言語横断検索における機械翻訳の利用
– 文書翻訳に基づく順位付けの精密化 –

藤井 敦　石川 徹也
図書館情報大学
{fujii,ishikawa}@ulis.ac.jp

1 はじめに

1990年代から，WWW（Web）などを介して外国語文書を電子的に入手する機会が急激に増えている．それらの文書を母国語キーワードで効率良く検索したり，また逆に多くの国の人に自分のWebページを検索してもらうためには，言葉の障壁を越えた検索処理が必要である．これは「言語横断情報検索（Cross-Language Information Retrieval: CLIR）」と呼ばれ，ACM SIGIRなどの国際会議，TREC [11] やNTCIR [10] などの情報検索評価ワークショップにおいて主要なテーマのひとつである．

検索質問と対象文書の言語的な違いを吸収し，いかにして単言語検索に帰着させるかが，CLIRの中心的課題である．この観点から，従来のCLIRは以下に示す3種類の方式に分類できる．

(a) 検索質問を文書言語に翻訳する方式 [1, 4, 12, 17].
(b) 対象文書を検索質問言語（ユーザ言語）に翻訳する方式 [9, 13, 16].
(c) 検索質問と対象文書を中間言語に変換する方式 [2, 5].

各方式の検索精度を比較した実験結果が報告されているものの [9, 13]，それらは検索質問と対象文書の言語の組合せや実験環境に依存するため，各方式の優劣について一概に結論づけることは難しい．

適合性フィードバックに基づく対話的な検索や，ユーザが検索文書を閲覧する際の簡便さを考慮すれば，検索時に文書がユーザ言語に翻訳されている点で方式(b)が好ましい．McCarley [9] は，方式(a)と方式(b)で計算されたスコア（すなわち，検索質問に対する各文書の適合度）を平均して文書の順位付けに利用することで，個々の方式の検索精度が改善できることを示している．しかも，この結果は英仏両方向の実験で確認されている．

以上まとめると，方式(b)はCLIRにおいて有効である．しかし，現状では方式(a)に基づくシステムが比較的多く，方式(b)に関する研究例はそれほど多くない．

その理由のひとつは，対象文書群を機械翻訳するコストが高いことである．例えば，Oard [13] はTREC-6言語横断タスク用データに含まれる48カ月分の新聞記事の独英機械翻訳に10マシン月を費やしている．この問題は，想定するユーザ言語の種類が多い場合や，Webページのように日常的に更新される文書群に対しては，より一層深刻になる．

本研究では，翻訳コストの問題を解消しつつ，文書翻訳方式の利点を生かすためのCLIR方式を提案する．具体的には，まず方式(a)に従って検索質問を文書言語に翻訳し，検索と順位付けを行う．次に上位N文書のみをユーザ言語に機械翻訳する．最後に，各翻訳文書について，ユーザ言語で書かれた本来の検索質問に対する適合度を計算し，それに基づいて文書の再順位付けを行う．

本手法は，方式(a)と(b)を統合したMcCarley [9] の手法に類似している．しかし，本手法には以下に示す利点がある．

- 翻訳対象が比較的少数の文書に限定されるため，機械翻訳に要するコストが低い．
- 文書翻訳はユーザの端末で個別に行えるため，文書データベースや検索用サーバの管理が楽である．
- 機械翻訳システムの選択や切替えが容易であり，また機械翻訳の訳質改善を早期にCLIRに活用できる．

翻訳機能を備えたWebブラウザが既に市販されており，これらに文書再順位付け機能を追加することで，本手法をWeb上のCLIRに応用できる．

2 システム構成

本研究で提案するCLIRシステムを図1に示す．現在，本システムは日英間で双方向に検索可能である．

システムの入力はユーザ言語で書かれた検索質問である．日本語の検索質問が句や文で書かれている場合は，

形態素解析器 [19] を用いて分割した後，品詞情報に基づいて内容語のみを検索キーワードとして抽出する．英語の検索質問に対しては，WordNet [3] に定義されている不要語（stopword）リストと品詞情報に基づいて，内容語のみを抽出する．

次に，検索キーワードを文書言語に翻訳するために，著者らが提案したキーワード翻訳法 [4, 17] を用いる．この手法では，「EDR 対訳辞書・専門用語対訳辞書」[18] と，独自に作成した語基辞書を用いて対訳候補を列挙する．さらに辞書未登録のカタカナ語に対しては，音韻的に等価な単語に英訳するための翻字処理（transliteration）を行う．そして，語の共起に基づく統計的手法を用いて訳語曖昧性を解消する．

文書検索では，翻訳されたキーワードを入力として外国語文書を検索し，さらにそれらを適合度に基づいてソートする．現在，目的に応じて Web 上の各種検索エンジンや「SMART」[14] などが利用可能である．

検索された外国語文書の上位 N 件を対象に，機械翻訳（MT）システムを用いてユーザ言語に翻訳する．ここで，N はユーザが要求している文書数や MT システムの翻訳時間などを考慮して適宜設定する．文書翻訳には，「Transer」日英／英日 MT システム[1] を用いている．なお，Transer は検索キーワード翻訳にも利用できる．

最後に，各翻訳文書について，本来のユーザ言語キーワードに対する適合度を計算する．そして，外国語キーワード／文書間で計算された適合度と組み合わせて，新しい適合度を計算し，その値に基づいて再順位付けする．

以下，3 章で文書翻訳に基づく文書再順位付けについて，4 章で「NACSIS テストコレクション」[7] を用いた日英検索実験についてそれぞれ説明する．

3 文書順位付けの精密化

図 1 における文書検索処理は，言い替えれば文書言語での単言語検索である．他方において，ユーザが入力した検索キーワードと機械翻訳された検索文書を用いれば，ユーザ言語での単言語検索が可能である．そこで，両方の検索処理で得られる適合度を統合して文書順位付けを精密化することが，本研究で提案する CLIR システムの特長である．日英／英日機械翻訳は完全に表裏一体ではなく，一方の翻訳では多くの訳語曖昧性を生じる場合でも，逆方向の翻訳は比較的簡単な場合もある．すなわち，両者を組み合わせることで誤訳などの不備を互いに補い合える可能性がある．

[1] http://www.nova.co.jp/

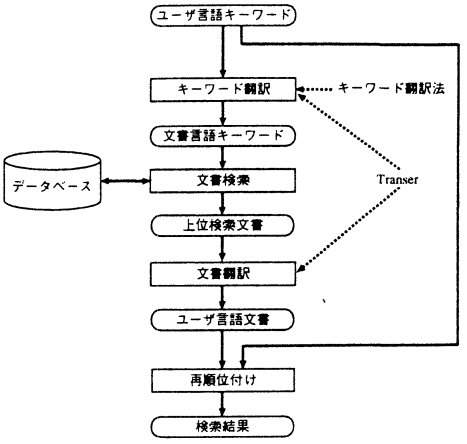

図 1: CLIR システム構成

通常の文書検索では索引付けを静的に行い，大規模なデータベースに対しても効率の良い検索を実現する．さらに，統計的に文書の適合度を計算するためには，各文書における索引語の頻度分布を調べる必要がある．しかし，機械翻訳後のユーザ言語での単言語検索では，すでに対象文書が限定されており，また即時性を重視することから，索引付けに依存しない手法が必要である．そこで，単純な表層マッチングによって，各文書における検索キーワードの出現頻度を計算する．

適合度の計算にはベクトル空間法を用いる．ただし，予備実験の結果，文書長を考慮しない場合に検索精度が向上したので，検索質問ベクトルと文書ベクトルの内積のみによって適合度を計算する．検索キーワードの重み付けには TF·IDF 法の一種である式 (1) を用いる．

$$(1 + \log(f_{t,d})) \cdot \log \frac{N}{n_t} \tag{1}$$

ここで，$f_{t,d}$ は文書 d におけるキーワード t の出現頻度であり，n_t は初期検索で得られた N 件のうちキーワード t を含む文書数である．

最後に，文書言語とユーザ言語で個別に計算した適合度（それぞれ $drel$，$urel$ とする）を統合する．ただし，使用する検索エンジンによっては，二つの適合度は異なる次元を持つので，算術平均よりも幾何平均を用いて新しい適合度 rel を計算する．これを式 (2) に示す．

$$rel = drel^{\alpha} \cdot urel^{\beta} \tag{2}$$

ここで，α と β は二つの適合度の影響を制御するパラメタであり，現在は，$\alpha = \beta = 1$ としている．しかし，今後は日英／英日翻訳の訳質を考慮しながら適切な値を設定する必要がある．

4 評価実験

4.1 概要

NACSISコレクション[7]を用いて，本システムを日英検索の観点から評価した．具体的には，TRECやNTCIRなどで行われているように，あらかじめ用意された検索課題に対して適合文書を出力し，11点補間なし平均適合率を尺度として評価を行った．

NACSISコレクションは，65学会の論文から収録した約33万件の抄録，日本語の検索課題39件（公式版），各検索課題に対する正解文書リストからなる．文書には，「文書番号」「論文タイトル」「著者名」「出典」「抄録」「著者キーワード」「学会名」のフィールドがある．各フィールドは日本語か英語，あるいはその両方で書かれている．各検索課題には，適合(A)または部分的適合(B)と判定された文書が対応付けされている．

このコレクションはNTCIR参加システムの評価に用いられており，検索課題中の「検索要求」フィールドを入力とし，A判定文書のみを正解と見なした場合の平均適合率は2.1–18.2%であった[10]．そこで，本実験の実験設定をNTCIRと等しくし，既存の結果との比較を容易にした．

文書検索には「SMART」[14]を利用し，日本語対訳を持つ英語抄録（約19万件）を対象に，「論文タイトル」「抄録」「著者キーワード」フィールドを用いて索引付けを行った．

4.2 実験結果

本システムにはいくつかの選択肢があるので，異なる組合せ方ごとに評価を行った．具体的には，文書検索における検索件数Nの値を変化させ，さらに検索キーワードの翻訳法として以下の4通りを比較した．

(1) 検索要求文をTranserで翻訳する．
(2) 検索要求文中の内容語をTranserで翻訳する．
(3) 検索要求文中の内容語をキーワード翻訳法[4]で翻訳する．
(4) 手法(2)と(3)の翻訳結果を統合する．

さらに，機械翻訳に基づく再順位付けの効果を調べるために，再順位付けを行い手法を下限に，人手による文書翻訳に基づく再順位付けを上限として比較した．人手による文書翻訳には，検索対象とした英語文書に対応する日本語文書を用いた．

以上の組合せに対する平均適合率を表1に示す．表1において，手法(1)-(4)は再順位付けを行わない結果であり，「+MT」と「+HT」はそれぞれ機械翻訳と人手による翻訳に基づく再順位付けを追加した結果である．また，4つの検索要求翻訳法に対する機械翻訳と再順位付けの平均CPU時間を表2に示す．

表1: 各手法の11点補間なし平均適合率（%）

	検索文書数（N）						
手法	50	100	200	400	600	800	1,000
(1)	9.49	10.17	10.74	11.01	11.12	11.19	11.24
+MT	13.42	15.44	16.78	17.05	17.34	17.51	17.85
+HT	16.66	19.01	20.70	21.73	22.30	22.59	22.97
(2)	9.53	10.20	10.85	11.13	11.23	11.31	11.34
+MT	14.51	15.63	16.98	17.19	17.43	17.66	17.64
+HT	16.19	18.19	20.17	21.05	21.65	22.03	22.17
(3)	12.15	13.01	13.55	13.85	13.94	13.99	14.03
+MT	15.27	17.18	18.58	19.58	19.78	20.06	20.17
+HT	17.22	19.15	20.97	22.12	22.41	22.79	22.95
(4)	12.29	13.05	13.76	14.05	14.16	14.21	14.26
+MT	16.75	17.75	19.11	19.61	19.77	19.85	20.04
+HT	18.14	19.68	21.42	22.42	23.01	23.19	23.56

表2: 文書翻訳と再順位付けのCPU時間（秒）

検索文書数	50	100	200	400	600	800	1,000
CPU時間	15.6	30.1	59.2	117.6	176.2	271.1	293.9

(Pentium III 700MHz)

4.3 考察

まず検索キーワード翻訳法を比較すると，検索文書数に関わらず，手法(1)から(4)に進むに従って平均適合率が良くなり，両者を異なる翻訳法を組合わせることが有効であることが分かった．手法(3)はEDR専門用語辞書に基づいており，Transerに比べると概して専門的な訳語を出力した．例えば，「複数データ」や「脊椎動物」に対して，前者は「multiple data」と「craniate」を，後者は「more than one data」と「vertebrate」をそれぞれ訳語とした．また，キーワード翻訳法は，Transerが翻訳できなかった「コラボレーション」や「モバイル」などのカタカナ語を翻字によって適切に英訳した．

次に，機械翻訳に基づいて再順位付けすることで，検索文書数に関わらず，各手法の平均適合率が向上した．TRECやNTCIRのように上位1,000件を出力した場合，手法(3)において20.17%となり，NTCIRで報告された値を上回った．

表1に示した平均適合率は，検索課題ごとに平均適合率を求め，それらをマクロ平均したものである．すなわち，数値を比較しただけでは，検索課題によらず一様に有意に向上したのか，それともある特異な検索課題に起因する偶発的な向上なのかを区別しにくい．そこで，検索課題39件をランダムサンプルと見なし，ウィルコクソン（Wilcoxon）符合順位検定を行った[6, 8, 14, 15]．

その結果，手法(1)-(4)の全てにおいて，再順位付けによる適合率の差は有意水準0.01で有意であった．

表2より，検索文書が1,000件の場合は文書翻訳と再順位付けに合わせて5分近くかかり，即時性の点からは改善の余地があることが分かった．しかし，検索文書が少ない場合は，再順位付けの効用を保持したまま，より高速な処理が可能なことを確認できた．

最後に，人手による翻訳を用いると，再順位付けの効果が例外無く顕著になることが分かった．機械翻訳の訳質を向上させることで，さらなる検索精度の向上が期待できる．また，文書翻訳による再順位付けを行うと，検索キーワード翻訳による適合率の差異は減少した．

5 おわりに

本研究で提案した日英／英日CLIRの特長は，検索された外国後文書をユーザ言語に機械翻訳し，翻訳結果に基づいて順位付けを精密化する点にあった．論文抄録を検索対象に用いて日英検索の実験を行った結果，文書翻訳に基づく再順位付けによって検索精度が有意に向上した．計算効率の改善は今後の研究課題である．また，ユーザ言語に翻訳された検索文書に基づいて対話的に検索する枠組についても今後検討する必要がある．

謝辞

Transer 機械翻訳システムは（株）ノヴァの許諾を，NACSISコレクションは学術情報センターの許諾を得て使用させて頂きました．

参考文献

[1] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64–71, 1998.

[2] 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 Artificial Intelligence*, pp. 708–714, 1997.

[3] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] Atsushi Fujii and Tetsuya Ishikawa. Cross-language information retrieval for technical documents. In *Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 29–37, 1999.

[5] Julio Gonzalo, Felisa Verdejo, Carol Peters, and Nicoletta Calzolari. Applying EuroWordNet to cross-language text retrieval. *Computers and the Humanities*, Vol. 32, pp. 185–207, 1998.

[6] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, 1993.

[7] Noriko Kando, Kazuko Kuriyama, and Toshihiko Nozue. NACSIS test collection workshop (NTCIR-1). In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299–300, 1999.

[8] E. Michael Keen. Presenting results of experimental retrieval comparisons. *Information Processing & Management*, Vol. 28, No. 4, pp. 491–502, 1992.

[9] J. Scott McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 208–214, 1999.

[10] National Center for Science Information Systems. *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

[11] National Institute of Standards & Technology. *Proceedings of the Text REtrieval Conferences*, 1992–1998. `http://trec.nist.gov/pubs.html`.

[12] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–81, 1999.

[13] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pp. 472–483, 1998.

[14] Gerard Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, 1971.

[15] Padmini Srinivasan. A comparison of two-poisson, inverse document frequency and discrimination value models of document representation. *Information Processing & Management*, Vol. 26, No. 2, pp. 269–278, 1990.

[16] 酒井哲也, 梶浦正浩, 住田一男, Gareth Jones, Nigel Collier. 機械翻訳を用いた英日・日英言語横断検索に関する一考察. 情報処理学会論文誌, Vol. 40, No. 11, pp. 4075–4086, 1999.

[17] 藤井敦, 石川徹也. 技術文書を対象とした言語横断情報検索のための複合語翻訳. 情報処理学会論文誌, 2000. (掲載予定).

[18] 日本電子化辞書研究所. EDR電子化辞書仕様説明書, 1995.

[19] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』version 1.5 使用説明書. Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学, 1997.