言語情報間の含意・矛盾関係の認識

乾 健太郎
奈良先端科学技術大学院大学

1 言説間の隠れた対立関係を見つける

ウィキペディア[1]の「マイナスイオン」の項目には次のような一節がある．

(1) マイナスイオン商品の解説や、健康本の著述の中には「マイナスイオンが疲労回復・精神安定を始めとする様々な健康増進効果をもたらす」と主張するものがあるが、これらの効果は客観的に証明されたものではない。

この記述によると，いっときブームを巻き起こした「マイナスイオン健康論」はかなり眉唾もののようだ．しかし，試しにウェブ（World Wide Web）の検索エンジンで「マイナスイオン」を検索してみると，次のようにその効用を謳うページがウェブの世界にはまだ数多く残っていることがわかる．

(2) 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。

立場が変われば，見方も変わり意見も変わる．こうした主張の対立は当然ながらウェブの世界にはごまんとある．ただし，それらのほとんどは読者からは陽に見えないことに注意する必要がある．たとえば，マイナスイオン健康論を謳っているのは専らマイナスイオン商品の販売サイトだが，彼らのページにはマイナスイオンの効用を否定するような，自分達に都合の悪い記述は当然載らない．ウェブはハイパーリンクによって情報が有機的に繋がり合うという素晴らしい特性を持っているが，ページの著者は自分に不都合な情報にわざわざリンクを張ったりしないので，右のような対立の多くはウェブの読者にとって「隠れた関係」でしかない．

ウェブ上に暗に存在するこうした言説間の対立関係あるいは類似関係を発見する計算機プログラムがあったらどうだろうか．ユーザが例えば先の (2) のようなマイナスイオン商品の効用を謳う販売サイトに足を踏み入れたとしても，(1) のような記述との隠れた対立関係を機械で予め発見できていれば，それを自動的にユーザに提示し，注意を促すことができる．そうやって，ウェブの玉石混淆の言説に対する多角的な視点をユーザに提示することができれば，情報の偏りや思いこみによる誤信を回避できる可能性がでてくるだろう．

[1] http://ja.wikipedia.org/

本稿では，こうした言説間の類似・対立関係を発見する技術を開発するという，筆者が現在携わっている研究課題 [10, 16] をとおして，近年研究者の関心を集めている含意関係認識の研究動向を概観する．

2 含意関係認識という課題

2 つの言説が類似あるいは対立関係にあるか否かを判断するには，両者の共通部分と差異を認識する必要がある．例えば，我々が (1) と (2a) の間に対立関係を見いだすのは，それらがともに「マイナスイオンに疲労回復効果がある」という共通の命題に関する言説でありながら，その命題に対する書き手の態度に (2a) は肯定的，(1) は否定的という差異があるからである．この意味で，我々の課題の根幹部分は，以下に述べる含意関係認識という言語処理の基本問題に帰着できる．

含意関係認識（Recognizing Textual Entailment; RTE）は，一対のテキストが与えられたときに一方が他方の記述から含意（あるいは推論）されるか否かを判別する問題で，2005 年から 3 年間続いた評価型ワークショップ Pascal RTE Challenge [2] をきっかけに注目を集め始めた研究トピックである．例えば，次の (3) ではテキスト t が仮説 h を含意するが，(4) の t は h を含意しない．(3) は YES で，(4) は NO と答えられる計算モデルを作ることが RTE の目標である．

(3) t. The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.
h. Cardinal Juan Jesus Posadas Ocampo died in 1993.（含意される）

(4) t. At the same time the Italian digital rights group, Electronic Frontiers Italy, has asked the nation's government to investigate Sony over its use of anti-piracy Software.
h. Italy's government investigates Sony.（含意されない）

こうした課題が言語処理研究者の関心を集めるのは，それが質問応答や情報抽出，複数文書要約，機械翻訳など，広範囲のアプリケーションに共通する言語処理の基本問題をうまく切り取っているからである [6, 13]．すなわち，言語には同じ情報を伝える言い回しがいくつも用

意されており，言語処理プログラムはどの言い回しとどの言い回しが同じ情報を伝えるかを判断できなければならないという問題である．例えば，(3) に答えるためには，“died” で伝えられる情報が別の言い回しの “killed” でも伝えられることが分からないといけない．逆に，もしこれに答えられるプログラムが作れれば，例えば次の (5) のような質問の答えを (3t) のテキストから正確に探し出せる質問応答システムを作れるようになる．

(5) When did Cardinal Juan Jesus Posadas Ocampo die?

容易に想像できるように，これは計算機にとって決して簡単な問題ではない．(3) に答えるためには，“kill” から “die” が含意されることの他にも，“kill” 事象の発生年が “1993” であることを t の解析時に認識する必要がある．(4) の場合は，“Italy” と “the nation” の照応関係や動詞 “ask” が与える内包的文脈の認識が必要である．こうした例が示唆するように，含意関係認識では，第一に個々のテキストからどれくらいリッチな意味的情報を引き出せるか，第二に “kill” と “die” の因果関係のような膨大な推論知識をどうやって集めるか，が鍵になる．以下，これら 2 つの問題について，我々の試みを例に動向を紹介する．

3 テキストからの意味的な情報の抽出

言語処理の分野では，1990 年代以降の大規模コーパスに基づく統計的言語処理の研究によって形態素･構文解析や固有表現抽出などの基礎的な解析技術が飛躍的な発展をとげた．しかしながら，前述の Pascal RTE Challenge では，こうした「浅い」処理から一歩意味に踏み込んだ解析技術の重要性が明らかになってきている．代表的な問題は，照応・省略の解消と命題の時間・ムード情報の解析である．

照応・省略解析はこれまで主に，(1) の「疲労回復・精神安定」と「これら」や (4) の “Italy” と “the nation” のような共参照関係の解析，あるいはテキスト中の述語の項を同定する述語項構造解析の文脈で研究されてきた．日本語では述語の必須格までがしばしば省略の対象になるため，省略やギャップの解消を含む述語項構造解析はとくに重要である．述語項構造解析は，例えば (6) の第 2 文から，「政府ガ 関係省庁ニ 協力ヲ 要請する」や「関係省庁ガ 計画ニ 協力する」のような述語と項の関係を抽出する問題である．

(6) 政府は低所得者を支援する計画を発表した。関係省庁の**協力**を**要請** する。

この例のように文を越えた省略関係や名詞化された述語の項構造までうまく同定できれば，含意関係認識に有用な情報となる．こうした照応・省略解析の研究は，日本語に限っても，京都テキストコーパス第 4.0 版や GDA コーパスなどの資源の蓄積とともに近年急速に発展を見せており [11]，近い将来形態素・構文解析と並ぶ基盤技術に成熟することが期待されている．我々のグループでも，その一歩として，日本語で最大規模の述語項構造・共参照タグ付きコーパス [3] を構築するとともに，オープンソースの述語構造解析器 SynCha[2]の開発を進めている [4, 12]．

言説間の対立関係を捉える上でもう一つ重要な情報は，命題に対する書き手の態度，すなわちムード情報である．例えば，(1) と (2a) の対立は「マイナスイオンの疲労回復効果」の真偽に対する書き手の態度の対立であった．ここで注意を要するのは，ムード解析の対象となる言語形式の範囲の決め方である．言語学におけるムード表現の分析対象はこれまで「らしい」,「まい」,「しなくてはならない」のような機能語や複合辞が中心だったが [15]，実際には命題の真偽に関する態度を表す言い回しはそれよりはるかに多様であり，次の例のような広義のムード表現を広範囲にカバーする必要がでてくる．

(7) a. 鯨の数は十分に回復している
b. 鯨の数は回復**からはほど遠い状況にある**

(8) a. マイナスイオンはトルマリンから生成します
b. トルマリンがマイナスイオンを放出する**とされるが、それはあり得ないことである**

この問題に対し我々は，従来の機能表現辞書（例えば松吉らの辞書 [14]）に加え，より多様なムード表現を識別するための計算モデルを新たに開発中である [8]．ただし，ムード情報の分類体系やタグ付きコーパスの作成方法など残された課題も多く，言語学研究との密な連携が必要であると考えている．

4 推論のための知識

冒頭の (1) と (2a) の対立関係を発見するには,「疲労回復」と「疲労を取り除く」が同じ事象を指しうることを知っていなければならない．(3) に答えるには，“kill” と “die” の因果関係の知識が必要であった．こうした言語上での推論の研究は 1980 年代に遡るが，当時の研究は十分な量の言語知識や常識的知識を用意する方法論がなかったため，実用規模には発展しなかった．しかし，当時では考えられない規模の言語データが入手可能になった現在，一定の水準に達した統計的言語解析技術を用いて大量の言語データから実用規模の推論知識を自動獲得できる可能性が見え始めている [7, 18]．こうした背景から我々のグループでは，基本的な推論知識の人手による整備と大規模コーパスからの知識獲得を並行して進めてきた [5]．

人手による整備では，まず国語辞典の語釈文から述語項構造間の基本的な意味関係を収集した．例えば，動詞「倒す」の語釈文「立っている物に力を加え傾け、横にする（岩波国語辞典）」からは,「X が Y を倒す ⇒ X が Y を横にする（上位下位関係）」の他に,「⇒ X が Y に力を加える（行為-手段関係）」,「⇒（行為前は）Y が立っていた（前提条件）」など，多様な意味関係が収集できる．すでに岩波国語辞典第 5 版の収録動詞（11469 語）について作業を終えており，8 種類の意味関係を合わせ

[2] http://syncha.sourceforge.jp/

て約 3 万 5 千件収集できている [17]．これによって，例えば次のような推論ができるようになった．

(9) a. 地熱が大気を暖める ⇒ 大気が暖かくなる
b. バターを焦がす ⇒ バターが黒くなる
c. 警官が犯人をとりおさえる ⇒ 犯人が警官に捕まる
d. 金属を沸かす ⇒ 金属に熱を加える

また，語彙概念構造による動詞意味分析の枠組み [9] に基づき，高頻度動詞約 4 千語，7 千語義について 5 階層からなる意味分類を行った [19]．最下層は約千クラスに分類されており，例えば「所属・客体変化」のクラスには「配属する，取り立てる，引き抜く，立てる，招く」などが属し，これらはすべて同じ項構造を持つ．これにより，語釈文だけでは捉えられない基本概念間の含意関係をカバーできる．こうした知識をうまく使えば，例えば (10a) のような表現を類義表現として，また (10b) を反義表現として認識することができ，類似・対立言説の自動発見への一歩となる．これらの資源の配布については http://cl.naist.jp/~inui/research/EKB/ を参照されたい．

(10) a. 分煙を 進める/行う/促進する/求める/推進する ； 分煙が 進む ； 分煙に なる
b. 分煙が 遅れる

一方，コーパスからの知識獲得については，次の 2 種類の手がかりを組み合わせる手法を開発し，成果を上げた [1]．1 つ目は，(11) の「〜たけれども〜ない」のような共起パターンで，こうしたパターンとの共起を調べれば，例えば因果関係のような意味関係を持つ用言対（「かける」と「通じる」）を集めることができる．第 2 の手がかりは，(12) のようにそれらの用言対が同じテキスト内で同じ項（この例では「電話」）を持つ事例の集まりで，意味的に関係の深い用言対がどの項を共有しやすいかがわかる．これらの手がかりを組み合わせると，例えば (13) のような推論知識を自動獲得することができる．現在のところ，5 億文規模のコーパスから行為-効果関係や行為-手段関係等の関係を 1 万対を超える規模，80 %以上の精度で獲得できることを確認している．

(11) a. かけ たら ...通じ た
b. かけ たけれども ...通じ ない
c. かけ 続けても ...通じ ない

(12) a. ...サンタバーバラに 電話 を かけ てくれて、... でも、なかなか 電話 が 通じ ないので、...
b. ... 司会者に 電話 を かけ てもらいます。... 電話 が 通じ るなり、...

(13) a. Xをかける（行為）⇒ Xが通じる（効果）
X＝｛電話，願い，魔法，...｝
b. Xにかける（行為）⇒ Xに通じる（効果）
X＝｛相手，彼女，...｝

今後は，人手で整備した前述の基本知識を知識獲得過程に利用し，獲得効率の改善をはかる実験を行う予定である．

5 まとめ

本稿では，ウェブに散在する言説間の隠れた類似・対立関係を発見するという課題を念頭におきながら，含意関係認識と呼ばれる言語処理の基本問題についてその研究動向を紹介した．照応・省略解析や時間・ムード解析といった一歩意味に踏み込んだ解析技術，また大規模な知識ベースに基づく頑健な推論を必要とするこの課題は，大量のコーパスを背景とする統計的言語解析の研究が成熟期を迎えた現在，我々が次に注力すべき重要課題の一つと位置づけられよう．こうした研究は言語学との連携に負うべきところも多い．本稿がその一助となれば幸いである．

謝辞

類似・対立言説の自動発見という課題設定は，(独) 情報通信研究機構・知識処理グループの諸氏との議論に多くを負っている．記して深く感謝する．

参考文献

[1] Shuya Abe, Kentaro Inui, and Yuji Matsumoto. Two-phased event relation acquisition: Coupling the relation-oriented and argument-oriented approaches. In *Proc. of the 23rd International Conference on Computational Linguistics (COLING)*, 2008. (to appear).

[2] 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.

[3] 飯田龍, 小町守, 乾健太郎, 松本裕治. 述語項構造と共参照関係のアノテーション: Naist テキストコーパス開発の経験から．言語処理学会第 13 回年次大会発表論文集, 2007.

[4] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 625–632, 2006.

[5] 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.

[6] 乾健太郎． 自然言語処理と言い換え． 日本語学, Vol. 26, No. 11, pp. 50–19, 2007.

[7] 乾健太郎, 鳥澤健太郎． WWW からの知識獲得：自然言語処理における新展開．日本語学, Vol. 27, No. 2, pp. 48–61, 2008.

[8] 乾健太郎, 原一夫. 経験マイニング：Web テキストからの個人の経験の抽出と分類. 言語処理学会第 14 回年次大会発表論文集, pp. 1077–1080, 2008.

[9] 影山太郎. 動詞の意味と構文. 大修館書店, 2001.

[10] 河原大輔, 黒橋禎夫, 乾健太郎. 主要・対立表現の俯瞰的把握 － ウェブの情報信頼性分析に向けて. 情報処理学会自然言語処理研究会, 第 2008-NL-186 巻, 2008.

[11] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proc. of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 176–183, 2006.

[12] Mamoru Komachi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Learning based argument structure analysis of event-nouns in Japanese. In *Proc. of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp. 120–128, 2007.

[13] 黒橋禎夫. 言語コンピューティング. 人工知能学会誌, Vol. 22, No. 5, pp. 711–718, 2007.

[14] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol. 14, No. 5, pp. 123–146, 2007.

[15] 森山卓郎, 仁田義雄, 工藤浩. モダリティ. 日本語の文法 3 . 岩波書店, 2000.

[16] 村上浩司, 松吉俊, 隅田飛鳥, 森田啓, 佐尾ちとせ, 増田祥子, 松本裕治, 乾健太郎. 言論マップ生成課題：言説間の類似・対立の構造を捉えるために. 情報処理学会自然言語処理研究会, 第 2008-NL-186 巻, 2008.

[17] 大西良明, 乾健太郎, 松本裕治. 事態間関係知識の整備と含意文生成への応用. 言語処理学会第 14 回年次大会発表論文集, pp. 1152–1155, 2008.

[18] 関根聡, 乾健太郎, 鳥澤健太郎. 提唱「コーパスベース知識工学」. 言語処理学会第 13 回年次大会併設ワークショップ「大規模 Web 研究基盤上での自然言語処理・情報検索研究」, 2007.

[19] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤. 意味の包含関係に基づく動詞項構造の細分類. 言語処理学会第 14 回年次大会発表論文集, pp. 1037–1044, 2008.