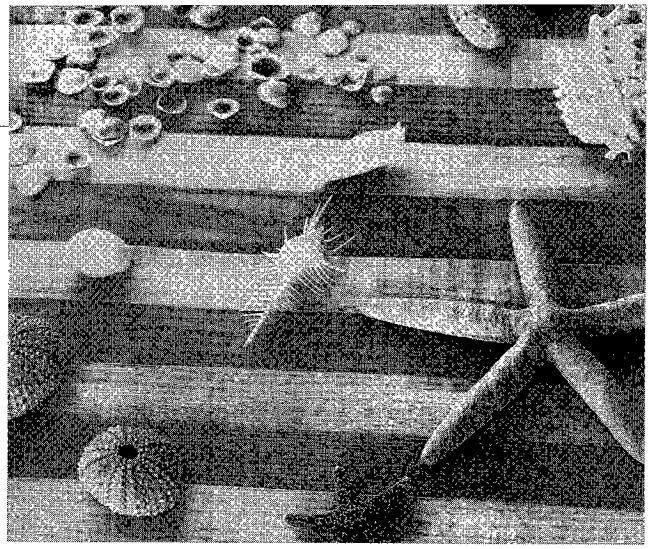

動クエリー作成では，自動クエリー作成の場合よりも逆転の可能性は上がるという．

異なるユーザモデルが，適合性に基づく評価に与える影響は，ごく限られたものであるのかもしれない．

1990年代にTRECから始まった評価ワークショップを通して，システムについていろいろなことが分かってきたのと同時に，評価方法についてもいろいろなことが分かってきた．これらの成果は，商用システムの実験室的でない評価にも反映されるだろう．

3.3 NTCIRへの参加から学んだこと

Fredric C. Gey カリフォルニア大学バークレイ校 *gey@ucdata.berkeley.edu*
Aitao Chen カリフォルニア大学バークレイ校 *aitao@sims.berkeley.edu*
Hailing Jiang カリフォルニア大学バークレイ校 *hjiang1@sims.berkeley.edu*
（著・訳）
岸田 和明 駿河台大学 *kishida@surugadai.ac.jp*

■NTCIRへの参加■

カリフォルニア大学バークレイ校は，NTCIRワークショップに参加し，日本語の単言語検索と，日本語と英語との言語横断検索の2つのタスクに取り組んだ．前者においては，バイグラム（bi-gram）[訳注1]による方法と辞書を用いる方法とを比較検討し，後者においては，NTCIRのコレクション中の日本語と英語が対訳になっている部分から抽出した対訳辞書と，専門用語を含まない辞書を使った機械翻訳とを比較した．ワークショップに提出した我々の論文[10]では，日本語検索および日本語－英語の言語横断検索のための複数の方法を比較している．具体的には，語分割[訳注2]のための2つの方法と，検索質問を翻訳するための2つの方法をテストした．この研究は，Text REtrieval Conference（TREC）[5]への参加を通じて行った，全文に対する単言語検索および言語横断検索に関する諸研究[9), 11), 14)]の上に成り立っているものである．

日本人以外でNTCIRに参加したいと考えている人々は2つの障壁に直面している．それは，日本語への不慣れさと語の境界の検出である．それまでヨーロッパ言語で研究を行ってきたグループは，語分割のアルゴリズムと諸研究とを学ぶ必要がある．バークレイのグループでは，以前に中国語の語分割の研究[9)]があることと，日本人の研究者が1998年から1999年までバークレイに滞在していたことによって，この障壁は難しいものではなかった．

■文献のランキング■

NTCIRの検索では，我々はすべて，以前にTREC-2で用いた方式[11)]を使って，文献の順位を算出した．TRECのテストコレクションに対する随時検索（ad hoc retrieval）において，この方式は，長い検索質問と人手で修正された検索質問とに対して頑健であることが示されている．他の言語に適用した場合でも（この方式はTRECの英語のコレクション上での訓練結果である），TREC-4のスペイン語，TREC-5の中国語[13)]，TREC-6とTREC-7のヨーロッパ言語（フランス語，ドイツ語，イタリア語）[12), 14)]のように，この方式はよく機能する．すなわち，このアルゴリズムは，語の境界の検出（語分割）さえ可能ならば，どの言語であろうとも，頑健であることをこれまで示してきたわけである．この方法はロジスティック回帰に基づく確率型検索モデルであり，詳細については，我々のワークショップの論文[10)]を参照してもらいたい．

訳注1　たとえば，「情報検索...」ならば，情報，報検，検索，索...のように重複した2文字ずつに分ける．
訳注2　テキストを語の単位に切り分けること．

■NTCIRの単言語検索タスク■

ほとんどの情報検索システムでは，文献と検索質問は複数の語で表現される．日本語の文章では，語の境界は明示されないので，索引作成の第1段階は，通常，日本語テキストの語分割になる．日本語の文章では，漢字，カタカナ，ひらがな，英字の4つの文字集合が用いられ，文章中ではこれらが混ぜこぜになっている．多くの場合，ひらがなは内容を表す語（content-bearing term）ではない傾向があるので，我々の索引作成からは除外した．この結果，語の断片は漢字かカタカナで構成されることになった．

語分割の方法の1つは辞書に基づく最長一致である．これは，テキスト中の文字列を辞書の見出し語と照合して，一致する見出し語の中で最長のものを語として取り出す方法である．一般に，語への分割で高い正確性を得るには，対象となるテキストに比べて，幅広い収録範囲を持つ辞書が必要である．我々は，漢字とカタカナからなる419,741個の見出し語を持つ辞書（というよりも語のリスト）を作成した．その多くは，テストコレクション中の日本語キーワードのフィールドから抽出した訳注3．この辞書を使った最長一致アルゴリズムによって，漢字とカタカナの固まり（chunks）を各語に分解した．この方法は，日本語－英語の横断検索において，日本語の検索質問を英語に翻訳する前段階として検索質問を切り分けるのにも用いている．

別の方法は，文献と検索質問とを，漢字かカタカナからなるバイグラムに分解することである．単言語検索においては，バイグラムによる語分割の方が辞書を使った語分割を上回った．前者の平均精度.4378に対して後者は.3329であり，約32％上回っている訳注4．

■言語横断検索タスク■

言語横断検索は通常，検索質問を翻訳するか，文献を翻訳するか，その両者を第3番目の言語に翻訳するかのいずれかによって実行される[15),16)]．検索質問は，機械翻訳システムか，対訳辞書を用いることにより翻訳できる．そして，その対訳辞書の収録範囲が言語横断検索システムの性能に大きな影響を与える．日本語－英語横断検索において我々がとったアプローチは，テストコレクション中で英語と日本語の両方の著者キーワードを持つ文献から対訳辞書を作成することであった．そして，それを使って，日本語の各検索語を英語における同等のものに写像して，そのすべての英訳に対して，英語のコレクション（ntc1-e0）を検索した．

我々の対訳辞書は，テストコレクションntc1-je0における日本語と英語の著者キーワードのフィールドから，それらの語をフィールド中に出現する順序で機械的にペアにまとめることによって構成した．

日本語の検索質問の英語への翻訳においては，はじめに，辞書に基づく最長一致を使って，検索質問中の語を切り分けた．そして日本語の各語に対して，最頻出の英語をその翻訳として採用した訳注5．言語横断検索の問題点の1つは，いくつの語を翻訳として採用するかである．このテストコレクションは技術的報告の抄録から構成されているので，我々は日本語の各索引語は一般に1つの英語の翻訳を持つのみと仮定した．ただし，その翻訳には複数の英単語が含まれる可能性がある．以上の方法による検索質問の英語の翻訳を，ntc1-e0コレクションの文献と照合した．また，言語横断検索用の検索質問を機械翻訳システム☆3を使って翻訳することも試みた．

日本語－英語横断検索における我々の最良の検索結果は.3755の平均精度を示し，これは，機械翻訳を使った場合の.1925に比べて約2倍である．この結果は，各著者が付与したキーワードから作成したコーパスベースの対訳辞書を前者が用いたことによるものと考えている．著者キーワードは，日本語と英語の高度に技術的な語彙の変換を可能とするものである．

■NTCIRコレクションに関する考察■

単言語の日本語検索に関しては，より単純なものがより良いという，ある意味では驚くべき結果を得た．すなわち，漢字とカタカナのバイグラムによる語分割が，辞書に基づく語分割を約30％上回った．これは，辞書の不完全さと句としての性質，すなわち，我々は長い複合語を意味ある構成要素に分解する手段を持たなかったことに起因している．しかし，言語横断検索では，句の語分割が翻訳のより高い精度に結びつく．

言語横断検索の性能は数多くの要因に影響を受けているようである．たとえば，対訳コーパスにおける翻訳の質，日本語の語分割の正確さ，文献のランキング方式の有効性，などである．翻訳の不完全性や非一貫性，あるいは英語のスペルミスが，対訳辞書の品質を劣化させ，その結果，言語横断検索の性能が悪化する．

訳注3　これは各文献の著者が付与した著者キーワードである．
訳注4　平均精度は検索実験においてよく使われる評価の指標である．全適合文献の平均精度を全検索質問について平均した値．
訳注5　著者キーワードの並び順だけに着目して対訳辞書を作っているので，当然，1つの日本語に対して複数の英語が対応することになるが，その対応の頻度を計数しておく．なお，著者キーワードは，日英のキーワードが同じ順序で一対一対応で並んでいるとは限らない．
☆3　日本語－英語の機械翻訳システムであるGAZELLEを使って検索質問を英語に翻訳してくれた，南カリフォルニア大学Information Science InstituteのKevin KnightとEd Hovyに深く感謝したい．

対訳テキストから構築された対訳辞書を使った検索質問の翻訳と，人手による翻訳との検索性能を比較してみるのも面白いだろう．人的な資源の問題から，我々はこれらの比較ができない．日本語の単言語での随時検索においては，バイグラムによる語分割が辞書に基づく語分割をかなり上回った．しかし，バイグラムによる方法が辞書に基づく方法よりも優れていると直ちに結論することはできない．なぜなら，我々の辞書は非常に限定されたものだからである．もし，文献のコレクション中に出現するすべての語を含んだ辞書が存在し，なおかつ語分割が完全ならば，情報検索においてどの方法がより良いかの結論を得ることができるだろう．

情報検索分野へのNTCIRの大きな貢献は，NTCIR-1コレクションそのものであり，これは次のような恩恵を研究者にもたらす．

- 日本語の検索質問と文献の適合判定付きのコレクション
- 日本語－英語の対訳コーパス
- 技術的・科学的な領域のコーパス

この3点目は特に重要である．情報検索の実験において幅広く利用されている科学分野のテストコレクションはクランフィールドコレクションであり，これは現在，すでに成立から40年を経過しようとしている．NTCIR-1コレクションは規模・範囲ともに，クランフィールドコレクションをはるかに凌駕している☆4.

参考文献

1) Witten, I. H., Moffat, A. and Bell, T. C.: Managing Gigabytes, Van Nostrand Reinhold, p.150 (1994).
2) Ozawa, T., Yamamoto, M., Umemura, K. and Church, K. W.: Japanese Segmentation Using Similarity Measure for IR, Proceedings of NTCIR Workshop 1, pp.89-96 (1999).
3) 山本英子, 梅村恭司, 小澤智裕, 山本幹雄, Church, K. W.: 一般化文字列類似度を用いた文字ベースの情報検索, IREXワークショップ予稿集, pp.95-100 (1999).
4) 小澤智裕, 山本幹雄, 山本英子, 梅村恭司: 情報検索の類似尺度を用いた検索要求文の単語分割, 言語処理学会第5回年次大会発表論文集, pp.305-308 (1999).
5) http://trec.nist.gov/
6) http://host.limsi.fr/RIAO/
7) Ellis, D.: New Horizons in Information Retrieval, London, Library Association (1990) (邦訳: 細野公男他訳, 情報検索論－認知的アプローチへの展望, 東京, 丸善 (1994)).
8) Voorheers, E. M.: Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness, SIGIR '98, pp.315-323, Melbourne, Australia (1998).
9) Chen, A., He, J., Xu, L., Gey, F. C. and Meggs, J.: Chinese Text Retrieval without Using a Dictionary, In Proceedings of ACM SIGIR '97 (1997).
10) Chen, A., Kishida, K., Jiang, H., Liang, Q. and Gey, F. C.: Comparing Multiple Methods for Japanese and Japanese-English Text Retrieval, In Proceedings of First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Tokyo, JAPAN (Sep. 1999).
11) Cooper, W. S., Chen, A. and Gey, F. C.: Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression, In The Second Text REtrieval Conference (TREC-2), pp.57-66 (Mar. 1994).
12) Gey, F. C., Jiang, H., Chen, A. and Larson, R.: Manual Queries and Machine Translation in Cross Language Retrieval and Interactive Retrieval at TREC-7, In The Seventh Text REtrieval Conference (TREC-7), pp.527-539 (1999).
13) Gey, F. C., Chen, A., He, J., Xu, L. and Meggs, J.: Term Importance, Boolean Conjunct Training, Negative Terms, and Foreign Language Retrieval: Probabilistic Algorithms at TREC-5, In The Fifth Text REtrieval Conference (TREC-5) (1996).
14) Gey, F. C. and Chen, A.: Phrase Discovery for English and Cross-Language Retrieval at TREC-6, In The Sixth Text REtrieval Conference (TREC-6), pp.637-648 (1998).
15) Hull, D. A. and Grefenstette, G.: Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval, In 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1996).
16) Oard, D. W.: A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, In Machine Translation and the Information Soup, Third Conference of the Association for Machine Translation in the Americas, pp.472-483 (1998).

それぞれのNTCIR

本編では，参加者側から，チームの紹介と評価プロジェクトに関する意見・要望を，大学，企業研究所，海外からの参加という立場でそれぞれ1チームにまとめていただいた．企業研究所と海外のチームはNTCIRにおいて上位にランクされた2チーム，大学チームは成果報告会での発表後，休み時間も質問攻めにあうほど手法に関心を持たれていたチームである．これらのチームはそれぞれ，検索モデルと語分割法に特徴的な工夫をしているという点では一致するが，参加目的・バックグラウンドは大きく異なっている．これは，今回の評価プロジェクトに広い範囲の研究者が参加したことを意味しており興味深い．また，参加の立場は異なっていても，評価プロジェクトで作成された情報検索評価コレクションの有用性と今後の順調な発展を期待している点では一致している．

(平成12年6月30日受付)

☆4 我々の研究はNSF (National Science Foundation) の情報・データ管理プログラムからの助成 (番号: IRI-9630765), およびDARPA (Department of Defense Advanced Research Projects Agency) からの助成 (番号: N66001-97-C-8541, AO-F477) を受けている．