1K3-1

形式意味論に基づく含意関係テストセット構築の方法論

A Methodology for Constructing Inference Problem Sets Based on Formal Semantic Studies

川添愛[*1] 田中リベカ[*2] 峯島宏次[*2*3] 戸次大介[*1*2*3]
Ai Kawazoe Ribeka Tanaka Koji Mineshima Daisuke Bekki

[*1]国立情報学研究所 National Institute of Informatics
[*2]お茶の水女子大学 Ochanomizu University
[*3]独立行政法人科学技術振興機構, CREST CREST, Japan Science and Technology Agency

This paper introduces JSeM test suite, a collection of inference problems with Japanese sentences for evaluation of semantic processing systems. The test suite groups inference problems by semantic phenomena, following the policy of the FraCaS test suite (the original version in Cooper et al. 1996 and the machine-readable version by Bill MacCartney). The test suite consists of the multilingual subset and the Japanese subset, to cover both the universal phenomena and Japanese-specific ones. This paper outlines the design policy and the methodology to construct the multilingual subset currently available online.

1. はじめに

形式意味論の主な研究対象の一つである文と文の間の推論関係は、近年、自然言語処理における含意関係認識タスクの対象としても重要性を増している。言語学コミュニティにおいては文間の推論の観察・分析が蓄積されており、言語現象とそれがもたらす推論の可能性についての知識がある程度共有されている。しかしその多くは暗黙の知識であり、また高度な専門性を必要とするため、現在のところそれらが含意関係認識の評価において有効に利用されているとは言い難い。

筆者らは、含意関係認識システムの評価に資することを目的とし、「日本語意味論テストセット」(Japanese Semantics test suite) を構築している。これは、日本語の意味論的な現象に基づく推論関係データセットである（ここでは「推論関係」という語を、いわゆる entailment だけでなく、presupposition、implicature なども含む推論関係一般を指すものとして使う）。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存在しない。

日本語意味論テストセットは、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。筆者らはすでに、前者をカバーした β 版を開発し、公開中である。本論文では構築済みの多言語サブセットを中心に、自然なデータを作成する方法論について論じる。

2. 背景

多くの場合、含意関係認識タスクの評価に使われる推論テストデータは 1) 正しいと仮定される「前提」、2) 前提から推論されるかどうかが問われる「仮説」、3) 前提から仮説が推論できるかどうかについての判断（true, false, unknown、あるいは yes, no, unknown）の三つの部分からなる。以下は、シェアドタスク PASCAL RTE challenge にて使用された RTE-4 のテストの例である。

T（前提）: In the end, defeated, Anthony committed suicide and so did Cleopatra, according to legend, by putting an asp to her breast.
H（仮説）: Cleopatra committed suicide.
（判断）: TRUE

近年、含意関係認識タスクの評価データにおいては、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。例えば RTE のデータについては、Bentivogli ら (2010)、 Sammons ら (2010) が前提と仮説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 年代に FraCaS consortium により、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 のテストを含む。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や世界知識は最大限に制限されている。これは PASCAL RTE などの実テキストから作られたテストセットとは違い、言語学者のチームによって構築された、英語の作例のテストセットであり、実質的に意味的な現象およびそれに対する知見のアーカイブとなっている。また、Bill MacCartney によって作成された機械可読な XML 版は 含意関係認識システムの評価に広く使われている (MacCartney and Manning 2007, 2008, Lewis and Steedman 2013, Tian et al.

連絡先: 川添愛, 国立情報学研究所, 東京都千代田区一ツ橋 2-1-2

2014 等)。すなわち、意味論研究の知見を、自然言語処理システムの評価に有効に活用している一例である。

FraCaS タイプのデータセットは、言語学の成果に基づいた信頼できる現象のみを扱っているという利点がある。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。また、FraCaS において扱われている意味論的な現象は、量化、複数性、照応、テンス、比較、命題的態度等、基本的かつ普遍性の高いものである。現在、Robin Cooper の主導により FraCaS を多言語化する MultiFraCaS project が進められており、英語版 FraCaS に一対一対応するペルシャ語や中国語のバージョンが構築されている。

3. 日本語意味論テストセットの構築

3.1 留意すべき問題

FraCaS で扱われている現象は、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。ただし、日本語のバージョンを作る上では困難が生じる。特に単なる翻訳では極めて危険で、不自然なばかりか、データの正確さが保証できなくなってしまう可能性がある。というのは、日本語と英語においては、対訳レベルの対応と現象レベルの対応とにずれが見られるからである。また、Bos (2008) によって指摘されているように、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。

3.2 構成

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。

3.3 フォーマット

テストセットの作成にあたって、筆者らは以下のフォーマットを採用している。

- `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, undef)
  - `inference_type` 属性：推論のタイプ
  - `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

以下、特徴的な点について述べる。

translation 属性と same_phenomena 属性

`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。

inference_type 属性による推論の分類

推論とは複合的現象であり、文 S_1 が S_2 を含意するというとき、そこには様々な言語的要因・文脈的要因が関与しているのがふつうである。日本語意味論テストセットでは、各テストデータに関与する意味論的現象を `phenomena` タグによって示すほかに、`inference_type` タグによって、前提と結論の間に成り立つ推論のタイプを明示している。これにより、「量化表現」「複数性」「否定」といった個別の言語現象による分類とは別に、「含意」「前提」「慣習的含み」といった異なるタイプの推論という軸からデータセット全体を切り分け、各推論のタイプごとにシステムの能力を評価することが可能になる。

ここでは、現代的な形式意味論・語用論の文脈でよく知られている推論の分類[*1]に基づいて、含意と前提 (presupposition) という代表的な 2 つのタイプの推論を区別する。

含意は、発話の中心的な内容 (at-issue content) を表し、他のタイプの意味・推論とは区別して、主張内容 (asserted content)、真理条件的内容 (truth-conditional content) などとも呼ばれる[*2]。次は含意の典型例である。

(1) jsem-id:10

P1	日本人研究者が一人ノーベル賞を受賞した。
H	ノーベル賞を受賞した日本人研究者がいた。

含意は、会話の含みなどの語用論的推論とは異なり、取り消し不可能である。また、含意は文を否定、モダリティ、条件文の前件、疑問、仮定といった文脈に埋め込むと消失するという特性を持つ。

これに対し前提は、発話の主眼ではなく、むしろ発話の背景にある内容 (backgrounded content) を表す。そのため、前提は通常、話し手と聞き手の共通了解になっている事柄や、特に議論の余地のない、発話の文脈において目新しさを伴わない内容を表している。(2) は比較表現「〜以上に」が引き起こす前提の例である[*3]。

---

*1 例えば、Chierchia & McConnell-Ginet (2000), Levinson (2000), Kadmon (2001), Potts (2005) などを参照。

*2 前提や語用論的な含意も含めた広義の含意関係（推論関係）とは区別して、「意味論的含意 (semantic entailment)」と呼ぶこともある。

*3 興味深いことに、「より」を伴う比較表現の場合、このような前提は生じない。例えば、「太郎は花子より早起きだ」は「花子は早起きだ」を含意せず、し

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

(2)　jsem-id:620

P1	太郎は花子以上に早起きだ。
H	花子は早起きだ。

(3) は叙実述語 (factive predicate) のひとつである「〜ことを嬉しく思う」が伴う前提である。

(3)　jsem-id:737

P1	太郎は花子が高校を卒業したことを嬉しく思った。
H	花子は高校を卒業した。

前提は、含意と同様に後続文による取り消しが不可能であるが、含意とは異なり否定、モダリティ、条件文の前件といった文脈に埋め込まれたときに消失せず、文全体から推論可能な内容として生き残る（すなわち、**投射**する）という特徴をもつ。

含意と前提の他に、代表的な推論のタイプとして、慣習的含み (Conventional Implicature) と会話の含み (Conversational Implicature) が挙げられる*4。慣習的含みは、前提と同様に発話の背景にある情報にかかわる推論であるが、前提とは異なる投射の性質を示すことが知られている (Potts 2005)。会話の含みは、語用論的推論の一種であり、含意、前提、CI とは異なり、文脈によって取り消し可能であるという特徴がある 。現時点のデータセットでは、推論現象として扱っているのは主に含意と前提だけであるが、慣習的含みや会話の含みに関しても今後データセットを拡張していく予定である。

3.4　構築プロセス

この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。原則として、1 名がテストを構築し、他の 1 名がチェッカーとしてテストをチェックする。チェックの際には、answer 属性の値を隠した状態で推論の可否を確認する。さらに 1) ターゲットとなる現象が適切に含まれているか（他の雑多な要因が入っていないか）、2) 明記されている以外の曖昧性がないか、3) 文が十分に自然であるかの三点について確認する。後述するベータ版の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た。ここでは翻訳の適切さ・自然さという要素も入ってくるため、以下のようなプロセスでの作成にあたった。

1. 真理条件的に等価な訳を作り、現象のタグ付けをする。
2. ターゲットとなる現象に関わる表現の異形の入った例を作成し、バリエーションを追加する。
3. 現象間の対応についてのサーベイ。
   - `answer, phenomena, note` にて反映させる。
   - 原文と異なる現象が入っている場合は、`phenomena` にて表示。
4. 訳が自然にならない場合は、以下の工夫をする。
   - 文型を変える（標準的な語順から分裂文にするなど）。
   - 省略と復元（節を要求する比較級で、隠れた名詞を補うなど）。
   - どうしても自然な例が作れない場合は、同じ現象を含む日本語例を独自に作る。
5. 訳が曖昧になる場合は、曖昧性を除去する。
   - 原文との意味の同一性を保つために新しい表現を追加できない場合は、`note` にて、意図されている解釈を記述する。

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、構築上留意した点を述べる。

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。例えば英語の 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

たがって、「太郎は花子より早起きだ」から「太郎は早起きだ」を推論することはできない。「より」「以上に」が伴う含意と前提の区別について詳しくは、Hayashishita (2007), Kubota (2012) を参照。

*4　会話の含みはさらに、一般的な会話の含み (Generalized Conversational Implicature, GCI) と個別的な会話の含み (Particularlized Conversational Implicature, PCI) に下位分類することができる。特に GCI は、「3 人の学生が来た」から「4 人以上は来なかった」を推論する例など、量化・数量表現にかかわる語用論的推論と深く関係しており、含意関係認識の重要なデータとなりうる推論現象である。

全体項目数		788
FraCaS 対訳項目	現象レベルで一致 (unknown 含む)	553
	現象レベルで不一致 (no)	71
対訳以外の項目（日本語例）		164

表 2　β 版の概要（2015 年 3 月時点）

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。例えば、“X won more orders than Y lost.” に対応する文としては、「X は Y が失った量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。「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。例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。例えば英語の 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。β 版は現在公開中である[*5]。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。

*5　https://researchmap.jp/community-inf/JSeM/

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] J. Bos. 2008. “Let’s not argue about semantics.” Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008):2835–2840, Morocco.

[3] G.Chierchia and S. McConnell-Ginet. 2000. Meaning and Grammar: An Introduction to Semantics. MIT Press.

[4] 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.

[5] N. Kadmon, 2001. Formal Pragmatics. Blackwell.

[6] M.Lewis and M.Steedman. 2013. “Combining distributional and logical semantics.” Transactions of the Association for Computational Linguistics, 1:179–192.

[7] 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.

[8] B.MacCartney and C.D. Manning. 2007. “Natural logic for textual inference.” In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 193–200.

[9] 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.

[10] C. Potts. 2005. The Logic of Conventional Implicatures. Oxford University Press.

[11] 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, Sweden.

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

[13] R.Tian, Y.Miyao, and T.Matsuzaki. 2014. “Logical inference on dependencybased compositional semantics.” In Proceedings of ACL, 79–89.

[14] 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.