NICT発の基盤的言語資源
―言い換え処理向け言語資源を中心に―

独立行政法人情報通信研究機構　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　**橋本　力**

PROFILE　京都大学情報学研究科研究員、山形大学理工学研究科助教を経て、情報通信研究機構主任研究員。博士（言語科学、情報学）。情報処理学会論文賞、言語処理学会論文賞、言語処理学会年次大会優秀発表賞等受賞。

ch@nict.go.jp　TEL 0774-98-6896

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）が必要である。本稿ではこのような言語データを総称して「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。

情報通信研究機構（NICT）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、計算機による自動言い換え処理に深く関わるものを中心に、つまり、同義関係や含意関係、矛盾関係といった意味的関係に関わるものを中心に、当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。

2 高度言語情報融合フォーラム ALAGIN

本稿で紹介する基盤的言語資源は高度言語情報融 ALAGIN (Advanced LAnGuage Information Forum, http://www.alagin.jp) の言語資源配信サイト（http://alaginrc.nict.go.jp）から入手できる。また、本稿で紹介しきれない多数の基盤的言語資源や言語処理ツールも上記サイトから入手できるので、興味を持たれた方は是非サイトをチェックしていただきたい。

ALAGIN 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。ALAGIN ではこの他にも、情報通信研究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。

3 体言に関する言語資源

3.1 基本的意味関係の事例ベース

本データベースは、約 1 億ページの Web 文書上において文脈の類似度［1］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。本データ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以下の表に示す。

表 1　基本的意味関係の事例ベースの分類

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／赤ん坊
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。

本データベースの特色は、普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、ユーザーが「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

本データベースは、文字レベルの編集距離の近い、語句の異表記対（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。

本データベースで収集対象としているのは「ギョウザ、ギョーザ」のように一つの文字だけが異なる語句対、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 節で述べた「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
- ゴミ置き場／ゴミ置場
- ギタープレー／ギタープレイ

・ツィンマーマン／ツィマーマン
・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

本データベースには、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［2］で作られた 18,797 の異表記対（我々が「準異表記対」と呼ぶものも含む）に加え、2,758 の非異表記対を含んでいる。

自動獲得した異表記対のデータは、小島らの手法［3］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。以下に例を挙げる。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」には有名作曲家が文脈類似語として収録されている。

・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) ...

文脈類似語は、因果関係などの意味的関係の獲得［4］や Why 型質問応答［5］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 [6] を参照されたい。

3.4　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 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（sysetID: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以下に示す。

表２　日本語 WordNet の意味関係の例

分類	例
上位概念	動物／変温動物
被構成要素	エアバック／自動車
因果関係	映写する／表れる
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 節で述べた通り、「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　用言に関する言語資源

4.1　日本語パターン言い換えデータベース

本データベースは、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、一文中で任意の名詞 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 を防ぐ > というパターンに対して、下表にあるようなパターンが、言い換えとしての尤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。

本データベースは 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、下図にあるように、「交通事故による経済的な損害に関して」という文の断片からは <A による > というパターンが抽出される。

表３　<A は B を防ぐ > の言い換えパターン

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

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［4］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。

パターン：<AによるB>

図１　係り受けパスからのパターン抽出

4.2　動詞含意関係データベース

本データベースは、含意関係が成立している動詞一語または述語フレーズのペア（正例。483,686 ペア）と含意関係が成立していない動詞一語または述語フレーズのペア（負例。278,043 ペア）の計 761,729 ペアを列挙したものである。含意関係が成立する動詞または述語フレーズのペアとは、一方の指す事態が成立するなら、同時かそれ以前に、もう一方の指す事態も成立すると言えるペアである。以下に正例、負例それぞれの例を挙げる。「→」は、正例の場合は、左の動詞または述語フレーズが右の動詞または述語フレーズを含意することを、負例の場合は含意しないことを意味する。

・正例
　▶チンする → 加熱する
　▶酔っぱらう → 飲む
　▶セリーグ優勝する → リーグ優勝する
　▶粉塵を吸入する → ほこりを吸い込む
　▶微生物が水中のよごれを酸化分解する → 物質を微生物が分解する
・負例
　▶読書する → 寛ぐ
　▶深煎りする → 挽く
　▶準優勝する → 優勝する
　▶お仏壇に手を合わせる → 手に数珠を掛ける
　▶お客さま向けのサービスだ → お勧めのサービスだ

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、ある動詞または述語フレーズの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。

本データベースの述語フレーズ含意関係のうち、正例 312,544 ペア、負例 24,996 ペアは、文献［7］の手法による自動獲得結果をそのまま収録したものである。つまりこれらについては人手による全数チェックはしていないが、サンプルを調査した結果、90% 以上が正例、負例に正しく分類されていることを確認した。動詞間の含意関係獲得法の詳細については文献［8］を参照されたい。

4.3　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、今年度末に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。

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（http://wisdom-nict.go.jp/）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。ユーザー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ーに提示しなくてはならない。

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［9］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。

5 おわりに

本稿では、言い換え処理向けのものを中心に、情報通信研究機構（NICT）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で構築した基盤的言語資源を紹介した。当研究室では、本稿では紹介しきれなかった多くの言語資源や言語処理ツールを開発、公開している。それらの中には商用利用可能なものも多数含まれる。興味を持たれた方は、是非、ALAGIN 言語資源配信サイト（http://alaginrc.nict.go.jp）をチェックしていただきたい。

参考文献

[1]. 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．言語処理学会第第 16 回年次大会発表論文集，pp. 84-87，2009.

[2]. 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．言語処理学会第第 16 回年次大会発表論文集，pp. 990-993，2010.

[3]. 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．言語処理学会第第 16 回年次大会発表論文集，pp. 928-931，2010.

[4]. 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.

[5]. Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Junichi Kazama, and Yiou Wang. Why question answering using sentiment analysis and word classes. In EMNLP, 2012.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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 EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning, pp. 619-630. 2012.