入れ子型固有表現に対する変分情報ボトルネック法の適用

原田宥都　渡辺太郎
奈良先端科学技術大学院大学
{harada.yuto.hq4,taro}@is.naist.jp

概要

大規模に事前学習された言語モデルは豊富な意味情報や構文情報を学習しているが，個々のタスクでそれらの全ての情報が必要であるとは限らない．本研究では，言語モデルの持つ情報のうち，特定のタスクで重要な情報を取り出す手法として「変分情報ボトルネック法」を用い，これを特に「入れ子型固有表現認識タスク」に適用するための手法を提案する．実験を通して，提案モデルでは入力特徴の次元を削減することで予測精度が向上することを確認した．また分析の結果，入れ子レベルの深さに応じて，固有表現の予測に重要な情報が異なることを示した．

1　はじめに

ELMo [1] や BERT [2] など，巨大な言語資源を用いて訓練された大規模な言語モデルは，特定のタスクに依存しない方法で学習が行われているが，それらは言語に関する汎用的な表現を獲得しているため，近年では様々なタスクにおける最先端の性能に貢献している．

汎用的な埋め込み表現は様々なタスクに役立つ一方で，次元数の大きいものが多く，特定のタスクにおいてはそのような大きな次元数は必要がない可能性もある．タスクを解く上で重要な情報のみを取り出し，適切な次元数へ圧縮することができれば，モデルが入力特徴にフィットしやすくなり，性能が向上することが期待される．また，残された情報を分析することで，そのタスクにおいてどのような情報が重要なのか，という知見を得ることができ，更なる性能向上に役立てることができる．

本研究では，自然言語処理タスクの一つである入れ子型固有表現認識を対象として，事前学習済み言語モデルに含まれる情報についての調査を行う．

入れ子型固有表現とは，一つの固有表現が，内部に複数の固有表現を含んでいるものを指す．図 1 に示した例のように，それぞれの固有表現が階層的な構造となっているため，より内側の入れ子レベルの固有表現はスパンが短く，より外側の入れ子レベルの固有表現はスパンが長いという特徴を持つ．本研究では，異なる入れ子レベルにおいて予測に必要な情報はそれぞれ異なるのではないか，という仮定を置き，各入れ子のレベルにおいて，予測に必要な情報を単語埋め込みから保存する手法を提案する．提案手法では，単語埋め込みから特定の情報のみを保存するために，変分情報ボトルネック法 [3] を用いる．

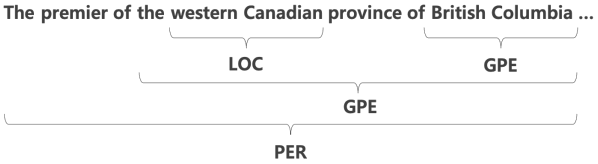


図 1　３層に入れ子になった固有表現の例

変分情報ボトルネック法とは，情報ボトルネックの原理に基づくアイデアであり，ニューラルネットワークの入力表現を，予測精度を保ちながら次元圧縮することのできる手法である．

本研究では，この変分情報ボトルネック法をエンコーダとして利用し，入れ子型固有表現認識に適用した．３種類の提案モデルについて実験による性能評価を行い，2 つの提案モデルがベースラインの性能を上回った．また学習したエンコーダの分析を通して，各入れ子レベルにおいて予測に重要な情報がどのように異なるかを可視化することができた．

2　関連研究

2.1　入れ子型固有表現認識

入れ子型固有表現認識（Nested Named Entity Recognition, Nested-NER）は，情報抽出のタスクの一つであり，[4] によって提案され，近年盛んに研究されている．例えば既存のアプローチとして，入れ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

子構造を明示的に捉えるようなモデル設計をした特徴ベースのアプローチが存在している [5, 6, 7, 8]．また，レイヤードモデルを用いたアプローチ [9, 10] も有効であることが分かっている．これらの研究では，入れ子を扱わない一般的な NER モデルを積み重ねることで，段階的に大きな固有表現を抽出することにより，入れ子になった固有表現を扱うことができる．[11] ではさらにレイヤードモデルにおける誤差の伝搬などの問題にも対処し，最先端の性能を達成している．

2.2 情報ボトルネック法

情報ボトルネック法（Information Bottleneck）とは，情報理論の分野に起源を持つ手法であり，近年では機械学習において利用されている．ある信号 $x \in X$ から別の信号 $y \in Y$ を予測する際に，予測に無関係である冗長な情報を取り除いた X の圧縮表現 T を見つけることが目的であり，次のように記述できる．

$$\mathcal{L}_{IB} = \beta I(X,T) - I(T,Y) \tag{1}$$

$I(-,-)$ は相互情報量を示している．これは，$I(X,T)$ をなるべく最小化しながら $I(Y,T)$ を最大化すること，つまり，圧縮表現 T が X についての情報を最大限破棄しつつも，T が Y を予測するために必要な情報は最大限保持するということを意味している．

2.3 変分情報ボトルネック法

[3] において (1) の効率的な変分推定が提案されており，以下のように導出されている．

$$\mathcal{L}_{VIB} = \beta \mathop{\mathbb{E}}_{x}[\mathrm{KL}[p_\theta(t|x), r(t)]] + \mathop{\mathbb{E}}_{t\sim p_\theta(t|x)}[-\log q_\phi(y|t)] \tag{2}$$

ここで $r(t)$ は $p(t)$ の変分近似、$q_\phi(y|t)$ は $p(y|t)$ の変分近似である。$p_\theta(t|x)$ と $q_\phi(y|t)$ は独自のパラメータセットを持つニューラルネット，つまりエンコーダとデコーダであることを意味している．

2.4 情報ボトルネック法の自然言語処理への応用例

情報ボトルネック法は，近年，自然言語処理の分野においても様々な問題解決やモデルの分析の手法として利用されている．構文解析 [12]， 要約 [13]，低資源の状況における言語モデルの正則化 [14]，Transformer モデルの各層の間でトークンの表現がどのように変化しているかの分析 [15]，といった事例がある．特に，変分情報ボトルネック法を構文解析タスクに適用した [12] は，特定のタスクのために単語埋め込み表現を専門化させるという汎用性の高い手法であり，本研究ではこれを参照し，変分情報ボトルネック法を入れ子型固有表現認識タスクに適用する．

3 提案手法

3.1 ベースライン

我々の入れ子型固有表現認識モデルは，基本的なニューラル・レイヤード・アーキテクチャである Layered-BiLSTM-CRF モデル [10] を参考にしている．図 2 に概要を示したが，それぞれの入れ子レベルに対応して系列ラベリングモデルを学習させ，それらをパイプライン的に積み重ねることで，入れ子型の固有表現を認識する．

ベースラインモデルにおいては，文中の単語は BERT によって単語埋め込み表現へマッピングされ，１層目の入力として用いられる．２層目以降においては，下の層の BiLSTM の出力を受け取り，入力として用いる．

3.2 変分情報ボトルネック法の適用

本研究では，テキスト x が，事前学習済み言語モデルによって得られた M 次元のベクトル表現として入力されるものとする ($x \in \mathbb{R}^M$)．また，圧縮表現 t も N 次元のベクトルであるとする ($t \in \mathbb{R}^N$)．x から圧縮表現 t へマッピングする VIB エンコーダ $p_\theta(t|x)$ は，[3] と同じく，平均 $\mu(x,\theta)$ 分散 $\sigma^2(x,\theta)$ を持つ N 次元の正規分布で与えられるとし，デコーダ $q_\phi(y|t)$ は softmax 関数を通して得られるカテゴリカル分布として，$f(t,\phi)$ とおく．$r(t)$ はデフォルトで正規分布 $\mathcal{N}(0,1)$ とする．この時，式 2 の右辺の第一項は $\mathcal{N}(\mu,\sigma^2)$ と $\mathcal{N}(0,1)$ との KL ダイバージェンスなので

$$\beta \mathrm{KL}[p,r] = -\frac{\beta}{2}(|\mu|^2 + N(\sigma^2 - log\sigma^2 - 1)) \tag{3}$$

となり，式 2 の右辺の第二項は，x, y をデータから，t を再パラメータ化トリック [16] によってサンプルすることで，

$$CrossEnt(f(\mu(x,\theta) + \sigma(x,\theta)\epsilon, \phi), y) \tag{4}$$

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

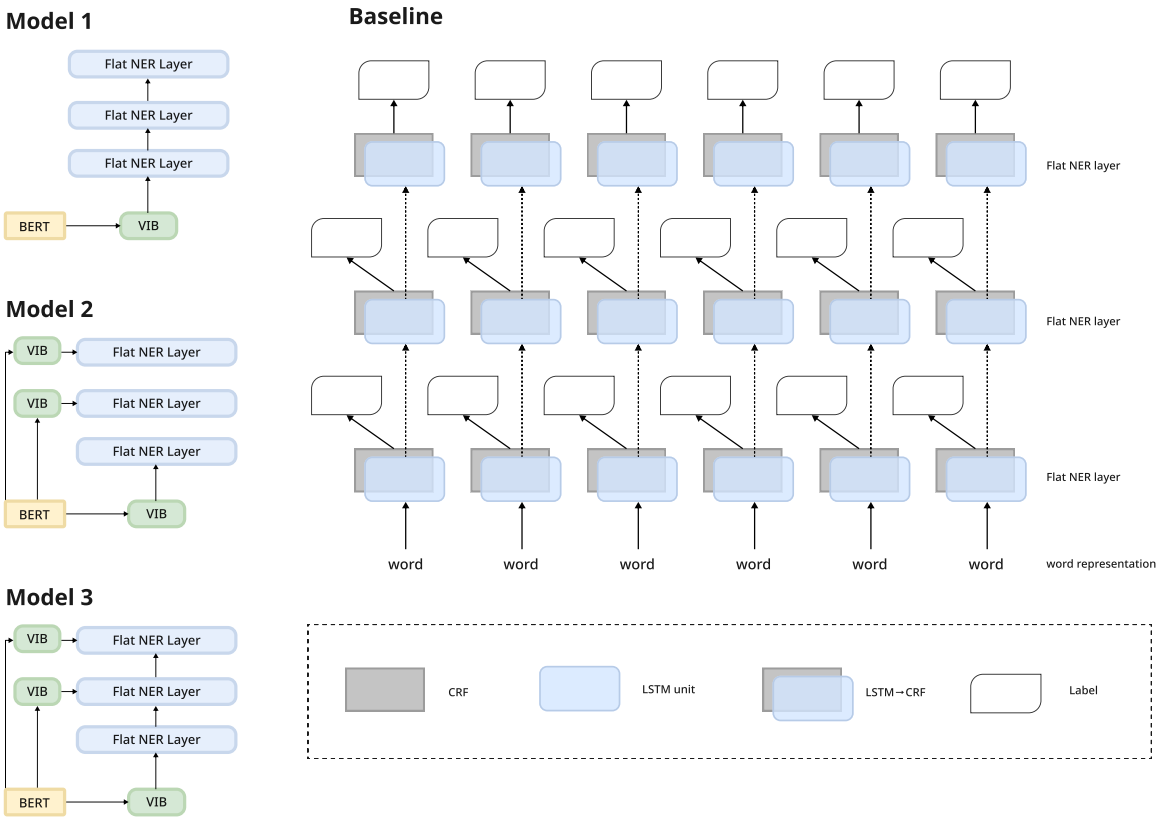


図 2　提案モデルのアーキテクチャの概要

となる．ここで $\epsilon \sim \mathcal{N}(0,1)$ である．

3.3　提案モデル

モデル１：低層モデル　低層モデルでは，VIB エンコーダによる埋め込み表現 r_1 を１層目にのみ入力するため，１層目の BiLSTM の出力 h_1 と２層目以降の出力 h_i は

$$\begin{aligned} h_1 &= BiLSTM(r_1) \\ h_i &= BiLSTM(h_{i-1}) \end{aligned} \tag{5}$$

となる．この場合，VIB エンコーダは，１層目の固有表現を予測するために必要な情報をなるべく残すように学習する．そのため，上の層になるほど VIB エンコーダの影響は薄れる．

モデル２：独立モデル　独立モデルでは，全ての層で VIB エンコーダによる埋め込み表現のみを用いるため，層 i における BiLSTM の出力は

$$h_i = BiLSTM(r_i) \tag{6}$$

となる．この場合，各層で異なる VIB エンコーダを学習するため，VIB エンコーダはその層において必要な情報のみを保存するように学習する．

モデル３：独立＋伝搬モデル　独立＋伝搬モデルでは，VIB エンコーダによる圧縮表現を全ての層に入力し，２層目以降は下の層 BiLSTM の出力を用いるため，１層目の BiLSTM の出力 h_1 と２層目以降の出力 h_i は

$$\begin{aligned} h_1 &= BiLSTM(r_1) \\ h_i &= BiLSTM(h_{i-1}, r_i) \end{aligned} \tag{7}$$

となる．この場合，２層目以降の VIB エンコーダは，下の層から伝搬してくる情報と，本当にその層で必要な情報の差分を保存するように学習する．

4　実験

4.1　実験設定

データセット　提案モデルを ACE2005 コーパス [17] 上で評価した．

入れ子型固有表現認識のデータセットとして用いられるコーパスとしては他にも GENIA コーパスなどが知られるが，深い入れ子レベルを持つ固有表現が多く存在しているため，ACE コーパスを採用している．また，学習用データとして用いるには数が少ないため，入れ子レベル 4 以上の固有表現は今回は

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

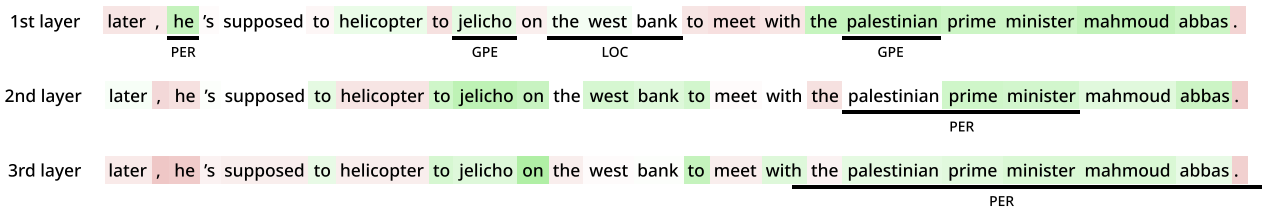


図 3　Integrated Gradients による VIB エンコーダにおける単語重要度の可視化の例

	1st layer	2nd layer	3rd layer
Baseline	76.38	59.46	41.41
Model 1	**79.22**	63.00	43.79
Model 2	76.08	52.60	20.61
Model 3	78.96	**66.85**	**49.93**

表 1　提案モデルの性能比較（F1-score）

用いなかった．コーパスの統計についての詳細は，表 A（付録）に示した．

パラメータ設定　BiLSTM-CRF のハイパーパラメータとして，バッチサイズを 100，隠れユニットの数を 200，学習率を 0.0045 とした．また VIB エンコーダについて，圧縮比率 β を 0.00001，学習率を 0.0001 とした．圧縮次元数 d は 384, 192, 96, 12 を試した．

4.2　結果と考察

各提案モデルの F1 スコアによる性能の比較を表 4.1 に示す．提案モデルは全て 192 次元に圧縮した場合の結果である．それぞれのベースラインと比較すると，モデル１（低層モデル）とモデル３（独立＋伝搬モデル）が，いずれもベースラインの性能を上回っている．特にモデル１において性能が改善されているのは，大きな次元数を持つ入力特徴適切な次元数へ圧縮することが，性能の改善に有用であるからだと考えられる．モデル２（独立モデル）は他のモデルに性能が劣るが，下の層から受け渡される情報がなくともある程度の精度は維持できるということがわかった．レイヤー２，レイヤー３においては，いずれもモデル３のスコアが最も高く，ベースラインを大幅に改善している．これは，それぞれの層において，下の層からの情報だけでは不足している部分を VIB エンコーダによる入力が補った影響であると考えられる．

Integrated Gradients[18] を用いて，モデル２で学習済みの VIB エンコーダの挙動を可視化した結果の一例を図 3 に示している．それぞれの VIB エンコーダは各層のラベルを予測するために必要な情報のみを残すように学習しており，テキストへのハイライトは入力単語の重要度を算出したものである．モデル１では特に名詞一単語，またそれを導くような冠詞に強く着目すること，モデル２では長めのスパンを持つ名詞句や，それらを導くような単語に強く着目すること，モデル３ではモデル２の特徴に加えて，助動詞や動詞などの文構造に関わる情報を捉え，それらの重要度を低く付けること，といったそれぞれの傾向があることを確認した．

また，Conductance[19] を計算することにより，VIB エンコーダが BERT の情報から破棄した部分と残した部分をヒートマップとして可視化した例を，図 B(付録) として示している．それぞれの列は約 50 例の単語を，行は VIB エンコーダに BERT 出力層が入力された直後の 768 のニューロンにおける特徴の重要度を表している．それぞれのレイヤーにおいてのみ着目される特徴がある，とする仮定は，このようなヒートマップを通して定量的に評価することが可能であり，これについては現在調査中である．

5　おわりに

本研究では，変分情報ボトルネック法を入れ子型固有表現認識に適用するための手法を提案した．提案モデルは，大規模な言語モデルによる埋め込み表現を適切に次元圧縮することで，ベースラインの性能を改善することができた．また，残された情報を Integrated Gradients や Conductance といった手法よって分析・可視化することで，タスクにとって必要な情報はどのようなものであるかについての知見を得ることができ，解釈性の向上に寄与した．

謝辞

本研究は JSPS 科研費 JP20K23325 の助成を受けたものである．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 2227–2237. Association for Computational Linguistics, June 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[3] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In **Proceedings of the International Conference on Learning Representations**, 2016.

[4] Jenny Rose Finkel and Christopher D. Manning. Nested named entity recognition. In **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing**, pp. 141–150. Association for Computational Linguistics, August 2009.

[5] Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. A neural transition-based model for nested mention recognition. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1011–1017. Association for Computational Linguistics, October-November 2018.

[6] Wei Lu and Dan Roth. Joint mention extraction and classification with mention hypergraphs. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 857–867. Association for Computational Linguistics, September 2015.

[7] Aldrian Obaja Muis and Wei Lu. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2608–2618. Association for Computational Linguistics, September 2017.

[8] Bailin Wang and Wei Lu. Neural segmental hypergraphs for overlapping mention recognition. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 204–214. Association for Computational Linguistics, October-November 2018.

[9] Beatrice Alex, Barry Haddow, and Claire Grover. Recognising nested named entities in biomedical text. In **Biological, translational, and clinical language processing**, pp. 65–72. Association for Computational Linguistics, June 2007.

[10] Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. A neural layered model for nested named entity recognition. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1446–1459. Association for Computational Linguistics, 2018.

[11] Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. Pyramid: A layered model for nested named entity recognition. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5918–5928. Association for Computational Linguistics, 2020.

[12] Xiang Lisa Li and Jason Eisner. Specializing word embeddings (for parsing) by information bottleneck. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2744–2754, Hong Kong, China, 2019. Association for Computational Linguistics.

[13] Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3752–3761. Association for Computational Linguistics, 2019.

[14] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. Variational information bottleneck for effective low-resource fine-tuning. In **International Conference on Learning Representations**, 2021.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[16] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In **2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings**, 2014.

[17] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. Ace 2005 multilingual training corpus. In **Linguistic Data Consortium, Philadelphia 57.**, 2006.

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research**, pp. 3319–3328, 2017.

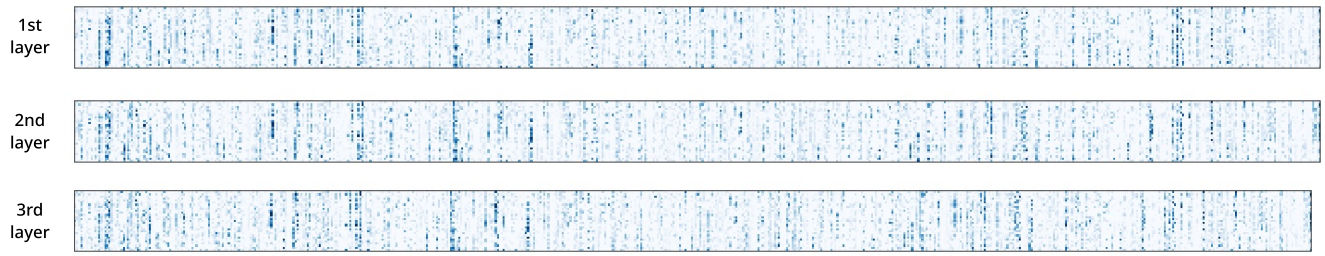
[19] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron? In **ICML**, 2019.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　付録：ACE2005 コーパスの統計的詳細

Item	Train	Dev.	Test
Documents	370	43	51
Sentences	9849	1221	1478
FAC	924	83	173
GPE	4725	486	671
LOC	763	81	69
ORG	3702	479	559
PER	13050	1668	1949
VEH	624	81	66
WEA	652	94	67
Outermost entity	18455	2285	2724
Nested level	6	4	5
Entities in level 1	19676	2429	2936
Entities in level 2	3934	448	505
Entities in level 3	731	85	102
Entities in level 4	90	10	10
Entities in level 5	7	0	1
Entities in level 6	2	0	0
Entity avg. length	2.28	2.33	2.28
Multi-token entity	10577	1323	1486
Overall entities	24440	2972	3554

B　付録：VIB エンコーダにおけるニューロンの重要度の可視化



This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).