日本語意味論テストセットの構築

川添 愛[1]　田中 リベカ[2]　峯島 宏次[2]　戸次 大介[2]
国立情報学研究所[1] お茶の水女子大学大学院[2]
zoeai@nii.ac.jp, tanaka.ribeka@is.ocha.ac.jp, mineshima.koji@ocha.ac.jp, bekki@is.ocha.ac.jp

1　目的

ある文の表す内容が真である時に別の文の内容が真であると推論される場合、前者と後者の間には含意関係が成り立つ。ここでは以下、前者を「前提」と呼び、後者を「結論」と呼ぶ。文と文との含意関係は、言語学においては意味論の中心的な説明対象の一つであり、また自然言語処理においては近年、意味処理タスクの中核となっている。

筆者らは、意味に関する理論および含意関係認識システムの評価に資することを目的とし、日本語の意味論的な現象に基づく含意関係データセットの構築に着手している。当データセットを「日本語意味論テストセット」(Japanese Semantics test suite) と呼ぶ。同様のテストセットとして英語では 1990 年代に FraCaS test suite (Cooper ら 1996) が作成されているが、日本語の現象を扱ったものはまだ存在しない。このテストセットは、FraCaS 等の他言語データとリンクする部分（多言語サブセット）と、日本語のみの部分（日本語サブセット）からなる。本稿では、このテストセットの概要と、FraCaS 対応部分の構築について述べる。

2　背景

言語理論の検証に利用できるテストセットに求められる第一の要件は、説明対象となる現象をミニマルに表していることである。含意関係に関して言えば、含意関係の有無の判断に影響する他の要因（世界知識や、対象となる言語現象以外の現象）が極力排除されていることが望ましい。そしてこの点は、含意関係認識タスクの評価データにおいても重要な要件である。近年は特に、個別の現象に対するシステムのパフォーマンスを測れるような評価データの重要性が広く認識されている。

たとえばシェアドタスク PASCAL RTE challenge においては、Bentivogli ら (2010)、Sammons ら (2010) が前提と結論の関係をより基本的な関係の連鎖として書き下す方法を提案し、RTE-5 データセットに対するアノテーションを行っている。日本語・中国語を対象とする NTCIR RITE においても、RITE-2 以降、前提-結論ペアの関係をより基本的な関係に限定したテストセット（UnitTest データ）が提供されている。

その他、SemEval-2014 Task1 では Compositional Distributional Semantics が対象とする語彙的・統語的・意味的現象に特化した SICK データセット (Marelli ら 2014) が提供されている。日本語に関しては、上述の RITE UnitTest の他に、小谷ら (2008) による京大 Textual Entailment 評価データがある。これは、一つの例の推論に関わる要因を一つあるいは二つに絞ったデータであり、推論の要因は大きく分けて「包含」「語彙（体言）」「語彙（用言）」「構文」「推論」の五つがある。文は比較的単純であるが、推論に語彙的知識や常識が必要となる例が多い。

FraCaS test suite は、1990 年代の FraCaS プロジェクトにおいて、自然言語処理システムおよび意味理論の推論能力を評価する目的で構築された。主に形式意味論が対象とする言語現象の関わる推論を中心に、346 のテストを含んでいる。一つの例が一つの現象についてのみテストできるよう意図されており、ターゲットとなる現象以外の要因は最大限に制限されている。Cooper ら (1996) のオリジナルのテストセットは、前提となるテキスト、yes/no 疑問文、それに対する答え（Yes, No, あるいは Don't know) から構成されていた。2007 年に Bill MacCartney により作成された XML 版では、各テストに「結論」となるテキストが追加される等の修正が加えられ、他の含意関係認識タスクデータに習ったフォーマットとなっている (http://www-nlp.stanford.edu/ wcmac/downloads/fracas.xml)。MacCartney ら (2008) は、自然論理に基づく含意関係認識モデルの評価にこのデータを利用している。現在では、Robin Cooper が中心となり、XML 版の FraCaS を多言語化するプロジェクト MultiFraCaS が進められている。

FraCaS タイプのデータセットの利点は、言語学の

成果に基づき、信頼できる現象のみを扱っているという点である。この点は、データの信頼性（含意関係に関して言えば、含意関係の有無の判断）が、言語学コミュニティによって保障されていることを意味する。また、ターゲットとなる現象とは無関係の世界知識や文脈、語彙による影響を極力制限しているため、発話状況や語彙等を入れ替えても成り立つことの多い、ある意味一般化されたデータである。さらに、扱われている意味論的な現象には、量化、複数性、照応、テンス、比較、命題的態度等があり、これらは日本語における既存のデータセットではあまりカバーされていない。よって、理論的な側面はさることながら、言語処理タスク用のデータという実用的な観点からも、日本語において同様のデータセットを構築する意義は十分にある。ただ、FraCaS タイプのデータにおいては文の自然さの実現が大きな課題の一つであり、これは日本語においても例外ではない。

3 日本語意味論テストセットの概要

3.1 テストセットの構成と内容

筆者らが構築している「日本語意味論テストセット」では、FraCaS の方針にならい、言語現象ごとにデータをまとめ、原則として一つの例を一つの現象（あるいは特定の現象間の相互作用）に対応させる。ただし FraCaS とは異なり、一つの現象に対応する例を複数用意する場合もある。

日本語意味論テストセットは FraCaS 対応部分を中心とする多言語サブセットと、日本語独自の現象を含む日本語サブセットからなる（各部の詳細は表 1 を参照）。ただし、日本語サブセットの項目も、「対訳」レベルでは FraCaS test suite の項目に関連付けられる場合がある（詳しくは後述）。コアとなる各現象が出そろった後、現象間の相互作用を示すようなデータを随時追加していく予定である。

（前提）すべてのイタリア人男性が偉大なテノール歌手になりたがっている。
（結論）偉大なテノール歌手になりたがっているイタリア人男性がいる。
（含意関係の有無：yes）

テストの例 1：量化子の conservativity（fracas-002 に対応）

（前提）どちらの理事もかつてはビジネスマンだった。
（結論）どちらの理事もかつては一流のビジネスマンだった。
（含意関係の有無：no）

テストの例 2：量化子の monotonicity（fracas-045 に対応）

3.2 使用するタグ

テストセットにて利用する主なタグは以下のとおりである。

- `problem`： テスト
  - `jsem_id` 属性：固有の ID
  - `answer` 属性：含意関係の有無（yes, no, unknown）
  - `phenomena` 属性： 現象の種類（複数指定可）
- `link`：他言語リソースとのリンク（多言語対応部分）
  - `resource` 属性： リンク先リソース名
  - `link_id` 属性： リンク先の対応項目 ID
  - `translation` 属性： リンク先の項目と対訳レベルで一致するか（yes,no,unknown）
  - `same_phenomena` 属性： リンク先の項目と現象レベルで一致するか（yes,no,unknown）
- `p`： 前提
- `h`： 結論
- `note`： コメント

link 要素の属性に translation と same_phenomena の二つを設けたのは、特に多言語サブセットについて、1) 意味論的な現象を含む文の対訳コーパス、2) 日本語と他の言語との間で共通する現象のアーカイブの二つの性格を与えることを意図してのものである。前者は主に自然言語処理用リソースとしての要件であり、後者は理論的な要件である。単純に他言語のテストセットを日本語に翻訳するだけでは、これら両方を満たすことは不可能である。後に述べるように、英語の項目の対訳ではあるが本質的に異なる現象を含むテストや、英語の項目の対訳ではないが同様の現象を示すテストを作成する場合があるため、ここでは「（リンク先の項目と）対訳レベルで同一視できるか」と「現象レベルで同一視できるか」とを明示的に区別する。

3.3 構築プロセス

テストセットの構築は、筆者ら 4 名（言語学者 3 名と言語学の素養のある大学院生 1 名）で行っている。多言語サブセットの構築においては、FraCaS を四分割し、各パートに対して 1 名が対応部分の構築を担当している。データ構築の上で最も留意する点は、1) 含意関係の判断において、ターゲットとなる現象以外の要因が入っていないか、2) 曖昧性がないか、3) 文が十分に自然であるかの三点である。ただし、やむを得

多言語サブセット	FraCaS に含まれる現象	一般量化子、複数性、照応、省略、形容詞、比較、テンス、動詞、命題的態度
日本語サブセット	FraCaS に含まれない現象	前提、フォーカス、量化子のスコープ、条件文、モダリティ、相互代名詞、分裂文、副詞関連、「同じ/別の」、CI 等
	日本語独自の現象	各種「は—が」構文、取り立て詞、「自分」、「の」照応等
	二つ以上の現象の相互作用	複雑な等位接続 (束縛変項照応の関わるもの等)、条件文とモダリティの相互作用等

表 1: 日本語意味論テストセットの構成

ず他の要因が入ったり曖昧性が残ったりする場合は、note 要素にその旨を明記する。また、多言語サブセットにおいて、FraCaS の項目の対訳で文が不自然になる場合は、同じ現象を含む自然な日本語の例を作ることにしている。チェック作業では、データを構築した者とは別の者がチェッカーとなり、answer 属性の値を隠した状態で含意関係があるかないかを確認し、さらに上記の三点についても確認する。

4 β版の構築

筆者らは、FraCaS 対応部分を中心に、日本語意味論テストセットのβ版を作成した。概要を表 2 に示す。以下、主な現象について、構築上の留意点を述べる。

一般量化子 各種量化表現の conservativity および monotonicity の関わる含意関係のテストを扱う。日本語の量化表現には、語彙的な多様性ならびに名詞句・格助詞との位置関係により、多くの異形がある。たとえば英語の every N に対応する表現として、「すべての N が」「N すべてが」「N がすべて」「どの N も」等がある。これらは含意関係に関しても英語と同様の振る舞いを見せるが、一部の形式（遊離数量詞等）の特殊性については多くの議論があるため、これらと英語の一般量化子が「現象レベルで一致するか」に関しては判断を保留している。また、英語の no N、neither N 等、monotone decreasing GQ の一部については、日本語には直接の対応物が存在しない。「誰も〜ない」「どちらも〜ない」等の形式の対訳は含めているが、現象レベルの一致はないものとしている。

照応 「彼（ら）/彼女（ら）」「それ（ら）」等の関わる表現の他に、英語の再帰代名詞を含むテストの「対訳」レベルの対応物として、「自分」の関わるテストも含めた。ただしよく知られているように、「自分」は英語の再帰代名詞の対応物ではなく、含意関係のテストにおいてもその違いは如実に表れるため、現象レベルでは対応させていない。また、束縛変項照応の例に関しては、日本語の「彼/彼女」では束縛変項として解釈しづらいことを考慮し、対訳とは別にソ系の指示詞を利用した日本語の例を作成した。

形容詞・動詞 英語の形容詞のテストへの対応物として、ここでは、「赤い」のような形容詞（イ形容詞）だけでなく、「大きな」のような形容動詞（ナ形容詞）や「本物の」のようないわゆる状詞も含めている。現象としては肯定的 (affirmative) な形容詞（「本物の」等）と非肯定的 (non-affirmative) な形容詞（「偽物の」等）の違いの関わる含意関係、比較クラスによって左右される含意関係等を扱っている。動詞の含意関係では、動詞のアスペクト分類にかかわるもの、および動詞の分配的読みと集団的読みにかかわるもの等を扱っている。

比較級 比較級には、「〜より」が句（名詞句）を要求する比較級 (phrasal comparatives) と節を要求する比較級 (clausal comparatives) とがあるが、日本語では節を要求する比較級は、英語に比べて作りにくい。たとえば、“X won more orders than Y lost.” に対応する文としては、「X は Y が失った 数 より多くの注文を得た」のように隠れた名詞を補うなどの配慮が必要である。ここでは「X 社は Y 社より 3000 台多くのコンピュータを売った」のような数量表現がかかわる比較級の含意関係も扱っている。

時制形式 FraCaS ではテンスと時間関係が関わる現象として、英語の過去形、完了形、未来形等の時制を扱っているが、英語と日本語のテンスのシステムは大きく異なるため、現象の同一性を考慮しつつ忠実な対訳を作ることが困難である。たとえば日本語には英語の現在進行形に相当する形式が存在せず、かわりにアスペクト形式の「テイル」を用いるが、「テイル」に

全体項目数		602
FraCaS 対訳項目	現象レベルで一致 (unknown 含む)	441
	現象レベルで不一致 (no)	62
対訳以外の項目（日本語例）		99

表 2: β版の概要（2015 年 1 月時点）

は（少なくとも）進行と結果残存の読みが存在する。この曖昧性を除去するためには「ずっと」「もう」等の副詞を付加する等のコントロールが必要である。

命題的態度　FraCaS においては、know 等の叙実動詞、manage 等の含意動詞、see 等の知覚動詞の表す態度の関わる推論が扱われている。日本語においては、動詞のタイプに加え、補文標識が「こと」「の」「と」のいずれであるかによっても補文内容が含意されるか否かが左右されるため、配慮が必要である。また、動詞の翻訳も注意を要する点である。たとえば英語の try に対して「〜しようとする」「〜しようと試みる」の二つの訳が考えられるが、Sharvit(2003) の指摘する “John tried to cut a tomato, #but there were no tomatoes to cut.” のような文の不自然さを踏まえれば、try を「〜しようと試みる」と訳すのがより適切である（「トマトを切ろうと試みたが、トマトがなかった」は英語同様不自然であるのに対し、「トマトを切ろうとしたが、トマトがなかった」は自然）。

5　おわりに

本稿では、日本語意味論テストセットの概要について述べた。このプロジェクトは開始したばかりであるが、FraCaS test suite に対応するセットをすでに構築済みで、2015 年 2 月上旬の公開を予定している。FraCaS 対応部分は、MultiFraCaS フォーマットでも提供する予定である。

今後は、コアとなる現象を一通りカバーしたのち、二つ以上の現象間の相互作用が関わるテストセットの構築に着手する。また、言語学コミュニティに広く声をかけ、各現象の専門家によってデータのチェックおよび作成が実現できるような環境を整えていく予定である。

参考文献

[1] L. Bentivogli, E. Cabrio1, I. Dagan, D. Giampiccolo, M. L. Leggio, B. Magnini. 2010. “Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference.” Proceedings of LREC 2010:3544-3549, Valletta, Malta.

[2] R. Cooper, D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jan, H. Kamp, D. Milward, M. Pinkal, M. Poesio, S. Pulman,T. Briscoe, H. Maier, and K. Konrad. 1996. “Using the framework.” Technical report,FraCaS: A Framework for Computational Semantics. FraCaS deliverable D16.

[3] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi and R. Zamparelli. 2014. “A SICK cure for the evaluation of compositional distributional semantic models.” Proceedings of LREC 2014, Reykjavik (Iceland): ELRA, 216-223.

[4] B. MacCartney and C. D. Manning.2008. “Modeling semantic containment and exclusion in natural language inference.” The 22nd International Conference on Computational Linguistics (Coling-08), Manchester, UK, August.

[5] M. Sammons, V. G. Vinod Vydiswaran, D. Roth. 2010. “Ask not what textual entailment can do for you...” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics:1199-1208, Uppsala, Sweden.

[6] Y. Sharvit. 2003.“Trying to be Progressive: the Extensionality of Try.” Journal of semantics 20.4: 403-445.

[7] 小谷通隆 , 柴田知秀, 中田貴之, 黒橋禎夫. 2008. 日本語 Textual Entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会第 14 回年次大会:1140-1143.