■2 群（画像・音・言語）- 10 編（自然言語処理）

8 章 情報抽出

【本章の構成】

本章では，用語抽出（8–1 節），固有表現抽出（8–2 節），評価表現抽出（8–3 節），知識獲得（8–4 節），言い換え（8–5 節），要約（8–6 節），質問応答（8–7 節）について述べる．

■2 群 - 10 編 - 8 章

8-1 用語抽出

（執筆者：岡崎直観）［2009 年 8 月 受領］

用語抽出とは，ある分野のコーパスから用語を自動的に抽出する処理である．抽出された用語は，専門用語辞書や索引の自動構築に用いられたり，機械翻訳，知識抽出，要約などの自然言語アプリケーションで，コーパスを特徴付ける語彙の基本単位として採用される．

社会や技術の高度化・細分化に伴い，各分野において無数の用語が作り出されており，用語管理の必要性が高まっている．例えば，生命医学分野では，米国立医学図書館（National Library of Medicine：NLM）が，生命医学分野の文献及び抄録のデータベース MEDLINE*を提供している．このデータベースは，2008 年度時点で約 1688 万件の文献を収録しており，過去 1 年間に追加された文献数は 67 万件にのぼる．生命医学文献には，遺伝子，タンパク質，化合物，病気などの概念を表す用語が大量に含まれるが，専門家に依頼して MEDLINE 全体の用語を管理するのは限界がある[1]．生命医学分野で最も規模の大きいシソーラスである UMLS†は，2009 年時点で 521 万件の用語を専門家の手で管理しているが，MEDLINE の文献数と比較すると用語の数が十分とは言い難い．そこで，自然言語処理技術を生命医学文献に適用し，タンパク質の関係（相互作用など）を抽出したり，遺伝子に関する記述を集約するには，既存の言語資源に記述されていない用語を補う必要がある．

8-1-1 用語とは何か？

用語を抽出するにあたって，「用語とは何か？」を定義することが望ましいが，この問いに正確に答えるのは難しい．文献[2]は，様々な研究者が専門用語をどのように規定しているかをまとめ，専門用語の理論について考察を与えている．残念ながら，専門用語は経験的対象であるため，「専門分野の概念を指示する語もしくは句」や「専門分野で作られた語で，一般的に通用しない語」などの定義は，専門用語の本質をつかむには不十分であると論じている．

文献[3]は，用語抽出手法の包括的な調査を行い，これらの手法で暗黙的に導入している用語の特質として，用語の言語学的な構造（Linguistic Structure），単位性（Unithood），用語性（Termhood）を挙げている．言語学的な構造とは，用語を内部で構成する要素（語や形態素）の構造のことであり，具体的には「名詞の連続」などの品詞パターンで実装されることが多い．単位性とは，用語を構成する要素の文法的な結合，もしくは隣接による結合の強さ及び安定性を表し，連語（Collocation）に関する統計的尺度などで数値化される．用語性とは，用語が分野特有の概念を代表する度合いを表し，頻度などの統計的尺度で計測される．

用語抽出は，連語（複合語）抽出，固有表現抽出，未知語処理などのタスクと関連が深い．用語は複数の語から構成されることが多いため，連語や複合語（Multi-Word Expression）の抽出と類似した手法が用語抽出に採用される．固有表現抽出は，特定の意味カテゴリー（人名，組織名，地名など）を持つ用語を抽出するタスクとみなすことができるが，分野を代表する度合（用語性）はさほど考慮されない．

---

* Medical Literature Analysis and Retrieval System Online （MEDLINE®): http://www.nlm.nih.gov/databases/databases_medline.html

† Unified Medical Language System （UMLS）: http://www.nlm.nih.gov/research/umls/

表 8･1　用語抽出における代表的な品詞パターン

品詞パターン	用語の例（日本語）	用語の例（英語）
(名詞)+	**クラウドコンピューティング**; **ヒト免疫不全ウイルス**	cloud computing; Human Immunodeficiency Virus
(形容詞)+(名詞)+	**黒い雨**	global warming
(名詞)+の (名詞)+	**公共の福祉**	—
(名詞)+(前置詞)(名詞)+	—	receptor of estrogen; time to progression; weight for age
(数詞)+(名詞)+	**一等親血縁者**;	first degree relative

8-1-2 用語抽出の手法

これまで，数多くの用語抽出法が提案されてきたが，ほとんどの手法は，　品詞のパターン等を用いてコーパスから用語の候補を抽出する，　抽出された候補に「用語らしさ」のスコアを付与する，という 2 段階で構成される．スコア付けされた用語の候補は，閾値や上位からの順位などの基準で用語の認定をしたり，重み付き用語リストとして用いられる．

(1) 用語の語構成

第 1 段階では，与えられたコーパスに出現する言語学的な構造に着目し，用語の候補を取り出す．この処理は，しばしば言語学的フィルタ（Linguistic Filter）と呼ばれ，品詞の出現パターンで実装されることが多い．表 8･1 に，用語抽出でよく用いられる品詞のパターンと，対応する日本語と英語の用語の例を示す．これらのパターンは，品詞が付与された単語列を受け取り，正規表現にマッチした部分単語列を用語の候補として出力する．例えば「(名詞)+」は，名詞の 1 回以上の連接を用語候補として取り出す．なお，コーパス中の文字列を品詞付き単語列に変換するには，日本語の場合は形態素解析，英語の場合は品詞タガーが用いられる．

第 1 段階のルールで，コーパス中の用語をかなり正確に抽出することができるが，一方で用語としてふさわしくない表現も取得してしまうことがある．例えば，「(名詞)+」というパターンを「ヒト 免疫 不全 ウイルス」という 4 つの名詞の連接に適用すると，「ヒト」「ヒト免疫」「ヒト免疫不全」「ヒト免疫不全ウイルス」「免疫」「免疫不全」「免疫不全ウイルス」「不全」「不全ウイルス」「ウイルス」という 10 個の用語候補が得られるが，「ヒト免疫不全」や「不全ウイルス」が用語として適切かどうかは疑問が残る．

(2) 用語抽出における統計的尺度

そこで，第 2 段階では用語候補 t の用語らしさを，単位性や用語性などの観点から評価し，スコア付けを行う．用語性を測る統計尺度としては，コーパス中における出現頻度が代表的である．すなわち，コーパス中で出現する回数 $f(t)$ が多い用語ほど，その分野を特徴付ける概念を指し示すと仮定する．また，特定の文書に偏って頻出する用語を評価する TFIDF も，用語性を測る尺度として用いられる*.

$$\mathrm{TFIDF}(t) = \frac{f(t)}{N} \cdot \log \frac{|D|}{\mathrm{df}(t)}$$

ここで，N はコーパス中に含まれるすべての用語候補の総出現回数，$|D|$ はコーパス中に含

* 情報検索などで用いられる TFIDF は，文書ごとに語のスコアを計算するが，用語抽出の場合はコーパス全体で共通のスコアを計算したいので，若干定義が異なる．

まれる文書数，df(t) は用語候補 t を含む文書の数である．

単位性を測る尺度としては，用語を構成する語の共起の強さを測る尺度が用いられる．用語が連語として定型的に振る舞うのであれば，その構成要素は互いに独立に出現するのではなく，高い共起性を持つはずである．共起の強さを計る指標として，t 検定，z スコア，カイ二乗検定，相互情報量，対数尤度比などが用いられる[4]．例えば，語 x と y の連語から構成される用語候補の t 検定スコアは，次式で近似される[5]．

$$\mathrm{TScore}(t) = \frac{p(t) - \mu(t)}{\sqrt{\sigma^2/N}} \approx \frac{f(xy) - f(x)f(y)/N}{\sqrt{f(xy)}} \quad (t \text{ が 2 語 } x \text{ と } y \text{ で構成される場合})$$

ここで，$p(t)$ は用語候補 t の出現確率，$\mu(t)$ は t の構成要素が独立に出現すると仮定したときの出現確率，$\sigma^2 = \mu(t)\,(1 - \mu(t))$ である．

C-Value スコア[6]は，C-Value 法は用語の用語性と単位性（安定性）を統合した尺度である．

$$\mathrm{CValue}(t) = \begin{cases} \log_2 |t| \cdot f(t) & (t \text{ が別の用語候補に含まれないとき}) \\ \log_2 |t| \cdot \left\{ f(t) - \frac{1}{|W_t|} \sum_{w \in W_t} f(w) \right\} & (t \text{ が別の用語候補に含まれるとき}) \end{cases}$$

ここで，$|t|$ は用語候補 t を構成する単語の数，W_t は t を含む用語候補の集合，$|W_t|$ は集合 W_t に含まれる用語候補の数である．C-Value スコアは，長い（語数の多い）用語を重視する係数 $\log_2 |t|$ を除けば，t の頻度 $f(t)$ から，t を含む用語候補 $w \in W_t$ の頻度の平均値を引いたものと解釈できる．したがって，用語候補 t が他の候補の一部としてよく出現する場合（例えば「不全ウイルス」は「免疫不全ウイルス」「肝不全ウイルス」の一部としてよく出現する），候補 t のスコアが減点される．

用語抽出法は，本節で紹介したもの以外にも数多く提案されており，様々なバリエーションが存在する．抽出したい用語の分野やタイプによって抽出手法も異なるので，これらの優劣を一般的に決めるのは難しい．しかし，特定の分野に限定した正解コーパス*を利用し，用語抽出法の比較も報告されている[7]．

■参考文献

1) S. Ananiadou, D. B. Kell, and J. Tsujii：“Text mining and its potential applications in systems biology,” *Trends in Biotechnology*, vol.24, no.12, pp.571-579, 2006.
2) 影浦峡：“「専門用語の理論」に関する一考察,” 情報知識学会誌，vol.12，no.1，pp.3-12, 2002.
3) K. Kageura and B. Umino：“Methods of automatic term recognition: a review,” *Terminology*, vol.3, no.2, pp.259-289, 1996.
4) C. D. Manning and H. Schütze：“*Foundations of Statistical Natural Language Processing*,” MIT Press, 1999.
5) K. Church, W. Gale, P. Hanks, and D. Hindle：“Using statistics in lexical analysis,” In Lexical Acquisition: Using On-line Resources to Build a Lexicon, Lawrence Erlbaum, pp.115-164, 1991.
6) K. Frantzi, S. Ananiadou, and H. Mima：“Automatic recognition of multi-word terms: the C-value/NC-value method,” *International Journal on Digital Libraries*, vol.V3, no.2, pp.115-130, 2000.

* 例えば，NTCIR TMREC コーパス（日本語の情報処理分野の論文抄録），GENIA コーパス（英語の生命医学分野の論文抄録）など．

7) M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto：“Terminology extraction: an analysis of linguistic and statistical approaches,” *Studies in Fuzziness and Soft Computing*, vol.185, pp.255-279, Springer Berlin, 2005.

■2 群 - 10 編 - 8 章

8-2 固有表現抽出

（執筆者：磯崎秀樹）[2009 年 6 月 受領]

固有表現（Named Entity）とは，人名・地名・組織名などの，いわゆる固有名詞のことである．しかし，「固有名詞」と言うと，名詞 1 語という印象が強い．実際には「国境なき医師団」のように複数語からなっていたり，名詞以外の語が含まれていることもあるので，専門用語としては「固有表現」という．映画・音楽・本などのタイトルも固有表現の一種であるが，これらは非常に多様である．

これらの固有表現を，文章の中から見つけ出して，それが人名か地名か組織名かなどのクラスに分類することを「固有表現抽出」と呼ぶ．日時・距離・温度などの数値表現も文章の重要な要素であるため，これらの数値表現もしばしば固有表現として扱われる．

固有表現抽出が注目されはじめたのは，MUC（Message Understanding Conference）という 1990 年代の情報抽出の会議からである．固有表現抽出の端的な応用例として「質問応答システム」がある．質問応答システムは，「東大寺を建立したのは誰か？」などの質問を受け付けて，その答を大量の文書データの中から探し出して答える．この場合は「誰」と尋ねられているので，「東大寺」や「建立」で検索した文書の中に含まれる人名を探し出せば，回答を絞り込むことができる．手軽に使える日本語固有表現抽出として，係り受け解析器 CaboCha[KM04] に組み込まれているものがある．

固有表現抽出を実現する手法として，当初は人間がルールを書き下していた．これは数値表現では特に有効であるが，人名や組織名になると，ルールを書き下すのが難しくなる．現在広く使われているのは，機械学習による方法である．この方法では，文章中のここが人名，ここは組織名，という「正解データ」を大量に作成し，「機械学習」で一般的な判定基準を得る．日本語の正解データとしては，1999 年に開催された IREX（Information Retrieval and Extraction Excersize）というワークショップの「CRL 固有表現データ」が標準的である．英語の正解データとしては，2003 年に開催された CoNLL（Conference on Computational Natural Language Learning）という会議のデータが標準的である．

固有表現抽出が難しいのは，以下のような原因による．

- 多義性：「ワシントン」や「中野」は地名としても人名としても使われる．店の名前かもしれない．文脈を見なければ分類が難しい．
- 未知語：文章はまず形態素解析されてから固有表現抽出されるが，形態素解析の辞書に登録されていない固有名詞も多く，形態素解析が誤るので，固有表現抽出も誤る．
- 複合語：固有表現には名詞以外のものが含まれていることがあるし，「省電力機構」のように一般名詞が連続しているときに，それが固有表現かどうか判断するのも難しい．

機械学習としては，「隠れマルコフモデル（HMM）」[BMSW97]，「決定木学習」[SGS98]，「最大エントロピー法」[Bor99, UMM+00]，「サポートベクトルマシン（SVM）」[YKM02, IK03]，「条件付き確率場（CRF）」[ML03] などの手法が用いられる．

例えば「佐藤田中会談」は，常識的には「佐藤」と「田中」の会談だろう．しかし「佐藤

田中（でんちゅう）」という人がいるかもしれない．固有表現の切れ目を明確にするために使われるのが IOB2 [SV99] や Start-End [SGS98]などの方法である．IOB2 では各固有表現の最初の語を後続の語と別クラスにする．Start-End では最後の語・単独の語も区別する．

機械学習では，各単語をベクトルとして表現して，数学の問題として解く．「佐藤田中会談」の「田中」は，直前が「佐藤」で直後が「会談」である．そこで，「分類対象の単語は「田中」」，「後続の単語はサ変名詞」，「直前の単語は漢字」などの多数の特徴を成分とするベクトルに変換して，各ベクトルができるだけ正しく分類されるように学習する．

しかし，これらの手法には，人間が大量に正解データを用意しなければならない，という問題がある．そこで「半教師あり学習（Semi-Supervised Learning）」[AZ05, SI08]や「教師なし学習（Unsupervised Learning）」[CS00, ECD+05] という手法が試みられている．

■参考文献

AZ05) R.K. Ando and T. Zhang：“A high-performance semi-supervised learning method for text chunking,” in *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, 2005.

BMSW97) D.M. Bikel, S. Miller, R. Schwartz, and R. Weischedel：“Nymble: a high-performance learning name-finder,” in *Proc. of the Conference on Applied Natural Language Processing (ANLP)*, pp.194-201, 1997.

Bor99) A. Borthwick：“*A Maximum Entropy Approach to Named Entity Recognition*,” PhD thesis, New York University, 1999.

CS00) M. Collins and Y. Singer： “Unsupervised models for named entity classification,” in *EMNLP/VLC*, pp.100-110, 2000.

ECD+05) O. Etzioni, M. Cafarella, D. DOwney, A.-M. Poposcu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates： “Unsupervised named-entity extraction from the web: an experimental study,” *Artificial Intelligence*, vol.165, no.1, pp.91-134, 2005.

IK03) 磯崎秀樹, 賀沢秀人：“固有表現抽出のための SVM の高速化,” 情処学論, vol.44, no.3, pp.970-979, 2003.

KM04) 工藤　拓, 松本裕治：“カーネル法を用いた言語解析における高速化手法,” 情処学論, vol.45, no.9, pp.2177-2185, 2004. http://chasen.org/˜taku/software/cabocha/.

ML03) A. McCallum and W. Li：“Early results for named entity recognition with conditional random fields, freature induction and web-enhanced lexicons,” in *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*, 2003.

SGS98) S. Sekine, R. Grishman, and H. Shinnou：“A decision tree method for finding and classifying names in Japanese texts,” in *Proc. of the Workshop on Very Large Corpora (VLC)*, 1998.

SI08) J. Suzuki and H. Isozaki：“Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data,” in *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pp.665-673, 2008.

SV99) E.F. Tjong, K. Sang and J. Veenstra：“Representing text chunks,” in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, pp.173–179, 1999.

UMM+00) 内元清貴, 馬　青, 村田真樹, 小作浩美, 内山将夫, 井佐原均：“最大エントロピーモデルと書き換え規則に基づく固有表現抽出,” 自然言語処理, vol.7, no.2, pp.63-90, 2000.

YKM02) 山田寛康, 工藤　拓, 松本裕治：“Support vector machines を用いた日本語固有表現抽出,” 情処学論, vol.43, no.1, pp.44-53, 2002.

■2 群 - 10 編 - 8 章

8-3 評価表現抽出

（執筆者：高村大也）［2017 年 10 月 受領］

テキストの評価分析 (sentiment analysis)[1, 2]とは，テキストに表現されている評価や感情をコンピュータで処理し分析することである．例えば，「携帯電話を買いました．デザインも素敵で，使い勝手もいいし，かなり気に入っています.」という文書の著者は，購入した携帯電話に対し，そのデザインや操作性という観点において，肯定的な評価を与えていることが分かる．テキスト評価分析の技術により，自由回答アンケートの自動分析及び支援，また企業のコールセンターに蓄積された顧客の意見の集約，あるいは書き手の喜怒哀楽に基づく電子メールの自動分類など，幅広い応用が可能になる．特に，ブログなどの CGM の普及により，大量のテキストデータから情報を抽出するニーズが高まっており，テキスト評価分析は近年注目を浴びている．

ここでは，テキスト評価分析における技術を，評価表現辞書を構築する技術，評価という軸で文書や文を分類する技術，評価情報の詳細を獲得する技術の 3 つに大別して概観する．

8-3-1 評価表現辞書の構築

各単語について評価極性（Semantic Orientation）という属性を考える．例えば，“美しい” の評価極性は肯定的（Positive），“汚い” の評価極性は否定的（Negative）だと考えられる．各単語の評価極性を記述した評価表現辞書は，文書中の意見的な箇所の発見や，文書や文の評価極性の判定などに利用でき，評価分析の基盤となる．評価表現辞書の自動構築手法は 2 種類ある．

一つは，文書データ中での共起を利用した手法である[3]．この手法は，評価極性が等しい単語同士は共起しやすいという傾向を用いている．このような傾向は，「食事が 美味しく て，うれしい」のような用例から生じていると考えられる．“優れている” などの肯定極性の種単語（評価極性が既に分かっている単語）と共起しやすい単語は肯定，“劣る” などの否定極性の種単語と共起しやすい単語は否定と判定する．共起の尺度には自己相互情報量などが用いられる．もう一つは，語彙ネットワークを用いた方法である[4]．シソーラスなどの語彙ネットワーク上で互いに近い位置にある単語同士は，評価極性も等しい傾向があると考えられる．これを用い，種単語の評価極性を語彙ネットワーク上の他の単語に伝播させ，評価極性を決定することができる．また，単純な伝播でなく，語彙ネットワーク上に PageRank アルゴリズムを適用して評価極性を決定する手法や，語彙ネットワークに確率モデルを想定し，より尤もらしい状態を算出することにより評価極性を決定する手法などがある．

8-3-2 評価を軸とした文書及び文分類

文書や文を，評価や感情という観点から精度良く分類する技術を紹介する．ここで扱う分類問題は，肯定評価クラスと否定評価クラスの二値分類や，これに中立評価クラスを加えた三値分類，強い肯定と弱い肯定などを分けて考える多段階の分類，複数の独立した感情クラスへの分類などがある．また，文書や文がそもそも主観的であるか，客観的であるかを判定する分類問題もある．

最も単純な二値分類手法は，評価表現辞書を利用し，肯定極性を持つ単語数と否定極性を

持つ単語数の大小により評価クラスを決定する手法である[4]．その際，単語の周囲に否定表現（“ない”，など）がある場合には，極性を反転させて数える．

現在は，教師付き学習に基づく手法[5]が主流となっている．単語や，部分単語列などを素性として，ナイーブベイズ分類器やサポートベクトルマシンなどを用いることが多い．文書は文から成ることを利用し，文書の評価極性と，それが含む各文の評価極性を同時に推定する手法もある[6]．また，評価表現辞書を文書及び文分類に有効に利用する方法として，評価表現辞書が与える各単語の評価極性が，文脈により反転されるかどうかを予測しつつ文分類を行う手法がある[7]．

8-3-3 評価情報の詳細を獲得する技術

応用先によっては，評価の極性だけでなく，評価の対象（Subject）や，着目している属性（Attribute），評価者（Opinion Holder）など，詳細な評価情報が必要になることがある．例えば，ある携帯電話の機種について，ユーザは操作性に不満を持っているのか，デザインに不満を持っているのかなどの情報は，将来の製品改良に重要であろう．これらをレビュー文書などのテキストデータから抽出するためには文脈情報が使われるが，詳細情報項目間の共起傾向を示す統計値を用いることで性能の向上が望める[8]．詳細情報項目の抽出を個別に行うのでなく，例えば，各言語表現が評価者らしいか，評価を表しているかなどの個別の予測と，2 つの言語表現が評価者と評価という関係を有しているかの予測を，整数計画問題を用いて同時に行うことも有効である[9]．このようにして抽出した詳細評価情報を提示することで，簡易的な評価要約（Sentiment Summarization）が実現できる[10]．

■参考文献

1) B. Pang and L. Lee：“Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol.2, pp.1-135, 2008.
2) 乾　孝司, 奥村　学：“テキストを対象とした評価情報の分析に関する研究動向,” 自然言語処理, vol.13, no.3, pp.201-241, 2006.
3) P. Turney：“Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proc. of Annual Meeting of the Association for Computational Linguistics, pp.417-424, 2002.
4) M. Hu and B. Liu：“Mining and summarizing customer reviews,” Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.168-177, 2004.
5) B. Pang, L. Lee, and S. Vaithyanathan：“Thumbs up? sentiment classification using machine learning techniques,” Proc. of Conf. on Empirical Methods in Natural Language Processing, pp.76-86, 2002.
6) R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar：“Structured models for fine-to-coarse sentiment analysis,” Proc. of Annual Meeting of the Association for Computational Linguistics, pp.432-439, 2007.
7) D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura：“Learning to Shift the Polarity of Words for Sentiment Classification,” Proc. of Int. Joint Conf. on Natural Language Processing, pp.296-303, 2008.
8) N. Kobayashi, K. Inui, and Y. Matsumoto：“Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining,” Proc. of Conf. on Empirical Methods in Natural Language Processing, pp.1065-1074, 2007.
9) Y. Choi, E. Breck, and C. Cardie：“Joint Extraction of Entities and Relations for Opinion Recognition,” Proc. of Conf. on Empirical Methods in Natural Language Processing, pp. 431-439 (2006).
10) I. Titov and R. McDonald：“A Joint Model of Text and Aspect Ratings for Sentiment Summarization,” Proc. of Annual Meeting of the Association for Computational Linguistics, pp.308-316, 2008.

■2 群 - 10 編 - 8 章

8-4 知識獲得

（執筆者：乾健太郎，鳥澤健太郎，関根　聡）[2009 年 8 月 受領]

自然言語理解を実現するためには言語に関する様々な知識が必要である．例えば，“ I saw a girl with a scarf ”という文を「私はスカーフを使って女の子を見た」ではなく「私はスカーフを付けた女の子を見た」と翻訳するためには，一般的に scarf は女性が身につけるものであり見るための道具ではないという知識が必要である．また，「太郎と花子が都会に出たが，彼だけ戻った.」という文において，「彼」が「太郎」を指すことを理解するためには，「太郎」は男性であり，「彼」は男性を指示する代名詞であるという知識が必要である．このような知識を人手で作成することはほとんど不可能である．言語で書かれた語彙に関する必要な知識の種類と量は限りなく広い．それは，分野や文脈依存性を考慮に入れたり，知識の優先順位を付けるためのスコアリングや確率の付与された知識を作成する必要があるためである．人手でこのような知識を作ることの限界は，1990 年前後の人工知能研究の限界をもたらし，その後，意味論の研究は停滞していたが，2000 年を過ぎた辺りから大量の文書データを基に言語解析のための語彙知識を作成しようという研究が盛り上がりを見せている．現在，WEB や新聞記事など大量の文書データが入手可能であり，コンピュータの処理能力とデータの蓄積量は向上しており，様々な研究が活発に進められている．この分野の研究の内容は 2 つの軸でまとめることができる．一つは，何の知識を獲得するか（対象），であり，もう一つはどうやって獲得するか（方法論）である．以下では，この 2 つの軸について重要な点をまとめてみる．

8-4-1 対象　知識の対象は，この世に存在し，文書に記述できるもの全てが対象になるわけであるが，主には，名詞で表現される「モノ」と動詞で表現される「コト」に分類される．それぞれの詳細を見ていくと，モノには以下のような語彙知識が必要なことが分かる．

- 同義表現：あるモノに関する表現が同一の指示対象を指すという知識
- 上位下位関係：バナナや桃が果物であるというモノの間の概念的な上位下位関係
- 修飾関係：主に名詞と形容詞の関係で表現されるモノに関する修飾関係
- 属性：東京が日本の首都であるといったモノの特有の属性
- 部分全体関係：主翼や胴体が飛行機の部分をなすといった部分全体関係

上記のモノに関する語彙知識は全体の一部であり，モノとモノの関係を考えると，モノのライバル関係，従属関係，そのほかモノの種類によって様々な知識がある．

コトの語彙知識には主に以下の 4 種類の関係が見出せる．

- 上位下位関係：「歩く」は「移動する」の一例であり，下位概念である．
- 部分全体関係：「就寝する」に対し「布団に入る」という部分的な事象が存在する．
- 因果関係：「雨が降る」と「洗濯物をしまう」といった事象間の因果関係

- 時間関係：「歯を磨く」「うがいをする」といった時間の前後関係

また，コトとモノの間の関係も重要な語彙知識であり，それには以下のようなフレーム上で必要な知識が議論されている．

- 格フレーム：特定の動詞がある特定の種類の名詞を主語や目的語，道具格など荷物という格フレームの知識であり，構文解析や照応解析で使われる[1]．
- 生成語彙論：モノにはモノの目的，機能を表す telic role（「本」には「読む」という telic role がある）とモノが生成にいたる agentive role（「本」には「書く」という agentive role がある）があるとする理論である[2]．
- スクリプト：コト，つまり出来事が生じる典型的な系列を記述した知識であり，テキストの意味解釈に有用とされた．例えば，「レストラン」でのスクリプトは「席につき，注文をし，食事をし，支払いをする」といったものである[3]．

8-4-2 方法論大量な文書から語彙知識を獲得する方法論として，獲得するための知識をどこに求めるかという問題に対する 2 つのアイデアと，代表的なアルゴリズムであるブートストラッピングについて説明する．

(1) 語彙構文パターン

特定の言語表現がある特定の意味を曖昧性少なく表現することがあるという性質を使って，その意味の知識を獲得する方法である．例えば「X や Y などの C」という表現からは，大多数の場合「C の下位概念の例としての X や Y」という知識を獲得することができる[4]．

(2) 文脈類似度

「同じコンテキストに現れる単語は意味的に似ている」という原則[5]に基づいて，同義語と類義語やパラフレーズなど類似のモノやコトに関する知識を獲得するために使われる方法である．

(3) ブートストラッピング

自分が欲しいと思っている表現の種や，自分が欲しいと思っている単語の種を与え，そこに類似した表現や単語を収集する方法がブートストラッピングの方法である．例えば，（モーツアルト，1756 年）という人名と誕生年の種ペアから，それらの単語が表れる表現を検索し，そこで得られた表現のうち，この関係を示す度合いの強い表現を選びそれを誕生年表現とする．次には，この誕生年表現に表れる人名と年表現のペアを獲得し，人名と誕生年の知識を大量に獲得するという方法である[6]．上記に説明した語彙構文パターンや文脈類似度を応用することもできる．

これらの方法以外にも，対訳文書や同じ報道をしている新聞記事などの既存の情報のアライメントを使って語彙知識を獲得する方法や，複数の単語が平均期待以上に共起している場合にはそれらの単語には何らかの関係があるとする方法，大量の文書を構文解析し，その中から繰り返し得られる情報には共通する知識が含まれると仮定して知識を獲得する方法などがある．また，このように教師なしで獲得した知識は精度が高くない場合が多く，人手によ

る整理が必要であるが，その労力を最小化するための方法として能動学習や WEB 上の大衆の知識を利用したクラウドコンピューティングなどの利用にも注目が集まっている．

■参考文献

1) C. Fillmore：“Frame semantics and the nature of language,” in Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, vol.280, pp.20-32, 1976.
2) J. Pustejovsky：“The Generative Lexicon,” MIT press, 1995.
3) R. Schank and R. Abelson：“Scripts, Plans, Goals and Understanding,” an Inquiry into Human Knowledge Structures (Chap. 1-3), L. Erlbaum, Hillsdale, NJ, 1977.
4) M. Hearst：“Automatic acquisition of hyponyms from large text corpora,” in COLING-92, 1992.
5) Z. Harris：“Distributional structure,” in: Katz, J. J. (ed.) The Philosophy of Linguistics. New York: Oxford University Press, 1985.
6) S. Brin：“Extracting Patterns and Relations from the World Wide Web,” in WebDB, 1998.
7) Y. Matsumoto and T. Utsuro：“Lexical Knowledge Acquisition,” in ’Handbook of Natural Language Processing,’ R. Dale, H. Moisl, H. Somers (eds.), Marcel Dekker, Part II, Chapter 24, pp.563-610, 2000.

■2 群 - 10 編 - 8 章

8-5 言い換え

（執筆者：佐藤理史）[2009 年 7 月 受領]

8-5-1 言い換えとは

「あることを表現するのに，色々な言い方がある」ことは，自然言語が持つ重要な特徴の一つである．色々な言い方（同じ意味内容を表す複数の表現）を結びつける変換を，言い換えまたはパラフレーズ（Paraphrase）と呼ぶ．このほかに，「言い換え」という語は，言い換えられた表現，あるいは，元の表現と言い換えられた表現のペアを指すこともある．ことばをことばで説明することを可能としているのは，言い換えの存在による．

自然言語処理が伝統的に扱ってきた問題は，表現の曖昧性の問題，すなわち，同一の表現が複数の異なる意味を表しうることに起因する問題（同形式異内容の問題）である．これに対し，言い換えでは，異なる表現が同一の意味内容を伝達しうることに起因する問題（異形式同内容の問題）を解くことが，中心的課題となる．

言い換えのタスクは，大きく，言い換え認識と言い換え生成の 2 つに分けられる．前者は，2 つの表現が意味的に等価かどうかを判定するというタスクである．後者は，ある表現から，それと意味的に等価な表現を生成するというタスクである．いずれの場合も，「意味的に等価」とは，近似的に等しいということであり，完全に等価であることを意味するわけではない．2 つの表現間には何らかの差異が存在することに注意を払う必要がある．

言い換えの現象は多岐にわたっており，十分に整理・分析されてきたとは言い難い．しかしながら，意味的等価をどのようなレベルで考えるか，どのような言語単位で言い換えが行われるか，という 2 つの軸は，言い換え現象を整理する際に有用と考えられている．

意味的等価性のレベルとしては，内包的意味の等価性，参照的意味の等価性，語用論的効果の等価性の 3 つのレベルが考えられる[1)]．内包的意味の等価性とは，おおよそ，2 つの表現の真偽値が一致すること（表現 A が真であれば表現 B も真となる．その逆も成り立つ）である．同義語の置換（例えば，「本」と「書籍」）や構文変換による言い換え（例えば，能動態・受動態変換）の多くが，ここに分類される．参照的意味の等価性とは，例えば，「筆者の主張」と「佐藤の主張」のように，参照対象が同一であることを指す．最後の語用論的効果の等価性は，「宿泊したいのですが」と「部屋は空いていますか」のように，その発話によって達成したい目的が同一である場合を指す．

一方，言い換えの言語単位としては，語，句，節（単文），文，段落などが考えられる．言い換えの単位が大きくなればなるほど，各種の要因が絡み合い，現象は複雑化する．

8-5-2 言い換え認識

情報検索や質問応答などのタスクでは，表現の一致ではなく意味の一致が求められる．例えば，「日英機械翻訳システム」という情報要求に対しては，「日英自動翻訳システム」や「日本語を英語に翻訳するプログラム」について書かれた文書がその情報要求に合致する．しかし，そのような表現を含んだ文書を出力するためには，それらの表現が「日英機械翻訳システム」の言い換えであることが認識できなければならない．質問応答では，含意と呼ばれる，より緩やかな関係の認識も必要となる．例えば，「村上春樹は多くの小説を書いた」は「村上

春樹は作家である」を含意する（前者が真であれば，後者も真となる）．これが認識できるようになれば，前者の記述に基づいて，「村上春樹は作家か？」という質問に正しく答えることができる．

語の言い換え認識は，語の同義性判定であり，シソーラスを利用することによりそれなりに判定できる（一般に，シソーラスは類義語であるか否かの情報を提供するが，同義語であるかどうかの情報は提供しない）．句の言い換え認識は，同義語の情報と句の構成規則を組み合わせることにより実現できると考えられる．節（単文）の言い換え認識は，ある種の意味表現に変換するか，あるいは，変換規則を適用して標準形に変換した後，等価性を判定することになろう．ただし，言い換えの単位が大きくなるにつれて，認識すべき意味的等価性が応用に強く依存してくる可能性がある．現在，含意の認識という問題設定が広く採用されているのは，比較的明確にその関係の有無が判定できるからであろう．

8-5-3 言い換え生成

言い換え生成は，いわば単言語内翻訳である．通常の 2 言語間の翻訳とは異なり，言い換え生成は，何らかの目的（なぜ言い換える必要があるのか）を必要とする．典型的には，文を短くする，分かりやすくする（難易度を下げる），文体を変更する，などがその目的となる．つまり，言い換え生成は，単に，入力された表現と意味的に等価な表現を生成するだけでは不十分であり，言い換えの目的に合致した表現を選択することが必要となる．しかしながら，現在まで研究では，そのようなレベルには達していない．

言い換え生成の実現法は，変換規則とその変換規則の適用の対象となる内部構造の種類によって分類できる．言い換えの単位を語に限定するのであれば，文を語の列に変換し，そのうちの幾つかを同義語で置き換えることでも言い換え生成が実現できる．句や節にまで言い換えの単位を広げる場合は，文を構文解析し，得られた構文構造に対して，その一部を書き換える変換規則を適用する方法が用いられる．もちろん，より深い解析処理を前提とすることもできる．特定の文法理論と意味理論を採用し，その枠組下で文を意味構造に写像すれば，同一の意味構造から生成しうる文の集合が言い換え集合（内包的意味が等価な文の集合）となる．このアプローチは，言い換え認識にも適用可能な方法であるが，現在の言語処理の技術レベルでは，頑健性および精度に問題がある．

ある表現から，それと同義と思われるような表現の候補を生成する機構の実現自体は，それほど難しくはない．しかし，実際に，多様な言い換え表現を生成し，かつ，過剰な候補を生成しないようにするのは，至難の技である．そのため，候補生成の後，それぞれの候補の適切さを評価する処理を設けることが多い．そこでは，意味差分，文法的適格性，コロケーション，表現の自然さなどを評価すべきであるが，そのいずれもが，どのように評価すればよいか，十分にはわかっていない．例えば，「書籍購入代金」の「書籍」を「本」に置き換えることによって「本購入代金」が生成されるが，この表現よりも「本の購入代金」や「本を購入するためのお金」の方が望ましいと判定する必要がある．

いずれにせよ，言い換え生成は，未成熟な技術であり，これからの研究を待たなければならない点が多い．人間に匹敵する言い換え生成能力実現への道は，まだ見えない．

■参考文献

1) 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.

■2 群 - 10 編 - 8 章

8-6 要約

■2 群 - 10 編 - 8 章

8-7 質問応答

（執筆者：福本淳一）［2009 年 8 月 受領］

質問応答技術とは，自然言語で記述された任意の質問文に対して新聞記事や Web 上の情報などの大量の文書情報から適切な回答を得るものである．質問応答に関する研究としては，積み木の世界の操作に関する質問を扱った SHRDLU や，月の岩石などのついての質問を扱った LUNAR といったシステムがあったが，これらは対象分野についてあらかじめ詳細に記述された知識があることを前提にした質問応答技術である．本節で扱う質問応答は，質問対象分野について詳細に記述された知識は前提としていないものである．

質問応答技術に関する研究は, 情報検索技術に関する評価型ワークショップである TREC（Text Retrieval Conference）footnotehttp://trec.nist.gov/において，1999 年の TREC-8 で Question Answering（QA）トラックが設定されてから注目を集めるようになた．TREC の QA トラックでは，ある事実に関する質問文に対して回答と思われるもの上位 5 件を返すものであり，回答文字列は対象となる文から抽出された 50 バイトもしくは 250 バイトであった．その後，TREC 2001 では回答が複数存在する質問に答えるリストタスク，一連の関連質問に答えるコンテキストタスク，TREC 2003 では事物の定義を答えるデフィニションタスクが追加され，対象となる質問の範囲が広がっている．日本においては，国立情報学研究所主催の NTCIR（NII-Test Collection for Information Retrieval）*ワークショップのタスクとして第 3 回から QAC（Question Answering Challenge）が設定された．QAC では，TREC と同様の 5 位までの順位付きで回答を返すもの，可能な回答をすべて返すリストタスク，ある話題に関して連続する関連質問を与えるコンテキストタスクが設定されている．

質問応答システムの回答抽出までの一般的な処理の流れは，まず，入力された質問文に対して回答としてどのような情報を求めているのかを示す質問タイプの認識と文書検索のためのキーワード抽出を行う質問文解析を行う．次に，質問文解析で得られたキーワードを用いて対象となる知識源から文書検索を行う．検索された文書に対しては，質問文解析で認識された質問タイプと適合する回答候補の抽出を行う．最後に各回答候補について質問文中のキーワードの距離などの情報を用いて回答候補の重み付けを行い，最も重みの高い回答候補順に上位何件かの回答を返す．例えば，「恐竜が絶滅したのはいつですか？」のような質問を質問応答システムに与えた場合「いつですか？」の表現から質問タイプとして時間情報を得るとともに「恐竜」「絶滅」がキーワードとなる．文書検索では，これらのキーワードを含む文書が情報検索技術を用いて検索され，検索結果の文書から質問タイプとして時間情報となるものとして，「約 6500 万年前」「1 億 5000 年前」「1 週間」などが回答候補として抽出される．そして，質問文から抽出されたキーワード情報と回答候補との関連性として，単語距離などの情報を用いて距離の合計が最も短いものを選択するなどの方法で回答の重み付けを行い，最も重みの高い候補として「約 6500 万年前」が得られる．

質問応答システムにおいて重要な技術として質問タイプの判定がある．従来の質問応答技術では，factoid 型質問と呼ばれるある事実に関する情報として，人名，組織名，地名，価格，数値などの「名称」を対象にした質問がある．これらの認識のためには，それぞれの質問タ

* http://research.nii.ac.jp/ntcir/

イプに応じた質問文の表現パターンを記述することで判定を行うものや学習による判定手法もある．質問タイプ情報は，回答抽出において回答候補を抽出するためにも用いられることから，固有表現抽出（8-2 節）の抽出タイプと同じタイプ分類が用いられる．質問タイプごとの質問表現を以下に示す．

- 人名に関する質問
  「〜したのは誰ですか？」「〜は何という名前ですか？」
- 場所名に関する質問
  「〜したのはどこですか？」「〜したの場所はどこですか？」
- 組織名に関する質問
  「〜したのはどこですか？」「〜は何という部署ですか？」
- 時間に関する質問
  「いつ〜しましたか？」「〜は何年でしたか？」
- 価格に関する質問
  「〜はいくらですか？」「〜価格はどのくらいですか？」
- 割合に関する質問「〜は何%ですか？」「〜はどのくらいの割合ですか？」

以上の factoid 型質問だけでなく，「〜はなぜですか」「どのようにして」「〜は何ですか」などの non-factoid 型質問と呼ばれる質問も質問応答技術の対象として扱われてきている．non-factoid 型質問とは，回答が事実に関する「名称」にとどまらず，文やパラグラフの範囲が回答となる質問である．例えば，「なぜ」という質問表現により，原因や理由を問う質問に対する回答抽出方法として，文間の意味的な関係を利用した手法がある．質問文「なぜ電車が遅れたのですか？」に対して，知識源から例文「信号が故障したため，電車が遅れました.」が得られた場合，例文中では「ため」という表現により，因果関係が示されており，一方が質問文とマッチし，他方が回答として選択され，回答「信号が故障したため」を得ることができる．また，「どのようにして」のように手順を問う質問の場合，まず，知識源から手順に関して記述されている部分を箇条書き表現や html タグ情報などから抽出し，抽出された手順表現と質問文から抽出されたキーワードとの距離などの情報を用いて回答を得る方法がある．「〜は何ですか」のようにある事柄についての定義や説明を問う質問については，知識源の文書から「〜とは... である.」のような表現パターンに注目することで言葉の説明表現を抽出し，それらとの照合などにより回答を得る方法がある．このように non-factoid 型質問と呼ばれる質問については，質問ごとに抽出パターンを準備することで回答を得ることができる．

質問応答システムの評価としては，本節の最初でも述べたとおり，TREC の QA トラック，NTCIR の QAC，CLEF*などの評価型ワークショップにおいて実際の評価が行われてきた．質問応答の評価方法としては，TREC において提案された手法として，与えられた質問に対して，優先順位をつけた回答候補を上位 5 位までを回答する．評価は，最も高順位にある正

* http://clef.iei.pi.cnr.it/

解文字列を対象とし，各質問に対して回答順位の逆数による得点付けを行い，その平均である MRR(Mean Reciprocal Rank) によって総合評価とするものが多く用いられている．また，質問の回答についての根拠情報も同時に評価対象としており，根拠情報として回答を得た文書番号が利用されている．また，複数の可能の回答をすべて答える質問応答タスクの場合，システムが返した回答すべてを対象とし，システムの出力の正解の割合である適合率 P と，正解のうちシステムがどれだけ答えたのかの割合である再現率 R から，F 値 $(= 2P * R/(P + R))$ によって各質問の評価値を計算し，その平均 MF（Mean F 値）で評価する．

■参考文献

1) J. Burger, C. Cardie and et.al. “Issues, Tasks and Program Structures to Roadmap Research in Question Answering (QA),” `http://www-nlpor.nist.gov/projects/duc/roadmapping.html`, 2001.
2) J. Fukumoto, T. Kato, and F. Masui：“An evaluation of question answering challenge (QAC-1) at the NTCIR workshop 3,” ACM SIGIR Forum，vol.38, Issue 1, pp.25-28, 2004.
3) T. Kato, J. Fukumoto, F. Masui, and N. Kando：“Handling Information Access Dialogue through QA Technologies - A novel challenge for open-domain question answering,” Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL，pp.70-77, 2004.
4) T. Kato, J. Fukumoto, and F. Masui ：“Are Open-domain Question Answering Technologies Useful for Information Access Dialogues? -An Empirical Study and a Proposal of a Novel Challenge-,” ACM Transaction on Asian Language Information Processing，no.3, pp.243-262, 2005.
5) 佐々木, 磯崎, 他：“質問応答システムの比較と評価,”, 電子情報通信学会, 言語理解とコミュニケーション研究会, NLC2000-24, pp.17-24, 2000.
6) 神門, 安達 他：“評価ワークショップによるテキスト処理研究−第 3 回 NTCIR ワークショップを例として,” 人工知能学会誌，vol.17, no.3, pp.312-319, 2002.
7) 加藤, 桝井, 福本, 神門：“リスト型質問応答の特徴づけと評価指標,” 情処研究会報告, NL-163-16, pp.115-122, 2004.
8) 福本, 桝井：“質問応答技術 -大量のデータをもとに任意の質問に答える-,” 情報処理，vol.45, no.6, pp.580-585, 2004.
9) 諸岡, 福本：“Why 型質問応答のための回答選択手法,” 信学技報, NLC2005-107 pp.7-12, 2006.
10) 加藤, 福本, 桝井, 神門：“情報アクセス対話に向けた質問応答技術の評価 ふたたび− NTCIR5 QAC3 での試み−,” 情処研究会報告, NL-172-8, pp.55-62, 2006.