自然言語処理と言い換え

乾 健太郎
奈良先端科学技術大学院大学

1 言語表現の多義性と同義性

言語は曖昧性であふれている．いや，正確には，言語を機械的に解析し，例えば翻訳するプログラムを作ろうとすると，言語は曖昧性であふれているように見える．

「彼女の手を握る」の「手」は英語の “hand” に訳せても,「他に手がない」の「手」に “hand” は使えない．「英語と数学を教える」と「親父と酒を飲む」はどちらも「〔名詞１〕と〔名詞２〕を〔動詞〕」の形をしているが，前者の「と」は等位関係を表す接続助詞，後者の「と」は随伴を表す格助詞である．どの場合の「手」がどの意味で，どの場合の「と」がどの意味を表すかは人間にとっては多くの場合易しい問題だが，計算機にとっては大問題である．

このように，言語を計算機で解析する際に，入力文に見かけ上いくつもの解釈があるように見えることを言語処理研究者は「曖昧性がある」あるいは「多義性がある」と言い，それらの解釈の中から書き手が意図した「真の」解釈を推定する問題を**曖昧性解消**あるいは**多義性解消**と呼んできた．言語処理研究を黎明期から長く牽引してきた機械翻訳では，何よりもまずこの多義性が問題になる．その意味で，言語処理研究の歴史はその大半が多義性との戦いの歴史だったと言える．

一方，機械処理の対象として見たときの言語にはもう一つ大きな問題がある．本特集号のテーマである**言い換え**の存在がそれである．例えば，次の２文は概ね同じ内容を伝えており，互いが互いの言い換えと考えられる．

(1) a. 「レ・ミゼラブル」の著者はV. ユーゴーだ
b. 「レ・ミゼラブル」はV. ユーゴーが書いた

言語にはこうした同じ情報を伝える，すなわち同義の言語表現がいくつも存在する．そのことがなぜどのように言語処理を難しくするのか．本稿ではこの視点から近年の言語処理研究の動向をながめ，言い換えに関する語彙資源の開発の現状と課題を論じる．日本語学と言語処理の研究交流の一助となれば幸いである．

2 同義性を認識するという問題

言い換えの存在が問題になるもっとも身近な例は文書検索であろう．Web の商用検索エンジンに代表される文書検索は，ユーザが入力した検索語を含む文書を網羅的に検索し，何らかの尺度で順次づけしてユーザに返す．「奈良先端大」で検索するときに,「奈良先端大」だけでなく「奈良先端科学技術大学院大学」や「NAIST」という文字列を含む文書も一緒に集めたいということになれば,「奈良先端大」,「奈良先端科学技術大学院大学」,「NAIST」が同義であるという知識を検索エンジンに持たせておく必要がある．多くの固有名詞や専門用語にこうした略称があり，さらに新しい名前が次々に造られていることを考えると，こうした同義語を網羅的に検索エンジンに与えるのは簡単な仕事ではない．

問題は略語のような単語の言い換えだけに留まらない．質問応答と呼ばれる問題を例にもう少し複雑なケースを考えよう．質問応答は，(2) のような質問文の答えを情報源である文書集合から探し出す課題である．

(2) 『坊ちゃん』の著者は誰ですか？

うまい具合に質問文と同じ言い回しの (3a) のような記述が情報源中にあれば，簡単な文字列照合で答えを見つけることができるが，実際にはそうでないことの方が多い．仮に情報源中に (3b) のような記述しかない場合でも，質問応答システムは正しい答えを見つけなければならない．

(3) a. 『坊ちゃん』の著者は夏目漱石です。
b. 夏目漱石は明治 39 年の春に『坊ちゃん』を雑誌「ホトトギス」に発表、...

すぐに気づくように，これも言い換えの存在に根ざす問題と考えることができる．(3b) から答えを探し出すには，(3b) が (3a) の内容を伝える別の言い回しであることが認識できればよい．ただし，(3b) は (3a) よりも多くの情報を伝えているので，厳密には両者は言い換えではない．(3b) が成り立てば (3a) も成り立つという意味で，両者は含意関係にあると言える（(3b) が (3a) を含意する）．すなわち，質問 (2) の答え (3a) を情報源の記述 (3b) から見つける質問応答課題は，その答えが情報源の記述から含意されるか否かを判別する含意認識問題に帰着できる．書名とその著者の関係だけをとっても，現実の文章では (4) ように様々な言い回しで表現される．質問応答では，こうした多様な表現の間の同義関係，あるいは含意関係を網羅的かつ正確に認識する技術が求められているのである．

(4) 《著者名》が「《書名》」を著す
《著者名》が「《書名》」を発表する
《著者名》の代表作「《書名》」
《書名》（《著者名》）

もう一つ興味深い例を挙げよう．与えられたトピックに関連する文書を集めて解析し，その要約を作成する複数文書要約と呼ばれる研究分野がある．ネット上に分散する情報を関連づけて分析する手段として今後重要性を増すと目される技術である．複数文書要約では，要約対象の文書が共通に含んでいる情報を見つけることがまず重要になる．例えば，要約対象の文書が (5) のような文をそれぞれ 1 つずつ含んでいたとしよう．

(5) a. マイナスイオンを取り込むと、酸が中和されて、弱アルカリ性に戻るので、疲労を取り除くことができます。
b. マイナスイオンはプラスイオンを抑え、身体の疲労を回復させるという、体に欠かせない大切な物質です。
c. マイナスイオンが健康増進に役立つという科学的証明はされていません。

(5a) の「マイナスイオンを取り込むと疲労を取り除くことができる」の部分と (5b)「マイナスイオンは身体の疲労を回復させる」の部分はほぼ同義である．また，これらの部分と (5c) の「マイナスイオンが健康増進に役立つ」には含意関係がある．もしシステムがこうした関係を認識できれば，(6) のような要約を生成できる可能性が出てくる．ここまでできれば大したものだろう．

(6) マイナスイオンは身体の疲労を取り除くと言われているが、科学的証明はされていないという意見もある。

もちろん，これらの問題を解くには現状の技術はまだまだ未熟である．しかし，こうした例から想像が広がるように，異なるテキストの間の同義性あるいは含意関係を認識する問題は，多義性の解消とともに多くの言語処理アプリケーションに共通する中心的な課題として近年認知されるようになり，ホットな研究トピックになりつつある [3, 8].

3 同義・含意関係知識の獲得

当面最大の問題は知識の収集である．前述のような問題を解決するには,「奈良先端大」と「NAIST」が同義であり，それぞれ「大学名」のインスタンスであること,「《書名》の著者は《人名》だ」や「《書名》は《人名》が書いた」が同義であり,「《人名》の代表作『《書名》』」はそれらを含意することを知識として計算機に与えておく必要がある．

3.1 既存の語彙資源からの抽出

既存のシソーラスの中には単語間の同義関係を直接的に記述したものがある．たとえば，EDR 電子化辞書[1]や WordNet[2]には，非常に細かい意味分類に基づく単語間の同義関係や上位下位関係が与えられている．例えば EDR では，次のような単語が同一の概念を指す語としてまとめられている．

(7) 相勤める，勤務する，勤労する，就役する，就労する，勤める，働く，労作する，労働する

また，国語辞典の語釈文も同義関係や上位下位関係の代表的な収集源である．たとえば，岩波国語辞典では「アイス」の第 3 語義の語釈に

(8) アイス (3) ＝ 「アイス　キャンデー」「アイスクリーム」の略。

とあり，ここから「アイス」の同義語として「アイスキャンデー」「アイスクリーム」が得られる．より一般的には，

(9) a. 家屋 ＝ 人が住むための 建物。
b. 書斎 ＝ 読書・執筆などをするための 部屋。

のように，語釈文の主要語から見出し語の上位語が得られる．語釈文は比較的統一された形式で記述されるため，単純な抽出ルールを用意するだけでも機械的に同義語や上位語を収集することができる [15].

語釈文からはもっと複雑な知識も収集できる．例えば,「倒す」の語釈文 (10a) は，(b) の上位下位関係だけでなく，(c)，(d) の手段-目的関係や (d) の行為-前提関係など，多様な関係を表していると解釈することもできる．

(10) a. 倒す ＝ 立っている物に力を加え傾け、横にする。
b. X が Y を倒す －上位→ X が Y を横にする
c. X が Y を倒す －手段→ X が Y を傾ける
d. X が Y を倒す －手段→ X が Y に力を加える
e. X が Y を倒す －前提→（行為前は）Y が立っている

こうした知識の抽出は完全な自動化は難しいとしても，ある程度人手をかければ可能であり [9]，こうした資源の整備と共有化が急がれる．

3.2 コーパスからの知識獲得

入手可能なコーパスの大規模化に伴って，コーパスから同義表現を獲得する試みも多数報告されている．これまでの方法は大きく，(a) パラレルコーパスから同義表現を獲得する試みと (b) 出現文脈の類似度に基づいてノンパラレルコーパスから類義語を獲得する試みに分けられる．

3.2.1 パラレルコーパスからの同義表現獲得

次の (a) と (b) のような対訳文[3]，すなわち同じ意味を持つ文を集めたものを対訳コーパス，あるいはパラレルコーパスと呼ぶ．

[1] http://www.iijnet.or.jp/edr/

[2] http://wordnet.princeton.edu/

[3] 例文 (11)，(12) は文献 [13] による．

(11) a. The athletic field *was swamped with* spectators.
b. 競技場は大勢の観客**で身動きができなかった**.
c. be swamped with ～ ⇔ ～で身動きができない

こうした対訳事例が大量にあれば，既知の語句の対応付け（例えば，“the athletic field” と「競技場」）とを自動的に行って，そこから (11c) のような新しい翻訳知識を自動的に獲得するできる場合がある．対訳コーパスからの翻訳知識獲得の試みについてはすでに多数の報告があり，ある程度成功を収めている．

翻訳知識は異言語間の同義表現と見なせるので，同一言語内の言い換えの獲得にも同様の方法が使えそうである．言い換えの場合，大量の言い換え事例の入手は翻訳の場合ほど容易でないが，それでもいくつか方法がある．

まず，(11b) と (12a) のように同じ原文に対して複数の翻訳がある場合は，それらを言い換え事例と見なすことができる．

(12) a. 競技場は大勢の観客で**膨れ上がった**
b. 《場所》が《人》で膨れあがる ⇔ 《場所》が《人》で身動きが出来ない

こうした複数の翻訳は，例えば『海底二万里』のように，同じ原著から何冊もの訳本がでている作品から得ることができる [1]．また，海外旅行用の旅行会話集なども同じフレーズに違う訳がいくつも付いていて，次のような面白い言い換え事例も集まってくる [10]．

(13) s. それ以上は安くなりませんか
t. それが最終的な値段ですか

さらに，例えば同じ事件を報道している複数の新聞社の記事のように，部分的に内容が重なる可能性の高い文章の集合も近似的にはパラレルコーパスと見なすことができ（コンパラブルコーパスと呼ばれる），有用な知識源として使える場合がある [12]．

3.3 出現文脈の類似度に基づく類義表現獲得

パラレルコーパスに頼らない方法もさかんに研究されている．その代表は出現文脈の類似度に基づく方法である．

よく知られるように，意味の近い単語は同じような使われ方をする傾向がある．試みに，手元の新聞記事約 30 年分のコーパスを使って名詞と動詞（格助詞と動詞の組）の共起頻度を調べると，例えば「着物」や「和服」は表 1 に挙げたような動詞とよく共起することがわかる．これを pLSI[4] と呼ばれる統計処理 [11] によって平滑化し，「着物」と「和服」各々についてそれと共起する動詞の出現確率を推定したものが表 2 である．「着物」と「和服」は概ね同義な語と考えられるが，両者の出現文脈の分布がほぼぴったり重なっている．このように，意味的に似ている語句はその出現文脈の分布も似ている傾向がある．**分布仮説**[6] と呼ばれるこの性質を利用すれば，出現文脈の分布の類似度から逆に語句の間の類似度を機械的に推定することができる．こうして計算される語句の意味類似度には**分布類似度**という術語が定着している．

パラレルコーパスからの知識獲得と違って，分布仮説に基づく知識獲得は大量に入手可能な生コーパスを知識源に使えるという利点がある．ただし，分布仮説はあくまでも傾向でしかなく，報告されている限りでは，同義語と類義語を区別できるほど分布類似度の解像度は高くない．例えば，前述の名詞と動詞の共起行列に戻って「着物」との出現文脈の類似度を調べると (14) のような語が上位に並ぶが，これらから同義語として「和服」だけを選ぶ方法は今のところない．

(14) ドレス，ブレザー，背広，ユニホーム，ワンピース，浴衣，リクルートスーツ，ジャケット，和服，ユニフォーム,...

また，この例からも想像されるように，分布類似度は表現間の意味的類似性を測る尺度に過ぎず，上位下位関係のような階層構造を得るにはさらに工夫が必要であり，今後の研究が待たれる．

4 意味の差異

これまで「同義」という概念をかなり無造作に使ってきたが，問題がそう単純でないことは明らかである．「言語は同義語を嫌う」と述べた Clark [2] の指摘のとおり，真に同義と考えられる表現は実際にはそれほど多くない．例えば，前節の EDR の同概念語の例 (7) を見ても，「勤務する」と「勤める」は同義語と言ってもよさそうだが，「勤労する」や「就労する」はそれらとは少し違ったニュアンスを持っている．

こうした類義表現間の意味の差異は，同義・含意関係を認識するタスクではすぐには問題にならないが，言い換えを生成する場合には最初から無視できない．言い換えの生成とは，与えられた文（あるいは文章）をそれと同じ内容の別の表現に変換する処理を指す．例えばネット上に急増する文書を高齢者や子供，外国人といった利用者の言語能力にあわせて読みやすい平易な文面になおすなど，機械翻訳と同様，コミュニケーション支援に多様な応用が期待されている領域である．

言い換え生成は，単純には，入力文（あるいは文章）の一部の語句をそれと同義な表現に置換することによって実現できる．しかし実際には，たとえ同義語といえどもいつでも置換できるとは限らない．たとえば，「随所」と「各地」は EDR 電子化辞書によると同概念語とされるが，厳密には意味が異なる．このため，(15a) の「随所」は「各地」に置換できるが，(15b) の「随所」は置換できない，というようなことが起こる．

(15) a. **随所**（→ **各地**）でがれきの山が生まれ，火災も発生し，死傷者も多数，確認されている．

[4] probabilistic latent semantic indexing

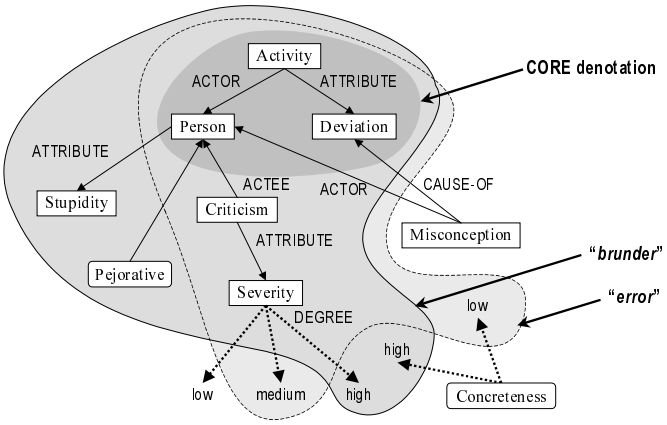

図 1: 類義語 “error” と “brunder” の意味記述 [4]

b. 片仮名交じりの文語体，しかも難解な言葉が**随所**（→ ∗**各地**）にあり，法学専攻の学生をすら悩ます現行刑法の法文が現代用語に書き換えられる．

こうした類義表現間の使い分けは現在の技術ではまだまだ難しいが，それでも二つの相補的な方向に研究が進んでいる．

第一の方向は，言い換え先の表現と周囲の文脈の繋がりの良さを統計的に評価する方法である．例えば，(16a) の「基盤」は，(b) の「土台」への言い換えは適当だが，(c) の「根底」は不適当である．

(16) a. 政党責任者が党の**基盤**を固める．
b. → 政党責任者が党の**土台**を固める．
c. → ∗ 政党責任者が党の**根底**を固める．

こうした判断は，3.3 で述べたような名詞と動詞の共起頻度に基づいて，例えば「根底」と「を固める」の共起のしやすさを統計的に調べることによってある程度機械的に行うことができる [5, 7]．

第二の方向は，類義表現間の意味や用法の差異を形式的に記述した語彙資源を構築し，語彙的選択のための制約として用いるアプローチである．例えば Edmond が提案する意味記述によれば，類義語 “brunder” と “error” の意味は，図 1 のように，どれくらい強い非難か（criticism の severity）とか馬鹿げた誤りかどうか（stupidity）といった属性によって区別される [4]．

日本語でも，例えば松吉らが開発した日本語機能表現辞書 [14] には各機能表現について難易度（５段階）と文体（常体，敬体，口語体，堅い文体）の情報が付与されている．松吉らの機能表現言い換えシステムは，この情報を利用して，例えば入力「見てくれるか」に対し，(17) のように多様な言い換え候補を生成するとともに，そこから指定の難易度や文体に合うものだけを選択することができる．

(17) 見て下さい，見てください，見てもらえるか，見てくれないか，見てちょうだい，見てもらえないか，見て下さるか，見ていただけますか

5 おわりに

言語処理は近年，大規模コーパスを用いた統計的手法の目覚ましい進歩によって形態素解析や構文解析などの浅い言語解析技術に飛躍的発展を見た．現在の研究は，そうした統計的手法の高度化を追求する方向と，意味の問題に一歩踏み込み本稿で紹介したような古くて新しい問題に再挑戦する方向に進んでいる．そうした文脈の中，言い換えの認識や生成は，応用横断的な有用性を持つだけなく含意や暗示的意味の問題も提供するなど，言語処理が意味の領域に向かうための格好の基本問題となっている．最後に紹介したような語彙意味資源の設計や開発の方法論など，言語学的研究に負うべきところも大きい．広く言語に関わる研究者の参画を期待したい．

謝辞

本稿の執筆は言い換えに関わる問題を再考する良い機会になった．機会を与えてくださった国立国語研究所の井上優氏に記して深く感謝する．

参考文献

[1] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 50–57, 2001.

[2] Eve Vivienne Clark. Conventionality and contrast: pragmatic principles with lexical consequences. In *Kittay and Lehrer* (*Eds.*), *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pp. 171–188. Lawrence Erlbaum Associates, 1992.

[3] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[4] Philip Edmonds. *Semantic representations of near-synonyms for automatic lexical choice*. PhD thesis, CSRI-399, Department of Computer Science, University of Toronto, 1999.

[5] 藤田篤, 乾健太郎, 松本裕治. 自動生成された言い換え文における不適格な動詞格構造の検出. 情報処理学会論文誌, Vol. 45, No. 4, pp. 1176–1187, 2004.

[6] Zellig S. Harris. Distributional structure. *Word*, Vol. 10, pp. 146–162, 1954.

[7] Diana Inkpen. A statistical model of near-synonym choice. *ACM Transactions of Speech and Language Processing*, Vol. 4, No. 1, pp. 1–17, 2007.

[8] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[9] 乾健太郎. 事態オントロジー：言語に基づく推論のためのコトに関する基本知識. 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp. 27–30, 2007.

[10] 大竹清敬. 用例に基づく換言: 中日旅行会話翻訳への適用. 言語処理学会第 9 回年次大会発表論文集, pp. 345–348, 2003.

[11] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 183–190, 1993.

[12] Yusuke Shinyama and Satoshi Sekine. Paraphrase acquisition for information extraction. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pp. 65–71, 2003.

[13] Satoshi Shirai, Kazuhide Yamamoto, and Francis Bond. Japanese-English paraphrase corpus. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS*) *Workshop on Language Resources in Asia*, pp. 23–30, 2001.

[14] 松吉俊, 佐藤理史. 体系的機能表現辞書に基づく日本語機能表現の言い換え. 言語処理学会第 13 回年次大会予稿集, pp. 899–902, 2007.

[15] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将. 国語辞典情報を用いたシソーラスの作成について. 情報処理学会自然言語処理研究会, NL-83, 1991.