# 5-5　基盤的言語資源

## *5-5 Fundamental Language Resources*

橋本　力　　呉　鍾勲　　佐野大樹　　川田拓也
HASHIMOTO Chikara, Jong-Hoon Oh , SANO Motoki, and KAWADA Takuya

**要旨**

基盤的言語資源とは、質問応答システムや情報分析システム等の言語情報処理システムのいわばビルディングブロックであり、大きく分けて言語処理ツールと言語データの２つに分類できる。それら基盤的言語資源の中には、計算機資源やマンパワー、予算や時間的制約等の観点から、組織によっては構築が困難なものも多い。一方で、様々な言語情報処理システムの間で様々な言語資源が共有可能であり、コミュニティ全体として研究を着実に進展させるためには、それら共有可能な言語資源を多数構築、公開することが重要である。情報分析研究室ではこれまでに、大規模な並列計算環境と、自然言語処理に精通した多数の研究者、経験豊富な多数の言語アノテータにより、組織によっては構築が困難と思われるものも含めて、大規模、高精度な基盤的言語資源を多数構築、公開してきた。本稿では、未公開のものも含めて、情報分析研究室がこれまでに構築してきた基盤的言語資源を紹介する。なお、本特集号 **5-4** [1] で解説された言語処理ツールについては割愛する。

Fundamental language resources are classified into natural language processing tools and natural language data, which are used as building blocks for natural language information processing systems such as question answering systems and information analysis systems. Various kinds of natural language information processing systems generally have necessary fundamental language resources in common. However, some fundamental language resources are difficult to construct for some organizations due to limited computational capability, limited manpower, budget constraint, or time constraint. Thus, it is important to construct and publish such fundamental language resources in order for the research community to make steady progress. We, Information Analysis Laboratory members, have constructed and published many fundamental language resources that are precise and have wide-coverage, some of which are difficult to construct for some organizations, with a large-scale high-performance computing environment, many researchers who are acquainted with natural language processing, and many richly-experienced linguistic data annotators. In this paper, we present fundamental language resources that we have constructed, including those that will be released in the near future. We do not present natural language processing tools that have described in **5-4** of this special issue.

**[キーワード]**
言語資源，辞書，コーパス，言語処理ツール，ALAGIN フォーラム
Language resources, Dictionaries, Corpora, Language processing tools, ALAGIN Forum

## 1　はじめに

情報爆発時代と呼ばれる今日、ビッグデータから必要とする情報をピンポイントで探しあてる質問応答システムや情報分析システム等の言語情報処理システムがその重要性を一層増しているのは明らかである。このような言語情報処理システムは一般に高度な「言語理解」能力を必要とする。例えば質問応答システムでは、「河津川で釣れるのは何？」という質問に対して、「河津川で鮎解禁」や「河津川にオオウナギがいる」、「河津川のアマゴは美しい」等の、「河津川で釣れる」とは

直接記述されていない回答候補の文を大量文書から検出し、「鮎」や「オオウナギ」、「アマゴ」を回答として読み取れなくてはならない。人間は多くの様々な言語知識を元に文や文章を「解析」した上で言語を理解しているものと考えられるが、計算機が言語を理解する場合も多くの様々な言語知識（言語データ）と解析器（言語処理ツール）が必要である。本稿ではこのような言語データと言語処理ツールを総称して「基盤的言語資源」と呼ぶ。

一般に、高度な言語情報処理システムを構築する際に必要な、つまりビルディングブロックとして用いられる基盤的言語資源は多岐に渡り、かつ、個々の言語資源の構築には技術、経験、知識のみならず、大規模な計算機資源やマンパワー等の莫大なコストを要することが多い。従って、組織によっては必要な基盤的言語資源を全て自前で用意するのが困難であり、このことがコミュニティ全体としての研究の着実な進展の障壁となっている。

ユニバーサルコミュニケーション研究所情報分析研究室は、Web から収集した膨大な文書集合と大規模な並列計算環境、経験豊富な多数の言語データアノテーター、言語情報処理に精通する研究者を擁しており、高度な基盤的言語資源の構築・配信を誇る。情報分析研究室ではこれまでに、コミュニティ全体で研究を着実に進展させることを目的として、質問応答システムや情報分析システム等、多様な言語情報処理システムにとって重要で、かつ、構築に大きなコストのかかるものを含む数多くの基盤的言語資源を構築、公開してきた。

本稿では、本特集号 **5-4** [1] で解説された言語処理ツールを除く、情報分析研究室がこれまでに構築してきた基盤的言語資源を、未公開のものも含めて紹介する。

表 1 と表 2 に **3** 以降で紹介する基盤的言語資源の一覧を示す。表 1 にある基盤的言語資源は、**2** で述べる高度言語情報融合フォーラム ALAGIN[*1] を通して ALAGIN メンバー限定で公開しているものである。一方、表 2 にある基盤的言語資源は、フリーウェアとして一般に公開されているものである。「種別」欄にある DB、Service、Tool はそれぞれデータベース、Web サービス、ツールを表す。

＊1　http://alaginrc.nict.go.jp/

**表 1　言語資源一覧：ALAGIN 会員限定**

| 名称 | 公開 | 種別 | 規模 |
|---|---|---|---|
| 日本語パターン言い換えデータベース | 2009 年 | DB | 約 25 億件 |
| 動詞含意関係データベース | 2009 年 | DB | 約 12 万ペア |
| 負担・トラブル表現リスト | 2009 年 | DB | 約 2 万件 |
| 文脈類似語データベース | 2009 年 | DB | 約 100 万語 |
| 上位語階層データ | 2009 年 | DB | 約 7 万語 |
| 単語共起頻度データベース | 2009 年 | DB | 約 100 万語 |
| カスタム単語集合作成サポートサービス | 2010 年 | Service | — |
| 日本語係り受けデータベース | 2010 年 | DB | 約 46 億件 |
| 基本的意味関係の事例ベース | 2010 年 | DB | 約 10 万件 |
| 日本語異表記対データベース | 2010 年 | DB | 約 160 万件 |
| 意味的関係抽出サービス | 2011 年 | Service | — |
| 京都観光ブログの評価情報付与データ | 2011 年 | DB | 約 1,000 記事 |
| 述語フレーズ含意関係データベース | 2012 年度末 | DB | 約 60 万ペア |
| 活性／不活性データベース | 2012 年度末 | DB | 約 1 万件 |
| 述語フレーズ矛盾関係データベース | 2012 年度末 | DB | 約 100 万ペア |
| 述語フレーズ因果関係データベース | 2012 年度末 | DB | 約 100 万ペア |

表2 言語資源一覧：フリーウェア

| 名称 | 公開 | 種別 | ライセンス | 規模 |
|---|---|---|---|---|
| 日本語 WordNet | 2009 年 | DB | NICT 独自 | 約 9 万件 |
| 上位下位関係抽出ツール | 2010 年 | Tool | GPL | — |
| 日本語 Wikipedia エントリの係り受けデータベース | 2011 年 | DB | CC BY-SA 3.0 | 約 8 億件 |
| Para-SimString | 2012 年度中 | Tool | Modified BSD, LGPL, or GPL | — |
| QE4Solr | 2012 年度中 | Tool | Modified BSD, LGPL, or GPL | — |

## 2 高度言語情報融合フォーラム ALAGIN

高度言語情報融合フォーラム ALAGIN（Advanced LAnGuage Information Forum）は、言語の「壁」を感じさせないコミュニケーションを実現する、スーパーコミュニケーション技術の普及・促進を目的としたフォーラムである。平成 21 年の設立以降、民間企業、大学、研究機関及び国の関係者が集結して、テキスト／音声の翻訳、音声対話システム、適切に情報を検索する技術や信憑性判定を含めた情報分析技術、高度情報検索技術、ならびにこれらの技術の前提となる今までにない規模の言語資源（辞書、コーパスなど）の研究開発、実証実験・標準化等を行い、その成果たるツールや言語資源を広くフォーラムの会員に提供すべく活動している。

本稿で紹介する言語資源と本特集号の **5-4** [1] で解説された言語処理ツールは、フリーウェアも含めて、ALAGIN の言語資源配信サイト*2 から入手できる。

ALAGIN ではこの他にも、当機構ユニバーサルコミュニケーション研究所の多言語翻訳研究室と音声コミュニケーション研究室で開発、構築されたツールやデータ類の配信も行っている。

なお、ALAGIN のより詳細な活動内容や会員数等については、本特集号 **8-1** [2] を参照されたい。

## 3 体言の意味的関係データベース

### 3.1 基本的意味関係の事例ベース

「基本的意味関係の事例ベース」は、約 1 億ページの Web 文書上において文脈の類似度 [3] が高い 2 語間の意味的関係を人手で分類し、ラベル付けした結果を収録したもので、102,436 語対が収録されている。例えば、「電子計算機」と「電算機」などの略記対、「患部」と「治療部位」などの異形同義対などが収録されている。「基本的意味関係の事例ベース」で扱われている語句対の意味的関係の種類全てを表 3 に示す [4]。

表3 「基本的意味関係の事例ベース」における意味関係の分類

| 分類 | 例 |
|---|---|
| 異表記対 | 問い合わせ／問合せ |
| 略記対 | つくばエクスプレス／TX |
| 異形同義語対 | 乳飲み子／赤ン坊 |
| 対義語対 | 乾麺／生麺 |
| 部分・全体語対 | たし算／四則計算 |
| 同類語対 | にわか雨／夕立 |

異表記対は、「問い合わせ」と「問合せ」など、読みが同じで、かつ、意味が同じ語対である。略記対は、「つくばエクスプレス」と「TX」など、一方の語が他方の語の短縮形あるいは略記の語対である。異形同意語対は、「乳飲み子」「赤ン坊」など、異表記対・略記対に該当しないもので、同一の事象・事物を示す語対である。対義語対は、「乾麺」「生麺」など互いに対義の語対である。部分・全体語対は、「たし算」と「四則計算」のように、部分を表す語と全体を表す語との語対である。同類語対は、「にわか雨」「夕立」など過度に抽象的でない共通の上位語をもつ語対である。

「基本的意味関係の事例ベース」の特色は、普

*2 http://alaginrc.nict.go.jp/

通名詞の意味的関係だけでなく、一般的なシソーラス（類語辞典）などには記載されることが稀な専門用語や固有表現の意味的関係を多数収録している点にある。例えば、サイテス／ワシントン条約、サンフランシスコ講和条約／対日講和条約、シナイ山／ホレブ、バックカントリースキー／山スキー、シナジー効果／相乗効果などといった異形同義語対が収録されており、これを利用することで、例えば、ユーザが「ワシントン条約」を検索キーワードとして入力した際に「サイテス」をキーワードとして自動追加し、より多くの検索結果を得ることなどが可能になる。

### 3.2 日本語異表記対データベース

日本語異表記対データベースは、文字レベルの編集距離の近い、日本語の語句の異表記対（あるいは「表記揺れの対」）の正例と負例を集めたものである。例えば、「ギョウザ、ギョーザ」、「ギョウザ、ぎょうざ」、「ギョウザ、餃子」は異表記対である。異表記対の典型的な用途としては情報検索における「検索式（query）の拡張」が挙げられる。例えば、ユーザーが検索に「餃子」と入力している時に、その検索条件を「餃子 OR ギョーザ OR ギョウザ OR ぎょうざ」に自動展開することが可能になる。

本データで収集対象としているのは「ギョウザ、ギョーザ」のように1つの文字だけが異なる単語対（すなわち、編集距離が1の異表記対）のみであり、「ギョーザ、餃子」のような編集距離が1以上の異表記対は収録していない。

**3.1** で述べた「基本的意味関係の事例ベース」に収録されている異表記対は編集距離による制限はないが、収録数は約3万である。一方、本データベースに収録されている異表記対は、編集距離が1のものに限ってはいるが、収録数は100万対以上である。

以下は、日本語異表記対データベースに含まれている異表記対の例を示している。

- 「Center、center」（大文字と小文字の違い）
- 「ゴミ置き場、ゴミ置場」（送り仮名の有無の違い）
- 「ギタープレー、ギタープレイ」（語末の「ー」と「イ」の違い）
- 「ツィンマーマン、ツィマーマン」（「ン」の有無の違い）
- 「ブルース・スプリングスティーン、ブルーススプリングスティーン」（「・」の有無の違い）

日本語異表記対データベースには、人手で作成した異表記対のデータとテキストから自動獲得した異表記対のデータが収録されている。人手で作成した異表記対のデータは、黒田らの手法[4]で作られた48,067の異表記対、10,730の準異表記対、そして2,758の同義異語対（非異表記対）を含んでいる。表4にその例を挙げる。

自動獲得した異表記対のデータは、小島らの手法[5]をもとにして作成されたものである。異表記対の自動獲得のため、まず、1億件のWeb文書に出現する語句（主として単語）から頻度上位1,000万以内の語句を抽出し、これらから成る全ての単語対のうち、編集距離が1のもののみを異表記対の候補とする。そして、上述した人手作成の異表記対を学習データとして用いて分類器を学習し、異表記対の候補を異表記対か否かに分類する。最後に95％以上の精度で獲得された約115万から153万の異表記対を日本語異表記対データベースに収録した。

### 3.3 文脈類似語データベース

文脈類似語データベースは、約100万の見出し語それぞれに対して、Web文書上での出現文脈が最も類似している名詞最大500語を類似度とともに列挙したものである。表5に例を挙げる。各文脈類似語の直後の数値は類似度を表す。「ルパン三世」にはアニメタイトルが、「チャイコフスキー」には有名作曲家が、「カラヤン」には有名指揮者が、「ストーンズ」には懐かしのバンドが文脈類似語として収録されているのが分かる。

文脈類似語は、因果関係などの意味的関係の獲得[6]やWhy型質問応答[7]などの自然言語処理タスクにおいて、その有用性が確認されている。例えば、「ガンの原因は何ですか？」のような病気の原因を求める質問の回答にはその病気と関連する有害物質やウィルス、身体の部位などを表す単語を含む場合が多い。言い換えれば、質問文に「ガン」あるいは「ガン」と類似する単語、つまり「ガン」の文脈類似語が含まれている場合、そ

**表4　人手作成の異表記対の例**

| 種類 | 例 |
|---|---|
| 異表記対 | 「第一週目、第1週目」、「4カ月後、四カ月後」、「Flash Player、Flash player」、「Center、center」「ゴミ置き場、ゴミ置場」、「割引き価格、割り引き価格」、「ギタープレー、ギタープレイ」、「ブルース・スプリングスティーン、ブルーススプリングスティーン」 |
| 準異表記対 | 「法違反、法律違反」、「補足給付、補足的給付」、「調査法、調査手法」、「株取得、株式取得」、「米本社、米国本社」、「手数料額、手数料金額」、「胴体下、胴体下部」、「満州軍、満州国軍」、「土曜・日曜、土曜・日曜日」、「依頼者、依頼者様」 |
| 同義異語対 | 「コンスタンティヌス、コンスタンティヌス帝」、「インテル、インテル社」、「シックスアパート、シックスアパート社」、「米アップル、米アップル社」、「Siemens、Siemens社」、「フィナンシャル・タイムズ、フィナンシャル・タイムズ紙」、「ビハール、ビハール州」、「北海道札幌、北海道札幌市」、「差別的、差別的だ」、「エリア外、エリア以外」 |

**表5　文脈類似語の例**

| 見出し語 | 文脈類似語 |
|---|---|
| ルパン三世 | ルパン3世（－0.229）　名探偵コナン（－0.259）　宇宙戦艦ヤマト（－0.265）　ケロロ軍曹（－0.28）　鉄腕アトム（－0.282）　ガッチャマン（－0.287）　デビルマン（－0.289）　サイボーグ009（－0.294）　新世紀エヴァンゲリオン（－0.295）　ヤッターマン（－0.305）　聖闘士星矢（－0.308）　セーラームーン（－0.308）　... |
| チャイコフスキー | ブラームス（－0.152）　シューマン（－0.163）　メンデルスゾーン（－0.166）　ショスタコーヴィチ（－0.178）　シベリウス（－0.18）　ハイドン（－0.181）　ヘンデル（－0.181）　ラヴェル（－0.182）　シューベルト（－0.187）　ベートーヴェン（－0.19）　ドヴォルザーク（－0.192）　ラフマニノフ（－0.193）　バルトーク（－0.198）　... |
| カラヤン | クレンペラー（－0.21）　バーンスタイン（－0.215）　トスカニーニ（－0.227）　フルトヴェングラー（－0.227）　ベーム（－0.23）　チェリビダッケ（－0.232）　アバド（－0.239）　ムラヴィンスキー（－0.242）　クーベリック（－0.245）　ヴァント（－0.254）　リヒテル（－0.256）　メンゲルベルク（－0.256）　ハイティンク（－0.265）　アーノンクール（－0.276）　... |
| ストーンズ | YMO（－0.215）　メタリカ（－0.223）　ビートルズ（－0.236）　ローリング・ストーンズ（－0.245）　エアロスミス（－0.268）　ツェッペリン（－0.277）　Beatles（－0.284）　ローリングストーンズ（－0.287）　QUEEN（－0.292）　ベンチャーズ（－0.294）　ビーチ・ボーイズ（－0.295）　ピンク・フロイド（－0.297）　レッド・ツェッペリン（－0.301）　ラモーンズ（－0.301）　ディープ・パープル（－0.301）　ニール・ヤング（－0.305）　ザ・フー（－0.306）　... |

の回答として適切な文には、有害物質を表す単語の文脈類似語や、ウィルスを表す単語の文脈類似語、体の部位を表す単語の文脈類似語が含まれる傾向がある。本データベースにより、このような質問文とその適切な回答の間の傾向を明示的に捉えることが可能になり、その結果、質問応答の性能を向上させることができる。

文脈類似語の自動獲得手法の詳細については、Kazamaら[3][8][9]を参照されたい。本データベースの構築で使用された文脈については、さらに本稿**5.1**も併せて参照されたい。

### 3.4　上位語階層データ

上位語階層データは、**6.1**で説明するフリーウェア「上位下位関係抽出ツール」によって日本語Wikipedia（2007/03/28版）から自動獲得した上位下位関係の上位語を人手で階層化したものであり、合計約69,000名詞句から成る階層的シソーラスである。このような上位語の階層化により、自動獲得した上位下位関係の間の意味的な関連性を推定することが可能になる。例えば、上位下位関係「黒澤明の映画作品→七人の侍」と「映画作品→ローマの休日」のそれぞれの上位語「黒澤明の映画作品」と「映画作品」は次のように階層化できる。

- ●作品→映画作品→黒澤明の映画作品
- ●作品→映画作品

つまり、「七人の侍」と「ローマの休日」は「作品」と「映画作品」という上位語を共有するため、同じ概念（つまり「映画作品」）に属すると推定することが可能となる。

上位語の階層化は、上位下位関係抽出ツールで

獲得した上位下位関係の上位語を形態素解析し、その結果から階層化に用いられる名詞句を抽出することによって行われる。例えば、「黒澤明の映画作品」からは「作品」、「映画作品」、「黒澤明の映画作品」が、「鹿児島県の市町村」からは「鹿児島県の市町村」、「県の市町村」、「市町村」が階層化のための名詞句として抽出される。そして、各々の名詞句が上位下位関係における上位語として適切であるか否かを人手で判定する。上位語の階層化についての詳細は黒田ら[10]を参照されたい。なお、本データはWikipediaから抽出した上位下位関係の上位語を日本語のWordNet[11]に接続するために使われ、その有効性が確認されている[12]*3。

### 3.5 単語共起頻度データベース

単語共起頻度データベースは、各単語に対して、それとの意味的関連を表す共起スコアの高い単語を、スコアの高い順に、スコアとともに列挙したものである。共起スコアとしてDice係数、DPMI[13]、共起頻度の3種類を用いた。共起スコアの元となる共起頻度は、約1億件のWeb文書を用いて、次の3つの条件のもとで計算した。

- 約100万語の全組み合わせについての文書内の共起
- 約50万語の全組み合わせについての近接4文内の共起
- 約50万語の全組み合わせについての1文内の共起

意味的関連の強い単語は、互いに共起しやすいため、単語共起頻度データベースを一種の関連語データベースとして使うことが可能である。例えば、「クリスマス」と「野球」それぞれのDice係数上位5語は以下のようになっており、関連の深い語ほど高いスコアが与えられているのが分かる。

「**クリスマス**」:「お正月」(0.172339)、「誕生日」(0.119606)、「サンタ」(0.113987)、「冬」(0.112612)、「年末」(0.110775)

「**野球**」:「サッカー」(0.362974)、「格闘技」(0.227781)、「プロ野球」(0.220464)、「ゴルフ」(0.210349)、「テニス」(0.208742)

なお、単語共起頻度データベースは、類推による単語間の意味的関係獲得[14]に用いられ、その有効性が確認されている。

### 3.6 負担・トラブル表現リスト

「負担・トラブル表現リスト」は、「災害」「心理的ストレス」「アスベスト汚染」など人間活動に負荷を与えたり、マイナス効果をもたらす問題や障害に関係する表現、20,115件を収録したデータベースである。データベースに収録されている負担・トラブル表現は、De Saegerら[15]の手法に基づきWeb文書から自動獲得されたものを人手で検証・分類したもので、各負担・トラブル表現には“病”、“被害”、“不正行為・違反”、“有害物質”などの分類ラベルが付与されている。例えば、“病”には「B型肝炎」、「インフルエンザ」、「クリプトコッカス症」などが、“被害”には「ケミカルハザード」、「サンゴ食害」、「サリドマイド薬害」などが、“不正行為・違反”には「スキミング」、「居眠り運転」、「権利侵害行為」などが、“有害物質”には「催眠ガス」、「酸性降下物」、「自動車排ガス」などが該当する。他の負担・トラブル表現と分類ラベルの例を表6に示す。

大規模な負担・トラブル表現リストの構築は、想定していなかった意外なトラブルを網羅的に検索することを可能とする。例えば、2011年3月11日から2011年6月17日までに発信された東日本大震災に関連したツイート、約320万件[16]に含まれる負担・トラブル表現を検索した場合、

**表6 負担・トラブル表現の例**

| 分類 | 例 |
|---|---|
| エラー | core dump、DBエラー、Out of Memory、アンダーフロー |
| 自然現象 | エルニーニョ、かまいたち、メイルシュトローム、黄砂 |
| 破損・損傷 | メルトダウン、ラインブレイク、液晶割れ、荷痛み |
| 有害生物 | レタス病害虫、アオコ、アクネ菌、ネキリムシ |

*3 Kurodaら[12]によると、Wikipediaから抽出した上位下位関係の上位語と日本語のWordNet synsetの間の対応率は元々約8%程度であったが、本データの階層化情報を用いることによってその対応率が約95%になった。

「停電」、「断水」など一般的に想定できる問題に関するツイートだけでなく、ライフラインが使用できない中で寒さ対策として使用されていた練炭によって発生した「一酸化炭素中毒」、避難所での生活を避け車内で避難生活を送ったことにより発生した「エコノミー症候群」など、いわゆる“災害関連死”や二次災害として生じたトラブルに関連するツイートも検出、特定することができる。このように、2万件を超える負担・トラブル表現リストは、想定が難しいトラブルを特定する際などに有効な言語資源となる。

### 3.7 日本語 WordNet

日本語 WordNet は、プリンストン大学で開発された Princeton WordNet 等に着想を得て開発されたもので、93,834 語を synset と呼ばれる同じ概念を示す語の集合にグループ化したものである。例えば、「行動」「営み」「行為」「活動」「営為」といった表現が 1 つの集合（synsetID: 00030358-n）としてグループ化されており、さらに、それに対する定義文として「人々が行う、あるいは起こす事」が、例文として「殺人と他の異常な行動の話があった」が収録されている。なお、日本語 WordNet には一部用言も収録されている。

日本語 WordNet は、同義語を 1 つの synset にグループ化するだけでなく、synset 間の上位下位関係（例えば、家具・椅子）、構成要素・被構成要素関係（例えば、脚・椅子）など synset 間の意味関係も収録している。日本語 WordNet で扱われている意味関係の一部とその例を表 7 に示す。

上位概念リンクは、「動物」と「変温動物」のように、一方の synset がもう一方の synset の上位概念であるような 2 つの synset の間に張られるリンクである。被構成要素リンクは、「自動車」と「エアバック」のように、一方の synset の表す対象がもう一方の synset の表す対象の構成要素となっている synset 間に張られるリンクである。因果関係リンクは、「映写する」と「表れる」のように、一方の synset の表す事態の成立が、もう一方の synset の表す事態を引き起こすような synset 間に張られるリンクである。含意リンクは、「吹っ掛ける」と「請求する」のように、一方の synset の表す事態が成立するなら、同時かそれ以前に、もう一方の synset の表す事態も成立するような 2 つの synset の間に張られるリンクである。なお、因果関係については **4.5** を含意については **4.1** を併せて参照されたい。

表 7 日本語 WordNet における synset のリンクの種類とその例

| 分類 | 例 |
|---|---|
| Hypernym（上位概念） | 動物・変温動物 |
| Meronyms（被構成要素） | エアバック・自動車 |
| Causes（因果関係） | 映写する・表れる |
| Entails（含意） | 吹っ掛ける・請求する |

日本語 WordNet は、Weblio 辞書の英和和英辞書[*4] をはじめ、様々な用途で利用されている。また、「基本的意味関係の事例ベース」と同様に検索クエリの拡張や言い換え認識などにも利用できる。なお、**3.1** で述べた通り、「基本的意味関係の事例ベース」は固有名詞や専門用語を多く収録しているのに対し、日本語 WordNet は一般的な単語を中心に収録している。つまり両者は相補的な関係にある。

## 4 用言の意味的関係データベース

### 4.1 動詞含意関係データベース

このテータベースは、含意関係が成立している動詞のペア（52,689 ペア）と含意関係が成立していない動詞のペア（68,819 ペア）の計 121,508 ペアを列挙したものである。含意関係が成立する動詞ペアとは、一方の動詞の指す事態が成立するなら、同時かそれ以前に、もう一方の動詞の指す事態も成立すると言えるペアである。例えば、「スタメン出場する」は「先発する」を、「チンする」は「加熱する」を、「あざ笑う」は「笑う」を、「酔っ払う」は「飲む」を、「借りる」は「貸す」を含意する。

含意関係は多くの言語情報処理システムにおいて重要な役割を果たす意味的関係である。例えば質問応答システムは、「昨日の巨人–阪神戦で先発したのは誰？」という質問に対し、Web 等の大量文書から「昨夜の阪神戦では巨人久保がスタ

＊4 http://ejje.weblio.jp/

メン出場」等の質問文とは文字列上大きく異なる文を回答として読み取れなくてはならない。この場合、「スタメン出場する」が「先発する」を含意するという知識が必須である。

本データベースの負例（含意関係が成立していない動詞ペア）は、正例（含意関係が成立している動詞ペア）とセットで、機械学習への入力として利用できる。つまり、ある動詞ペアの間に含意関係が成立するかどうかを識別するモデルを学習する際の学習データとして使用することができる。

正例と負例はそれぞれ4種類に下位分類されている。以下では各分類を例とともに説明する。正例と負例は全て、橋本らの手法[17][18]により自動獲得した結果を人手でチェックしたものである。なお、以下では動詞ペアの左側の動詞、つまり含意する側の動詞を「動詞1」と呼び、右側の動詞、つまり含意される側の動詞を「動詞2」と呼ぶ。

#### 4.1.1 正例群

正例群の総ペア数は52,689ペアで、動詞1の総異なり数は36,058、動詞2の総異なり数は8,771である。

**含意が成り立つ類義／上位下位関係** 動詞1と動詞2の間に含意が成立し、かつ、類義関係あるいは上位下位関係（動詞2が動詞1の上位概念）が成立している動詞ペアである。ただし、次に述べる「文字列上包含関係にあり、含意が成り立つ類義／上位下位関係」は含まれていない。ペア数は33,802、動詞1の異なり数は18,128、動詞2の異なり数は7,650である。以下に例を挙げる。

- 「挑戦する→チャレンジする」
- 「チンする→加熱する」
- 「同乗する→乗る」
- 「組み立てる→作る」
- 「代用する→使う」

**文字列上包含関係にあり、含意が成り立つ類義／上位下位関係** 含意が成り立つ類義／上位下位関係にあてはまる動詞ペアのうちの、動詞1が動詞2を文字列上包含している動詞ペアである。ペア数は15,599、動詞1の異なり数は15,367、動詞2の異なり数は2,440である。以下に例を挙げる。

- 「あざ笑う→笑う」
- 「セリーグ優勝する→リーグ優勝する」
- 「流れ出る→出る」
- 「そそり立つ→立つ」
- 「一部免除する→免除する」

**前提関係** 動詞2が動詞1の前提条件になっている動詞ペアである。上の2種類の含意関係は動詞1の事態と動詞2の事態が同時に起こるが、「前提関係」では、動詞2の事態が動詞1の事態に時間的に先行する。ペア数は2,846、動詞1の異なり数は2,227、動詞2の異なり数は711である。以下に例を挙げる。

- 「酔っぱらう→飲む」
- 「稲刈する→田植する」
- 「乗捨てる→乗る」
- 「離職する→働く」
- 「首席卒業する→学ぶ」

**作用反作用関係** 動作主体が異なる、一方が作用でもう一方が反作用と言える2つの動詞から成るペアである。一方、上の3種類の含意関係はいずれも、動詞1と動詞2の動作主体が同じである。ペア数は442、動詞1の異なり数は336、動詞2の異なり数は328である。以下に例を挙げる。

- 「借りる→貸す」
- 「受取る→手渡す」
- 「教える→学ぶ」
- 「売る→買う」
- 「預ける→預かる」

#### 4.1.2 負例群

負例群の総ペア数は68,819ペアで、動詞1の総異なり数は14,658、動詞2の総異なり数は7,077である。

**含意、反義、予測関係ではない関連語ペア** 含意関係、あるいは以下で述べる反義関係、予測関係のいずれにも当てはまらないが、何らかの関連が認められるペアである。ただし、次に述べる「文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア」は含まれない。ペア数は68,306、動詞1の異なり数は14,168、動詞2の異なり数は7,006である。以下に例を挙げる。

- 「通勤する→走る」
- 「読書する→寛ぐ」

●「ブログ巡りする→休む」
●「農業体験する→住む」
●「押し黙る→俯く」

**文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア**　含意、反義、予測関係ではない関連語ペアのうちの、動詞1が動詞2を文字列上包含している動詞ペアである。ペア数は294、動詞1の異なり数は290、動詞2の異なり数は101である。以下に例を挙げる。

●「冴渡る→渡る」
●「準優勝する→優勝する」
●「怒り出す→出す」
●「歌い上げる→上げる」
●「解毒する→毒する」

**反義関係**　反義関係にあるペアである。ペア数は51、動詞1の異なり数は46、動詞2の異なり数は42である。以下に例を挙げる。

●「閉める→開ける」
●「反比例する→比例する」
●「失う→得る」
●「下げる→上げる」
●「飛び去る→飛来する」

**予測関係**　含意関係とは言えないが、動詞1の事態が起こるなら、その後動詞2の事態が起こる可能性が高いと言えるようなペアである。ペア数は168、動詞1の異なり数は154、動詞2の異なり数は121である。以下に例を挙げる。

●「紅葉する→落葉する」
●「深煎りする→挽く」
●「入会希望する→入会する」
●「印刷プレビューする→印刷する」
●「受験する→進学する」

### 4.2　述語フレーズ含意関係データベース

このデータベースは、含意関係が成立している述語フレーズのペア（正例）と含意関係が成立していない述語フレーズのペア（負例）を列挙した近日公開予定の言語資源であり、約60万ペアの収録を予定している。動詞含意関係データベースが単語間の含意関係を扱うのに対し、述語フレーズ含意関係データベースはフレーズ間の含意関係を扱う。以下に例を挙げる。

●「すべての債務を免除される→債務の支払責任を免除してもらう」
●「地球全体の平均気温が上昇する→地球規模で気温が上昇していく」
●「粉塵を吸入する→ほこりを吸い込む」
●「インシュリンの量が不足する→インスリンの作用が弱くなる」
●「現金でトレードする→お金で取引する」

述語フレーズ含意関係も動詞含意関係と同様、多くの言語情報処理システムにおいて重要な役割を果たす。例えば質問応答システムは、「細胞を老化させる原因は何？」という質問に対し、Web等の大量文書から「DNA損傷が細胞を酸化させる」等の質問文とは文字列上大きく異なる文を回答として読み取れなくてはならない。この場合、「細胞を酸化させる」が「細胞を老化させる」を含意するという知識が必須である。

また、動詞含意関係データベースと同様に、本データベースは正例と負例の2つに大きく分けられる。負例は正例とセットで機械学習への入力として利用できる。つまり、ある述語フレーズペアの間に含意関係が成立するかどうかを識別するモデルを学習する際の学習データとして使用することができる。

正例と負例は全て、橋本らの手法[19][20]により、Web上の定義文から自動獲得した結果から構築した。そのうちの一部は人手でチェックした上で、残りは自動獲得結果をそのままデータベースとして公開する予定である。

本データベースでは、意味的構成性の観点からフレーズペアを「完全に構成的なフレーズペア」と「部分的に構成的なフレーズペア」に分類している。前者は、ペアをなす2つのフレーズの間で、どの内容語も相手方のフレーズに同義か同義に近い内容語が存在するようなフレーズペアである。例えば「合鴨を水田に放す→田にアイガモを放す」は、どの内容語も相手方のフレーズに同義語が存在するので「完全に構成的なフレーズペア」である。後者の「部分的に構成的なフレーズペア」は、相手方のフレーズに同義か同義に近い語を持たない内容語が少なくとも1つ存在するようなフレーズペアである。例えば「地震の揺れを建物に伝わりにくくする→建物自体の揺れを小さくする」は、「地震」「伝わる」「小さい」が相手方のフレーズに同義あるいは同義に近い語が存

在しないので「部分的に構成的なフレーズペア」である。

意味的構成性の高いフレーズペアは含意関係にあることの自動認識が意味的構成性の低いフレーズペアに比べて容易であると考えられる。つまり本データベースは含意認識の難易度に応じて述語フレーズペアを分類していると見なせる。

以下に、「完全に構成的なフレーズペア」と「部分的に構成的なフレーズペア」の例を挙げる。

●完全に構成的なフレーズペア
- 「生薬をいくつも組み合わせる→いくつもの生薬を組み合わせる」
- 「エネルギーが光になる→エネルギーが光となる」
- 「個人情報の取り扱い方法を定める→個人情報の取扱い方法を定める」
- 「インターネット上のマナーのことだ→ネットワーク上のエチケットのことだ」
- 「介護サービス計画を作成する→ケアプランを作成する」
- 「文科省が推進している→文部科学省が推進する」
- 「アメリカで考案される→米国で生まれる」
- 「コンピューターに記憶させておく→PC に保存しておく」
- 「パワーが宿る→力を秘めている」

●部分的に構成的なフレーズペア
- 「かみ合わせや歯並びを回復する→噛み合わせを復元する」
- 「悪性細胞が認められる→がん細胞が発生する」
- 「シワやシミを解消する→しわなどを改善する」
- 「無線 LAN アクセスポイントを共有する→アクセスポイントを公開する」
- 「オートバイで旅行する→バイクで走る」
- 「会員間でクルマを共同利用する→クルマを複数の人間で共同利用する」
- 「電気エネルギーを使用している→エネルギーを電気でまかなう」
- 「情報共有を図る→コミュニケーションを取る」
- 「もずくやコンブに含まれている→海藻類の中に含まれる」
- 「コレステロールや中性脂肪の割合が高い→脂質の値が高い」

### 4.3 活性／不活性データベース

活性／不活性データベースは活性／不活性テンプレートを列挙した、今年度末公開予定の言語資源であり、活性／不活性テンプレート約 1 万を収録する予定である。活性／不活性とは、我々が文献[21][22]で提案した意味的極性で、「が発生する」や「を防ぐ」などの「助詞＋述語」（以下、テンプレートと呼ぶ）を以下の「活性」、「不活性」、「中立」の 3 つに分類する。

**活性テンプレート**　項（主語や目的語等）の指す対象の主たる機能、効果、目的、役割、影響が準備あるいは活性化されることを含意する。（例:「を引き起こす」、「を使う」、「を買う」、「を進行させる」、「を輸入する」、「が増える」、「が可能になる」）

**不活性テンプレート**　項の指す対象の主たる機能、効果、目的、役割、影響が抑制あるいは不活性化されることを含意する。（例:「を防ぐ」、「を捨てる」、「を治療する」、「が減る」、「を破壊する」、「が不可能になる」）

**中立テンプレート**　活性でも不活性でもないもの。（例:「を考える」、「を探す」、「に比例する」）

例えば、「地震を引き起こす」は「地震」の影響が活性化されることを、「津波を防ぐ」は「津波」の影響が不活性化されることを含意する。

活性／不活性は文献[23][24]にあるようないわゆる評価極性（good/bad）とは独立である。例えば「が上達する」も「を発症する」も活性だが前者のみが good で、「を治療する」も「が頓挫する」も不活性だが後者のみが bad である。

活性／不活性テンプレートには様々な利用法が考えられるが、本稿では述語フレーズ矛盾関係データベース（**4.4**）と述語フレーズ因果関係データベース（**4.5**）の構築への応用について述べる。

活性／不活性データベースは、我々の開発した手法[21][22]により自動獲得したものを人手チェックすることで構築した。以下に活性／不活性データベースに収録予定の活性と不活性のテンプレートの例を挙げる。

●活性テンプレートの例
- を高める
- を誘発する
- を組織する
- を犯す
- を正常化する
- を充填する
- で煮る
- が高揚する
- が豊富だ
- に達する

●不活性テンプレートの例
- を麻痺させる
- を騙す
- を顰める
- を非難する
- を静める
- に逆らう
- が衰退する
- が脱線する
- が脆くなる
- で失敗する

### 4.4 述語フレーズ矛盾関係データベース

このデータベースは、「癌を破壊する⊥癌を進行させる」や「運転を助ける⊥運転を妨げる」のように矛盾関係が成立している述語フレーズのペア（正例）と、「癌に罹る⊥癌を研究する」のように矛盾関係が成立していない述語フレーズのペア（負例）を列挙した、今年度末公開予定の言語資源である。正例負例あわせて100万対前後の述語フレーズペアを収録する予定である。本データベースの述語フレーズは全て、「癌を破壊する」のように、名詞、助詞、述語それぞれ1語ずつから構成されるものである。全ての「助詞＋述語」は活性テンプレートあるいは不活性テンプレート（**4.3**）である。

矛盾関係が成立する述語フレーズペアとは、一方の述語フレーズの表す事態ともう一方の述語フレーズの表す事態とが同時には成立し得ないペアである。このようなペアに加えて、我々が「準矛盾関係」と呼ぶ述語フレーズペアも正例としてデータベースに収録した。準矛盾関係にある述語フレーズペアとは次の条件を満たすペアである。

1．一方の述語フレーズの表す事態ともう一方の述語フレーズの表す事態とは同時に成立しうる。
2．しかし、一方の事態、あるいは両方の事態の示す傾向が極限まで強まると、2つの事態は同時には成立し得ない、つまり、矛盾する。

準矛盾関係にある述語フレーズペアの例として「緊張感を伴う⊥緊張感を緩和させる」が挙げられる。緊張感を緩和させたとしても、依然として緊張感を伴っていることは往々にしてある。つまり両者は同時に成立し得るため、純粋な矛盾関係とはいえない。しかし、緊張感を伴うという事態の傾向が極限まで強まり、かつ、緊張感を緩和させるという事態の傾向が極限まで強まれば、両者は同時には成立し得ない。言い換えれば、極限の緊張を感じている事態と、緊張感が完全に緩和しきった事態は矛盾関係にあると言える。つまり、「緊張感を伴う⊥緊張感を緩和させる」は我々が言うところの準矛盾関係にある述語フレーズペアである。

矛盾関係と準矛盾関係にある述語フレーズペアの例を以下に挙げる。

●矛盾関係
-「アンバランスを是正する⊥アンバランスを生じさせる」
-「円安が止まる⊥円安が進行する」
-「騒音がひどくなる⊥騒音は減少する」
-「酸味がます⊥酸味が消える」
-「原発をなくす⊥原発を増やす」
-「ユーロが下落する⊥ユーロが強くなる」
-「ウイルスが死滅する⊥ウイルスが活性化する」

●準矛盾関係
-「痛みが発症する⊥痛みを減らす」
-「アクセスが生ずる⊥アクセスを抑制する」
-「放射能が放出される⊥放射能が減る」
-「シェアを有する⊥シェアが低下する」

述語フレーズ矛盾関係は多くの言語情報処理システムにおいて重要な役割を果たす。例えば、NICTで開発したWISDOM[*5]をはじめとするWeb情報分析システムは、Web文書中に書かれ

＊5　http://wisdom-nict.jp/

ているテキスト情報の間の矛盾を自動認識しなくてはならない。ユーザからの問い合わせが「原発停止による自然環境への影響は？」で、あるWeb文書に「放射能汚染の可能性のある原発を停止することで、自然環境を守ることができる」とあり、別のWeb文書に「原発停止により火力発電の割合が増え、$CO_2$増加により、自然環境を悪化させる」とある場合、Web情報分析システムは、2つのWeb文書に書かれている見解の矛盾を自動認識し、対立意見を整理してユーザに提示しなくてはならない。

動詞含意関係データベース、述語フレーズ含意関係データベースと同様に、本データベースも正例と負例の2つに大きく分けられる。負例は正例とセットで機械学習への入力として利用できる。つまり、ある述語フレーズペアの間に矛盾関係あるいは準矛盾関係が成立するかどうかを識別するモデルを学習する際の学習データとして使用することができる。

正例と負例は全て、橋本らの手法[21][22]により自動獲得した結果から構築した。自動獲得結果の適合率は、スコア上位100万ペアで約70%である。この矛盾関係獲得手法は、同じく橋本らの手法[21][22]で自動獲得した活性／不活性テンプレートを用いたものである。具体的には、「癌を破壊する⊥癌を進行させる」のように、1つの名詞と、極性が反対の活性／不活性テンプレート対（「を破壊する」は不活性、「を進行させる」は活性）から成る述語フレーズペアを自動獲得した。

### 4.5 述語フレーズ因果関係データベース

このデータベースは、「タバコを吸う⇒肺癌になる」のように因果関係が成立している述語フレーズのペア（正例）と、「タバコを吸う⇒会社に行く」のように因果関係が成立していない述語フレーズのペア（負例）を列挙した、今年度末公開予定の言語資源である。正例負例あわせて100万対前後の述語フレーズペアを収録する予定である。本データベースの述語フレーズは全て、「肺癌になる」のように、名詞、助詞、述語それぞれ1語ずつから構成されるものである。**4.4**で述べた「述語フレーズ矛盾関係データベース」と同様、全ての「助詞＋述語」は活性テンプレートあるいは不活性テンプレート（**4.3**）である。

以下に本データベースに収録予定の因果関係述語フレーズペアの例を挙げる。

- 「基礎代謝を高める⇒脂肪燃焼力を高める」
- 「学習意欲を高める⇒自己学習を促進する」
- 「輸出が増える⇒GDPが増加する」
- 「血行を促進する⇒新陳代謝を助ける」
- 「視界が良くなる⇒作業効率が向上する」
- 「大地震が発生する⇒メルトダウンを起こす」
- 「熱効率が良い⇒暖房効果を高める」
- 「インフレを起こす⇒円安が進行する」
- 「体力が落ちる⇒免疫力が下がる」
- 「国債先物急落を受ける⇒金利が上昇する」

本データベースにおける因果関係が成立する述語フレーズペアとは、左のフレーズの意味する事態、動作、状態が成立する場合としない場合を比べた時、成立する場合のほうが、右のフレーズの意味する事態、動作、状態の成立する可能性が高くなるフレーズペアを指す（左のフレーズの意味する事態、動作、状態は、右のフレーズの意味する事態、動作、状態とほぼ同時か、あるいはそれ以前に成立するものとする）。つまり、本データベースにおける因果関係は、左のフレーズの意味する事態、動作、状態が成立すれば、必ず右のフレーズの意味する事態、動作、状態が成立する、ということを保証するものではない。例えば、本データベースでは、大地震が発生する場合と発生しない場合とを比べると前者のほうがメルトダウンを起こす可能性は高いので、「大地震が発生する⇒メルトダウンを起こす」を因果関係として認めているが、これは大地震が常にメルトダウンに繋がるということを意味するものではない。

この他、本データベースを構築するにあたって、我々が一般性基準と真偽未決着基準と呼ぶ、本データベースに因果関係として収録するか否かに関する2つの基準を設けた。前者は、一般性が極端に低い因果関係はたとえ因果関係抽出元のコーパスに因果関係らしく書かれてあっても本データベースに含めない、というものである。例えば「新年会には市川さんが来るからベジタリアンメニューにしましょう」とコーパスに書かれてあっても「市川さんが来る⇒ベジタリアンメニューにする」は極端に一般性が低いと考えられるため、本データベースに因果関係として収録し

ない。後者の真偽未決着基準とは、真偽が科学的に未決着な因果関係は、その因果関係の妥当性を支持する記述がWebに1つでも見つかれば、本データベースに因果関係として含めるという基準である。例えばWebに「黒烏龍茶を飲むと脂肪の吸収が抑えられるそうです。」と書いてあれば、「黒烏龍茶を摂取する⇒脂質吸収を抑制する」を本データベースに因果関係として収録する。

つまり、本データベースを使用する上で注意すべきことは、本データベースに収録されている述語フレーズペアが因果関係としての妥当性を保証するものではない、ということである。Webに明記されていることを人手で確認した述語フレーズペアであっても、それはWebに書かれていることを確認しただけであり、因果関係として真に妥当であるかどうかを保証するものではない。

本データベースは正例と負例の2つに大きく分けられる。負例は正例とセットで機械学習への入力として利用できる。つまり、ある述語フレーズペアの間に因果関係が成立するかどうかを識別するモデルを学習する際の学習データとして使用することができる。

正例と負例は全て、文献[21][22]にある2種類の因果関係自動獲得手法の結果から構築した。1つはWebに書かれている因果関係を自動抽出する手法（以後、因果関係抽出法と呼ぶ）であり、もう1つは、Webには書かれていないが妥当である可能性の高い因果関係を自動生成する手法（因果関係仮説生成法と呼ぶ）である。因果関係抽出法は、「犯罪が増加すると不安が高まる」等のように、1つの活性／不活性テンプレート（例えば「が増加する」「が高まる」）と1つの名詞から成るフレーズ2つが順接接続（例えば「～と」）とともにWeb上の1文中で共起している場合に、その2フレーズを因果関係「犯罪が増加する⇒不安が高まる」として抽出する。自動獲得の適合率は、スコア上位50万ペアで約70%である。一方、因果関係仮説生成法は、抽出された因果関係（例えば「犯罪が増加する⇒不安が高まる」）の各フレーズを、それと矛盾するフレーズ（例えば「犯罪が増加する⊥犯罪を減らす」、「不安が高まる⊥不安が無くなる」。**4.4**を参照）で置換することで、因果関係の仮説（例えば「犯罪を減らす⇒不安が無くなる」）を自動生成する。なお、因果関係仮説のうちWebの1文中に書かれているものは出力から除外する。つまり、本データベースには、Webに書かれているものだけでなく、Webには書かれていないが妥当である可能性が比較的高い因果関係も仮説として収録されている。自動獲得の適合率は、スコア上位100万ペアで約57%である。以下に、本データベースに収録予定の因果関係仮説の例を挙げる。なお、括弧内に、仮説の元となった、Webに記載されていた因果関係を示す。

- 「ストレスが減少する⇒不眠が改善される」（「ストレスが増加する⇒不眠が続く」）
- 「デフレを阻止する⇒税収が増加する」（「デフレが進む⇒税収が減る」）
- 「楽しみが増大する⇒ストレスが減少する」（「楽しみが減る⇒ストレスが高まる」）
- 「犯罪を減らす⇒不安が無くなる」（「犯罪が増加する⇒不安が高まる」）
- 「塩素を減らす⇒バクテリアは増殖する」（「塩素を発生させる⇒バクテリアを死滅させる」）
- 「需要が拡大する⇒失業を減少させる」（「需要が減る⇒失業が増える」）
- 「疲れを軽減する⇒免疫を増強する」（「疲れがたまる⇒免疫が弱まる」）
- 「調子があがる⇒トラブルを防げる」（「調子が悪くなる⇒トラブルが起きる」）

### 4.6 日本語パターン言い換えデータベース

Webをはじめとする大規模な文書データから知識を獲得する際に、同じような意味を持つ、言い換え可能な文を認識することができれば、より多くの知識を得ることができる。「日本語パターン言い換えデータベース」は、文の係り受け解析の結果を利用して「AはBが豊富です」のような、1文中で任意の名詞AとBを結ぶパターンに対して、言い換えが可能な別のパターンを収めたデータベースである。例えば〈AはBが豊富です〉、〈AはBを防ぐ〉、〈AでBを喜ばせる〉というパターンに対して、それぞれ以下の表8～10にあるようなパターンが、言い換えとしてのもっともらしさを表すスコアとともに本データベースに収録されている。

「日本語パターン言い換えデータベース」は

5,000 万 Web 文書から獲得したパターンを言い換えの対象としている。パターンは係り受け解析の結果となる構文木の中で、一定の出現頻度を超える名詞 A と B をつなぐ係り受けパスに含まれる単語からなる。例えば、図 1 にあるように、「交通事故による経済的な損害に関して」という文からは〈A による〉というパターンが抽出される。

パターン間の類似度は、パターンの変数 A、B の位置に出現する名詞対の出現分布から計算される。詳細については文献[6]にある「SC（Single Class）」手法の記述を参照されたい。この手法は教師なし学習に基づく自動獲得手法であるため、本データベースに収録されている言い換えパターン全てが正確であるということは保証されない。

本データベースに関連して、我々は、Kloetzer ら[25]が提案した教師あり学習に基づく手法による自動獲得結果から、パターン間の含意関係のデータベースを現在構築中である。自動獲得結果のスコア上位 1,000 万ペアの適合率は約 70％である。以下に Kloetzer らの手法で獲得したパターン間の含意関係の例を挙げる。

- 「A を生み出す B → A を作る B」
- 「A に出向く B → A に行く B」
- 「A に上程されていた B → A に B を提出する」
- 「A を B に変更→ A を B にする」
- 「B に光る A → B に輝く A」
- 「A を乗り換えられる B → A を変更できる B」
- 「B の材料を生かした A → B の素材を使った A」
- 「A を担いだ B → A を背負った B」
- 「A が奉られている B → A を祀る B」
- 「B を強化する A → B を育てる A」

**表 8 〈A は B が豊富です〉の言い換え（スコア上位 5 パターン）**

| パターン | 言い換えスコア |
|---|---|
| 〈A は B が豊富〉 | 0.0549719888 |
| 〈A には B が豊富に含まれています〉 | 0.0382925298 |
| 〈A は B も豊富です〉 | 0.0377786173 |
| 〈A は B を多く含む〉 | 0.0336538462 |
| 〈A は B も豊富〉 | 0.0331325301 |

**表 9 〈A は B を防ぐ〉の言い換え（スコア上位 5 パターン）**

| パターン | 言い換えスコア |
|---|---|
| 〈A が B を防ぐ〉 | 0.0224161276 |
| 〈A は B を予防する〉 | 0.0186121788 |
| 〈A で B を防ぐ〉 | 0.0175963197 |
| 〈B を防ぐ A〉 | 0.0175141447 |
| 〈A は B を防止する〉 | 0.0132786565 |

**表 10 〈A で B を喜ばせる〉の言い換え（スコア上位 5 パターン）**

| パターン | 言い換えスコア |
|---|---|
| 〈A を B 様にご提供していきたい〉 | 0.0430107527 |
| 〈B 様に A を提供して参りました〉 | 0.0337078652 |
| 〈A を B 様に提供し続けること〉 | 0.0337078652 |
| 〈B 様に A を提供出来るように〉 | 0.0337078652 |
| 〈B 様に A を提供出来るよう〉 | 0.0333333333 |

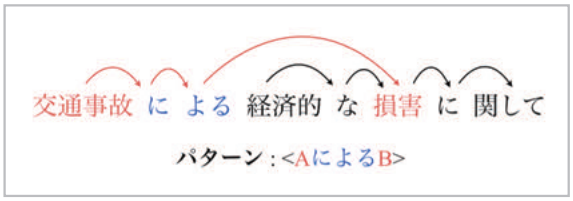


**図 1 係り受け解析結果からのパターン抽出**

## 5 係り受けデータベース、コーパス

### 5.1 日本語係り受けデータベース、日本語 Wikipedia エントリの係り受けデータベース

「日本語係り受けデータベース」「日本語 Wikipedia エントリの係り受けデータベース」は、大量の日本語文書を係り受け解析した結果から係り受け関係を抽出し、その頻度を収録したものである。表 11 に例を示す。

「日本語係り受けデータベース」は、Web 6 億文書のデータから、「関サバを食べる」や「関サバのお造り」等のように、2 文節から成る係り受け関係を抽出したもので、約 46 億件の係り受け関係とその頻度が収録されている。

「日本語 Wikipedia エントリの係り受けデータベース」も「日本語係り受けデータ」と同じ

表 11　係り受けデータベースにおける係り受け関係の例とその頻度

| データベース | 係り受け関係 | 頻度 |
|---|---|---|
| 日本語係り受け | 関サバを食べる | 20 回 |
| 日本語係り受け | 関サバのお造り | 7 回 |
| 日本語係り受け | 野球を観戦する | 40 回 |
| 日本語係り受け | 野球のボール | 20 回 |
| Wikipedia 係り受け | 風と共に去りぬを借りる | 12 回 |
| Wikipedia 係り受け | 三保の松原の景色 | 6 回 |
| Wikipedia 係り受け | 瞬間湯沸かし器で一酸化炭素中毒事故 | 8 回 |
| Wikipedia 係り受け | 星の王子さまを読む | 3,643 回 |

Web 文書を用いて係り受け関係を抽出したものであるが、「日本語係り受けデータ」が名詞の部分が 1 文節のものの係り受け関係だけを収録しているのに対して、「日本語 Wikipedia エントリの係り受けデータベース」では Wikipedia の記事のタイトル（エントリ）の内、2 文節以上のもの（例:「三保の松原」「風と共に去りぬ」）を含む係り受け関係とその頻度が収録されている。つまり、「日本語 Wikipedia エントリの係り受けデータベース」は 2 文節以上から成る係り受け関係を収録している。「日本語 Wikipedia エントリの係り受けデータベース」は、「日本語係り受けデータベース」に不足していた、複数文節から成る固有表現を含む係り受け関係を補うものと見なせる。

「日本語係り受けデータベース」「日本語 Wikipedia エントリの係り受けデータベース」は、「文脈類似語データベース」（**3.3**）をはじめとする、係り受け関係の頻度等をもとに構築される多くの言語資源にとって不可欠である。例えば「文脈類似語データベース」の構築では、出現文脈の類似する名詞をまとめ上げることによって、表 5 にあるようなアニメタイトルを表す名詞群、有名作曲家を表す名詞群、有名指揮者を表す名詞群、懐かしのバンドを表す名詞群等を自動獲得しているが、その出現文脈として本係り受けデータベースの情報が利用されている。出現文脈として用いられているのは、Web 文書における各名詞の係り先である。表 12 には「関サバ」と「関アジ」の係り先、つまり文脈類似語データベースにおける出現文脈の一部を挙げている。「関サバ」や「関アジ」といった魚を意味する単語にとって特徴的と考えられる出現文脈である「の刺身」、「の活造り」、「の干物」、「がおいしい」といった係り先の出現頻度が、「関サバ」と「関アジ」の両単語において高いことが分かる。言い換えれば、両者の出現文脈が類似していることが分かる。

表 12　「関サバ」と「関アジ」の係り先とその出現頻度

| 係り先 | 「関サバ」 | 「関アジ」 |
|---|---|---|
| の刺身 | 106 回 | 92 回 |
| の活造り | 12 回 | 11 回 |
| の干物 | 15 回 | 10 回 |
| を仕入れる | 4 回 | 4 回 |
| を使う | 10 回 | 14 回 |
| を堪能 | 4 回 | 6 回 |
| がおいしい | 25 回 | 10 回 |
| を食する | 2 回 | 7 回 |
| は有名だ | 9 回 | 14 回 |
| に劣らない | 4 回 | 10 回 |

### 5.2　京都観光ブログの評価情報付与データ

近年、Web を始めとする情報媒体の発達により、様々な人々が、多様な話題について意見や評価を公に発信することができるようになった。それに伴い、大量の文書から人々の意見を抽出し、集約する技術の研究が盛んになってきている。京都観光ブログの評価情報付与データは、こうした意見分析技術開発の基盤となる機械学習の学習用コーパスとして構築された。本データは「京都観光ブログ」と「京都観光ブログの評価情報付与データ」から構成される。

京都観光ブログとは、観光ドメインに特化した

日本語ブログ記事のデータベースである。執筆者は47名で合計1,041記事（1記事あたり平均約480文字）が含まれる。データ作成にあたっては、データの著作権はNICTが有するという条件の下で執筆者を募り、実際の京都観光に基づいた記事作成を依頼した上で行われている。各執筆者は我々が立ち上げたブログサイト（非公開）上で記事を作成している。

京都観光ブログの評価情報付与データとは、京都観光ブログから、文献[26][27]にある一定の基準に従って、評価情報（評判、意見）を人手で抽出したものである。さらに抽出された評価情報には、評価保持者、評価表現、評価対象などが付与されている。表13に記事の例を、表14に付与された評価情報の例を示す。アノテーション項目の詳細については文献[27]を参照されたい。

表14で挙げられているように、「きれいだ」のような主観的な意見だけでなく、「世界遺産に登録されている」など、客観的な記述であっても、それがトピックとなる観光名所などの利点や欠点が述べられているような記述であれば、抽出対象としている点が本データの特徴である。

従来の自然言語処理向けの学習用コーパスは新聞記事から作成されていた。しかし、ブログをはじめとするConsumer Generated Mediaは新聞記事等とは異なり、くだけた文体、口語表現、顔文字等が多用されるため、新聞記事から作られたデータで学習したシステムでは高い精度が期待できない。従って、ブログ等の自動解析技術の精度を向上させるためには、本データのような、ブログ記事から作成した学習データの整備が極めて重要である。

# 6 ツール、Webサービス、検索システム

## 6.1 上位下位関係抽出ツール

上位下位関係抽出ツールは、Sumidaらの手法[28]をもとにしてWikipediaダンプデータから上位下位関係となる単語対を抽出するツールである。上位下位関係とは、「YはXの一種（1つ）である」と言える下位語Yと上位語Xの関係と定義される。以下では上位下位関係を「X→Y」と表す。また、本ツールが出力する上位語、下位語はいわゆる「単語」にとどまらず「志摩市のスポーツイベント」のような複合的な言語表現も含む。

上位下位関係を結ぶ単語対、つまり上位下位関係候補の抽出では、図2に示したようにWikipedia記事の階層構造と定義文、カテゴリタグを用いた。

**階層構造**：記事のタイトル、セッションタイトル、箇条書きなどからなる階層構造から上位下位関係の候補を抽出する。図2（a）では、

表13 ブログ記事の例

| ID | タイトル | 記事 |
|---|---|---|
| 30 | 上賀茂神社 | せっかく来たので上賀茂神社も見ることに。ここは世界遺産にも登録されているのだとか。京都で最も古い神社の一つだそうです。バス停を下りてすぐの鳥居を抜けると、緑の空間が広がっています。そこにいくつか桜の木がありました。しだれ桜がきれいだった（以下略） |

表14 評価情報の例

| トピック | ID | 抽出文 | 評価表現 | 評価タイプ | 評価保持者 | 評価対象 | 対象関係 |
|---|---|---|---|---|---|---|---|
| 上賀茂神社 | 30 | ここは世界遺産にも登録されているのだとか。 | 世界遺産にも登録されているのだとか | メリット＋ | ［不定］ | ［上賀茂神社］ | 同一 |
| 上賀茂神社 | 30 | 京都で最も古い神社の一つだそうです。 | 京都で最も古い神社の一つだそうです | メリット＋ | ［不定］ | ［上賀茂神社］ | 同一 |
| 上賀茂神社 | 30 | しだれ桜がきれいだった。 | しだれ桜がきれいだった | 感情＋ | ［著者］ | ［上賀茂神社］ | 同一 |

「チーズ→プロセスチーズ」、「チーズ→ナチュラルチーズ」などが上位下位関係の候補として抽出される。

**定義文**：記事の第1文は定義文と見なせるが、そこから「～とは、～の一種。」などのパターンを用いて上位下位関係の候補を抽出する。図2（b）では、「食品→チーズ」が候補として抽出される。

**カテゴリタグ**：記事タイトルと記事のカテゴリタグからなる上位下位関係の候補を抽出する。図2（c）では、「発酵食品→チーズ」が候補として抽出される（「チーズ→チーズ」は上位語とその下位語候補が同一であるため除外する）。

抽出した全候補に対して、上位下位関係を表すか否かをSVMにより判定する。この判定には、上位語候補と下位語候補における形態素などの語彙的特徴、候補が現れたWikipedia記事の階層構造などの構造的特徴、そして上位語と下位語候補に関連するWikipediaのinfobox名、infoboxの属性などのWikipedia infoboxによる意味的特徴を素性として利用した。本ツールの上位下位関係獲得アルゴリズムの詳細についてはOhら[29]とSumidaら[28]を参照されたい。

本ツールにより2012年5月3日版の日本語Wikipediaから精度90％程度で抽出できた上位下位関係は約720万対であった。表15に、Wikipedia記事の階層構造、定義文、カテゴリタグから抽出された上位下位関係の数とその上位下位関係における上位語と下位語の異なり数を示す。表16に抽出した上位下位関係の例を挙げる。

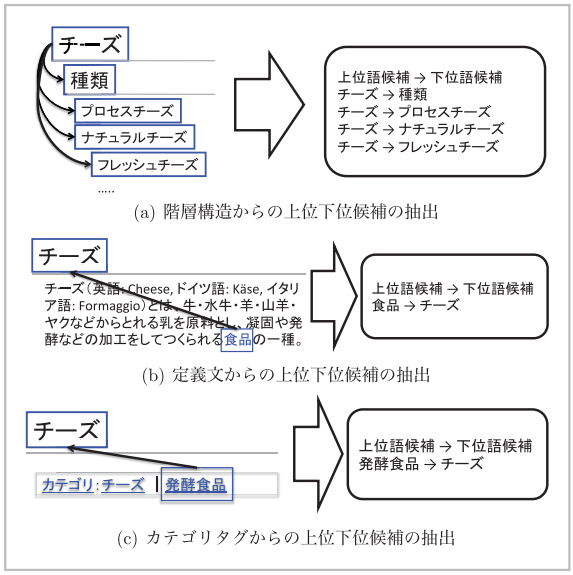


**図2 Wikipedia記事からの上位下位候補の抽出**

## 6.2 カスタム単語集合作成サポートサービス

我々はこれまで開発、構築してきた言語処理技術や言語資源を一般のユーザが容易に利用できるようにしたWebサービスを開発し、公開してい

**表15 2012年5月3日版の日本語Wikipediaから抽出した上位下位関係の数**

| 抽出先 | 上位下位関係数 | 上位語異なり数 | 下位語異なり数 |
|---|---|---|---|
| 階層構造 | 5,256,876 | 153,871 | 2,670,341 |
| 定義文 | 384,733 | 40,849 | 373,580 |
| カテゴリタグ | 1,766,485 | 63,876 | 652,284 |
| 合計 | 7,217,525 | 237,593 | 2,931,627 |

**表16 抽出した上位下位関係の例**

| 上位語 | 下位語 |
|---|---|
| 仏像 | 七面大明神像 |
| ジャズフェスティバル | BAY SIDE JAZZ CHIBA |
| 楽器 | カンテレ |
| 文房具 | スティックのり |
| 神楽団体 | 川平神楽社中 |
| プログラミング言語 | prolog |
| 戦争映画 | ハワイ・ミッドウェイ大海空戦 |
| 日本映画 | 歌う若大将 |
| AOCワイン | ラ・グランド・リュー　ブルゴーニュ |
| ゲーム | ファイナルファンタジーXI |
| テレビ時代劇 | 江戸の渦潮 |
| 放送事業者 | 西日本放送 |
| トラス橋 | 川島大橋 |
| 政治制度 | 直接民主制 |
| 病気 | セレン欠乏症 |
| 発電方式 | 太陽光発電 |
| 火力発電所 | ジェネックス水江発電所 |
| 羽毛恐竜 | シノサウロプテリクス |
| 都市 | バンクーバー |
| 市立中学校 | 伊佐市立大口南中学校 |
| 黄色顔料 | インディアンイエロー |
| 研究所 | 情報通信研究機構 |

る。ここで紹介する Web サービスを用いることによって、特別な知識がなくても任意の同じ種類に属するような単語集合や、因果関係のようなある種の意味的関係を持つ単語対を容易に獲得することができる。前者は「カスタム単語集合サポートサービス」として公開しており、後者は「意味的関係抽出サービス」として公開している。以下で「カスタム単語集合作成サポートサービス」について紹介し、**6.3** で「意味的関係抽出サービス」を紹介する。

カスタム単語集合作成サポートサービスは、意味的に類似する単語の集合（単語クラス）を作成するサービスである。単語クラスは多くの自然言語処理システムにおいて重要な役割を果たす。例えば、検索システムにおけるクエリの拡張や、キーワード連動型広告システムにおけるキーワード候補の自動提案などが応用例としてあり得る。

カスタム単語集合作成サポートサービスでは、約 1 億ページの日本語 Web 文書から、統計的な手法により、半自動的に大量の単語クラスを効率良く作成することができる。単語クラスを構成する単語の候補は、Web に出現する 1,000 万語である。手法の詳細に関しては文献 [30] を参照されたい。

本サービスにより、例えば以下のような単語クラスを取得することができる。

- 「お寺・神社」クラス
  - 「金閣寺」、「東大寺」、「正倉院」、「上賀茂神社」、「銀閣寺」、「三十三間堂」、「法隆寺」、「平等院」、「清水寺」、「日光東照宮」、「善光寺」、「厳島神社」、「平安神宮」、「中尊寺」、「出雲大社」、「白馬寺」、「飛鳥寺」、「明月院」、「浅草寺」、「三千院」、「薬師寺」、「南禅寺」、「室生寺」、「竜安寺」、「長谷寺」、「四天王寺」、「東福寺」、「唐招提寺」…
- 「釣り道具」クラス
  - 「釣り竿」、「餌」、「ルアー」、「針」、「おもり」、「テグス」、「天秤」、「リール」、「竹竿」、「玉網」、「ルアーロッド」、「フライロッド」、「釣り糸」、「タコテンヤ」、「ランディングネット」、「毛針」、「アンカーロープ」、「人工餌」、「さびき」、「ジグ」、「エギ」、「テキサスリグ」、「ワーム」、「餌木」、「カットテール」、「仕掛」…

本サービスは、図 3 にある Web ブラウザを用いたインターフェースにより、インタラクティブ

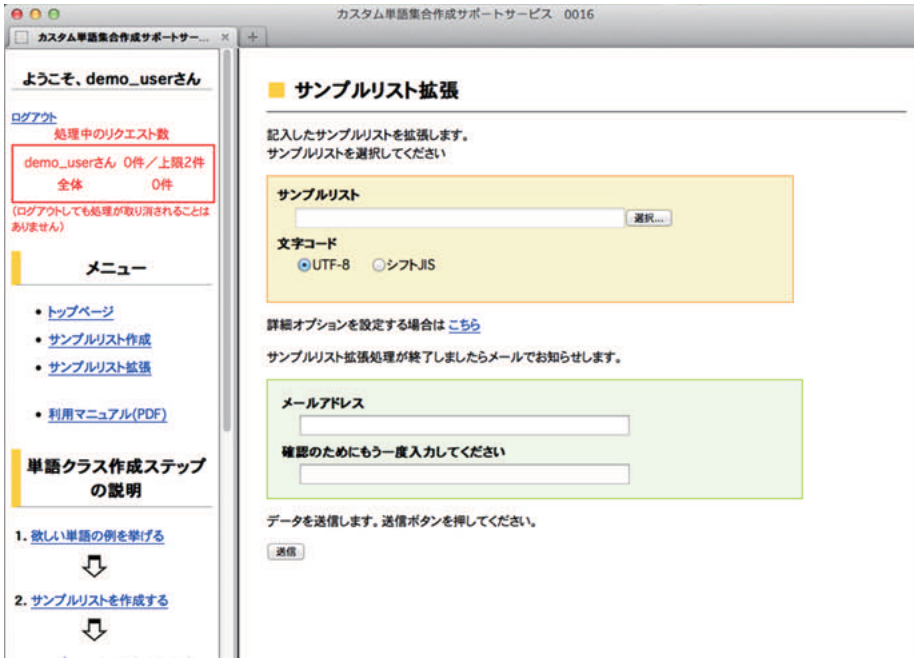


図 3 カスタム単語集合作成サービスのインターフェース

に単語クラスを作成できる。インターフェース上に示された指示に従って作業を進めるように設計されており、特別な知識を前提とせずに、所望の単語クラスを作成することができる。

### 6.3 意味的関係抽出サービス

意味的関係抽出サービスは「原因–結果」関係、「トラブル–予防策」関係、「音楽家–曲名」関係、「地名–名物」関係、「ヒーロー–敵」関係などの何らかの意味的関係を持つ単語対を抽出する Web サービスである。本サービスでは、Web 6 億文書から、統計的な手法により、半自動的に特定の意味的関係を持つ単語対を効率良く大量に抽出することができる。表 17 に「原因–結果」関係と「トラブル–予防策」関係に該当する単語対の例を挙げる。

本サービスを使ってユーザの所望する意味的関係を得るためには、その意味的関係を表す少数の言語パターンを指定するだけでよい。例えば、因果関係にある単語対を調べたい時は、「A が B の原因になる」「B の原因である A」を入力とすれば、「A によって起こる B」「A で B が発生」など、同じ因果関係で結びつきやすいパターンが自動的に学習される。このように、多くの人がすぐには思いつきにくい言語パターンを含めて、大量の類似パターンを学習し、最終的には全類似パターンを用いて、ユーザの求める意味的関係を持つ単語対が獲得される。

本サービスは自動的に学習した多様なパターンから膨大な意味的関係を獲得するため、本サービスにより、通常の Web 検索等では見落とす可能性の高い「意外ではあるが有用な情報」の発見が期待できる。

「カスタム単語集合作成サポートサービス」と同様、本サービスは、図 4 にある Web ブラウザを用いたインターフェースにより、インタラク

表 17 「原因–結果」関係と「トラブル–予防策」関係の例

| 原因–結果 | トラブル–予防策 |
|---|---|
| 連鎖球菌 – 化膿性関節炎 | 情報漏えい – 暗号化ソフトウェア |
| EB ウィルス – 伝染性単核球症 | 不正アクセス – ファイヤーウォール機能 |
| ツボカビ – カエルツボカビ症 | 床ずれ–エアマット |
| 断層–直下型地震 | 鳥害–防鳥ネット |


図 4 意味的関係抽出サービスのインターフェース

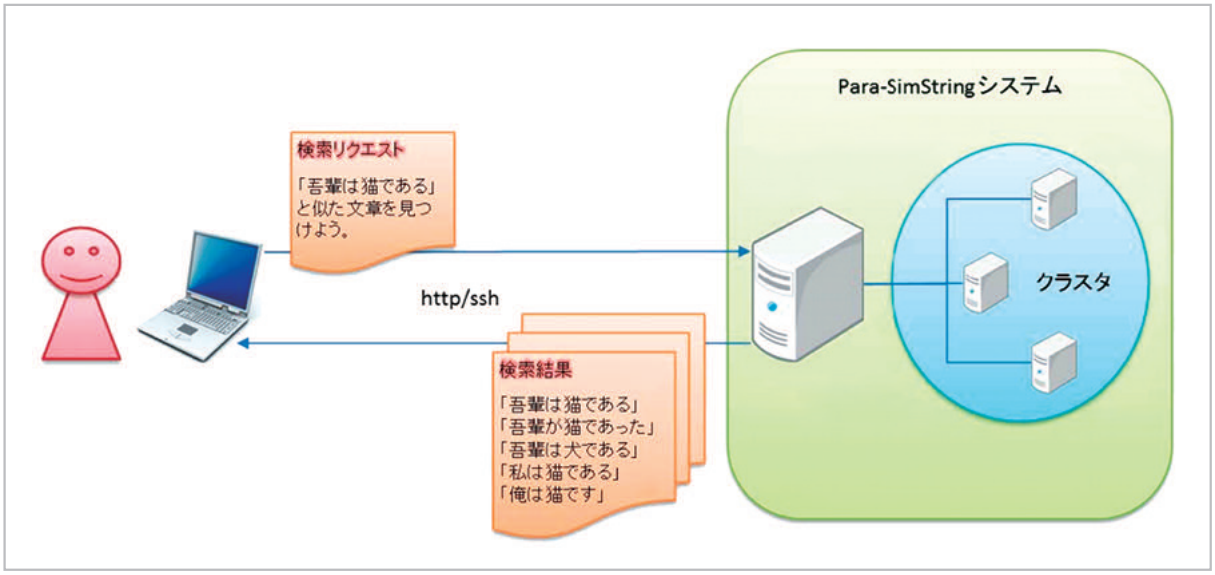


図5 Para-SimStringの入出力とシステム構成のイメージ

ティブに単語クラスを作成できる。インターフェース上に示された指示に従って作業を進めるように設計されており、特別な知識を前提とせずに、所望の意味的関係単語対を抽出することができる。

### 6.4 類似文字列並列検索システム：Para-SimString

文書は人間にとって最も身近なコミュニケーション手段の1つであるが、自然言語で書かれているため、同じ情報が複数の異なる表現、つまり言い換え表現で表される場合が頻繁に生じる。そのことが効率的な文書情報管理を妨げる一因であると言える。言い換え表現の自動認識技術は活発に研究されているが、大量の文書中の言い換え表現を高速で認識できる技術は今のところ存在しない。Para-SimStringは、表層上ある程度以上類似する言い換え表現にターゲットを絞り、また、並列処理技術を導入することで、大量の文書から言い換え表現を高速かつ柔軟に検索する手段を提供するものである。

より正確には、Para-SimStringは、クラスタマシン上に分散配置された大量の文書集合から、ユーザが入力したクエリ文字列と表層上類似する一行を並列、かつ、高速に検索するプログラムである。例えばクエリ文字列が「消費税の増税を閣議決定した」の場合、大量の文書から、「消費税増税を閣議で決定」「消費税率増を内閣が決定した」等の一行が（もし存在すれば）取得される。つまり、クエリ文字列と完全には一致しないが、ほぼ同じ意味を表していると考えられる表層上類似する文字列を網羅的に検索できる。

本システムの特長は、索引付けと検索の処理を並列で行える点にある。この特長は対象の文書集合が膨大である場合特に有効であり、さらに並列計算環境が利用可能であればこの特長をさらに活かすことができる。

なお、索引付けと検索処理のコアエンジンとして、オープンソースソフトウェアのSimString[*6]を利用している。

図5にPara-SimStringの入出力とシステム構成のイメージを示す。

### 6.5 Solr用クエリ拡張システム：QE4Solr

企業や大学等の組織内に蓄積されている文書の検索には、その組織の専門性、特殊性に対する理解が必要となる場合が多い。例えば、人工知能に関連する学科内の文書を検索するには、「AAAI」と「Association for the Advancement of Artificial Intelligence」、「アメリカ人工知能学会」が同じ対象を指しうることを理解していなくてはな

*6 http://www.chokkan.org/software/simstring/index.html.ja

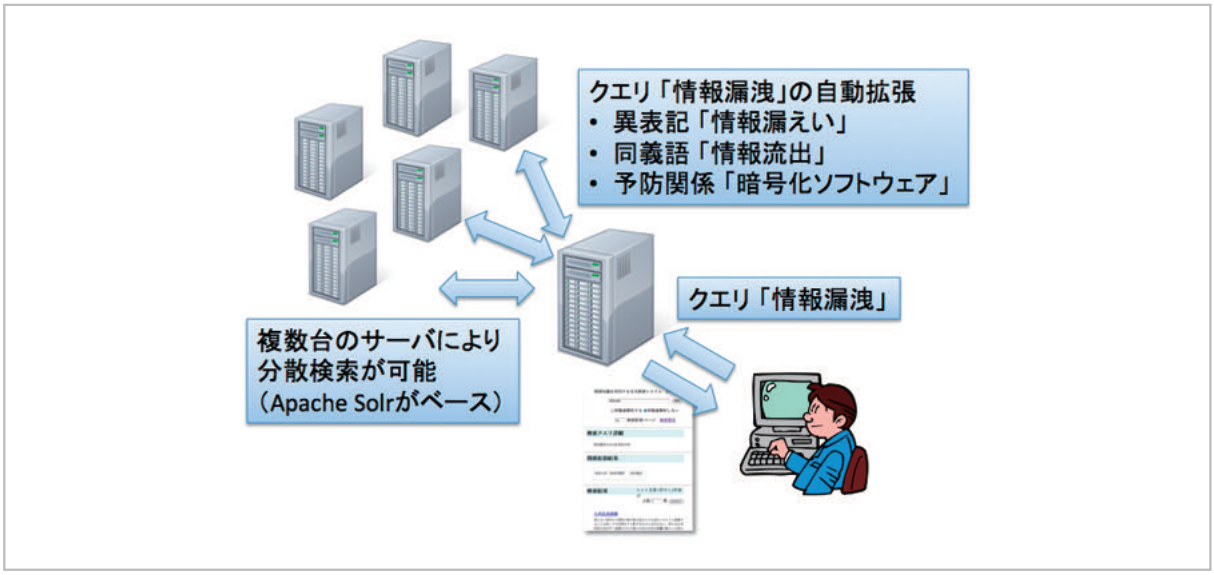


図 6　QE4Solr によるクエリ拡張とシステム構成のイメージ

らない場合があるだろう。QE4Solr はオープンソースの文書検索システム Apache Solr 用に開発されたクエリ拡張システムで、知識ベースを検索要求の自動拡張に柔軟かつ容易に利用できるのが特長である。本システムに、組織の専門性、特殊性に関する知識を明示化した知識ベースを組み込むことで、当該組織の特性にマッチした知的な検索を実現できる。あるいは、大量の異表記、同義語、意味的関係に関する知識ベースを本システムに組み込むことで、検索漏れの防止や、意外だが有用な情報の発見が期待できる。

そのような知識ベースは、本稿で述べてきた各種データベースやカスタム単語集合作成サポートサービス、意味的関係抽出サービス等の Web サービスにより容易に構築可能である。

本システムは索引付けと検索の並列処理が可能であり、Web アーカイブ等の大規模な文書データも効率的に処理できる。

図 6 に QE4Solr によるクエリ拡張とシステム構成のイメージを示す。

## 7　おわりに

本稿では、未公開のものも含めて、ユニバーサルコミュニケーション研究所情報分析研究室がこれまでに構築してきた基盤的言語資源を紹介した。

基盤的言語資源は高度に知的な言語情報処理システムのビルディングブロックであり、日本の ICT 技術の発展をその根底で支える重要なインフラストラクチャーである。しかし、その構築には、大規模並列計算環境や、言語データアノテーションの経験が豊富な多数のアノテータ、言語情報処理に精通する多数の研究者という、組織によっては賄うことが到底困難な多大なコストを要する。

当研究室のミッションの 1 つは、構築に多大なコストを要するものも含めて、高品質な基盤的言語資源を継続的に構築、配信し、それによって言語情報処理技術をはじめとする日本の ICT 技術の着実な発展に寄与することである。実際我々は、これまでの我々の活動により、基盤的言語資源の整備をここ数年で飛躍的に進展させたと自負している。

人間の知性に迫る言語情報処理システムの構築には、質、量ともにさらに大幅な飛躍が基盤的言語資源に求められるが、当研究室には本稿には掲載しきれなかった未公開の基盤的言語資源が多数存在し、それらが言語情報処理の世界にブレークスルーをもたらす可能性は高いと考えている。読者諸賢には今後の我々の研究成果に是非とも期待していただきたい。

## 参考文献


1 風間淳一，王軼諶，川田拓也，“基盤的言語処理ツール，”情報通信研究機構季報，本特集号，5-4, 2012.

2 内元清貴，鳥澤健太郎，隅田英一郎，柏岡秀紀，中村哲，“高度言語情報融合フォーラム（ALAGIN），”情報通信研究機構季報，本特集号，8-1, 2012.

3 風間淳一，デサーガステイン，鳥澤健太郎，村田真樹，“係り受けの確率的クラスタリングを用いた大規模類似語リストの作成，”言語処理学会第 16 回年次大会発表論文集，pp. 84–87, 2009.

4 黒田航，風間淳一，村田真樹，鳥澤健太郎，“Web データに対応できる日本語異表記対の認定基準，”言語処理学会第 16 回年次大会発表論文集，pp. 990–993, 2010.

5 小島正裕，村田真樹，風間淳一，黒田航，藤田篤，荒牧英治，土田正明，渡辺靖彦，鳥澤健太郎，“機械学習と種々の素性を用いた編集距離の小さい日本語異表記対の抽出，”言語処理学会第 16 回年次大会発表論文集，pp. 928–931, 2010.

6 Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata,“Large scale relation acquisition using class dependent patterns,” In ICDM '09: Proceedings of the 2009 edition of the IEEE International Conference on Data Mining series, pp. 764–769, 2009.

7 Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang,“Why question answering using sentiment analysis and word classes,” In EMNLP, 2012.

8 Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa,“ A bayesian method for robust estimation of distributional similarities,” In Proceedings of The 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 247–256, 2010.

9 Jun'ichi Kazama and Kentaro Torisawa,“Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations,” In ACL-08: HLT: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 407–415, 2008.

10 黒田航，李在鎬，野澤元，村田真樹，鳥澤健太郎，“鳥式改の上位語データの人手クリーニング，”言語処理学会第 15 回年次大会発表論文集，2009.

11 Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki, “Enhancing the japanese wordnet,” In The 7th Workshop on Asian Language Resources, 2009.

12 Kow Kuroda, Francis Bond, and Kentaro Torisawa,“Why wikipedia needs to make friends with wordnet,” In Proceedings of The 5th International Conference of the Global WordNet Association (GWC-2010), 2010.

13 Patrick Pantel and Deepak Ravichandran,“Automatically labeling semantic classes,” In HLT-NAACL '04: Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 321–328, 2004.

14 土田正明，Stijn De Saeger，鳥澤健太郎，村田真樹，風間淳一，黒田航，大和田勇人，“単語分布類似度を用いた類推による単語間の意味的関係獲得法，”情報処理学会論文誌，Vol. 52, 2011.

15 Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama,“Looking for trouble,”In Proceedings of The 22nd International Conference on Computational Linguistics, pp. 185–192, 2008.

16 風間淳一，Stijn De Saeger，鳥澤健太郎，後藤淳，István Varga，“災害時情報への質問応答システムの適用の試み，”言語処理学会第 18 年次大会，pp. 903–906, 2012.

17 Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama,“Large-scale verb entailment acquisition from the web,”In Proceedings of EMNLP, pp. 1172–1181, 2009.

18 橋本力，鳥澤健太郎，黒田航，デサーガステイン，村田真樹，風間淳一，“WWW からの大規模動詞含意知識の獲得，”情報処理学会論文誌，Vol. 52, No. 1, pp. 293–307, 2011.

19 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi,“Extracting paraphrases from definition sentences on the web,”In Proceedings of ACL/HLT, pp. 1087–1097, 2011.

20 橋本力，鳥澤健太郎，デサーガステイン，風間淳一，黒橋禎夫，“Web 上の定義文からの言い換え知識獲得，”言語処理学会第 17 回年次大会，pp. 748–751, 2011.

21 橋本力，鳥澤健太郎，デサーガステイン，呉鍾勲，風間淳一，“もう一つの意味的極性「活性／不活性」と知識獲得への応用，”言語処理学会第 18 回年次大会，pp. 93–96, 2012.

22 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama,“Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web,” In Proceedings of EMNLPCoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (to appear), 2012.

23 Peter D. Turney,“Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews,”In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 417–424, 2002.

24 Hiroya Takamura, Takashi Inui, and Manabu Okumura,“Extracting semantic orientation of words using spin model,”In Proceedings of the 43rd Annual Meeting of the ACL, pp. 133–140, 2005.

25 Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Motoki Sano, Jun Goto, Chikara Hashimoto, and Jong Hoon Oh,“Supervised recognition of entailment between patterns,”言語処理学会第 18 回年次大会発表論文集，pp. 431–434, 2012.

26 川田拓也，中川哲治，森井律子，宮森恒，赤峯享，乾健太郎，黒橋禎夫，木俵豊，“Web テキストにおける評価情報の整理・分類およびタグ付きコーパスの構築，”言語処理学会第 14 回年次大会，pp. 524–527, 2008.

27 川田拓也，中川哲治，赤峯享，森井律子，乾健太郎，黒橋禎夫，“評価情報タグ付与基準，”2009. http://www2.nict.go.jp/univ-com/isp/x163/project1/eval_spec_20090901.pdf

28 Asuka Sumida and Kentaro Torisawa,“Hacking Wikipedia for hyponymy relation acquisition,”In IJCNLP '08: Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 883–888, Jan. 2008.

29 Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa,“Bilingual co-training for monolingual hyponymy-relation acquisition,”In ACL-09: IJCNLP: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 432–440, 2009.

30 Stijn De Saeger, Jun'ichi Kazama, Kentaro Torisawa, Masaki Murata, Ichiro Yamada, and Kow Kuroda,“A web service for automatic word class acquisition,”In Proceedings of the 3rd International Universal Communication Symposium, pp. 132–138. ACM, 2009.

（平成 24 年 6 月 14 日 採録）

橋本　力（はしもと　ちから）
ユニバーサルコミュニケーション研究所
情報分析研究室主任研究員
博士（言語科学、情報学）
自然言語処理
ch@nict.go.jp

呉　鍾勳（Jong-hoon Oh）
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（工学）
自然言語処理
rovellia@nict.go.jp

佐野大樹（さのもとき）
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（言語学）
言語学
msano@nict.go.jp

川田拓也（かわだたくや）
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（文学）
言語学
tkawada@nict.go.jp