は次のとおり。

(1) 各キュエリタームの訳語候補リストのうち、あるタームの同義語や異表記を仮想的に1つの訳語候補としてまとめる。仮想的な訳語候補の共起頻度は各要素の平均をとる。

(2) キュエリタームのグループ分け、および、各グループへの GDMAX 適用順序のスケジューリングを、次の優先順序で決定。

  (a) 一定数以上の訳語候補を持つタームは、曖昧性が多いタームと見 なして、GDMAX にかけるのを最後にまわす。

  (b) キュエリタームエクスパンションで拡張されたタームより、元のキュエリタームの訳語選択を優先。元のキュエリタームに対する GDMAX が進んだ時点で、拡張されたタームを1つずつ選択し、それと既に訳語候補の組合せが減少した元のキュエリタームとで GDMAX を実行。

  (c) 式1の左辺のように正規化した共起頻度の値が大きいターム同士を1つのグループにまとめ、各グループごとに独立に GDMAX を適用する。

(3) スケジューリングに従って、グループごとに GDMAX を実行。GDMAX 法により選択された訳語候補以外の候補を順次、訳語候補リストから除いていくことで、訳語候補の組合せ数を押さえる。

4 評価実験

本提案手法の評価実験を、文部省学術情報センター (NACSIS) が作成したワークショップ用のデータ「NACSIS テストコレクション」[4] を用いて行った。

図1 に、検索結果の Precision-Recall 曲線を示す。図中「拡張＆段階 GDMAX」とあるのが、本提案手法による結果、「段階 GDMAX のみ」がキュエリタームエクスパンションなしで段階 GDMAX のみ用いたもの、「CROSS」が機械翻訳による対照実験、「人 (拡張あり)」が人手により理想的な訳語と拡張するタームを与えたものである。各手法により訳語が選択された後は、同条件で既存の検索エンジン SMART [3] に与えた。

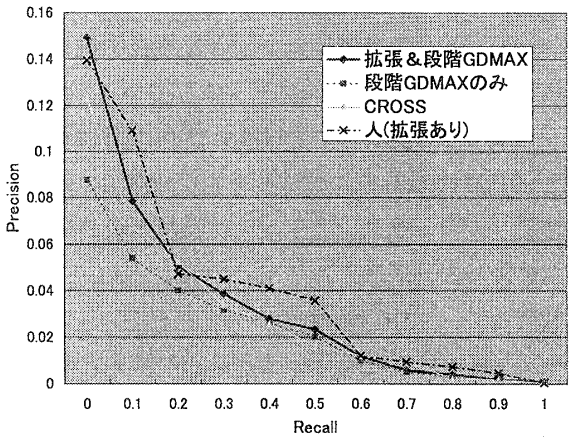

図 1: Precision-Recall 曲線

また Sun UltraSPARC-II(296MHz, 524Mbyte) 上での実測計算時間は、本提案手法が 6 分 17 秒 (NTCIR 全 53 課題、検索時間含めず)、それに対して GDMAX のみの場合が 8 時間計算しても終了しなかったため中止、という結果になった。

SMART の調整が不十分なため、全体的に低めの検索精度だが、以上から、本手法の有効性を確認できた。今後の課題としては、各種パラメータの自動的な調整や、他のキュエリタームエクスパンション手法について検討していきたい。

参考文献

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

[2] 奥村明俊 他, "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", 自然言語処理, Vol.5, No.4, 1998.

[3] G. Salton, "The SMART Retrieval System : Experiments in Automatic Document Processing", Prentice-Hall(1971).

[4] 学術情報センター, http://www.rd.nacsis.ac.jp/ ntcadm/index-ja.html