# 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析*

聶 添[†] 守谷 一朗[†] 井上 祐輔[†] 今田 貴和[†] 李 雪山[†]
宇津呂 武仁[†] 河田 容英[‡] 神門 典子[§]
筑波大学大学院 システム情報工学研究科[†] (株) ログワークス[‡] 国立情報学研究所[§]

## 1 はじめに

21 世紀の情報社会では，政府機関や企業などにとって，グローバル化により，自国の情報だけではなく他国の情報も重要となっている．近年のインターネットの普及により，非常に多くの人がウェブサイトを閲覧して情報を収集している．そうしたウェブ閲覧者の多くは，自らの関心事項について，Google，Yahoo!，Baidu といった検索エンジンを用いてウェブ検索を行っている．ここで，ウェブ検索者・ウェブ閲覧者が，検索エンジンを用いて他国の情報を得ることはそれほど容易なことではない．そこで，本研究においては，ウェブ検索者の関心事項に着目することにより，ウェブ上の情報を多言語 (日本語・中国語) 間で比較・対照分析し，他国の情報の収集を支援するとともに，言語間の差異発見の過程を支援するアプローチをとる．本論文では，特に，文献 [3] において質問回答事例，および，ウェブから収集したノウハウ知識に対して，日中間で比較対照分析を行う手法を提案する．

本論文の全体の流れを図 1 に示す．文献 [3] の手法においては，まず，質問回答サイトから収集した質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページの混合文書集合に対してトピックモデルを適用することにより，話題のまとまりを生成する．次に各話題を「ノウハウ知識」,「ノウハウ以外の知識」,「意見」,「その他」の 4 つに分類することで，ノウハウ知識を選定する．本論文では,「就活」および「結婚」を検索対象として収集されたノウハウ知識に対して日中間で比較対照分析を行った．その結果，検索対象「就活」において，面接に関して，日本特有のノウハウ知識として「敬語の使い方」,「就活メイク」，中国特有のノウハウ知識として「就活面接ショー番組」，日中共通のノウハウ知識として「就活生の服装」等が収集された．

## 2 質問回答事例の収集

日本側の質問回答事例のデータとして，Yahoo!知恵袋[1] から提供されている 2004 年 4 月 1 日 ~2009 年 4 月 7 日の 5 年間の質問回答事例のデータ (質問: 16,257,413 件，回答: 50,053,894 件) を用いた．本論文では，カテゴリ名，質問タイトル，質問本文のいずれかに検索対象 $q$ が含まれている質問を抽出し，その質問に対する回答本文全てを結合し，一つの質問回答事例 $d_q$ を作成した．各検索対象 $q$ あたりの質問回答事例の文書集合を $D_q = \{d_q^1, \ldots, d_q^k\}$ と定義する．

一方，中国側の質問回答事例は，2014 年 11 月 ~2015 年 1 月の期間に，Baidu(百度) 知道[2] から収集した．2015 年 1 月の時点で，Baidu 知道に掲載されていた解決済質問数は 354,412,701 件であった．

各検索対象において，日中それぞれの言語において収集した質問回答事例の数を表 1 に示す．

## 3 検索エンジン・サジェストを用いたウェブページの収集

各検索エンジン会社においては，ウェブ検索者の検索ログが蓄積されており，多数のウェブ検索者が検索したキーワードに対して，検索者が強い関心を持つ語を抽出し，検索エンジン・サジェストとして提示するサービスを提供している．ここで，検索エンジン・サジェストとして提示される語は，検索対象に対して，多数のウェブ検索者が AND 検索の形で二つ目以降に入力した語を情報源として抽出されたものである．そこで，本論文では，検索エンジン・サジェストには，ウェブ検索者の関心事項そのものが反映されていると考え，

*Comparative Analysis of Know-How Knowledge collected from Question Answer Examples and Web between Japanese and Chinese

†Tian Nie, Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Xueshan LI, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Yasuhide Kawata, Logworks Co., Ltd.
§Noriko Kando, National Institute of Informatics

[1] http://chiebukuro.yahoo.co.jp/
[2] http://zhidao.baidu.com/

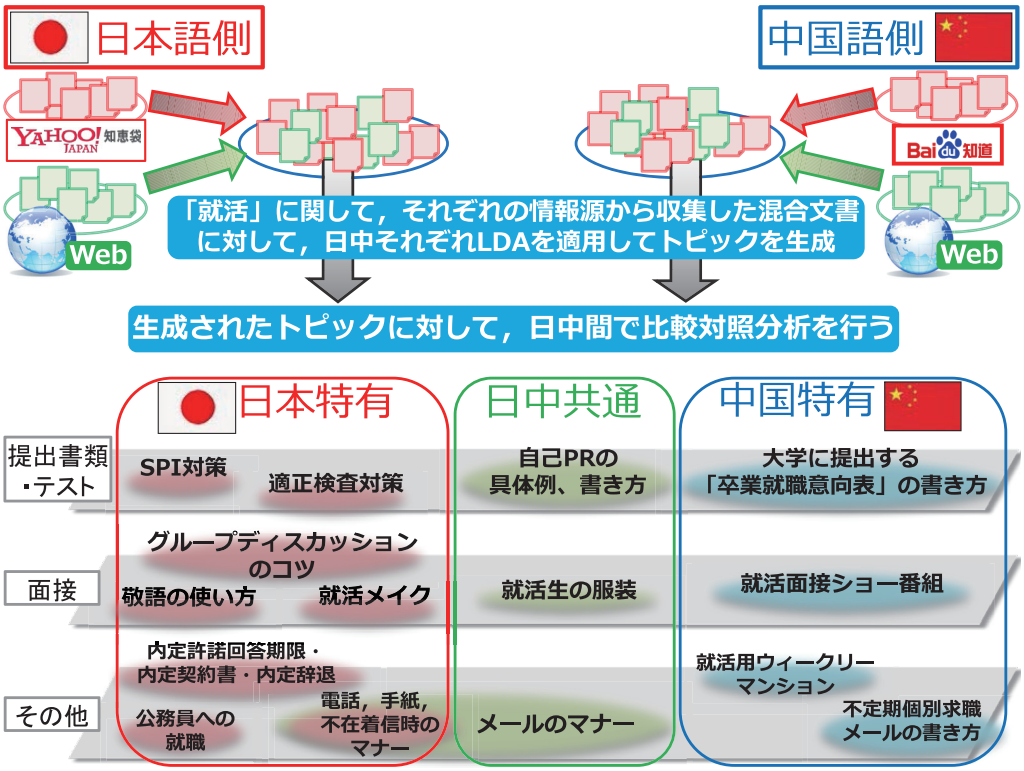


図 1: 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析の流れ (検索対象：「就活」の抜粋)

ウェブ検索者の関心事項を収集する目的で，検索エンジン・サジェストを収集する．

日本語側においては，Google[3]検索エンジンに対して，一検索対象当り 100 通りの文字列を指定し，最大 1,000 語のサジェストを収集する．100 通りの文字列とは具体的には，五十音，濁音，半濁音及び「きゃ」や「びゃ」などの開拗音である．例えば検索窓に「就活ね」と入力すると,「ネクタイ」や「ネクタイ 結び方」などがサジェストとして掲示されるので，それらを収集することにより，934 個のサジェストを収集した．

中国語側においては，Google 検索エンジンに対して，一検索対象当り 28 通りの文字列を指定し，最大 280 語のサジェストを収集する．28 通りの文字列とは具体的には，中国語のピン音の部首である．例えば検索窓に「求职(就活) j」と入力すると,「简历(エントリーシート)」などがサジェストとして掲示されるので，それらを収集することにより，209 個のサジェストを収集した．

各検索対象において，日中それぞれの言語において収集したサジェストの数を表 1 に示す．

ここで，ある検索対象に対して収集されたサジェストの集合を $\mathbb{S}$ とすると，$s \in \mathbb{S}$ となるサジェスト $s$ に対して，検索対象との AND 検索により上位 $N$ 件以内に検索されるウェブページ $p$ の集合を $\mathbb{P}(s, N)$

表 1: 各検索対象におけるサジェスト数，および，混合文書集合の記事数

| 検索対象 | 言語 | 質問回答事例数 | ウェブページ | | 質問回答事例数＋ウェブページ数 |
|---|---|---|---|---|---|
| | | | サジェスト数 | ページ数 | |
| 就活 | 日本語 | 11,366 | 934 | 13,221 | 24,587 |
| | 中国語 | 754 | 209 | 3,054 | 3,808 |
| 結婚 | 日本語 | 35,426 | 956 | 14,409 | 49,835 |
| | 中国語 | 753 | 248 | 4,085 | 4,138 |

(ただし，本論文においては，$N = 20$ とする) とし，各検索対象あたりのウェブページの文書集合 $D_w$ を $D_w = \bigcup_{s \in \mathbb{S}} \mathbb{P}(s, N)$ と定義する．なお，ウェブページの収集には Yahoo! Search BOSS API [4] を用いた．

## 4 トピックモデルの適用

2 節および 3 節で収集した質問回答事例の文書集合 $D_q$ とウェブページの文書集合 $D_w$ の混合文書集合 $D_{qw} = D_q \bigcup D_w$ を作成する．である．各検索対象における混合文書集合の記事数を表 1 に示している．本論文では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [1] を用いる．LDA を用いたトピックモデルの推定においては，語 $w$ の集合を $V$ として，語 $w(w \in V)$ の列によって表現

[3] https://www.google.com/

[4] http://developer.yahoo.com/search/boss

表 2: ノウハウ知識の話題数

| 検索対象 | 大分類の数 | トピック数 (ノウハウ知識/LDA 適用時) | | 話題数 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 日本語側 | 中国語側 | 日本特有 | 中国特有 | 日中共通 | 合計 | |
| | | | | | | | 日本語側 | 中国語側 |
| 就活 | 6 | 33/50 | 24/30 | 40 | 15 | 日:10，中:16 | 49 | 31 |
| 結婚 | 4 | 26/50 | 25/30 | 24 | 19 | 日:11，中:12 | 35 | 31 |

された文書の集合と，トピック数 $K$ を入力として，各トピック $z_n$ $(n = 1, \ldots, K)$ における語 $w$ の確率分布 $P(w|z_n)$ $(w \in V)$，及び，各文書 $b$ におけるトピック $z_n$ の確率分布 $P(z_n|b)$ $(n = 1, \ldots, K)$ を推定する[5]．本研究では，各文書に対して確率が最大のトピックを一意に割り当てることにより，各文書を分類することとした．記事集合を $D$，トピック数を $K$，1 つの文書を $d\,(d \in D)$ とすると，トピック $z_n\,(n = 1, \ldots, K)$ の記事集合 $D(z_n)$ は以下の式で表される．

$$D(z_n) \quad = \quad \left\{ d \in D \;\middle|\; z_n = \underset{z_u\ (u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

## 5　ノウハウ知識の収集

4 節の手順に従い，各トピックに割り当てられた確率上位 20 件の記事を分析したところ，トピックによっては，いずれかの情報源に偏るものがあった．そこで，今回の分析では，情報源ごとに確率上位 10 件の記事を分析し，そのうち 3 件以上同一とされる話題があった場合に，そのトピックの話題として抽出した[6]．これにより各トピックの情報源毎に最大 3 つの話題を抽出した．なお，話題分析の際には，各トピックにおける確率 $P(w|z_n)$ の高い語 $w$ とトピック及びウェブページに割り当てられたサジェストを参照して分析を行う．

次に，各トピックから得られた各話題を 1) ノウハウ知識，2) ノウハウ以外の知識，3) 意見，4) その他，の 4 種類に分類する．このうち，「ノウハウ知識」は，やり方についての情報など閲覧した人の行動につながるものである．具体的にはレシピサイト，方法や手順が書かれているもの，対策やマナー，コツなどがノウハウ知識にあたる．本論文では，ユーザの行動につながる知識は全てノウハウ知識であるとみなした．収集されたノウハウ知識の話題数を表 2 に示す．「ノウハウ以外の知識」は，それを見てもユーザの行動に影響を与えない情報である．例えば，「芸能人の結婚」がこれにあたる．「意見」は，多くの人の意見を求める相談や，自分の意見を主張しているものである．例えば，「就活中の友人，恋人とのつきあいかたについて」や「結婚後の嫁姑の問題」がこれにあたる．「その他」は，上記 3 つのいずれにも分類できないものである．例えば，「結婚占い」がこれにあたる．

## 6　ノウハウ知識の日中間対照分析

前節で収集されたノウハウ知識に対して，日中間で比較対照分析を行ったところ，各検索対象における，日本特有のノウハウ知識の数，中国特有のノウハウ知識の数，日中共通のノウハウ知識の数はそれぞれ表 2 に示す結果となった[7]．

検索対象「就活」においては，図 1 に示すように，日本特有のノウハウ知識として「SPI 対策」，「敬語の使い方」，「内定許諾回答期限・内定契約書・内定辞退」等 40 個のノウハウ知識が収集された．中国特有のノウハウ知識としては，「大学に提出する卒業就職意向表の書き方」，「就活面接ショー番組」，「就活用ウィークリーマンション」等 15 個のノウハウ知識が収集された (図 1 における中国特有のノウハウ知識の詳細な説明を表 3 に示す)．日中共通のノウハウ知識としては，「自己 PR の具体例，書き方」，「就活生の服装」，「メールのマナー」等，日本語側 10 個，中国語側 16 個のノウハウ知識が収集された[8]．

また，検索対象「結婚」においては，図 2 に示すように，日本特有のノウハウ知識として「女性から男性への逆プロポーズ」，「結婚祝い電報の文例」，「配偶者贈与について」等 24 個のノウハウ知識が収集された．中国特有のノウハウ知識としては，「結婚前の健康診断」，「結婚式当日の新居でのいたずら」，「新郎から新婦への結婚保証書」等 19 個のノウハウ知識が収集さ

[5]推定のためのツールとしては，GibbsLDA++を用いた．LDA のハイパーパラメータである $\alpha$，$\beta$ としては，$\alpha = 50/K$，$\beta = 0.1$，Gibbs サンプリングの反復回数は 2,000 を用いた．

[6]異なるトピックから同一の話題が収集される場合においても，本論文の分析の範囲においては，別の話題として数えた．

[7]日本語側の「メール，電話，手紙，不在着信時のマナー」のノウハウ知識のうち「メールのマナー」の部分は，日中共通のノウハウ知識として，日中共通の話題のうちの日本語側 10 個のうちの 1 つとして数えた．一方，残りの「電話，手紙，不在着信時のマナー」の部分は日本特有のノウハウ知識 40 個のうちの 1 つとして数えた．このため，日本特有の 40 個と日中共通の日本語側の 10 個を加えた計 50 個が，合計欄の日本語側 49 個よりも 1 つ多くなっている．

[8]中国語側のノウハウ知識「エントリーシート」，「自己分析」，および，「面接の対策」は複数のトピックから収集されたため，日本語側に比べて中国語側の話題数が多くなっている．

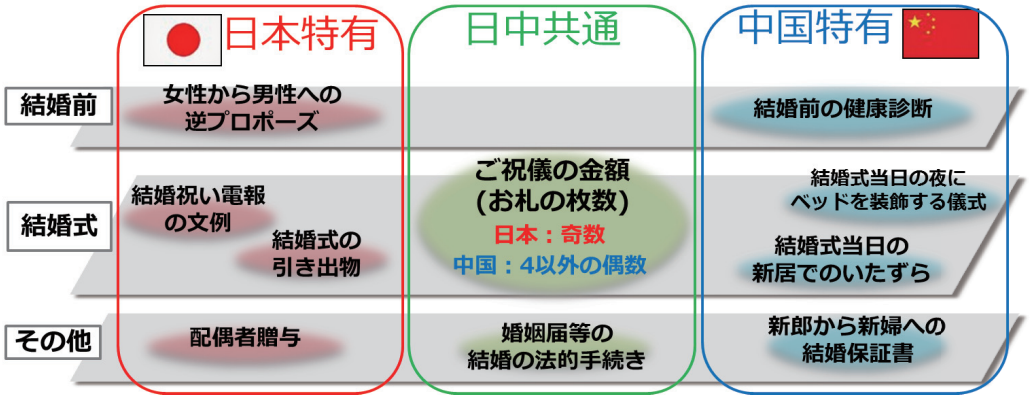


図 2: 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析の例 (検索対象：「結婚」の抜粋)

表 3: 中国特有のノウハウ知識の詳細説明

| 検索対象 | 話題 | 説明 |
|---|---|---|
| 就活 | 大学に提出する「卒業就職意向表」の書き方 | 就活生が自分の技術や就職希望企業について「卒業就職意向表」を記述して大学に提出する．大学側は企業による訪問面接の参加の斡旋をしてくれる． |
| | 就活面接ショー番組 | 就活生と企業がスタジオで実際に面接をしてその場で選考を行う番組 |
| | 就活用ウィークリーマンション | 主に就活生が利用する廉価型のウィークリーマンション |
| | 不定期個別求職メールの書き方 | 希望する企業に対して就活生が不定期の時期に自己アピールおよび選考依頼のメールを送る際のメールの書き方 |
| 結婚 | 結婚前の健康診断 | 中国では，結婚前に健康診断を受ける慣習がある． |
| | 結婚式当日の新居でのいたずら | 結婚式当日，親しい友人が新居に集まり新郎新婦にいたずらをする． |
| | 結婚式当日の夜にベッドを装飾する儀式 | 結婚式当日の夜に，赤い寝具を用いてベッドを装飾するという中国独特の儀式を行う． |
| | 新郎から新婦への結婚保証書 | 結婚の際，新郎が守るべき約束を結婚保証書に明記し渡す． |

れた (図 2 における中国特有のノウハウ知識の詳細な説明を表 3 に示す)．日中共通のノウハウ知識としては，「ご祝儀の金額 (お札の枚数)」，「婚姻届等の結婚の法的手続き」等，日本語側 11 個，中国語側 12 個のノウハウ知識が収集された[9]．

## 7 関連研究

文献 [5] においては，特定の話題について，日本語ブログ記事，および，中国語ブログ記事を収集し，日中両国の文化間差異の発見を支援する方式を提案している．しかし，ブログを情報源とする場合，日中両国の文化間の差異をウェブ検索者の視点から効率よく収集することが容易でないという問題があった．その他，文献 [4] においては，日中質問回答サイトを対象として，トラブル情報の比較対照分析を行い，日中両国の文化間の差異発見過程を支援する方式を提案している．また，文献 [2] においては，日中検索エンジン・サジェストを用いて，ウェブ検索者の関心事項に着目することにより，ウェブ上の情報から国・文化・言語間の差異発見過程を支援する方式を提案している．

[9]中国語側のノウハウ知識「ご祝儀の金額 (お札の枚数)」は複数のトピックから収集されたため，日本語側に比べて中国語側の話題数が多くなっている．

## 8 おわりに

本論文では，質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページに着目することにより，ウェブ上の情報を多言語 (日本語・中国語) 間で比較・対照分析し，他国の情報の収集を支援するとともに，言語間の差異発見の過程を支援する方式を提案した．

## 参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] 陳磊, 井上祐輔, 守谷一朗, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. トピックモデルを用いたウェブ検索者の関心の日中間対象分析. 言語処理学会第 21 回年次大会論文集, 2015.

[3] 守谷一朗, 井上祐輔, 今田貴和, 聶添, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集. 第 7 回 DEIM フォーラム論文集, 2015.

[4] 聶添, 新井翔太, 宇津呂武仁, 河田容英. 日中質問回答サイトの比較対照分析および文化間差異発見支援. 第 27 回人工知能学会全国大会論文集, 2013.

[5] 鄭立儀, 小池大地, 宇津呂武仁, 河田容英, 神門典子. 日中ブロガー・コミュニティの収集・俯瞰・対照分析. 情報処理学会研究報告, Vol. 2013–DBS–157/2013–IFAT–111, , 2013.