2P1-3in1

自然言語理解ユニットテストと意味表現の検討

An Investigation of Unit Tests and Semantic Representations for Natural Language Understanding

菅原朔 *1 Saku Sugawara　横野光 *2 Hikaru Yokono　相澤彰子 *3*1 Akiko Aizawa

*1東京大学 The University of Tokyo　*2株式会社富士通研究所 Fujitsu Laboratories Ltd.　*3国立情報学研究所 National Institute of Informatics

In this paper, we propose a new design methodology for comprehensive achievement test for natural language understanding (NLU) systems. We first investigate the expected NLU skills of several existing NLU tasks and enumerate all the grammatical elements and reasoning types used in these tasks. Based on the analysis, we expand the concept of "toy tasks" proposed by Weston et. al. and formulate "unit tests" where each test is defined as a QA-style task consisting of contextual statements and a query using only a specific grammatical element and a reasoning type. We also describe the prerequisite properties of the semantic representations needed for the unit tests, and discuss the advantages and disadvantages of commonly used representations. Then, we focus on Abstract Meaning Representation (AMR) and demonstrate how the AMR can be extended for the unit test.

1. はじめに

計算機による自然言語理解の実現を目的として、これまでに多くのタスクが考案されてきた。そこでは例えば長文の読解や常識的知識の運用など高度な能力が要求されるが、それらを解決するために提案されたシステムはドメイン依存的・タスク依存的になってしまうという難点があると言える。

前述の目的のためには、「システムが人間と同等の自然言語理解能力を備えていると言えるために解けなければならない問い」という観点から包括的に細分化されたテスト（本稿ではこれをユニットテストと呼称する）を整備し、そのテストによる詳細な評価と分析を通してシステムを設計・改良することが有効な方針であると考えられる。

また、語彙的に豊富かつ十分な規模のユニットテストを構築するためには、独立したテスト生成器によって生成が自動化されることが望ましい。その際、テストが要求する事項を十全に表現し処理できる意味表現がテスト生成器の内部に構成されていると、テストの正答出力が円滑に行える。一方でテストを解くシステムの側は入力された自然言語文に応じた意味表現を構築して何らかの処理を施すことにより解答を出力するが、この意味表現が（テスト生成器の場合と同様に）十分な表現力と形式的な汎用性を備えていなければ正答に至ることはできない。したがって、ユニットテストの内容として問われるべき事項を整理することに次いで、それらの事項を満たす意味表現を検討することが必要となる。

本稿の構成を説明する。**2** 節では、既存の自然言語理解タスクを概観しながら、自然言語理解の能力を構成すると考えられる要素や技能をユニットテストについて述べる（詳細は菅原 (2016) を参照のこと [24]）。**3** 節では、 Schubert による議論 [18][19] を参考にしながら理想的な意味表現が備えるべき要素を整理し、意味表現に関する既存研究がそれらの要素を満たしているか分析する。**4** 節では、既存の意味表現を例として取り上げ、**3** 節の分析をもとに具体的にどのような改良や拡張が行われるべきか考察する。

連絡先: 菅原 朔, 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻, sakus@is.s.u-tokyo.ac.jp

2. 自然言語理解ユニットテストの設計

2.1 既存の自然言語理解タスク

自然言語処理分野ではこれまでに様々な言語理解タスクが提案されている。まず言語理解能力を個々に掘り下げて問うタスクとして、含意関係認識を問う Recognizing Textual Entailment (RTE)[7] や、常識的な知識を問う Choice of Plausible Alternatives (COPA)[16] や Winograd Schema Challenge (WSC)[13] 、小学生程度の理科や数学の能力を問う Arist Challenge[6] 、談話関係を問う Shallow Discourse Parsing (SDP)(CoNLL 2015 shared task)[23] などが挙げられる。

総合的に文章理解を問うタスクとしては、 課題文とその要約・説明文の穴埋めからなる DeepMind Q&A Dataset (DMQA)[10] や The Children's Book Test (CBT)[11]、課題文と質問文・選択肢からなる Machine Comprehension Test (MCTest)[15] や Question Answering for Machine Reading Evaluation (QA4MRE)[21] などがある。穴埋めを解決するタスクについては一定の正答率が実現されている（例示した二つのタスクでは解答の候補が文中に含まれているか別途提示されている）ものの、選択式のタスクでは低い正答率に留まっており、幅広く要求される事項に対して汎用的・統合的に解答できるようなシステムが構築できるかどうかが焦点となる。

また、質問応答の形式で自然言語理解の基礎的な能力をテストすることを目的としたタスクとして、 Facebook bAbI tasks[22] がある。これは 20 種の小タスクで構成され、それぞれの小タスクは数千からなる自動生成された文脈の単文と質問文で構築されており、質問は人間の大人であれば簡単に解くことができるような難易度になっている。

2.2 ユニットテスト設計のための項目整理

本研究では前項で紹介した bAbI tasks をユニットテストの原型として扱い、これらを拡張・改善する方針で議論を進める。拡張のために、 bAbI tasks が文法事項を対象にしていない点に着目し、自然言語理解において次の二段階を区別して整理を進める。すなわち、

- 言語表現を認識する段階（文法要素の認識）
- 認識した情報を組み合わせる段階（技能的処理）

表 1: 文法要素の整理

品詞	文法要素
名詞/一般名詞	数、定性 (冠詞)
名詞/人称代名詞	格、人称、性、数
限定詞・代名詞	指示、疑問、関係、程度など
形容詞	叙述・限定用法、比較表現
動詞	態、様相、時制と相、文型
前置詞	時間指示、空間指示など
副詞	語句修飾、文修飾
接続詞	等位接続、従属接続
文/法	仮定、命令、疑問

表 2: 技能的処理の整理

技能的処理	主要タスク
列挙・数え上げ	bAbI
数理的処理	Arist
共参照解析	COPA, WSC, bAbI
論理推論	Arist, bAbI
類推・比喩の認識	SDP, MCTest
時間空間関係の認識	SDP, MCTest, bAbI
含意関係の認識	RTE, SDP, DMQA, CBT
因果関係の認識	Arist, SDP, bAbI
複文の理解	SDP, MCTest, QA4MRE
常識的知識の運用	COPA, WSC
外部知識の運用	Arist, MCTest, QA4MRE

である。前者の文法要素とは、特定の機能や性質を表現するために用いられる品詞や構文、文法範疇などを指す（表 1。作成には Aarts (2011)[1] などを参考にした。 bAbI tasks はこれらを部分的に含んでいる)。後者の技能的処理とは、認識した情報を語句や節の単位で何らかの関係のもとに結びつける操作を指す（表 2。それぞれの技能が必要なタスクを例示した)。

2.3 ユニットテスト設計の指針と具体例

ユニットテスト設計の指針として次の点が考えられる。

1. 単一の文法要素と技能の組み合わせを一単位とする
   → 個々の文法要素が漏れなく認識されているかを技能的な処理が必要とされる関係的な理解の上で問う
2. 内容語に変化を持たせた文を並列させる
   → タスクの意図を損なわない範囲で内容語を（辞書を用いて）ランダムに選択し、内容に多様性を持たせる
3. 人間が高い精度で解ける難易度にする
   → n-gram や bag-of-words 的な解決を防ぎつつ、人間が無理なく正答できる明瞭な問題とする

次に、具体的に作成したユニットテストの例を挙げる [*1]。

- 文法: 一般名詞の数 + 技能: 数値計算
  Context:
  Bill bought ten apples.
  Sylvia bought an apple.
  Jeff bought eight apples.
  Q: How many apples did boys buy? A: eighteen
- 文法: 人称代名詞の格、性 + 技能: 共参照解析
  Context:
  Mary had the red hat.
  Fred had the blue hat.
  Mary gave her hat to him.
  Q: What did Mary give to Fred? A: red hat

文法要素と技能の組み合わせに対して文のテンプレートが定義できれば、機能的に同一なクラスの語の集合を用意することにより語彙的な拡張を行いながらもデータセットの構築が自動化できると考えられる。上記の具体例では、一般名詞や人物名詞、同一の意味と見なせる動詞をタグにしたものをテンプレートとし、そのタグに対応した辞書からランダムに語を選択することにより、異なる文を自動的に作成することが可能になる。

*1 bAbI tasks では冠詞に関する厳密な規定がなく、単数のものは the に統一されている。本例でも区別が必要ない限りそれに倣う。

表 3: 意味表現が備えるべき要素 (Schubert, 2015) と本研究（表 1・表 2）の対応

Schubelt (2015)	本研究（表 1・表 2）
自然言語と同等の表現力	文法要素
一般性 (Genericity)*	外部知識
典型的パターン*	外部知識
語句や文の具象化 (Reification)*	共参照解析、複文
自然言語と容易に相互変換可能	-
参照関係を利用可能	共参照解析
意味的な直観に合致	-
推論に利用可能	論理推論
形式的な解釈が可能	-
特徴ごとの集約が可能	時間空間、列挙、数理、類推
規則や含意関係が学習可能	含意、因果、常識

3. 意味表現の要件定義と既存研究の分析

システムは自然言語の文や語を入力として受け取り、何らかの変形・補完を行って内部表現として保持し、タスクの要求を満たす情報を処理や探索によって抽出して出力する。この一連のプロセスにおける内部的な表現のことを意味表現と呼び、自然言語処理や計算言語学では長く研究が進められている。

1 節で述べたように、本研究のユニットテスト生成においても意味表現は重要な役割を果たす。前項の例で示したようにユニットテストは複数の単文からなる文脈と質問文、さらにその正答を一組にして構成されるが、自動的に正答を用意するためには文脈生成と同時にその文の意味表現も構成し処理を行わなければならない。その際の意味表現もテストを解くシステム側と同等のものである必要がある。

3.1 意味表現が満たすべき要件

前節で提示したユニットテストの項目を満足するために意味表現が備えていなければならない内容・形式について分析する。 Schubert は意味表現が備えるべき要素として表 3 のような項目を挙げている（各項目が満たされているとき実現されると見なせる表 1・表 2 の要件を対応付けて右行に示した）[18][19][*2]。まず、いくつかの項目について説明を加える。

- 一般性 (Genericity)
  ある事柄についての知識が典型的な場合正しく適用できる性質（例えば「りんごは赤い」という知識は典型的には正しいが、青りんごという例外が存在する）
- 典型的パターン
  複数の出来事や振る舞いについての決まったパターン（例えば「レストランで食事をする」という表現は「店に行く」「メニューを決める」「料理を食べる」「支払いをする」「店を出る」という一連の典型的な動作を暗黙のうちに意味している)。いわゆるスクリプト的知識と同様である
- 語句や文の具象化 (Reification)
  動詞や形容詞、文の名詞的な表現（例えば Beauty is subject. や That exoplanets exist is now certain.）を指す
- 意味的な直観に合致
  同義語や対義語などの一般的な語彙知識と矛盾しないこと

*2 *印を付けた項目は「自然言語と同等の表現力」の一部として詳述されているが [19]、本研究との対応付けの都合上独立した項目として列挙している。

- 形式的な解釈が可能
  シンボルが示す対象や真理条件を明示的に扱えること。これは事実と信念の区別や様相表現を適切に解釈することも要請する
- 特徴ごとの集約が可能
  特定のカテゴリーや時間、空間、数値、像などの情報を整理し何らかの目的のために処理できること

表 3 を踏まえ、本研究では意味表現が備えるべき要素を次のように整理した。

a 文法要素を網羅的に表現できること
b 自然言語と一貫した相互変換が可能であること
c 共参照関係や具象化された埋め込みが表現可能であること
d 論理推論が可能であること
e 表現が示す対象や真理条件が明示可能であること
f 時間や空間に関わる情報を関係づけて処理できること
g 数理的処理のための抽象化が可能であること
h 対象の特徴や性質ごとにカテゴリーを構成し、選択・集約して処理できること
i 語彙や事象、それに含まれる特徴の関連づけやパターンについて学習と利用が可能であること

個々の項目について補足する。b は、「同一の文脈下において」という制約が必要である。同一の文であっても、前後の文脈によっては対応する意味表現は異なるからである。f は、時間と空間の情報が文法要素によって表現される重要な要素だと考えられるため独立した要件として挙げた。g は、数理的処理は意味表現自体が行うものではないが、自然言語によって表現された数学的対象を処理可能な形に変換する作業は意味表現によってなされるべきであると判断し挙げた。h は類推やカテゴリーに関わる要件であり、対象となる表象の持つ情報が外部知識（あるいは直接的な知覚情報として）として補われることを要請する。i は知識や学習に関わる要件であり、語句や事象の生起関係を経験として蓄積し（談話関係の場合は概念化を経由しつつ）統計的な判断に利用できることを要請する。常識のようなより一般化された規則は、h の要件と組み合わせて学習されるものと見なすことができる。

3.2 意味表現の既存研究の分析

本項では意味表現に関する既存研究として代表的だと思われるものを列挙し、各項目について前項で提示した意味表現が備えるべき要件という観点から代表的な利点・欠点を分析する（付記した英字は前項の要件を指示している）。

First-order Logic[2]
- 概要: 述語、連言・選言、量化子、同値関係
- 利点 (d): 述語のみの表現に留めるなら十分な表現力を持ち、論理的な推論が容易
- 欠点 (a): 様相表現や文修飾など、一階の論理式では表現できない文法要素には対応できない。また、たとえば一般化された量化表現（most や few などの程度表現）についても十分な表現力を持たない

Discourse Representation Theory (DRT)[12]
- 概要: 談話構造付きの一階述語論理
- 利点 (d): 一階述語論理の利点に加え、共参照関係や量化子のスコープが構造化されている（これを拡張した Segmented Discourse Representation Theory (SDRT)[3] ではさらに文間の談話関係が構造化されている）
- 欠点 (a): 一階述語論理の欠点と同様である

Semantic Networks[20]
- 概要: 事物をノード、述語や関係をエッジで表したグラフ
- 利点 (d, f): 効率的に知識を関連付けて推論に利用できる
- 欠点 (a, e): 選言や一般化された量化表現が表現できない。また、モデル論的な解釈を持たないため、表現される内容が個別的な事実か普遍的な事実かを区別できない

Conceptual Meaning Representations[17]
- 概要: 少数に限定された動詞や因果関係と意味役割からなる抽象化された表現
- 利点 (d, i): 抽象化された定義語によって推論が容易になり、記述されるべき規則が少なくなる
- 欠点 (a, h): 個々の表現の細かな違いが削ぎ落とされてしまい、類義語の区別を要求する問いなどに対応できない

Abstract Meaning Representation[4]
- 概要: 述語（派生表現を含む）のフレームを基礎として意味的な修飾を与えた表現
- 利点 (f, h): 事実記述として適度に抽象化され扱いやすく、変換のための規則も学習可能である。日付・場所の情報や、意味的な関係を部分的に付与している
- 欠点 (a, e): 意味表現としては緩く、時制や事実/仮説の区別ができない

Extended Natural Logic[14]
- 概要: 語句レベルの係り受け関係に量化子と極性の情報を付与したもの
- 利点 (b, d, i*): もとの文を保持したまま極性の反転や量化子を扱い、含意関係認識が容易（i は部分的に実現）
- 欠点 (c, e): 共参照関係や様相表現に対応できない

Montague-style Intensional Logics[8]
- 概要: 内包性や含意関係のために拡張された述語論理
- 利点 (d, e): 形式的であり、様相表現などの一階述語論理が欠点としていたものを補っている
- 欠点 (c): 複雑な形式であるために非直観的である。また、複数の文にまたがる共参照関係や量化表現を解決できない例がある。これらの問題は動的意味論 （例えば動的述語論理 [9] など）の文脈に受け継がれている

以上のように、論理的な推論に適した形式 (d) を備える一方で、多くの意味表現が共通して文法要素の表現 (a) について欠点を抱えていることがわかる。概して文法要素を十分に表現するための複雑さは、推論のための柔軟な形式を損なってしまうと考えられる。また、単文だけでなく複数の文の情報を捉えることができる形式化 (c, h など) まで考慮できている意味表現もほとんど検討されていないと言える。

4. 既存の意味表現の具体的検討

本節では、**3** 節で提示した要件に基づいて既存の意味表現にどのような改良がなされるべきか具体的に議論する。本稿では Abstract Meaning Representation (AMR) を取り上げ、**2** 節で提案したユニットテストを実際に解く際に不足する表現・形式を指摘する。

前項で確認した通り、AMR は述語のフレームを基礎として抽象化がなされた意味表現であり、利点としてその項構造としての簡潔さと、それに付与された日付や場所の情報や概念的な関係が利用できることが挙げられる。一方で欠点として、処理の高速化のために時制や冠詞の情報を捨象しているという文法要素的な表現力の弱さや、文を独立にしか処理できない点、記述における事実・仮説（あるいは過去・未来など）を区別できない点などが指摘されている [5]。

さて、**2** 節で提示したユニットテスト例の 1 件目を AMR で表現すると、文脈の 1 文目は次のようになる。

```
Bill bought ten apples.
(b / buy
  :arg-0 (p / person :name (n / name :op1 "Bill"))
  :arg-1 (a / apple :quant 10))
```

しかし 2 文目を処理する際、冠詞や名詞の単数・複数が考慮されないので、数量としての 1 が落とされてしまい、

```
Sylvia bought an apple.
(b / buy
  :arg-0 (p / person :name (n / name :op1 "Sylvia"))
  :arg-1 (a / apple))
```

となって数値計算に必要な情報を失ってしまう。したがって、

```
Sylvia bought an apple.
(b / buy
  :arg-0 (p / person :name (n / name :op1 "Sylvia"))
  :arg-1 (a / apple :quant 1))
```

として数値の情報を加えるようパーサを拡張しなくてはならない。質問文は、

```
How many apples did boys buy?
(b / buy
  :arg-0 (b2 / boy)
  :arg-1 (a / apple :quant amr-unknown))
```

のように未知の項 amr-unknown を用いて表現するが、ここでも名詞の複数性や時制の情報が失われている。正しい計算のためには、少なくとも boy の複数性とその指示対象を示すためのシンボル（変数）が必要であり、例えば

```
Bill bought ten apples.
(b / buy
  :arg-0 (p / person :name (n / name :op1 "Bill")
                    :var v1)
  :arg-1 (a / apple :quant 10))
How many apples did boys buy?
(b / buy
  :arg-0 (b2 / boys :plural +
                    :var v4
                    :ref v1,v3)
  :arg-1 (a / apple :quant amr-unknown))
```

のように属性値を与える必要がある。システムはこの表現に従い、フレームが一致してかつ:ref の対象となっている個々の AMR のりんごの個数を計算することになる。

5. おわりに

本稿では、自然言語理解能力を評価するためのユニットテストの設計指針と具体例を述べ、意味表現が備えるべき要素を整理するとともにその整理に従って意味表現を改善する例を示した。ユニットテストを既存の言語理解タスクへと応用するためには、語彙の拡張やテストの複合、その依存関係について明確な規定が求められる。また意味表現については、各要件を満たす具体的な仕様を決定し、自然文が過不足なく表現できるかを実際に確認する必要がある。今後は以上の検討を進めながら、ユニットテストと意味表現の開発を進める予定である。

参考文献

[1] Bas Aarts. *Oxford modern English grammar*. Oxford University Press, Oxford New York, 2011.

[2] James Allen. *Natural Language Understanding*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1995.

[3] Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, 2003.

[4] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation (amr) 1.0 specification. Parsing on Freebase from Question-Answer Pairs. " In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, pp. 1533–1544, 2012.

[5] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Peter Clark. Elementary school science and math tests as a driver for ai: Take the aristo challenge! In *AAAI*, pp. 4019–4021, 2015.

[7] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pp. 177–190. Springer, 2006.

[8] David R Dowty. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*, Vol. 7. Springer Science & Business Media, 1979.

[9] Jeroen Groenendijk and Martin Stokhof. Dynamic predicate logic. *Linguistics and philosophy*, Vol. 14, No. 1, pp. 39–100, 1991.

[10] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[11] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

[12] Hans Kamp. A theory of truth and semantic representation. *Formal semantics-the essential readings*, pp. 189–222, 1981.

[13] Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[14] Bill MacCartney and Christopher D Manning. An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*, pp. 140–156. Association for Computational Linguistics, 2009.

[15] Matthew Richardson, J.C. Christopher Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193–203. Association for Computational Linguistics, 2013.

[16] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[17] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

[18] Lenhart K Schubert. Semantic representation. In *AAAI*, pp. 4132–4139, 2015.

[19] Lenhart K Schubert. What kinds of knowledge are needed for genuine understanding? In *IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015)*, 2015.

[20] John F. Sowa, editor. *Principles of semantic networks: explorations in the representation of knowledge*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1991.

[21] Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. Overview of qa4mre main task at clef 2013. *Working Notes, CLEF*, 2013.

[22] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: a set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[23] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T Rutherford. The conll-2015 shared task on shallow discourse parsing. *CoNLL 2015*, 2015.

[24] 菅原朔, 横野光, 相澤彰子. 自然言語理解ユニットテストの検討. 言語処理学会第 22 回年次大会論文集, pp. 111–114, 2016.