Vol.2009-DD-71 No.1
2009/6/5

複数国ニュースサイトの比較対象分析システム NSContrast における評価手法に関する検討

吉　岡　真　治[†1]

本研究では、複数国のニュースサイトにおける興味の違いを比較対照分析するためのシステム NSContrast を提案している。この NSContrast では、各国のニュースサイトの日本語版の記事だけではなく、原語版の機械翻訳結果も組み合わせたデータベースをもとに、共起語解析やバースト解析を行うことにより、その違いを分析する。本発表では、この NSContrast における評価手法についての検討結果を分析実験の結果と共に報告する。

Study on Evaluation Method of NSContrast: Comparative News Article Analysis System for Characterizing the Difference among Multiple Countries

Masaharu YOSHIOKA [†1]

In this research we have already proposed a news site analysis system NSContrast that characterizes the differences among news sites from multiple countries. In this analysis system, news articles database are constructed by using following three types of news sites data; Japanese news articles written in Japanese, foreign ones written in Japanese and machine translated version of foreign ones written in foreign language. In this paper, we discuss how to evaluate the effectiveness of the NSContrast and briefly introduced the result of user experiment study.

†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

1. はじめに

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。このような環境の変化に伴い、我々は、国内のメジャーなマスメディアから得られる情報だけではなく、様々な特徴を持った複数のニュース源から、世界のニュースについて多角的に調べることが出来るようになった。この結果、諸外国に支店を持つような国際的な企業でなくても、外国の国でおこる可能性のある問題やその問題の自国への影響に関する分析を行うことができるようになっている。

このようなサイトから発信される様々な情報を容易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。これらのサイトを使うことにより、同じトピックに関係する様々なニュース記事が同時に閲覧できたり、異なる国におけるトップ記事の違いなどを比較することが可能となっている。

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[1),2)]。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。このシステムを一般的なユーザに、短時間、使ってもらった上でアンケートを実施したところ、約半数のユーザから、各国の違いが分かる情報が得られたという評価を得た[1)]。

しかし、テキストマイニングのシステムとしての評価としては、このような短時間の利用では不十分であると考えられる。本稿では、本システムの構築目的に対応した評価を行うための実験方法についての検討結果を報告すると共に、その実験結果について報告する。

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[1),2)]。

⋆1 http://news.google.co.jp
⋆2 http://press.jrc.it/NewsExplorer/

© 2009 Information Processing Society of Japan

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[3] 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)。

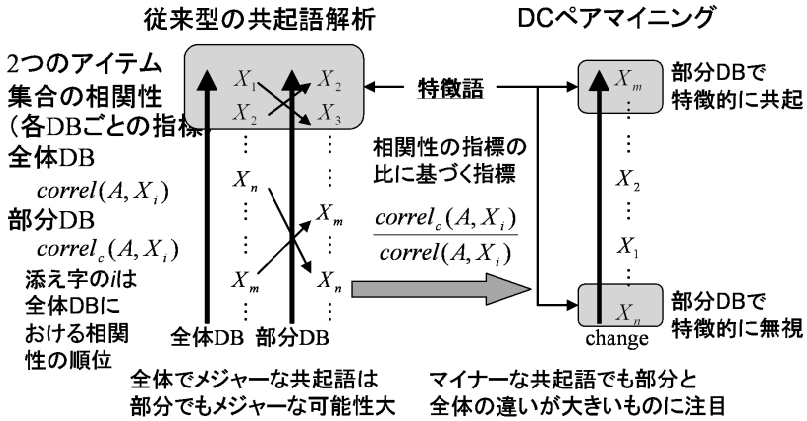

図 1 相関性の変化に注目した特徴的キーワード分析

2.2 複数言語によるニュース記事データベースの構築

本システムでは、様々な国のニュースサイトから日本語の記事を集めるだけでなく、現地語 (今回の場合は中国語、韓国語) で記述されている記事からニュース記事データベースを構築し、特徴語分析を行う。

しかし、この特徴語分析では、翻訳システムにおける翻訳間違いや表記の違いが分析結果に大きく影響を及ぼすので、その影響を吸収する必要がある。よって、本システムでは、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ている[2]。

まず、Wikipedia を用いた日中翻訳辞書の作成法について簡単に述べる。Wikipedia には、英語版・日本語版・中国語版などの様々な言語にエントリが存在し、そのエントリの中で同じ対象を説明しているものの間に言語間リンクというリンクが設定されている。このリンクは、異なる言語における同一の対象の表記法を示していると考えられるので、このリンクを用いた翻訳辞書の作成方法が提案されている[4]。本研究では、中日の翻訳で一番大きな問題であった外来語の固有名の名称 (例えば、オバマ大統領) に関する中日翻訳辞書を作成した。

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[5),6)] 用いて、中国語の機械翻訳システム (高電社製) 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。

韓国語については、表音文字を主体として記述される言語であるため、同表記異議語が多い。そのため、このように作成した辞書を優先的に適用する際に、文脈の判定など、より高度な処理をしないと、悪影響を及ぼす可能性がある。この点を考慮して、韓国語に関しては、クロスランゲージ社の Webtranser の翻訳結果を、そのまま利用することとした。

さらに、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、全ての転送を網羅的に利用すると、「紀元前 469 年→紀元前 5 世紀」、「相川有→日本の漫画家 あ行」といったリストに含まれるデータをリストへ転送するようなエントリが存在するため、転送の種類を整理して、不要な転送を行わないようにした。

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

© 2009 Information Processing Society of Japan

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 NSContrast の概要

NSContrast では、前小節で述べたニュース記事のデータベースに対して、以下の機能を提供し、ニュース記事の分析を支援する。

- 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。
- 各国のニュースサイト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。
- 時区間を分割した共起語解析による興味の変遷の分析
  国毎のニュースサイトの比較対照をするかわりに、記事を特定の時区間 (例えば、3 日ごと) で分割して、その期間の記事とそれ以外の記事を比較対照することにより、トピックの遷移を分析する。
- 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[7] 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。

2.4 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、トピック語を含むニュース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。牛肉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。

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。本システムでは、特徴語間の共起関係の強さ、各国のニュースサイト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 に先ほどの牛肉のトピックにおいて、日本でバーストしている 6/19〜23 日の記事を利用した分析結果を示す。

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。青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。赤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、同じ牛肉のトピックでも、日本の記事は食肉偽装の記事が多いのに対し、韓国では、米国の輸入牛肉の問題に関するデモの議論が多い事が分かる。

© 2009 Information Processing Society of Japan

Vol.2009-DD-71 No.1
2009/6/5

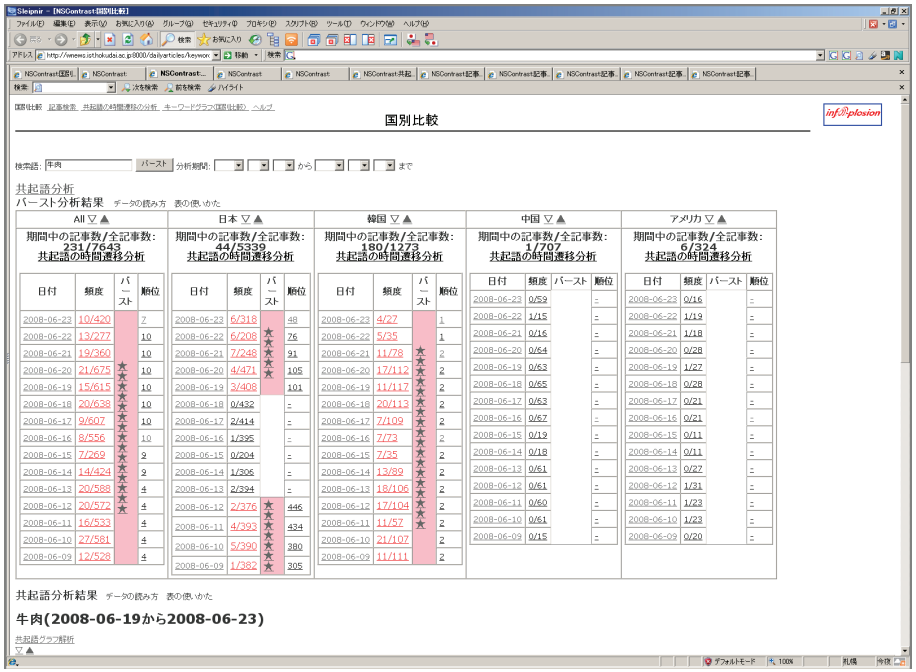

図 2　バースト解析結果の国別比較

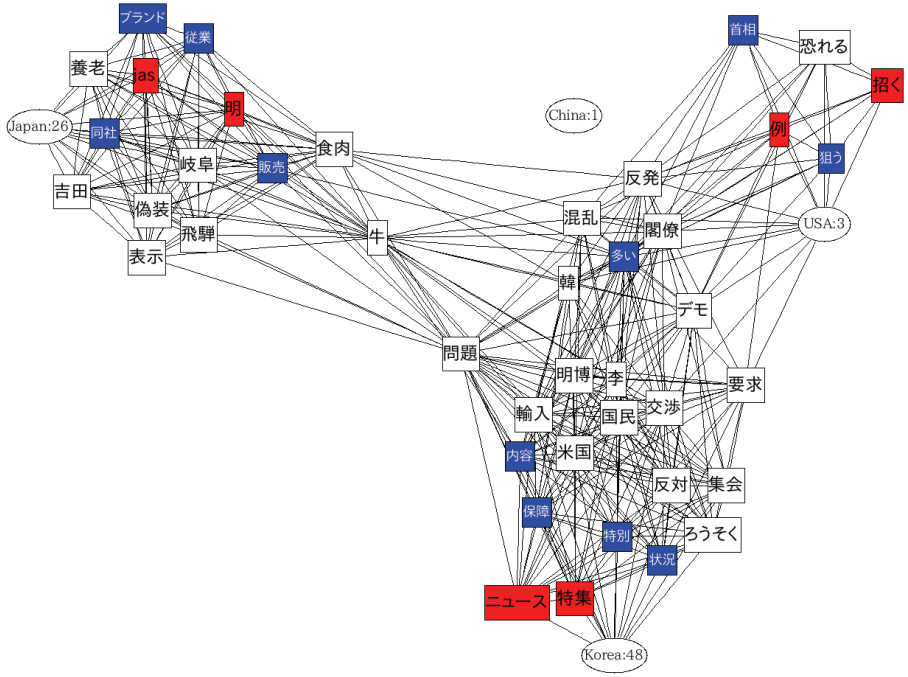

図 3　特徴的な共起語間の関係グラフ

© 2009 Information Processing Society of Japan

Vol.2009-DD-71 No.1
2009/6/5

3. 時区間を分割した共起語解析による興味の変遷の分析

さらに、6/9-23 日の記事を 3 日ごとに分割し、興味の変遷を分析する。図 4 に牛肉に関する韓国の記事の特徴語の時間遷移の分析結果を示す。各時区間におけるメジャーな共起語は、米国、輸入など、全ての期間でほぼ同じであるが、相関性の変化の高いキーワードをベースに記事検索を行うと、

- 9 日から 11 日はデモに関連した兵役義務の話
- 15 日から 17 日は自動車貿易との関係
- 18 日から 20 日は、狂牛病による死亡事例の放送が嘘であった事が判明
- 21 日から 23 日は、江省の進展具合いに注目

といった興味の変遷を示すキーワードを抽出することができた。

図 4 時区間分割による共起語の時間遷移の分析

3. NSContrast の評価

NSContrast のようなテキストマイニングシステムでは、一般的なユーザによる短時間の操作による評価では十分ではない。本システムの有用性を検証するためには、ある程度の背景知識を持ったユーザに対して、本システムの利用が想定されるような課題 (異なる国の報道の違いを分析) を与えた上でシステムの機能についての評価を行うことが適切であると考えた。

よって、各国間の報道の違いに関する知識を持っていると考えられる日本に在住する日中韓の少なくとも 2ヶ国以上の言語に精通している外国人の方に、システムを利用してレポートの作成課題を実行していただいた。

3.1 ユーザの作業方針

各ユーザに対しては、作業とその記録に関して以下の方針を与え、各国のニュースサイトの報道の違いに関するレポートの作成を依頼した。

(1) トピックに対する報道の違いに関する詳細については、以下の情報を記述してください。
  (a) 日本・韓国・中国における報道の違いについて気づいた事
  (b) 個別のニュースサイトや日本語版と韓国語版の違いなど、ニュースサイトによる違い (任意)

(2) 上記の違いを気づくために利用した情報

  (a) どの記事を参照したのか？
  (b) 報道の違いを特徴づけるようなキーワードはあるか？キーワードがあったとすると、どういうキーワードか？
  (c) システムの機能は、報道の違いを発見するのに役立ったのか？役立ったとすると、どの機能が役に立ったのか？
  (d) 他のシステムを利用した場合には、どのようなものを利用したか？
  現地語のニュースサイト自体を検索するシステムは用意していません。朝鮮日報には検索システムがあるようです。新華社 (中国) の場合は、Google で「キーワード ＋ site:news.xinhuanet.com」 で検索をすると新華社 (中国) のサイトにおける検索を行うことができます。翻訳システムの出力間違いで、特定の単語に関連するトピックが、全然扱われていないようにシステムが判断している場合があります。もし、不自然な結果のように思えた場合には、現地語のデータを確認してみてください。

(3) システムに対する評価
  (a) システムが表示したもので役に立った情報があれば、その情報
  (b) 出力結果が役に立たなかったもの

© 2009 Information Processing Society of Japan

また、追加の情報として、システム周りに関する以下のコメントを収集した。

(1) データの問題 (本文部分の切出ミスなどがあれば、教えてもらいたい)
(2) 翻訳システムの問題 (上記であげたような分析の問題になるような誤訳などがあれば、そのリストなど)

3.2 分析課題の設定

ある程度の記事数が必要なこと、報道の差があることが想定されることなどを前提にして、以下のようなカテゴリー毎に分析課題を設定することとした。

(1) 国際的なトピック
例：サミット、オリンピック、金融危機 (サブプライム、為替、株価)、北朝鮮
(2) 韓国中心の話題
例：李明博大統領、狂牛病、サムスン
(3) 中国中心の話題
例：四川大地震、チベット、台湾
(4) 日本中心の話題
例：麻生太郎、ニート
(5) 日韓の話題
例：靖国、領海、竹島
(6) 日中の話題
例：食物汚染 (残留農薬・メラミン)
(7) 韓中の話題

3.3 分析結果と評価

前節で述べた分析課題を考慮して、被験者には、以下の共通課題と、自由課題を与えて、分析レポートの作成、その作業過程の記録、システムへのコメントを得た。

- 共通課題
  金融危機・北朝鮮・オバマ大統領、麻生太郎
- 自由課題
  WBC、円高、おくりびと、失業

各々の被験者の作成したレポートから、各国のニュースサイト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た。また、共通課題を設定したものの、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」という共通課題からは、「通貨」、「米大統領」、「AIG」などに関する分析など、被験者の興味に応じて、分析の切り口が変わってくることが確認された。一方、「北朝鮮」の課題では、ちょうど、分析の実験を行う時期がミサイル (人工衛星) の発射が話題になっていたこともあり、だいたい同じような興味で分析が行われていた。

このようにユーザごとの分析結果の違いが起きる原因についてユーザからの意見をもらいながら分析したところ、分析課題について、大きく次の二つに分類があることが確認された。

事象分析 特定の事故、災害、国際的な出来事が起きた場合に、その事象がどのように報道されているかを分析する。
分析対象となる事象に関する記事数の差や、取り扱う観点の違いなども明確であり、ユーザごとの分析結果の違いが少なくなることが考えられる。

サーベイ分析 主に、長期的な記事群に対して、多観点から記述されている記事群の量的な違いなどを分析する。
長期間の記事を分析する場合には、全てのニュースサイトにおいて、網羅的な観点から記事が掲載されることから、その中のどの観点に注目して分析するかは、ユーザの興味に依存するため、ユーザごとの分析結果が違う場合が考えられる。

ただし、この二つの分析課題は、排他的なものではなく、「金融危機」などの課題に対しても、「リーマンショック」の時期に限定して分析を行うと、事象分析的な側面が強くなり、「北朝鮮」についても、長期の記事を用いて分析すると、サーベイ分析の側面が強くなると考えられる。

次回以降の実験を行う際には、この二つを区別する必要があると考えている。特に、サーベイ分析を行う場合には、分析の候補となる多観点のうち、興味を持った観点に注目することが多いため、観点に関する制約を与える、あるいは、観点の候補のリストを作るといった作業を行う必要があると考えている。

また、分析に用いる方法としても、事象分析には、ニュースアグリゲーションサイトや EMM News Explorer などで用いられている単純な記事数などをベースにした比較が有効であるが、サーベイ分析には、内容分析[8] のようなアプローチを適用していく必要があると考えている。

最後に、システムに関するコメントを以下に紹介する。

- 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

© 2009 Information Processing Society of Japan

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。
- 共起語解析の問題

  共起語解析に利用する単語に人や企業といった固有名に限定するといった制約をかけられる方が良い。
- 記事の構造の利用

  タイトルや見出しに出てくる単語に注目して分析を可能にした方が良い。

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ニュース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。

また、今後は、記事の構造の利用や共起語解析に関するコメントを考慮したシステムの改良も並行して行っていきたい。

4. 結　　言

本稿では、これまでに提案してきた、ニュース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、このようなテキストマイニングシステムを評価する方法についての検討結果とその実験結果について報告した。

この実験結果から、本システムで扱うような分析タスクには、世界的な事故や出来事を扱うような事象分析のタスクと多観点から記述された記事群を比較して分析を行うようなサーベイ分析のタスクがあることを確認すると共に、このタスクの違いがユーザごとの分析結果の違いの大きさに影響を与えることが考察された。今後の分析実験の課題設定の際には、この違いを考慮した課題設計をする必要がある。

また、ユーザから得られたコメントをベースにしたシステムの改良を行うと共に、更なる分析実験の設計と施行を繰り返し、システムの改良と有効性の検証を行っ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。また、被験者実験に協力していただいた皆様から、有益なコメントを頂いた。ここに記して謝意をあらわす。

参 考 文 献

1) 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).
2) 吉岡真治：多言語ニュースの対照分析のための Wikipedia 活用手法の研究，2009 年度人工知能学会全国大会講演論文集 (2009). (to appear).
3) 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).
4) Erdmann, M., KotaroNakayama, T.H. and Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia, *Proc. of International Conference on Database Systems for Advanced Applications (DASFAA) (Mar. 2008)*, Springer-Verlag GmbH, pp.686–689 (2008). LNCS 4947.
5) Bramantoro, A., Tanaka, M., Murakami, Y., Schäfer, U. and Ishida, T.: A Hybrid Integrated Architecture for Language Service Composition, *ICWS '08: Proceedings of the 2008 IEEE International Conference on Web Services*, Washington, DC, USA, IEEE Computer Society, pp.345–352 (2008).
6) 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.
7) Kleinberg, J.: Bursty and hierarchical structure in streams, *Proceedings of the 8th*

© 2009 Information Processing Society of Japan

Vol.2009-DD-71 No.1
2009/6/5

ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, New York, NY, USA, ACM Press, pp.91–101 (2002).

8) Krippendorff, K.: *CONTENT ANALYSIS: An Introduction to Its Methodology*, Sage Publication (1980). (三上俊治, 椎野信雄, 橋元良明 (訳) メッセージ分析の技法「内容分析」への招待, 勁草書房, 1989).

© 2009 Information Processing Society of Japan