自然言語理解ユニットテストの検討

菅原 朔 †　横野 光 ‡　相澤 彰子 ‡†

† 東京大学大学院 情報理工学系研究科　‡ 国立情報学研究所

sakus@is.s.u-tokyo.ac.jp, {yokono, aizawa}@nii.ac.jp

1 はじめに

計算機における自然言語理解や知識運用を実現するため、これまでに多くのタスクが提案されてきた。それらは例えば長文の読解や常識的知識の運用、文間の談話関係などを問い、言語理解における高度な能力を要求するものである。これらのタスクを解決するために構築されたシステムの中には高い成果を出すものもあるが、依然として人間と同程度の能力を実現するには至っておらず、またそれらのモデルはほとんどがタスク依存的・ドメイン依存的なのが現状である。

自然言語理解を可能にするシステムを設計するという目的においては、システムが少なくとも解けなければならない問題を細分化することができれば、詳細な性能評価が可能になると考えられる。それを受けて本稿では、自然言語理解に必要な能力を個別的に確認するためのユニットテストを構築することが有効な方針であると考え、既存の自然言語理解タスクを参考にしながら、自然言語理解の能力を構成する要素・技能が何であるか整理し、どのようなユニットテストを構築すればよいのかを議論する。

まず、自然言語理解のために考案された比較的高度なタスクをいくつか取り上げ、それらを解決するために必要な能力が何であるか確認する（**2** 節）。続いて、自然言語理解における基礎的なスキルを測るために考案されたタスク群である Facebook bAbI tasks を概観する（**3** 節）。最後に文法要素・技能的処理の 2 つの観点から **2**, **3** 節で確認した内容を整理して、ユニットテストを設計する指針を検討したのち、具体例を示す（**4** 節）。

2 言語理解タスクの概観

本節では、人間が読むような実際的な文書の言語理解を問うことを意図したタスクである MCTest、QA4MRE (CLEF 2013)、Shallow Discourse Parsing (CoNLL 2015 shared task) の 3 つを取り上げ、それぞれが求める言語理解の能力について確認する。

2.1 MCTest

MCTest (matchine comprehension of text)[1] は、平易な文章についてその理解を問う多岐選択式のタスクである [5]。データセットはドメインを限定しない子供向けの平易な文章とその内容に関する多肢選択式の問題（文書 660, 質問各 4）から構成される。正答のためには、複数の文章の内容を追跡して出来事の時間的な連続性や因果関係などをエピソード的に保持することが必要となる。ただし外部のドメイン的な知識や高度な語彙は必要とされないよう考慮されており、問題の文書だけで内容が完結している。

2.2 QA4MRE 2013

QA4MRE (Question Answering for Machine Reading Evaluation) 2013[2] は、CLEF (Conference and Labs of the Evaluation Forum) における高度な文章の理解を問う多肢選択式タスクである [6]。データセットは分野の異なる 4 つのトピックごとに用意された 800〜2000 語程度の長さの 4 つの文書からなり、それぞれの文書は 15 個の質問をもつ。文書の内容はドメイン限定的であるが、回答に必要な情報が文書外の背景知識（文章集合）として提供され、データセットで完結するように配慮されている。

QA4MRE で特徴的なのは、半数以上の質問で正答ために文書中の事実と背景知識を組み合わせることが求められる点である。例えば “Who is the wife of the person who won the Nobel Peace Prize in 1992?” という文は、the person を結ぶ形で 2 つの知識を必要とする。1 つは “wife of the person” であり、もう 1 つは “the person who won Nobel Peace Prize in 1992”

[1] http://research.microsoft.com/en-us/um/redmond/projects/mctest/index.html

[2] http://nlp.uned.es/clef-qa/

である。ここでは前者が文書中に、後者が背景知識に含まれており、両者を適切に組み合わせなければ正解を導くことができない。

2.3 Shallow Discourse Parsing

Shallow Discourse Parsing (CoNLL 2015 shared task, Xue 2015)[3] は、Penn Discourse Tree Bank (PDTB) を素材として、近傍にある 2 文が持つ談話関係を特定するタスクである [9]。データセットは接続詞や談話内容を示す語によって明示的なヒントが与えられているケースとそのような語の補助がない非明示的なケースに区別されている。

タスクで設定された談話関係の候補は 15 種類あり、同期や継続、原因、結果、条件、対比、譲歩、例示などが含まれる。これらを推定するためには、事象間における時間関係の比較や因果関係の認識、構造的な類推、含意関係の知識などが必要である。

3 Facebook bAbI tasks

Facebook bAbI tasks[4] は、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群である [7]。12 カテゴリー 20 種の小タスクで構成され、それぞれの小タスク（本稿では以降ユニットテストと呼称する）の訓練データとテストデータは文脈設定の単文と 1000 文ずつの質問からなる。データセットの構築には Bordes らの手法が利用され、限定的な語彙（150 語程度）ではあるもののタスクのための文章列を自動生成する試みがなされている [3]。

タスクは次の通りである（日本語は筆者による補足説明）。

1. Single Supporting Facts
   - 単一の情報の認識
2. Two or Three Supporting Facts
   - 複数の情報の関連付け
3. Two or Three Argument Relations
   - 複数の引数をとる述語の理解
4. Yes/No Questions
   - Yes/No 疑問文の理解
5. Counting and Lists/Sets
   - 事物の数え上げ、事物の列挙
6. Simple Negation and Indefinite Knowledge
   - 否定の理解、選言の理解
7. Basic Coreference, Conjunctions and Compound Coreference
   - 曖昧さのない単純な照応の解決
   - 連言の理解、連言を含む照応の解決
8. Time Reasoning
   - 時間指標詞による順序関係の把握
9. Basic Decuction and Induction
   - 基礎的な演繹・帰納推論
10. Positional and Size Reasoning
    - 空間指標詞による位置関係の把握
    - 量的性質の比較推論
11. Path Finding
    - 方角的な位置関係と移動の推論
12. Agent's Motivations
    - 人物の状態の追跡

次にいくつか具体例を引用する。

2. John is in the playground.
   John picked up the football.
   Where is the football? A: playground

6. Fred is no longer in the office.
   John is either in the classroom or the playground.
   Is John in the classroom? A: maybe
   Is John in the office? A: no
   Is Fred in the office? A: no

7. John and Mary went back to the hallway.
   Then they went to the bathroom.
   Where is John? A: bathroom

8. This morning Mary went to the kitchen.
   Yesterday Mary travelled to the cinema.
   This afternoon Mary went back to the office.
   Where was Mary before the kitchen? A: cinema
   Where was Mary before the office? A: kitchen

11. The kitchen is north of the hallway.
    The den is east of the hallway.
    How do you go from den to kitchen? A: west, north

言語理解のために構築されたシステムはこれらの質問にできるだけ少ない規則と学習で正答することが期待される [7]。質問の内容が高度なタスクに直接結びつくわけではないものの、人間の大人（あるいは子供でも）はこれらの質問に簡単に正答できる。したがって、各タスクごとに対応した個別のシステムではなく単一のシステムで正答できなければ言語理解の能力を持っているとは言い難いと考えられる。

Weston らは Memory Networks で実験を行い、タスクを個別に学習した場合は 20 種のうち 16 種のタスクで、全タスクをひとつのシステムで学習した場合は 13 種のタスクで、それぞれ正答率が 95%を上回る結果となったと報告している [8]。ただし学習では「どの行に正答があるか」という教師が与えられており、限定的な語彙における語順的な最適化がなされている可能性もあるため、語彙を拡張した場合やタスクを複合した場合に正答できるかといった詳細な分析が必要だと考えられる。

[3] http://www.cs.brandeis.edu/~clp/conll15st/index.html
[4] http://fb.ai/babi

表 1: 文法要素の整理

文法要素	小項目	bAbI
名詞・数/定性	代名詞の格/人称/性/数を含む	△
名詞・量化	代名詞的用法を含む	×
動詞・項構造	文型、自動詞/他動詞、動詞句	△
動詞・時制/相	両者の組み合わせ	×
動詞・態	受動態/能動態	○
動詞・語彙	移動/譲渡/状態/使役など	△
動詞・法	仮定法 (条件法)/命令法	×
動詞・様相	助動詞	×
形容詞・比較	原級/比較級/最上級	△
副詞	頻度/時間指示/空間指示	△
前置詞	時間指示/空間指示	△
文・疑問	一般疑問/特殊疑問など	○
文・否定		○
文・コピュラ		○
接続詞・語/句	連言・選言	○

4 ユニットテストの検討

4.1 文法的・技能的整理

前節で概観した Facebook bAbI tasks は、タスクの種類としては多様であったものの、語彙的には **2** 節で確認したような言語理解タスクよりも限られており、文法事項を網羅的に確認するものではない。

そのため、ユニットテストを検討するにあたり、本稿では自然言語理解において次の二段階を区別して整理を試みる。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）

である。前者の文法要素とは、何らかの機能や内容を表現するための文法範疇や品詞、構文などを指す。後者の技能的処理とは、語句や節を何らかの関係のもとで結びつける操作を指し、具体的には従属的接続関係の把握や因果関係認識、共参照解析、推論能力などが該当する。

たとえば、**3** 節の例 7 では、まず文法要素として “John and Mary” という連言を認識し、その後で技能的な処理として “They” が指す対象の共参照解析を行うことが必要になる。

文法要素は Aarts (2011）等を参考にして表 1 のように整理し（本稿では英文法を対象とした。また、形式的な文法事項から外れる特殊な用語法・構文等は含めなかった）[1]、bAbI の列は、bAbI tasks が各文法要素を含んでいるかどうかを表している（△は小項目を部分的に含むことを示す）。

技能的処理（以下、“技能” と略記する）は、**2**, **3** 節を参考にして表 2 のように列挙し、各技能を必要とするタスクを付記した。“bAbI*” は bAbI tasks において問われているものの小項目に記載した要素を網羅していないことを示している。なお、本稿では 3 種の言語理解タスクの分析に終始したため、この表 2 ですべての技能がカバーできているかについてはさらに他のタスクを分析し検討する余地があると思われる。

表 2: 技能的処理の整理

(M=MCTest, Q=QA4MRE, S=Shallow Discourse Parsing)

技能的処理	小項目	タスク
数値計算	四則演算	M
数え上げ/列挙	数詞/情報の保持	bAbI
共参照解析	代名詞の格/人称/性/数	bAbI*
推論	演繹/帰納/仮説推論	bAbI*
類推	比喩/対比	S
時間空間認識	指標詞/時制	bAbI*/M/S
含意関係認識	具体化/抽象化/換言	Q/S
因果関係認識	原因/結果など	S
複文の理解	関係詞・接続詞	Q/S
談話関係認識	接続詞/内容判断	S
知識運用	文書中の事実/外部知識	bAbI*/M/Q

4.2 ユニットテスト設計の指針

自然言語理解を行うシステムは、長期的な目標として談話的・対話的な状況下で人と同じ振る舞いをすることが期待される。しかし現時点でそのような理想的な振る舞いを前提とした対話的タスクを構築するのは、応答としてどのような発話であれば正解とみなすか（もしくは、誤っていると思われる発話の何が具体的に誤っているのか）を規定するのが困難であり、今のところ性能評価に適さないと考えられる。

bAbI tasks が企図しているように、ユニットテストを質問応答タスクとして設計するのはシステムの性能評価が比較的簡単になるという利点があるためである。しかしながらテキストベースのテストにおいて問題設定や質問文を適切に設定するのは容易とは言えない。たとえばある文法要素の理解を単文の文脈で確認したとしても、その文法要素を文の連なった複雑な文脈で参照して技能の遂行に利用できるかどうかは不明瞭である [5]。つまり、効率的なユニットテストを設計するためには、要素が表現する内容の文脈的な構造化・結合を問うべきであり [6]、テストは文法要素・技能ごとではなくひとつの文法要素と何らかの技能の組み合わせを一単位として構築されるのが望ましいと考えられる。このような「文法要素を充分に考慮した文脈を

[5] たとえば **3** 節の例 8 において、それぞれの文が時間的にいつであるかを問うだけでは、複数の時間指示詞の順序関係が理解できているかどうかはわからない。

[6] 仮に現実の対象物と文章を視覚的に結びつけることができる設定下なら、その対象の描写に用いる文法事項を個別的に問うことが可能だと考えられる。これに関連して、Antol らは画像を対象とした QA タスクを公開している [2]。

用いて技能を遂行する」見方は、Winograd Schema Challenge に関する井之上・杉浦らの指摘と類似的であり [4] [10] [11]、たとえば外部知識をコーパスから獲得し推論規則として適用する際には、その知識が妥当となるような文脈的情報が充分に考慮されていなければならない。そうした文脈情報として考えられる要素をより一般化して列挙することが本稿の目的のひとつである。

個々のユニットテストは文脈として主語や目的語などを入れ替えた表現を並列させることで、 n-gram や bag-of-words 的な単純な解決を防ぐことが可能になる。その上で **3** 節で触れたように「自然言語を解する大人がほぼすべて確実に解ける」ような曖昧性のない質問であることが必要であり、難易度のバランスには注意を払わなければならない。

以上のことをまとめると、ユニットテスト設計の指針として次のことが要点となる。

- 単一の文法要素と技能の組み合わせを一単位とする
- 内容語に変化を持たせた文を並列させる
- 人間の大人が高い精度で解ける難易度

4.3 ユニットテストの例

ここまでの議論を踏まえてユニットテストの具体例を挙げる。形式は bAbI tasks に揃え、内容は **2** 節で参照したタスクや Winograd Schema Challenge を参考にしている。

- 名詞・数 + 含意関係認識 + 数値計算
  Bill bought ten apples.
  Sylvia bought an apple.
  Jeff bought eight apples.
  How many apples did guys buy? A: eighteen
- 名詞・格/性 + 共参照解析
  Mary had the red hat.
  Fred had the blue hat.
  Mary gave her hat to him.
  What did Mary give to Fred? A: red hat
- 動詞・様相 + 知識運用 + 推論
  Tom is a student.
  Students must study math everyday.
  Daniel must study French everyday.
  What must Tom study everyday? A: math
- 動詞・時制/相 + 複文の理解 (関係詞)
  John was running with his dog who was white.
  John is running with his dog who is black.
  Was the dog who John was running with black?
  A: No
- 名詞・量化 + 因果関係
  Arachne is a spider.
  All spiders have to discard their old skin so that they can grow a new skin.
  Why must Arachne discard her old skin? A: to grow a new skin.

5 おわりに

本稿では、既存の自然言語理解タスクと Facebook bAbI tasks を参考にして、文法要素と技能的処理という観点から理想的なユニットテストが備えるべき項目を整理した。テストとして文の形を確定させれば、機能的に同一なクラスの語のセットを用意して語彙的な拡張を行うことで、データセットの作成を自動化することも可能だと考えられる [7]。今後は具体的なユニットテストの構築を進め、テストが言語理解システムの性能向上に寄与するか意味表現の観点から考察を行う予定である。

参考文献

[1] B. Aarts. *Oxford Modern English Grammar*. The world's most trusted reference books. OUP Oxford, 2011.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] Antoine Bordes, Nicolas Usunier, Ronan Collobert, and Jason Weston. Towards understanding situated natural language. *Proc. of the 13th Intern. Conf. on Artif. Intel. and Stat*, Vol. 9, pp. 65–72, 2010.

[4] Hector J. Levesque. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[9] 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.

[10] 井之上直也, 杉浦純, 乾健太郎. 共参照解析のための事象間関係知識の文脈化. 言語処理学会第 20 回年次大会論文集, pp. 717–720, 2014.

[11] 杉浦純, 井之上直也, 乾健太郎. 共参照解析における事象間関係知識の適用. 言語処理学会第 20 回年次大会論文集, pp. 713–716, 2014.

[7] bAbI tasks では、移動を表す動詞として “go” や “journey” 、 “travel” 、 “move” が同じものとして扱われており、これらを入れ替えることでデータセットに多様性を持たせている。