節埋め込みの意味論に動機づけられたプロービング

船蔵颯[1]　櫻川貴司[1]　峯島宏次[2]
[1]京都大学大学院　[2]慶應義塾大学
funakura.hayate.28p@st.kyoto-u.ac.jp　sakura@i.h.kyoto-u.ac.jp
minesima@abelard.flet.keio.ac.jp

概要

本稿では、「Transformer に基づく事前学習済み言語モデルは節埋め込みの意味論において重要な情報をエンコードしているか？」という問いに答えるために、節選択性に応じた述語分類と意味タグ付与という 2 つのタスクを設計し、事前学習済み BERT に対するプロービングを実施した。実験の結果、BERT の出力する埋め込みベクトルを用いた節選択性の予測が一定程度可能であることが明らかになったが、よりミクロな視点からの調査が課題として残っている。

1　はじめに

本稿では、「Transformer に基づく事前学習済み言語モデルは節埋め込みの意味論において重要な情報をエンコードしているか？」という問いに答えるために、節選択性に応じた述語分類と意味タグ付与という 2 つのタスクを設計し、事前学習済み BERT に対するプロービングを実施した。

節埋め込みの意味論は、英語の *know* や日本語の「知る」のように、目的語位置に節を取ることのできる述語を主対象とした形式意味論の領域である。この領域で特によく議論される節タイプとして、that 節に代表される平叙節と、wh 節に代表される疑問節とがある。節埋め込み述語はそれが取ることのできる節のタイプにより大きく以下の三つに分類される [1]。

Responsive　平叙節と疑問節の両方（例: *know*）
Anti-rogative　平叙節のみ（例: *believe*）
Rogative　疑問節のみ（例: *wonder*）

このように、目的語位置に取ることのできる節のタイプが述語によって異なることを本稿では**節選択性**と呼ぶ。異なる言語の述語であっても意味が類似していれば取りうる節タイプが一致する傾向にある [2, etc.] ことや、節選択制が推論と強く関連する [3, etc.] ことから、節選択性は意味論の問題として盛んに議論されている [4, 5, 6, etc.]。

ニューラルネットワークをはじめとする機械学習モデルが節選択性の知識をエンコードしているならば、節埋め込みの意味論に資する知見あるいは仮説をそのモデルから得られるかもしれない。また、ニューラルネットワークの持つ知識を解明するプロービングについては既に多くの研究がある [7] が、節選択性に関する調査は著者の知る限りで存在しない。本研究はこのような背景から冒頭の問いを立て、一つのケーススタディとして事前学習済み BERT の英語モデルを用いたプロービングを実施した。

具体的なタスクとしては、節選択性に応じた述語分類と、節選択性を反映する形に拡張された意味タグ付与という二つを設計した。前者については 2 節、後者については 3 節にて、それぞれの実験設定と結果を述べる。

2　実験 1: 節タイプ分類タスク

本実験では、文中に出現する節埋め込み述語を三つのカテゴリに分類するタスクを定義し、Hewitt ら [8] によって提案されたコントロール実験を行なった。以下では、このコントロール実験について説明し、そして使用したデータセットとモデルの設定について述べる。

2.1　コントロール実験

Hewitt ら [8] によって提案された実験手法は、目下の言語学的タスクの学習に加えて、そのタスクの分類ラベルがランダムに書き換えられたタスクの学習を行うというものである。後者のタスクをコントロールタスク (control task) と呼ぶ。二つのタスクで学習を行なった結果、それぞれのタスクに対する正解率の差を求めることができるが、この差を

selectivity と呼ばれる。この手法の背景には、**モデルの言語学的な特性をよりよく反映するプローブは selectivity が大きい**、という直観がある。

コントロールタスクは、素朴に言えば、**ある程度の構造はあるが、言語学的な知識とは無関係のタスク**である。コントロールタスクを定義する際、分類ラベルセットは元タスクと同一であり、書き換え後のラベルは語のトークンではなくタイプごとに決定される。すなわち、コントロールタスクはラベルの付与方法自体はランダムでありつつ、語のタイプが同じであれば同じラベルが付与されるという意味での構造を持っている。

2.2 データセット

本実験用のデータセットは Parallel Meaning Bank 4.0.0 (以下 PMB4) の英語パートをベースに作成した。このデータセットは [9, 10] で提案されたデータセットの現行最新版であり、自然言語文に対して組合せ範疇文法の導出木、意味タグ系列、DRS による意味表示といった情報がアノテーションされている。

PMB4 に含まれるデータは、全てのアノテーションが人手で確認済みのサンプル群、アノテーションの一部が人手で確認済みのサンプル群、アノテーションの全てが人手では未確認のサンプル群という三つに分かれる。本稿ではこれらをゴールドデータ、シルバーデータ、ブロンズデータとそれぞれ呼ぶ。我々は PMB4 の英語ゴールドデータ 10813 文と、シルバーデータ 136491 文から、節埋め込み文（9158 文）を抽出した。より具体的には、語彙レベルの統語範疇に以下の統語範疇のいずれかが含まれる文を抽出した（ただし $x \in \{b, dcl, ng, pss, pt\}, y \in \{dcl, em, qem\}$）。

- $(S_x \backslash NP)/S_y$
- $((S_x \backslash NP)/S_y)/NP$
- $((S_x \backslash NP)/S_y)/PP$

そして、各文中の節埋め込み述語に対して、節選択性に応じた分類ラベルを付与した。ラベルは RE, AR, RO であり、それぞれ responsive, anti-rogative, rogative に対応する。以下は分類ラベルが付与されたサンプルの例である。

(1) a. I haven't decided$_{\mathsf{RE}}$ yet whether I will attend the party.
b. I think$_{\mathsf{AR}}$ he ate about 10 oysters.
c. Tom asked$_{\mathsf{RO}}$ Mary if she had been to Boston.

また、データセット内に出現する語タイプごとにランダムな分類ラベルを付与することで、コントロールタスク用のデータセットを作成した。作成したデータセットの、各語タイプと分類ラベルの対応は付録中の表 5 の通りである。

2.3 モデルの設定

先述のデータセットを用いて、本実験では事前学習済み BERT の出力する文脈を考慮した埋め込みベクトル列のうち、節埋め込み述語に対応するベクトル（$\in \mathbb{R}^{768}$）を RE, AR, RO のいずれかに分類するモデル（以下プローブと呼ぶ）の学習を行なった。BERT のパラメータは `bert-base-uncased` のものを採用した。このモデルに対応するトークナイザはサブワードトークナイザであるのに対して、本タスクは単語レベルの分類タスクである。本実験では、トークナイザによって複数のサブワードに分割された語については、その語を構成する先頭のサブワードに対応する埋め込みベクトルをプローブへの入力とした。プローブとしては 1 層ニューラルネットワーク、決定木、ロジスティック回帰の三つを採用した。データは 学習データ: テストデータ = 8 : 2 に分割し、学習データを用いた 5 分割交差検証によるコントロール実験を実施した。学習データ、テストデータそれぞれの分類ラベルの分布は付録中の表 10 の通りである。また、1 層ニューラルネットワークの学習における各種ハイパーパラメータは付録中の表 6 の通りに設定し、決定木とロジスティック回帰の各種ハイパーパラメータは scikit-learn 1.3.2 の `DecisionTreeClassifier` クラスおよび `LogisticRegression` クラスのデフォルト値を採用した。

後述の通り、最も selectivity が高かったのは 1 層ニューラルネットワークであった。このプローブについては、表 6 の設定下で学習データ全体を用いた学習を行い、テストデータに対する推論結果を確認した。

2.4 結果

コントロール実験の結果と、学習データ全体で学習した 1 層ニューラルネットワークのテストデータに対する推論結果について以下に述べる。

プローブ	FCNN	DT	LR
元タスク	98.3(0.003)	81.4(0.015)	98.7(0.003)
コントロールタスク	95.6(0.005)	79.3(0.001)	97.0(0.004)
Selectivity	**2.7**	2.1	1.7

表 1　コントロール実験の結果。1 層ニューラルネットワークを FCNN、決定木を DT、ロジスティック回帰を LR とそれぞれ表記する。5 分割交差検証により各分割での正解率を算出し、平均値 (標準偏差) の形式で掲載している。平均値と標準偏差はそれぞれ、小数第 2 位以下、小数第 4 位以下を四捨五入している。

	RE	AR	RO
RE	1137	18	0
AR	20	659	0
RO	3	0	96

表 2　学習データ全体で学習した 1 層ニューラルネットワークの、テストデータに対する混同行列。縦軸が正解ラベル、横軸が予測ラベルに対応する。

コントロール実験　コントロール実験の結果を表 1 に示す。調査した範囲で最も高い selectivity を示したのは 1 層ニューラルネットワークであった。2.1 節で述べた直観を受け入れるならば、BERT の埋め込みベクトル内で表現される節選択性に関する情報に対して、より敏感なのは 1 層ニューラルネットワークであるといえる。

テストデータに対する推論　学習データ全体を用いて学習した 1 層ニューラルネットワークの、テストデータに対する推論結果を混同行列として表に示す。表の示す通り、正解ラベルが RE である事例に対して 18 件、AR である事例に対して 20 件、RO である事例に対して 3 件の予測誤りが確認された。そして、正解ラベルが RE である事例 18 件中 6 件が語タイプ doubt に対する予測誤り、正解ラベルが AR である事例 20 件中 6 件が語タイプ admit に対する予測誤り、正解ラベルが RO である事例 3 件中 3 件が語タイプ ask に対する予測誤りであり、いずれも特定の述語に誤りが偏る傾向が見られた（予測誤りの分布について詳細は付録中の表 7-9 を参照のこと）。今後は、これらの予測誤りの要因や、予測を誤った事例に共通する性質について調査する必要がある。

3　実験 2: 意味タグ付与タスク

本実験では、[10] で提案された意味タグ付与 (semantic tagging) タスクをベースとし、節埋め込み述語の節選択性への理解を問う系列ラベリングタスクの設計とモデルの学習を実施した。後述するように、このタスクは 2.1 節の 3 値分類タスクと比較して、節埋め込み述語に対するより精緻な理解を要する。以下では、本実験のタスク定義を行い、そして使用したデータセットとモデルの設定について述べる。

タグ全体	新規タグのみ
91.3(0.002)	60.7(0.145)

表 3　5 分割交差検証の結果。5 分割交差検証により各分割での正解率を算出し、平均値 (標準偏差) の形式で掲載している。平均値と標準偏差はそれぞれ、小数第 2 位以下、小数第 4 位以下を四捨五入している。

3.1　意味タグ付与タスク

意味タグ付与タスクは、文中の各語に対して意味タグ (semantic tag) を付与する系列ラベリングタスクである。意味タグは各語の意味論的特徴を反映することが目指されている点で、品詞ラベル、固有表現ラベルといった他の系列ラベルと異なる。2.1 節でも使用した PMB4 に収録されている各文には意味タグ系列がアノテーションされている。以下はその例である。

(2)　$\mathrm{I}_{\mathrm{PRO}}\ \mathrm{know}_{\mathrm{ENS}}\ \mathrm{that}_{\mathrm{SUB}}\ \mathrm{she}_{\mathrm{PRO}}\ \mathrm{is}_{\mathrm{NOW}}\ \mathrm{beautiful}_{\mathrm{IST}}\ .\mathrm{NIL}$

[10] では 73 種類の意味タグを含むタグセットが提案されており、その中で一般動詞に割り当てられるタグは以下の 5 種類である。

EXS　時制なし単純形
ENS　現在形
EPS　過去形
EXG　進行形
EXT　完了形

3.2　データセットの作成

我々は前小節で挙げた 5 つのタグの末尾に RE, AR, RO のそれぞれを付記した 15 種類のタグをタグセットに追加した。そして、PMB4 のゴールドデータに含まれる 170 の節埋め込み文に出現する節埋め込み述語に付与されている意味タグを、その述語の節選択性に応じて変更した。[1)] 上記の変更を受けたデータセットに含まれる文の例を以下に示す。

(3)　a.　$\mathrm{I}_{\mathrm{PRO}}\ \mathrm{know}_{\mathrm{ENS\text{-}RE}}\ \mathrm{that}_{\mathrm{SUB}}\ \mathrm{she}_{\mathrm{PRO}}\ \mathrm{is}_{\mathrm{NOW}}\ \mathrm{beautiful}_{\mathrm{IST}}$

1)　本実験では、意味タグのアノテーション誤りによる影響の大きさを考慮して、ゴールドデータのみを使用した。

	ENS-AR	ENS-RE	EPS-AR	EPS-RE	EPS-RO	EXS-AR	EXS-RE
ENS-AR	6	0	0	0	0	0	0
ENS-RE	0	4	0	0	0	0	0
EPS-AR	0	1	1	1	0	0	0
EPS-RE	0	2	1	3	0	0	0
EPS-RO	0	0	0	0	2	0	0
EXS-AR	0	0	0	0	0	1	0
EXS-RE	0	1	0	0	0	0	0

表 4　学習データ全体で学習した 1 層ニューラルネットワークの、テストデータに対する混同行列。縦軸が正解ラベル、横軸が予測ラベルに対応する。テストデータに出現しないラベルは省略している。

$_{\cdot \mathrm{NIL}}$

b. $\mathrm{He}_{\mathrm{PRO}}$ $\mathrm{knows}_{\mathrm{ENS}}$ $\mathrm{none}_{\mathrm{NOT}}$ $\mathrm{of}_{\mathrm{REL}}$ $\mathrm{us}_{\mathrm{PRO}}$ $_{\cdot \mathrm{NIL}}$

このように、意味タグが変更されたのは実際に節を取る述語のみであり、(3-b) のような文脈に出現する述語は意味タグの変更を受けていない。つまり、**変更後のデータセットによって定義される意味タグ付与タスクは、節選択性に応じた節埋め込み述語の分類だけでなく、その述語が実際に節埋め込み文脈に出現するかどうかの峻別も要求する**。この意味で、本タスクは 2.1 節の 3 値分類タスクと比較して難易度の高いタスクであるといえる。

3.3　学習設定

本実験では、事前学習済み BERT の出力する文脈を考慮した埋め込みベクトル（$\in \mathbb{R}^{768}$）を入力とし、その語の意味タグを予測するプローブの学習を行なった。プローブとしては 1 層ニューラルネットワークを採用した。2.1 節と同じく、BERT のパラメータは `bert-base-uncased` のものを採用した。意味タグ付与も単語レベルで予測を行うタスクであるから、複数のサブワードに分割された語については、その語を構成する先頭のサブワードに対応する埋め込みベクトルをプローブへの入力とした。また、学習は 2.1 節と同様に 5 分割交差検証と、学習データ全体を用いた学習の 2 通りを実施した。学習データ、テストデータそれぞれの新規意味タグの分布を付録中の表 11 に示す。また、1 層ニューラルネットワークの学習における各種ハイパーパラメータは 2.1 節と同じく表 6 の通りに設定した。

3.4　結果

交差検証の結果と、学習データ全体で学習した 1 層ニューラルネットワークのテストデータに対する推論結果について以下に述べる。

交差検証　交差検証の結果を表 3 に示す。表には、正解ラベルが新規タグである語に限定して算出した正解率も併記している。新規タグのみに限定した場合の平均正解率は、タグ全体での平均正解率よりも 30 ポイント以上低く、分割ごとのばらつきも比較的大きい結果となった。この結果につながった要因は、再アノテーションを受けたサンプル数の不足をはじめ様々なものが考えうるが、要因の解明は今後の課題である。

テストデータに対する推論　学習データ全体を用いて学習した 1 層ニューラルネットワークの、テストデータに対する推論結果を混同行列として表 4 に示す。予測誤りの傾向や要因の調査が今後の課題である。

4　結論

本稿では節選択性の観点から設計した二つのタスクについて、事前学習済み BERT モデルの出力する埋め込みベクトルを入力とするプローブの学習・評価を行なった。実験 1 で調査した限りでは、1 層ニューラルネットワークが selectivity の観点から最良であることが明らかになったが、プローブが予測に用いている情報が具体的にどのようなものかは未だ明らかではない。実験 2 にて構築したプローブは一定の傾向を持って予測を行うことは明らかになったものの、詳細なエラー分析は未着手である。いずれの結果についても、今後はよりミクロな視点での分析が必要であると言える。また、本研究では一貫して BERT の出力層をプローブへの入力として調査したが、節埋め込みの意味論と強く関連する情報が複数ある層のどこで表現されているかは自明ではなく、今後検討すべき問いである。

謝辞

本研究は JST CREST、JP-MJCR2114 の支援を受けたものです。

参考文献

[1] Utpal Lahiri, et al. **Questions and answers in embedded contexts**, Vol. 2. Oxford University Press on Demand, 2002.

[2] Jane Grimshaw. Complement selection and the lexicon. **Linguistic inquiry**, Vol. 10, No. 2, pp. 279–326, 1979.

[3] Aaron Steven White and Kyle Rawlins. The role of veridicality and factivity in clause selection. In **Proceedings of the 48th annual meeting of the north east linguistic society**, Vol. 3, pp. 221–234, 2018.

[4] Lauri Karttunen. Syntax and semantics of questions. **Linguistics and philosophy**, Vol. 1, No. 1, pp. 3–44, 1977.

[5] Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. **Inquisitive Semantics**. Oxford University Press, 2019.

[6] Wataru Uegaki. The semantics of question-embedding predicates. **Language and Linguistics Compass**, Vol. 13, No. 1, p. e12308, 2019.

[7] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 842–866, 2021.

[8] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. **arXiv preprint arXiv:1909.03368**, 2019.

[9] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 242–247. Association for Computational Linguistics, 2017.

[10] Lasha Abzianidze and Johan Bos. Towards universal semantic tagging. In **Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers**, pp. 1–6, Montpellier, France, 2017.

付録

分類ラベル	語タイプ
RE	admit, allow, assure, bet, claim, forget, observe, resent, realize, reveal, see, show, suggest, wish
AR	adjudge, believe, cable, confirm, convince, expect, guess, hear, hope, pretend, promise, suspect, whisper, wonder
RO	ask, assume, confess, decide, doubt, dream, feel, find, joke, know, notice, regret, reply, say, sense, tell, think, watch

表 5 2.2 節補足資料: コントロールタスク用データセットにおける語タイプと分類ラベルの対応

ハイパーパラメータ	値
バッチサイズ	8
学習率	1×10^{-3}
エポック数	30
最大系列長	128
重み減衰	0.01
乱数シード	0

表 6 2.3 節補足資料: 1 層ニューラルネットワークのハイパーパラメータの設定。ここに記載されている値以外は Transformers 4.36.2 の `Trainer` クラスのデフォルト値を採用した。

語タイプ	誤りに占める割合
doubt	0.33 (6/18)
realize	0.22 (4/18)
find, say, suggest	0.11 (2/18)
know, tell	0.05 (1/18)

表 7 2.4 節補足資料: 正解ラベルが RE である事例に対する誤りの分布

語タイプ	誤りに占める割合
admit	0.30 (6/20)
assure	0.15 (3/20)
assume, promise	0.10 (2/20)
claim, convince, feel, joke, reply, suspect, think	0.05 (1/20)

表 8 2.4 節補足資料: 正解ラベルが AR である事例に対する誤りの分布

語タイプ	誤りに占める割合
ask	1 (3/3)

表 9 2.4 節補足資料: 正解ラベルが RO である事例に対する誤りの分布

分類ラベル	学習データ内の出現回数	テストデータ内の出現回数
RE	4715	1155
AR	2663	679
RO	371	99

表 10 2.3 節補足資料: 節タイプ分類データセット内のタグ比率

意味タグ	学習データ内の出現回数	テストデータ内の出現回数
EXSRE	11	1
ENSRE	21	6
EPSRE	23	8
EXGRE	0	0
EXTRE	1	0
EXSAR	3	1
ENSAR	22	7
EPSAR	18	4
EXGAR	1	0
EXTAR	0	0
EXSRO	0	0
ENSRO	3	0
EPSRO	6	2
EXGRO	0	0
EXTRO	0	0

表 11 3.3 節補足資料: 意味タグデータセット内のタグ比率（節選択に関連するもののみを抜粋）