Subjective BERT: self-attentionによる「おいしいね」「おいしそうだよ」の意味理解

Subjective BERT: Understanding the Meaning of *Oishii ne.* 'I Want to Make Sure that You also Feel Delicious.' and *Oishi souda yo.* 'I Want to Tell You that It Looks Delicious.' by Self-Attention

岡 夏樹†, 松島 茜†, 萬處 修平†, 深田 智†, 吉村 優子‡, 川原 功司*
Natsuki Oka, Akane Matsushima, Shuhei Mandokoro, Chie Fukada, Yuko Yoshimura, Koji Kawahara
†京都工芸繊維大学, ‡金沢大学, *名古屋外国語大学
Kyoto Institute of Technology, Kanazawa University, Nagoya University of Foreign Studies
nat@kit.ac.jp

概要

言葉と画像（視覚情報）だけの結びつけを越えたより豊かな言葉の意味理解を目指して、言葉と様々な主観的感覚（視覚を含む）の間の関係を学習させたいと考えた。このために、言語と画像に加えて様々な主観的感覚を入力とする self-attention モデルである Subjective BERT を提案し、特に、機能語（終助詞や助動詞）の獲得に注目して、「おいしい・ね」「おいし・そうだ・よ」などの発話理解を試みた。計算機シミュレーションの中間結果を報告する。

キーワード：機能語, 終助詞, 助動詞, 主観的感覚, 言語獲得, 深層学習

1. はじめに

深層学習の一種である LSTM (Long short-term memory)を用い、大量の翻訳対から end-to-end で（人が多段の中間表現を設計することなく、入力から出力までを一つのニューラルネットワークで）学習することで Google の機械翻訳の質が急に向上した（Wu 2016）ことは記憶に新しい。その後、self-attention を計算原理とした単純なネットワーク構造を用いることにより、翻訳（Vaswani 2017）や自然言語理解（Devlin 2018）の性能がさらにもう一段高まった。これは、self-attention により形成される暗黙的な階層構造により、階層構造を作る明示的な文法規則なしで、階層性を持つ言語の理解と生成が実用的なレベルでできるようになったことを意味する。以上の成果は文字情報だけからの学習により達成されたが、最近では、言語に加えて画像も入力して self-attention で処理することにより、言語と画像を結び付けた理解や生成を目指した研究も活発である（Lu 2019, Chen 2019, Radford 2021）。

我々は、言語と画像（視覚情報）だけの結びつけを超えたより豊かな言葉の意味理解を目指して、言葉と様々な主観的感覚（視覚を含む）の間の関係を学習させたいと考えた。ここで主観的感覚とは、以下の総称としてかなり広い意味で用いる：

(i) 外受容感覚（視覚、聴覚、味覚、臭覚、触覚）
(ii) 固有感覚（身体各部の運動や位置などの知覚）、
(iii) 内受容感覚（内臓や血管等の状態の知覚）、
(iv) 心の状態（願望、選好、推量など）

心の状態は、過去と現在の感覚運動経験から立ち現れるものであるが、言葉の意味理解に必要な入力だと考え主観的感覚に加えた。本論文では、言語と画像に加えて様々な主観的感覚を入力とする self-attention モデルである Subjective BERT を提案し、特に、機能語（終助詞や助動詞）の獲得に注目して、「おいしい・ね」「おいし・そうだ・よ」などの発話理解を試みた。

2. Subjective BERT

Subjective BERT は、言葉と様々な主観的感覚（視覚を含む）の間の関係を学習させるために、BERT（Devlin 2018）の入力を言葉だけでなく様々な主観的感覚を含むように拡張したものである。本体の構造は BERT と同じものを使用する。

Subjective BERT への入力は、離散的な識別子(id)である必要がある。このため、本来であれば、感覚器等でとらえたアナログ信号に何らかの認識処理を施し、その結果の id を Subjective BERT に入力する、とすべきところであるが、本論文では簡略化のため、元信号を何らかの手段で検出することができ、かつ、この信号の認識処理が別途用意できたものとして、id を入力した以降のシミュレーションを行う。各入力信号の検出・識別過程も含めたシミュレーション実験を行うことは今後の課題とする。主観的感覚情報を計算機で処理できるように検出・認識することには困難が伴うと考えられるため、全体のシミュレーションを近い将来完成することは現実的でない可能性がある。しかし、この

検出・認識の後の部分に限定したとしても、そのシミュレーションを行い、機能語の理解のどの部分がself-attention に基づく計算で実行可能かを検討することには、人の知能を理解する上で意義があると考える。

Subjective BERT では BERT（Devlin 2018）同様、[CLS]トークンと[SEP]トークンを用いるが、[SEP]トークンの役割は異なる。Subjective BERT では、各モダリティからの入力を[SEP]で区切る。また、言語入力は一時刻前の発話と現時点の発話を連接したものとするが、それらの間も[SEP]で区切る。

token embedding と position embedding は BERT（Devlin 2018）と同じ方法を採用した。ただし、視覚入力と主観的感覚入力では特に順序はないため、同一の position embedding を使う。また、BERT の segment embedding に代えて、modality embedding をどのモダリティからの入力かを区別するために使用する。さらに、time embedding を一時刻前の発話と、現時点の発話や感覚入力とを区別するために導入した。これら 4 種類の embedding を加算したものが Subjective BERT に入力される。

3. シミュレーション実験の設定

シミュレーション実験の設定は次の通りであった。養育者が子どもに話しかける場面を想定し、Subjective BERT に対して以下の入力を与えた。

- １時刻前の発話と現時刻の発話：発話は、｛リンゴ・だ・よ、リンゴ・だ・ね、バナナ・だ・よ、バナナ・だ・ね、おいしい・よ、おいしい・ね、おいし・そうだ・よ、おいし・そうだ・ね、食べ・たい・ね、おなかすいた・ね｝の 10 通りのいずれか、または、発話無し、であった。「・」は Subjective BERT に入力される際の単語区切りを示す。子どもの見ているものや子どもの主観的感覚と整合する発話を養育者はすると想定して、入力する発話を選択した。整合性の制約をいくつか例示する：「リンゴだね」は、子どもがリンゴを見ているときだけ発話する；「おいしいね」「おいしそうだね」「おなかすいたね」等の発話は子どもがそう感じていると思われるときだけ発話する；「おいしいよ」「おいしそうだよ」は、子どもがそう感じてないと思われるときだけ発話する。
- 現時刻の画像：画像はリンゴとバナナについて、それぞれ、おいしそうな（食べるとおいしい）ものと、子どもにとってはおいしそうに見えない（が食べるとおいしい）もの（青りんごやシュガースポットの入ったバナナ）、合わせて 4 種類のいずれかとした。
- 現時刻の推量：上記 4 種類の果物の画像からの味覚（おいしいリンゴの味、おいしくないリンゴの味、おいしいバナナの味、おいしくないバナナの味）の推量結果の 4 種類＋（今実際に味覚を感じているので）推量していない状態の合計 5 種類のいずれかを入力した。なお、このような具体的な推量結果の入力に代えて、推量という抽象的な心の働き（助動詞「そうだ」に対応）の有無を入力とする場合のシミュレーションも実施予定である。
- 現時刻の味覚：子どもの食べる動作に伴い味覚が生じると想定し、おいしいリンゴの味、おいしいバナナの味、無し（食べてないとき）、の 3 種類のいずれかとした。
- 現時刻の空腹感：子どもの食べる動作からの時間経過で空腹感が決まると想定し、空腹感有り／無しのどちらかとした。
- 現時刻の願望：おなかがすいた時、または、おいしそうな果物を見た時に、リンゴを食べたい、または、バナナを食べたい、のどちらかの願望が生じると想定した。願望がない状態と合わせて合計 3 種類のいずれかとした。

4. Subjective BERT の学習と評価

事前学習として、BERT（Devlin 2018）で提案された masked LM を用いた。手順は次の通り：

1. 入力の全トークンの 15%をランダムに選び予測トークンとする。
2. 予測トークンの 80%を[MASK]トークンとし、10%をランダムなトークンと置き換え、10%を元のトークンのままとする。
3. クロスエントロピーロスを用いて、予測トークンを予測することを学習する。

本研究ではファイン・チューニングは行わず、事前学習済みのモデルに対して、任意のモダリティの情報から任意のマスクされたモダリティの情報を予測するクロスモーダル情報予測を行うことにより、マルチモーダル情報におけるモダリティ間の関係性をどの程度学習できているかを評価した。

5. 実験結果と考察

主観的感覚を入力に加えたことにより、「おなかすいたね」という発話を受けて「おなかすいた」という内容語に空腹感を対応付ける学習もできるが、本研究の主要な新規性は、機能語（「よ」「ね」のような終助詞や「そうだ」のような助動詞）の意味を獲得するところにあるため、以下では機能語の獲得に絞って論じる。

まず、終助詞「よ」と「ね」の働きがどう学習されたかを記す。「おいしいよ」という入力（対話相手の発話）に対しては、一定の条件下で次の時刻でのおいしいという感覚が予測されるが、現時刻の同感覚は伴わないことが学習され、一方、「おいしいね」という入力に対しては、一定の条件下で現時刻のおいしいという感覚が想起できるが、次時刻の同感覚は必ずしも伴わないことが学習された(図 1)。この結果は、終助詞「よ」で表現された発話者の意図（情報を伝える）や、「ね」で表現された意図（同じ感覚を持っていることの表明を求める）に即した情報処理の一部（「よ」を伴う発話に含まれる内容語に対応する情報はその発話時点でなく次の時点で観察される可能性が高いことを理解。また、「ね」を伴う発話に含まれる内容語に対応する情報はその発話時点で観察される可能性が高いことを理解。）ができるようになったことを示していると言ってよいと考える。ただし、すべての「よ」「ね」を伴う発話に対してこのような結果が一貫して得られてはいないため、学習データ、学習方法、評価方法の再検討を現在行っている。

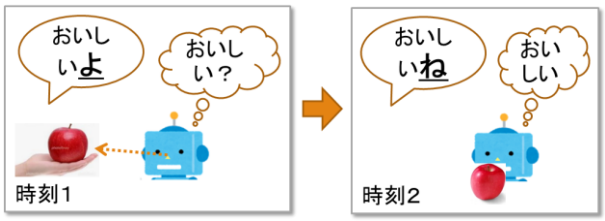

図 1　終助詞「よ」「ね」の使用場面

次に、助動詞「そうだ」については、「おいしそうだ」が他の感覚からの味覚の推測である（図 2）ことを学習できたと解釈できる結果を得た。具体的には、「おいしいね」という発話入力に対して、味覚と推量をマスクして予測させると、味覚の存在を比較的高い確率で予測し、逆に、「おいしそうだね」という発話入力に対して、味覚と推量をマスクして予測させると、味覚の欠如と推量の存在を比較的高い確率で予測した（表 1，表 2)。ただし、これらの表からも分かるように、「おいしいね」と「おいしそうだね」を比較すれば、想定通りの傾向を持っていることが分かるが、十分な学習ができているとは言えない結果であるため、こちらについても、学習データ、学習方法、評価方法の再検討を現在行っている。

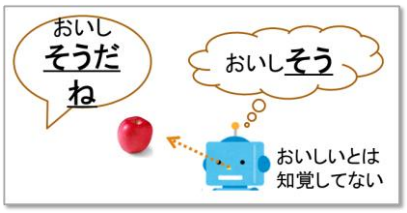

図 2　助動詞「そうだ」の使用場面

6. 結論と今後の展望

self-attention を用いたモデルにより、終助詞「よ」「ね」や助動詞「そうだ」の意味の一部をとらえることができる感触を得た。学習方法と評価方法を見直すことにより、確かに学習可能だという結果を得ることが次の目標である。このためには、学習データの規模をある程度大きくすることも必要かもしれない。

終助詞「よ」「ね」の用法には、聞き手が知らないと目される情報を伝える、聞き手と共有していると目される情報について同意を求める、などのように、相手の知識や心の状態についての情報が必要になるものがある。したがって Subjective BERT でこれを扱うためには、相手の動作や観察可能な状態を入力として加え、相手の心の状態を推測できるように拡張する必要がある。

また、Subjective BERT に強化学習モジュールを追加して、ロボットが何を見るかや、食べるなどの行動を獲得できるようにして、行動の学習と Subjective BERT の学習を並行して進めることも計画している。

現在のモデルでは、時間の区切りを、1 時刻前の発話、現時刻の発話、現時刻の間隔、のように恣意的に与えているが、この時間の扱いを設計者が完全に決めるのでなく、様々な粒度での扱いを可能にするか、学習システム側で時間の区切りを作り出せるようにすることも重要な課題である。この問題は（松尾 2021）でも指摘されている。

謝辞

本研究は JSPS 科研費 JP20H05004 の助成を受けたものである。

文献

[Wu 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean, Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, arXiv:1609.08144 [cs.CL], 2016.

[Vaswani 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, arXiv:1706.03762 [cs.CL], 2017.

[Devlin 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs.CL], 2018.

[Lu 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, arXiv:1908.02265 [cs.CV], 2019.

[Chen 2019] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, Jingjing Liu, UNITER: UNiversal Image-TExt Representation Learning, arXiv:1909.11740 [cs.CV], 2019.

[Radford 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv:2103.00020 [cs.CV], 2021.

[松尾2021] 松尾豊, 深層学習と人工知能, 認知科学, 第28巻, 第 2 号, pp. 299-307, 2021.

表１　各推量の予測確率（推量と味覚を MASK）

入力(発話)	発話時刻	おいしそうなリンゴに対する推量	おいしくなさそうなリンゴに対する推量	おいしそうなバナナに対する推量	おいしくなさそうなバナナに対する推量	推量なし
おいしそうだね	現時刻	0.25	0.10625	0.14375	0.08125	0.3625
おいしいね	現時刻	0.125	0.05625	0.11875	0.06875	0.15625

表２　各味覚の予測確率（推量と味覚を MASK）

入力(発話)	発話時刻	リンゴの味	バナナの味	味覚なし
おいしそうだね	現時刻	0.0	0.0	0.3625
おいしいね	現時刻	0.2375	0.2375	0.15625