多言語ニュースの対照分析のためのWikipedia活用手法の研究

Utilization of Wikipedia Information for Comparative Analysis on Multilanguage News Sites

吉岡 真治*1
Masaharu YOSHIOKA

*1北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

The News Site Contrast (NSContrast) system analyzes multiple news sites using the concept of contrast set mining and can extract the terms that characterize the differences in topics of interest for each country. However, because NSContrast does not pay attention to synonyms, it tends to extract terms with different synonyms as candidate characteristic ones. To avoid this synonym problem, Wikipedia is introduced as a source for synonym identification. We give some experimental results for this New NSContrast system.

1. 緒言

現在、インターネットを通じて、世界中のニュースサイトから情報を獲得することが可能になっている。このようなサイトから発信される様々な情報を用意に閲覧するためのシステムとして、Google news*1のようなニュースアグリゲーションサイトや、世界中から発信される情報を用いて、世界中の報道の量的な違いを分析する EMM News Explorer *2などのシステムが提案されている。

我々も、このような各国のニュースサイトの興味の違いに注目してニュースサイトの記事を比較分析するシステム NSContrast を提案している [吉岡 09]。このシステムでは、共起語解析の際に、トピック語と共起語の国ごとの相関性の変化に注目することにより、各々の国では、それほどメジャーではないものの、他の国との興味の違いを表すキーワードを提示する。また、本システムを実際に利用して、複数のユーザによる利用実験を行ったところ、入力したトピックを特徴づけるキーワードが抽出可能であるといった意見が得られた反面、機械翻訳システムの誤訳による影響が大きいということが指摘された。

特に、中国語においては、ニュースで良く出てくる人名などの固有名詞の翻訳間違い (例えば、中国語でオバマは、「奥巴馬」などと表記されるが、辞書に項目がないと、人名として訳されない) や、異表記の問題が分析結果に影響を与えるという問題点が指摘された。

本論文では、これらの問題点を解消するために、Wikipedia の言語間リンクならびに転送の情報を用いることにより、翻訳間違いや異表記の影響の低減を諮る方法を提案する。また、提案したシステムを利用した分析実験の結果についても報告する。

2. NSContrast:ニュースサイト比較分析システム

本節では、本研究で提案している複数ニュースサイトの比較分析システム NSContrast について、その分析手法の基礎となる相関性の変化に基づく特徴語分析の手法と、システムの概要について述べる。

連絡先: 吉岡真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7107, yoshioka@ist.hokudai.ac.jp

∗1 http://news.google.co.jp

∗2 http://press.jrc.it/NewsExplorer/

2.1 相関性の変化に基づくニュースサイトごとの特徴語分析

トピックに対応するような文書群を分析する方法として、文書群中に特徴的に現れる (例えば、文書群と相関性が高い) キーワードを抽出しリストアップする方法などが多く利用される。しかし、このような文書群に特徴的なキーワードのみに注目した場合には、個々のニュースサイトごとの特徴が現れるのではなく、ほとんどのニュースサイトが共通に興味を持つようなキーワードが現れ、個々のサイトごとの特徴を見つけ出すことは困難である (図 1)。

これに対し、本研究で提案するニュースサイトの分析手法では、コントラストセットマイニングの考え方に基づく相関性の変化に注目した解析 [Taniguchi 06] を行う。具体的には、相関性の大きなキーワードに注目するのではなく、特定のニュースサイトにおけるキーワードと文書群の相関性とそれ以外のニュースサイトにおける相関性の比をとり、その比が大きいもの (そのサイトでは、それなりに注目を浴びているトピックを表すが、他のサイトではあまり述べられていないキーワード)、その比が小さいもの (そのサイトでは、他のサイトに比べて、ほとんど無視されているトピックを表すキーワード) を特徴的なキーワードとして抽出する (図 1)。

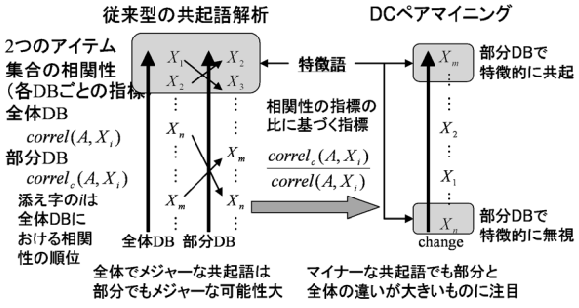

図 1: 相関性の変化に注目した特徴的キーワード分析

2.2 システムの概要

NSContrast では、日本語のニュースサイトから獲得した記事ならびに、中国・韓国の現地語ニュースサイトを機械翻訳した記事から構成されるニュース記事データベースに対して、下記の機能により記事の分析を行う。

- 情報検索システム
  サイト名や期間を限定して検索を行う。
- ニュース間の比較対照分析システム
  相関性の変化に基づく特徴語抽出を行う。システムは、各国ごとに、相関性の高い特徴語、相関性の変化の高い特徴語・相関性の変化の低い特徴語をリストとして表示する。

- 特徴語間の共起関係の可視化
  上記の特徴語間で共起関係の強いものの間にリンクを設定したグラフを作成し、バネモデルによりレイアウトすることによって、関連トピックの可視化を行う。また、通常のキーワードに加え、ニュースサイトが属する国のノードを作成し、特徴語がどの国の記事に多く現れているかを考慮したリンクを設定する。この国のノードと特徴語ノードの距離により、国ごとの興味の違いを可視化する。

- バースト分析
  単語の出現頻度の変化をもとに、特定の期間において注目を得たトピック語ならびに注目された期間を分析する手法であるバースト分析 [Kleinberg 02] を行うことにより、特徴的なキーワードと期間の情報を提供する。バースト分析の結果は、全てのニュースサイトの記事ごとに行うだけでなく、各国ごとに行い、その結果を比較することができる。

本システムでは、表 1 のサイトから定常的にニュース記事を収集し、毎晩、定常的にデータを更新することにより、朝になると前日までの記事を利用した解析が可能なように設定を行った。表 1 の記事数は、2008 年 1 月 1 日からの記事数の総計と 1 日あたりの平均記事数 (2009 年 4 月 19 日現在) を示す。

表 1: 利用したニュースサイトと記事数

サイト名 (国) URL (http:// は略)	記事数 (総計/1 日平均)
朝日新聞 (日) www.asahi.com/	58344/122
日経新聞 (日) www.nikkei.co.jp/	69638/146
読売新聞 (日) www.yomiuri.co.jp/	48675/102
CNN(米) www.cnn.co.jp/	9542/20
朝鮮日報 (韓) japanese.chosun.com/	24389/51
中央日報 (韓) japanese.joins.com/	18842/39
人民網 (中) j.peopledaily.com.cn/	18775/38
朝鮮日報 (韓:機械翻訳) www.chosun.com/	102009/214
新華社 (中:機械翻訳) www.xinhuanet.com/	367459/773

2.3 システムによる解析事例

本システムの機能を具体的な解析事例と共に説明する。

1. バースト解析による報道量の違いの分析

ユーザは、まず、興味のあるトピック語をシステムに入力する。システムは、トピック語を含む新聞記事を検索すると共に、記事の出現頻度をベースとしたバースト解析の結果を表示する。原油をトピックキーワードとし、6 月 23 日時点での解析結果を図 2 に示す。

この表では、日本のニュースサイト、韓国のニュースサイト、中国のニュースサイト、アメリカのニュースサイトを各々ひとまとめとし、各々の期間での検索語を含む記事数を表示している。また、日付に対応した欄が赤く (薄く) 塗り潰されており、バーストという欄に★がついている期間が検索語がバーストしている期間である。この表から各国における注目度の違いを理解することができる。

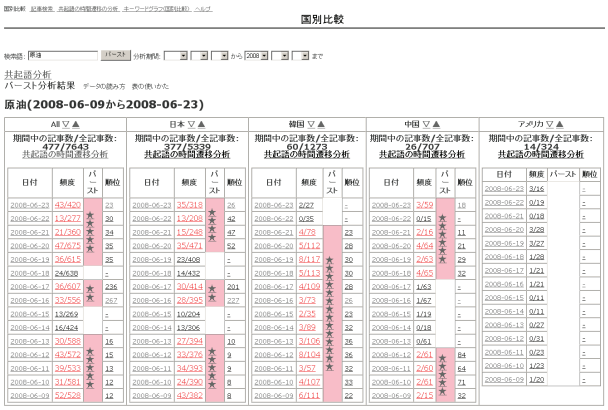

図 2: バースト解析結果の国別比較

2. 共起語解析による国ごとの特徴の分析

次に、これらのバースト情報に基づいて、共起語の対照比較を行う。昨年度のシステムで用いたような表形式の表現では、特徴語として抽出された語の間の関係が不明確であったため、本システムでは、特徴語間の共起関係の強さ、各国の新聞と特徴語の共起関係の強さをもとにリンクを設定したバネモデルによる共起語の可視化を行った。図 3 に先ほどの原油のトピックにおいて、全体的にバーストしている 6/19〜23 日の記事を利用した分析結果を示す。

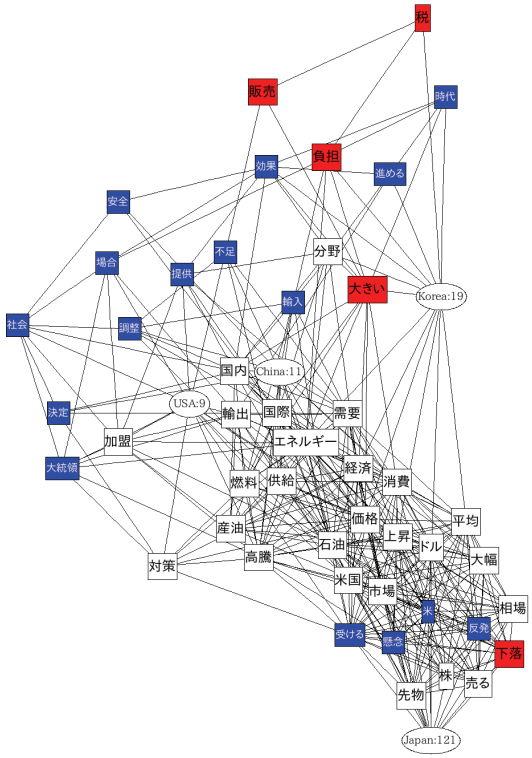

図 3: 特徴的な共起語間の関係グラフ

このグラフでは、相関性の変化が大きい共起語がどのようなメジャーなトピックに関係しているかを可視化するために、各国ごとに、相関性の変化が大きい共起語、相関性の高い共起

語を各々10 個ずつ選んで表示している (各国語との重なりがあるので、全体では、80 個より少ない)。白で表示されたものは少なくとも一つ以上の国で、相関性の高い語でメジャーなトピックを表す語である。濃い背景に白字で表された語は、相関性の変化が高い語で、特定の国の近くに配置される。濃い背景に黒字で示された語は、相関性も高く、相関性の変化も高い語で、その国においてメジャーでかつ、他の国ではメジャーでない特徴語を示している。図 3 からは、日本では、相場が下落し始めたことを、いち早く報道しており、韓国では、税負担の可能性が議論されているといった違いを見ることができる。

2.4 ユーザ実験：レポート作成課題の実践

本システムの有効性を検証するために、NSContrast の作成意図を考慮した利用シナリオを想定した実験を行った。具体的には、特定の事象に関する各国の報道の違いをレポートとしてまとめるという作業のために、本システムを利用するというシナリオを想定した実験を行った。

本実験で用いた機械翻訳システムは、中国・韓国の現地語の新聞に対し、クロスランゲージ社の Webtranser を適用して作成を行った。また、機械翻訳システムが与える影響を分析するためには、元々の言語の記事との対応関係を理解できることが望ましいので、今回の実験では、日本語に加え、中国語もしくは韓国語の少なくともどちらかは理解できる 4 名の方を被験者として実験を行った。

各々の被験者の作成したレポートから、各国の新聞の報道の違いを見つけるためのキーワードが本システムの特徴語分析から発見できたことが報告されると共に、下記に記す問題点が指摘された。

- 表記の違いの影響を強く受ける。
  特定のニュースサイトの記事において、他の新聞と異なる特有な表現が含まれる場合や、翻訳システムが一貫して間違いを行った場合に、その語が特徴語として抽出されやすい。ただし、この特徴語を利用することにより、記事の再検索のためのキーワードが得られる場合がある。
- 共起語間の関係の可視化について
  共起語の可視化自体は有効であるが、国との関係については、必ずしも、適切な配置であるとはいえない。

3. Wikipedia の活用による分析の精緻化

前節で述べた問題点を解消するためには、翻訳システムにおける翻訳間違いや表記の違いの吸収を行う必要がある。

本研究では、Wikipedia の言語間リンクを用いた翻訳辞書の作成と、日本語 Wikipedia の転送を用いた表記の違いの吸収を行った。

3.1 Wikipedia を用いた日中カタカナ語翻訳辞書の作成と利用

機械翻訳システムにおいて、辞書をメンテナンスすることは、非常にコストがかかる作業であるため、時事的な単語、例えば、「オバマ」や「マケイン」などといった、それまでのニュースにはあまり現れてこなかった人名・固有名などの翻訳については、辞書を整備が追い付いていないことが多く、結果として、適切な翻訳がなされないという事例が多く見受けられた。

このような問題に対して Wikipedia を利用して、翻訳辞書を作成する試みが提案されている [Erdmann 08]。特に、Wikipedia は、時事的な内容についての更新が頻繁に行われることから、今回の目的に、非常に適した言語リソースであると考えられる。

本研究では、まず、その有用性を検証するために Wikipedia の言語間リンクを利用した中日翻訳辞書を作成した。ただし、今回作成する辞書は、既存の辞書が不得意な分野を補強することが目的であるため、既存の辞書のエントリーを上書きする可能性が高い一般語については、辞書を作成する必要がない。また、漢字で表記される日本人の人名などは、中国語においても同じ漢字が用いられる事が多いため、翻訳の間違いが起こる可能性が少ない。これらのことを考慮して、日本語においても外来語であるカタカナ表記の語に関する翻訳辞書の構築を以下の手順で行った。

1. 言語間リンクを持つカタカナ語エントリの抽出
   日本語の Wikipedia の dump データからカタカナ語でありかつ中国語への言語間リンクを持つエントリーを抽出する。
   例：日本語「バラク・オバマ」→中国語「巴拉克・奥巴馬」のようなリンクを抽出
2. 姓と名の分離
   カタカナ語の内、人名については、「・」で分割されていることが多い。このような場合に、「・」で分割される対象の数が多い場合には、姓は姓どおし、名は名どおしで対応関係をつける。
   例：先の言語間リンクから、中国語「巴拉克」→日本語「バラク」、中国語「奥巴馬」→日本語「オバマ」という辞書のエントリを生成
3. 中国語における転送情報の利用
   中国語は日本語のように表音文字を持たないため、時事的に有名になった人名に対して、複数の漢字の表記が与えられることが多くある。この問題を処理するために、中国語における転送を利用して、表記のぶれに対応した辞書を生成
   例：中国語「歐巴馬」→中国語「奥巴馬」という転送と先の辞書を組み合わせ、中国語「歐巴馬」→日本語「オバマ」という辞書のエントリを生成

また、辞書のサイズを小さくするために、獲得したニュース中に現れない標記については、辞書のエントリーにいれないこととした。また、一般語をタイトルとするような固有名 (例：中国語「朋友」→日本語「フレンズ」：ドラマなどの名前) については、間違ってはいないが、読みにくい翻訳結果を生成する原因になる。このような辞書エントリーを削除するために、作成した辞書のエントリーについて、ニュース中に高頻度で現れるものについては、目視で確認し、不要と思われるエントリーについては削除することとした。この結果、2666 件の中国語表記に対する日本語表記を獲得した。

ここで作成した辞書を言語グリッドにおける辞書連携サービス [Bramantoro 08, NIC] と組み合わせることにより、その辞書内容を翻訳結果に反映させることとした。

3.2 転送を用いた表記の違いの吸収

次に、日本語の表記の違いの影響を減らすための Wikipedia の活用法について検討した。Wikipedia では、様々な表記のバリエーションが存在するエントリに対しては、代表表記によるエントリーの作成と、代表表記のエントリーへの転送が設定される。本研究では、この転送の情報を用いることにより、表記のバリエーションを代表表記にまとめる方法を利用した。

ただし、この転送について分析を行うと、以下のような転送の分類があることが確認された。

1. 表記の違いに関する転送
   例：JR 東海→東海旅客鉄道
2. 具体事例から抽象概念への転送
   例：みじん切り→切る (調理)

3. リストに含まれるデータからリストへの転送
   例：紀元前 469 年→紀元前 5 世紀、相川有→日本の漫画家 あ行
4. 関連する概念への転送例：接弦定理→円 (数学)、円周→円 (数学)

今回のような共起語解析のための利用を考えると、1,2 のように、同じ (あるいは類似している) 概念を、一つの表記にまとめる方法は有用であると考える。また、4 の関連する概念については、積極的に利用する必要はないと思われるが、共起語解析の対象に含めても大きな問題はないと考える。一方、3 のようなリストをひとまとめにすることは、あまり有用ではないと考える。

よって、転送のリストから 3 に相当するものを削除して、それ以外の項目については、転送先のエントリーにまとめて共起語を登録することとした。

3.3 ユーザ実験と考察

今回、新しく作成したデータに対して、前回 [吉岡 09] と同じ被験者に、レポート作成課題を与えるとともに、システムを利用していただき、コメントを頂いた。前回は、各ユーザにシステムを自由に使ってもらうことを前提として、レポート課題にあまり制限をかけなかったが、今回は、以下の共通課題を設定した。

- 金融危機・北朝鮮・オバマ大統領

これらの結果、ユーザは、共起語解析によって得られた特徴語やバースト解析の結果から、自分の興味に合うものを選択し、様々な観点から各国の違いの分析を行っていることが確認された。例えば、金融危機では、「通貨」、「米大統領」、「AIG」などに関する分析が行われた。

今後、定量的な分析を行うためには、このようなユーザの興味の違いが存在することを仮定した分析を行う必要があると考えられる。

また、今回の Wikipedia による表記の違いの吸収と翻訳の質の向上を含む翻訳システムについてのコメントとしては、以下のようなものがあった。

- 中国語の翻訳間違いに関する指摘が減少するものの、韓国語の翻訳間違いに関する指摘が存在した。
  韓国語については、オバマが「オオバ」＋「麻衣」といったような訳になってしまうという問題などが指摘された。また、中国語についても、カバレッジの問題から、複数名の人名についての翻訳間違いなどが指摘された。
- 表記の違いを吸収したが、文章中で、どの単語がそれに対応しているのかが分かりにくい。
  現在のシステムでは、表記の揺れなどを考慮した単語インデックスは作成しているが、文書中のどの語がそれにたいおうしているのかと言うことが明示的に示されていない。その結果、検索結果との対応付けが分かりにくい問題が発生した。
- 機械翻訳の質の問題
  現状の機械翻訳の結果は、まだ、不十分であるというコメントが多く得られた。

これらの問題点を解消するために、今後は、今回提案した Wikipedia によるカタカナ固有名詞の中日翻訳辞書の作成方法を他の言語にも適用し、その有効性を検証する必要がある。ただし、韓国語の場合は、表音文字であることから、同表記異議語が多く、文脈依存での訳し分けが必要な場合などが想定される。このような状況を想定した全体としてのシステム構築が求められる。

また、表記の違いの吸収の副作用については、文書中の表記と代表表記との関係を示すことなどにより、ユーザに分かりやすく提示する方法を考えたい。

また、機械翻訳システムの質の向上については、直接本研究の課題ではないが、読むべき価値のある新聞記事を探せているかどうかについては検討すべきであると考えている。

4. 結言

本稿では、これまでに提案してきた、新聞記事を分析する複数ニュースサイトの比較分析システム NSContrast の目的と機能について紹介を行うと共に、共起語解析によるテキストマイニングにおける自動翻訳システムの影響などについて問題点を述べた。また、この問題点を解決するための方法として、Wikipedia の言語間リンク、転送を使う方法を提案した。さらに、実験により、その有効性を確認した。

今後は、韓国語や英語など他の言語においても、同様のアプローチがとれるかどうかなどを検証していきたいと考えている。

謝辞

本研究を進めるにあたり、世界ニュース研究グループ (中川先生 (東大)、宇津呂先生 (筑波大学)、福原先生 (東大) ら) との有意義な議論を行った。また、言語グリッドグループのサービスを利用することにより翻訳作業を行った。ここに記して謝意をあらわす。

参考文献

[Bramantoro 08] Bramantoro, A., Tanaka, M., Murakami, Y., Schäfer, U., and Ishida, T.: A Hybrid Integrated Architecture for Language Service Composition, in *ICWS '08: Proceedings of the 2008 IEEE International Conference on Web Services*, pp. 345–352, Washington, DC, USA (2008), IEEE Computer Society

[Erdmann 08] Erdmann, M., Kotaro Nakayama, T. H., and Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia, in *Proc. of International Conference on Database Systems for Advanced Applications (DASFAA) (Mar. 2008)*, pp. 686–689, Springer-Verlag GmbH (2008), LNCS 4947

[Kleinberg 02] Kleinberg, J.: Bursty and hierarchical structure in streams, in *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 91–101, New York, NY, USA (2002), ACM Press

[NIC] NICT 言語グリッドプロジェクト：言語サービス利用マニュアル

[Taniguchi 06] Taniguchi, T. and Haraguchi, M.: Discovery of Hidden Correlations in a Local Transaction Database based on Differences of Correlations, *Data Engineering Applications of Artificial Intelligence*, Vol. 19, No. 4, pp. 419–428 (2006)

[吉岡 09] 吉岡 真治：NSContrast:世界ニュース比較分析システムの実験的評価, 言語処理学会第 15 回年次大会発表論文集, pp. 494–497 (2009)