IPA

2008 年度上期未踏 IT 人材発掘・育成事業(未踏ユース)採択案件評価書

1. 担当PM

筧　捷彦 PM(早稲田大学　基幹理工学部　情報理工学科　教授)

2. 採択者氏名

チーフクリエータ:　川場　真理子(筑波大学大学院システム情報工学研究科・博士前期課程 2 年)
コクリエータ:　中崎　寛之(筑波大学大学院システム情報工学研究科・博士前期課程 1 年)

3. プロジェクト管理組織

株式会社 メルコホールディングス

4. 委託金支払額

2,995,161 円

5. テーマ名

多言語ブログにおける文化間ギャップ発見システム

6. 関連Webサイト

なし

7. テーマ概要

近年、多くの人が海外に出かけ、また多くの外国人が日本を訪れるようになった。また、それに伴って、異文化交流の機会も増えてきた。そのような際に、自分の知って

いる知識や考えと相手が知っている知識や考えのギャップに驚く人も多いのではないだろうか。

従来、何かのトピックにおける海外での意見や情報は、海外の情報を買ってきたり、特派員を派遣しなければ得られないものであった。しかし、インターネットの爆発的普及により、世界中の情報を日本にいながら得られるようになった。また、ブログや SNS、Wikipedia に代表される、Web2.0 的コンテンツの登場により、多くの人々が手軽にコンテンツを作成することが可能になった。特にブログは世界中の人によって書かれ、多くの意見や情報が日々更新されている。ブログから同一トピックにおける日本と海外の比較を行うためには、英語と日本語両方の言語のブログを読む必要がある。しかし、両言語でトピックについて特徴的記事が書かれたブログを取得するのは困難であり、またどのような観点でブログを読むべきかを明確にしておく必要がある。そこで、我々は、ユーザがトピックを入力すると、日本語と英語両方のブログから関連するキーワードや文を取得し、ユーザに提示するシステムを作成する。関連するキーワードや文章があれば、特徴的な意見はどのようなものか、どのような観点でブログが書かれているか、などを知る手助けが出来る。

たとえば、「遺伝子組み換え食品」について検索したとする。そのとき、検索結果のウェブページだけでなく「加工食品」「健康」「農薬」「品質管理」「体に悪い」などのキーワードや「どういう影響が出てくるか分からないから怖い」といった文章がまとめて表示される。同様に英語で「Genetically modified organism」に関連する「cloned food」や「bio-technology」「environmental」「genetic engineering」などのキーワードや「In the U.S., the use of GM crops is already widespread. As new discoveries are made, bio-engineers could be the world's first-line defense against hunger」のような文章が表示される。

すると、日本では遺伝子組み換え食品に対して否定的な意見が多く、外国では否定だけでなく、食料危機の問題に対する遺伝子組み換え技術への期待などの肯定的な意見も見られる、ということが分かる。

そして、これらのキーワードや文章を比較することで、日本語と英語で遺伝子組み換え食品に対して、どのような意見の違いがあるのかを発見することが出来る。

我々はユーザが何かトピックについて検索すると、そのトピックに特徴的なキーワードや段落などを抽出し、日本語と英語両方を一度に提示するシステムを作成する。

日本と海外の意見を得るための情報源としては、ニュースやウェブページなど様々な物が考えられるが、今回は、日々増え続け、かつ主観的な意見、客観的な考察の両方が多く記述される、ブログを使用する。

また、関連するキーワードなどを抜き出したり、段落を抽出したりするためにはあらかじめトピックにおけるブログを検索しておかなければならない。

そのためトピックのリストをあらかじめ用意しておく必要があるが、人間の思いつく限り

の細かさであらゆる分野を網羅してあり、さらに、トピックが整理され、体系化されている必要がある。
また、日本語と英語で両方の訳を得られるものが良い。よって、トピックのリストとして Wikipedia のカテゴリ体系を利用する。
Wikipedia は世界中で利用される Web 百科事典として有名であり、現在日本語で 40 万記事、英語で 200 万記事ある。
また、日々増え続けているために新語などにも対応できるという利点もある。

8. 採択理由

プロジェクトのタイトルは、なかなか意味深長というべきか、曖昧性が高いというべきか、いろいろにとれてしまうものである。実際は、同じテーマを扱っている日本語 vs.英語のウェブ文書を材料として「文化間ギャップ」を調べるためのシステムを作る、というのがその本題である。このときウェブ文書が日本語と英語のものを比較しているものの、比較されているものは、日本での認識と、他国（アメリカとかイギリスとかに限らず、中国とか韓国とか）での認識との違いである。
すでに提案者は、さまざまな形で下調べをし、データを集めている。それらを元にして、「文化間ギャップ」を的確に比較し発見するシステムを組み立てたい、という。システムを作ることに関しては、これからやらなければならないことがたくさんある。計画は、材料となるウェブデータとして Wikipedia を使うことにして、一つのタイトルに対して、そこに書かれている内容から関連キーワードを抽出するとともに、ブログサイト中からそのタイトルを含むもの検索してきて同様に関連キーワードを抽出する。これら抽出された関連キーワードの集合に対して、言語間での翻訳対応を調べることで、差異を検出しよう、というのである。
手作業で行ってきたことをシステム化するとはいっても、現実的には、「同じテーマ」を扱っていることの判定、違いが生じていることの自動的なまとめ方など、いろいろとシステムとしての工夫が必要になる。なんといっても、提案者自らがそのツールを必要としているし、その上でやりたいことを山ほどもって目を輝かしているので、そうした工夫を重ねて新しい「発見」を助けてくれるツールが、便利に仕上がってくるに違いない。

9. 開発目標

ユーザの入力するトピックに対して，文化間のギャップを発見する手がかりとなるような，キーワードや文章を提示するシステムを開発することを目標とする。
ユーザが使いやすいシステムにするためには，トピックの選択が容易にできること，

また，文化間ギャップのあるキーワードをすぐに探し当てることができることの 2 点があげられる。そこで，本システムは，トピックを分類し，文化間での異なりがどの程度あるのかを発見しやすい検索システムの製作を目指した。

10. 進捗概要

文化間ギャップに関心をもつユーザに，文化間ギャップを発見する手がかりとなる情報を提示するシステムを作りたい，という目的は明確であった。また，Wikipedia には，それぞれの言語（国）によるものがあるが，多くのトピックついては，どの言語によるものも解説を上げている。したがって，基本的な用語などの対比を見るのに Wikipedia が利用すればよかろう。それらの用語を検索語として検索をすると，対応する用語を含んだそれぞれの言語でのブログが集まってくる。これらを適宜さばいてやれば，日本語圏，英語圏でのそれらの用語の扱い方の異同が見えるに違いない。こうしたアイディアももっていた。

しかしながら，それらをどういく形でシステムとすれば，利用者にとって，文化間ギャップを見つけるのに役に立つ手がかりを得るのに便利になるか，という構想をまとめ，システムにする作業にはなかなかはじまらなかった。提案時には，すでに日英の対比について，それなりの作業が進んでいたし，文化間ギャップが（手作業によって）見つかった例も示したので，開発者自らがこうした文化間ギャップ探しに興味（趣味）をもち，自ら提案する形の支援システムを使ってその興味を満たしたいという気持ちが強いと見たのだが，少し予想がはずれたようである。

プロジェクトレビューを何度か行ったが，なかなかことが進まなかった。

最終的には，目標とした形のシステムを作り上げることができたが，ついにそれを使ってこんな面白い文化間ギャップが見つかりました，という報告はなかった。残念である。不特定多数のためにこんなシステムがあればいい，という話はいろいろある。しかし，それだけでいいシステム作りをするのはなかなか難しい。適切が利用者がいてくれること，これが鍵の一つである。適切な利用者は，自分自身であったかまわないし，多くの場合，自分自身が利用者である（利用したがっている）ことが重要である。このあたりのインセンティブが少し低かったように思える。

11. 成果

ユーザが入力したトピックに対して，ブログから得られた共起語をキーワードとして提示し，またその共起語に関連するブログ記事をランキングして提示するシステムを開発した。

このシステムでは，ユーザが文化間ギャップのあるトピックを探し出すために，トピッ

クをカテゴリ分けし，分野ごとにトピックを調べることを可能になっている。さらに，調べたいトピックが決まっている場合にすぐにそのトピックの情報を得られるように，トピックの検索を行う機能を実装し，また，共起語から調べたいトピックをたどることのできる共起語検索機能も実装されている。

共起語の提示に関しては，共起語マップを作成し，各共起語が日本語と英語のどちらで多く語られているか，どの程度の頻度で語られているかを，ユーザが直感的に知ることのできるように図表の形で示してくれる。

このシステムが対象とするトピックは，Wikipedia に日本語のエントリも英語のエントリもあるトピックに限定されている。これは，基本の情報源として Wikipedia を利用していることからくる制限である。

このシステムでは，さらに，扱うトピックを制限して，日本語，英語共にヒット数が 1 万から 50 万の範囲にある Wikipedia エントリ 6000 のうちから人手で選び出した約 200 トピックだけに限っている。これは，事前の調査から，エントリ名を検索トピックとしてブログサイト検索を行ったときに，検索ヒット数が 1 万～50 万の範囲のものに，多くのブログ記事がありそうなトピックが集中していることがわかっていたからである。しかし，システムの実装に必要なデータを 6000 トピック分収集するには時間が足りなかったので，人手で選定した約 200 トピックを対象とした。

また，Wikipedia エントリは，それぞれ親となるカテゴリを持つ。選定した 200 トピックについても，それを基に適宜手を加えた形でカテゴリ分けして提示することで利用者の便宜を図っている。このシステムの入り口では，そのカテゴリ分けを使って，特定のトピックを選び出す画面が開く。

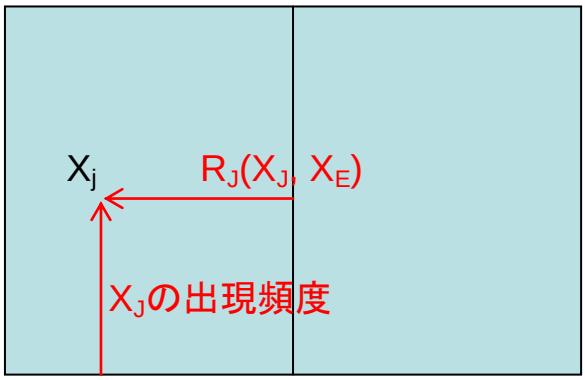
このシステムでは，トピックに限らず，日本語ブログ，英語ブログに現れる用語に基づいて種々の情報提供を行う。そのとき，システムは対応する日本語での用語と英語での用語との対を知っている必要がある。この対応する用語対を知るのに，Wikipedia に記録された言語間リンクと，オンライン辞書の英辞郎を使っている。

このシステムでは，選ばれたトピックの日本語・英語のそれぞれの単語を検索キーとして Yahoo!検索 API を使ってブログ記事を集めてくる。その上で，そのトピックに対する Wikipedia 記事でのキーワードが多く現れている順に上位 10 個のブログを選ぶ。そのそれぞれのブログ記事について，日本語の記事であれば形態素解析を行って名詞句を抽出し，一般語を取り除いた後で，頻度の高い方から 10 個を選んでその共起語とする。英語の記事では，1 単語，2 単語連語，3 単語連語を共起語としている。

これらの共起語について，つぎの値を求める。日本語の共起語 X_J に対しては，その日本語ブログにおける出現確率 $P_J(X_J)$ を求める。つぎに X_J に対応する英語の用語 X_E の英語ブログにおける出現確率 $P_E(X_E)$を求める。同様に，英語の共起語 Y_E に対しても英語での出現確率 $P_E(Y_E)$と対応日本語 Y_J の日本語での出現確率 $P_J(Y_J)$を求める。その上で，次の出現確率比を求めた。

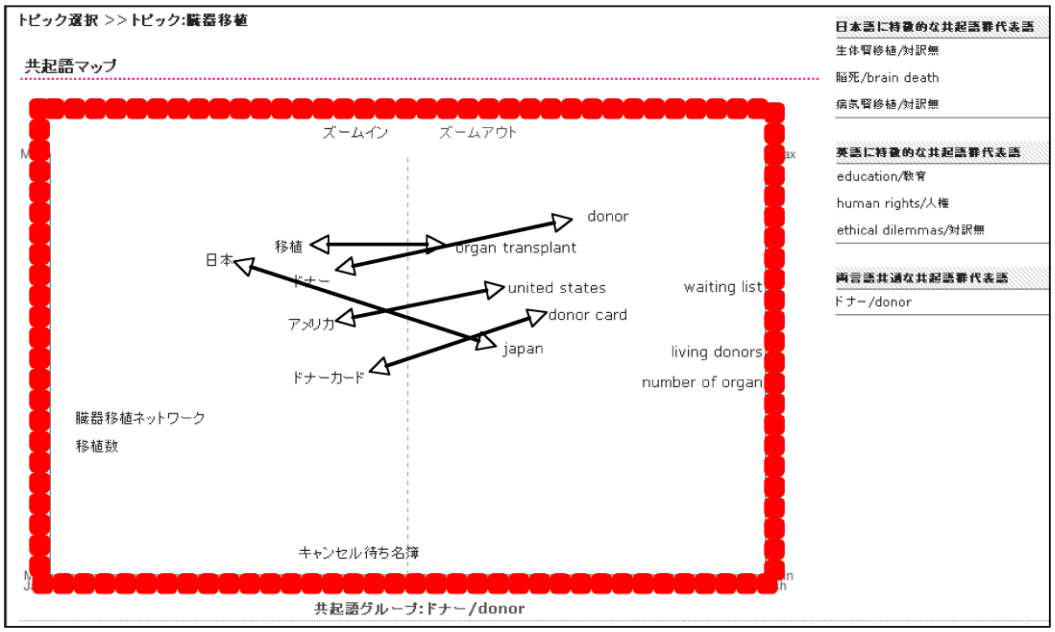
$$R_J(X_J, X_E) = P_J(X_J) / P_E(X_E), \quad R_E(Y_E, Y_J) = P_E(Y_E) / P_J(Y_J)$$

この出現確率比を使って，共起語（日本語）をつぎの図のように位置づける。



つまり，日本語 X_J は，出現頻度が高いほど上に位置し，出現確率比が高いほど左に位置するように配置するのである。言い換えれば，上にあるほど日本語ブログに多く現れていることを示し，左にあるほど日本語ブログに特徴的に現れていることになる。右半分には英語の共起語について同様の方式で配置する。

　このように，そのトピックに関する日本語ブログでの共起語の状況が左半分に，英語ブログでの共起語の状況が右半分に，それぞれ表示される。



これを見て，日本語・英語のブログでの共起語の状況を対比してみて，何か感じるものがあれば，その共起語を直接クリックすると，その共起語を内在させているブログのリストが開く。このリストには，共起語の出現回数の多い方から順にブログへのリンクが並んでいて，クリックすることで当該のブロクが開く。

現在のデータは約 200 トピックと大変少ない。また，共起語マップに関してもノイズが多い部分もある。そこで，今後はトピック数を増やしつつ，共起語マップの性能を上げる予定であるという。さらに，インタフェースを通じて多くの人が意見を出し合えるよう，コメント欄，Wiki 等の機能をインタフェースに実装する予定である。
システムは 2009 年度 3 月末をめどにウェブ上に公開する予定である。また，公開後はユーザからのフィードバックを元によりよいシステムとすべく，改良を加えていく予定であるという。

12. プロジェクト評価

プロジェクトの計画書に書いた事柄は，一通りのものができた。しかしながら，そのでき栄えは，まだ，プロトタイプを作ってみました，という段階にある。なにより，ブログという，時々刻々世界中で書込みがおこっている動的なものを対象としているだけに，

あらかじめシステム作成時に組み込まれたデータだけで動く，というシステムの設計は適切ではなかろう。オープンエンデッドになっている必要がある。

また，ぜひ自らこのシステムの利用者になって，“あ，こんな文化間ギャップが見つかった”“こんなのも見つかった”と楽しまない限りは，改善していく意欲もわかないだろう。ぜひぜひ，システム公開時には，そのホームページに，このシステムを使って面白い文化間ギャップをいくつか例示してあるようにしてほしい。

13. 今後の課題

開発者が自らその報告書に書いているように，このシステムを一般公開してほしい。それには，なにより多くのトピックが提供できるようにデータの増加を図らなければユーザが多くはつかないだろうし，道具として何度も使ってくれるようにならないだろう。そう考えると，さらには，システムの作りを変えて，時々刻々データを自動的に取り込む機能をつくるなり，現時点での状況を素早くデータに加えて表示するようにするなりのことを行う必要があるだろう。

そして，何より，開発者自身がこのシステムを使い込んで，面白い文化間ギャップを見つけてみて欲しい。それがシステムの改善を行うきっかけになるし，収録してあるトピックの件数を増大したくなるきっかけになる。そして，いずれは多くのユーザを獲得できることにもつながるに違いない。