文字系列情報による性能への影響から ニューラルモデルが有する言語的な傾向を見出せるか

黒澤 友哉　谷中 瞳
東京大学
{kurosawa-tomoya, hyanaka}@is.s.u-tokyo.ac.jp

概要

本研究では，文字系列情報を付加的に利用するニューラルモデルにおいて，文字系列情報による性能への影響度を定義し，その値がどのような言語的な傾向を反映しているかを検証する．実験では 48 の言語を対象に，品詞タグ付けと依存構造解析の 2 つのタスクにおいて，文字系列情報による性能への影響度を調査する．文字の数が多いほど情報量も多いことから，仮説として「文字系列情報が与える性能への影響度は，その言語の平均単語長が長いほど大きい」と立て検証する．実験の結果，ラテン文字を用いる言語において仮説を支持する値が得られた一方で，検定の結果，依存構造解析においては仮説を示すことができなかった．

1　はじめに

文字は，記述された自然言語を構成する最小の要素である．我々人間は文字を読み，意味を理解する．ニューラルモデルは基本的に，文字の特定の並びを一つのトークンとして扱う．一部のモデルは，文を構成する文字についてもトークンと同様埋め込みを計算するなどして，文字の情報を利用するものがある．ここではその文字の情報を**文字系列情報**と呼び，文を構成する文字の列として定義する．文字系列情報を利用したモデルは，それを利用しなかった場合よりも高い性能が，様々な言語とタスクで報告されている [1, 2, 3]．しかし，文字系列情報による性能上昇は，それを利用した場合と利用しなかった場合のスコアの差として計算されており，モデルがトークン（単語）から得られる情報に比べ，文字系列情報をどの程度利用したかが明らかになっていない．そのため本研究では [4] にならい，文字系列情報を付加的に利用するニューラルモデルにおける文字系列情報の影響の度合いを調査する．

最近は大規模汎用言語モデルの発展が目覚ましく，そのようなモデルは様々な言語において高い性能を達成している．ただし，高性能を達成できるのは十分なデータが存在する言語のみに限られ，低リソース言語においてはモデルの構築が難しいことが知られている [5]．いくつかの先行研究では，言語類型論などに基づいた言語的な特徴を利用したり，ニューラルモデルから言語的な特徴や傾向を見出したりする試みがある [6, 7]．後者の試みは，人間が観測し得ない傾向を観測できることができ，例えば言語転移を取り入れるモデルの学習に利用できることから，近年注目されているアプローチである [5]．しかし，後者では主に言語間で共通な特徴に着目されており，言語間で異なる特徴や傾向を見出す試みが多くない．

そこで本研究では，文字系列情報を付加的に利用するニューラルモデルにおける文字系列情報が性能に与える影響の度合い（**影響度**）を複数のタスクで評価する．そして，その評価結果を観察し，言語間で異なる言語的な傾向を明らかにする．文字系列情報は文字の数が多いほど情報量も多いという点に着目し，仮説として**平均単語長が長い言語ほど，文字系列情報による性能への影響度が大きい**と立て，この仮説を実験で検証する．対象とするタスクは，基礎タスクに分類される品詞タグ付けと依存構造解析とする．これらは系列ラベリングタスクであり，文字系列情報の比重が単語より大きく，その影響が数値に反映されやすいことを利用するという狙いがある．

2　関連研究

2.1　文字系列情報の利用

文字系列情報を利用するニューラルネットワークモデルには二種類あり，文字系列情報のみを利用す

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

るモデルと，それを付加的に利用するモデルに大別される．前者のモデルは，RNN や双方向 LSTM [8] などを利用して文字埋め込みを得る [9, 10]．最近では，注意機構 [11] を用いて文字埋め込みを計算するモデルも提案されている [12]．

文字系列情報を付加的に利用するモデルには，主に特定のタスクを解くものが多い．文字系列情報を付加的に利用する動機の一つとして，学習時に出現しなかった語彙 (out of vocabulary) が扱いやすくなるという点がある [13]．文字系列情報のみを利用するモデルの機構と学習済みの文字埋め込みを使用することで，品詞タグ付け，依存構造解析，意味解析，翻訳タスクなど，様々なタスクで高い性能を達成している [1, 2, 3, 14]．

2.2 多言語モデルにおける言語的な特徴や傾向の扱い

複数の言語にわたり共通のタスクを解くようなモデルには，World Atlas of Language Structures (WALS, [15]) [1)] という言語類型論から得られた言語的な特徴を保有するデータベースなどを利用し，その情報を利用するものがある．[6] は WALS から得た主語，動詞，目的語の順序などの言語的な特徴を埋め込み情報の一部として与えている．近年では学習済み多言語モデルの開発が活発化したことで，それらから言語的な特徴や傾向を見出す研究も出てきた．[7] は学習済み多言語モデルが有する，言語間で共通の統語的現象を明らかにした．また，[16] はバイト対符号化 (BPE) の結果から言語特有の傾向を見出した．本研究は，ニューラルモデルを利用して言語的な傾向を見出す点で [16] に類似しているが，ニューラルモデルの性能から言語的な傾向を見出す点は [16] と異なったアプローチである．

3 手法

文字系列情報を付加的に用いるニューラルモデルにおいて，その文字系列情報がモデルの性能に与える影響度を次のように定義する．

(i) 学習データと正しい文字系列情報を用いてモデルを学習する．

(ii) (i) の学習済みモデルを，評価データと正しい文字系列情報を用いて評価する．得られた値を x とおく．

(iii) (i) の学習済みモデルを，評価データと，(i) の学習時に出現しなかった文字のみから構成される文字系列情報を用いて評価する．得られた値を y とおく．

(iv) (ii) と (iii) の結果を比較し，文字系列情報の**影響度**を，正しい文字系列情報を与えて評価したときの性能と，正しい文字列と学習時に出現しなかった文字のみから構成される文字系列情報を与えて評価したときの性能差の比として定義する．この値は $\frac{x-y}{x}$ と計算される．

学習済みモデルの評価時に，学習時に出現しなかった文字のみから構成される文字系列情報を与えることで，そのモデルは学習時に観測した正しい文字系列情報を利用してラベルを予測することができなくなる．そのため影響度 $\frac{x-y}{x}$ は，その学習済みモデルにおける正しい文字系列情報の貢献の割合とみなすことができる．

1) https://wals.info/

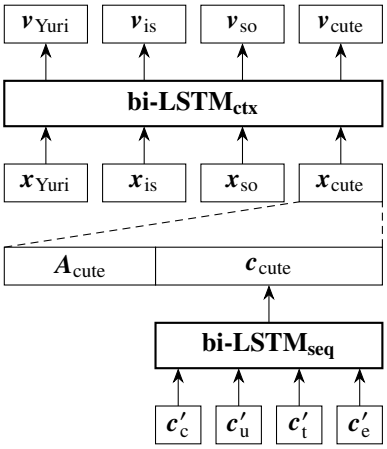


図 1 モデルの構造．A_{cute} は文字単語ベクトル以外のものを表しており，品詞タグ付けにおいては単語埋め込み w_{cute}，依存構造解析においては単語埋め込み w_{cute} と事前学習済みの単語埋め込み p_{cute} である．

4 実験

4.1 モデル

本実験では，品詞タグ付けと依存構造解析のそれぞれにおいてほぼ同等の構造を有するニューラルモデルを用いる．品詞タグ付けにおいては [1] のモデルを，依存構造解析においては [2] のモデルをそれぞれ用いる．各モデルの共通する構造を図 1 に示す．いずれのモデルにおいても，単語で使用される各文字 i の埋め込み c'_i を入力とし，それぞれの方向で最後の隠れ層における表現を利用する LSTM である sequence bi-LSTM [8] により文字から得られる単語ベクトル（以下，文字単語ベクトル）c_j を各単語

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

表 1 使用する言語のコード，データ数と平均単語長.

言語	データ数	平均単語長	言語	データ数	平均単語長	言語	データ数	平均単語長
ar	7,664	4.30	fr	16,447	4.65	nl	13,735	5.17
be	333	6.01	ga	1,020	4.52	no	20,045	4.92
bg	11,138	5.23	gl	4,003	4.86	pl	8,227	5.86
ca	16,678	4.70	got	5,400	5.17	pt	9,368	4.40
cop	320	3.70	grc	13,919	5.06	ro	9,524	5.23
cs	87,913	5.49	he	6,216	3.80	ru	5,030	5.98
cu	6,337	4.48	hi	16,647	3.88	sa	190	6.24
da	5,512	4.96	hr	8,889	5.54	sk	10,604	5.26
de	15,894	5.76	hu	1,800	6.27	sl	8,000	5.19
el	2,521	5.55	id	5,593	5.90	sv	6,026	5.45
en	16,622	4.47	it	13,884	4.71	ta	600	7.65
es	16,013	4.73	ja	8,232	1.75	tr	5,635	6.21
et	4,091	5.73	ko	6,339	3.26	uk	1,706	5.25
eu	8,993	6.51	la	2,273	5.70	ur	5,130	3.61
fa	5,997	3.93	lt	263	6.02	vi	3,000	4.37
fi	15,136	7.38	lv	3,972	5.68	zh	4,997	1.66

j について得る．数式では以下のように表される．

$$\boldsymbol{c}_j = \text{bi-LSTM}_{\text{seq}}(\boldsymbol{c}'_{1:m}) = \text{LSTM}_{\text{f}}(\boldsymbol{c}'_{1:m}) \circ \text{LSTM}_{\text{b}}(\boldsymbol{c}'_{m:1})$$

なお，$\text{LSTM}_{\text{f}}, \text{LSTM}_{\text{b}}$ はそれぞれ順方向，逆方向の LSTM で，引数の最後のベクトルを入力する個の LSTM の隠れ層における表現を指す．品詞タグ付けモデルにおいては，文字単語ベクトルは単語埋め込み $\boldsymbol{w}_j$ と結合される．一方，依存構造解析モデルにおいては，文字単語ベクトルは学習可能な単語埋め込み $\boldsymbol{w}_j$ に加え，事前学習済みの単語埋め込み $\boldsymbol{p}_j$ と結合される．そして，結合された単語ベクトル $\boldsymbol{x}_j$ を入力として，それぞれの方向で特定の位置の隠れ層における表現を利用する LSTM である context bi-LSTM により各単語の文脈ベクトル $\boldsymbol{v}_j$ を得る．数式では以下のように表される．

$$\boldsymbol{v}_j = \text{bi-LSTM}_{\text{ctx}}(\boldsymbol{x}_{1:n}, j) = \text{LSTM}_{\text{f}}(\boldsymbol{x}_{1:j}) \circ \text{LSTM}_{\text{b}}(\boldsymbol{x}_{n:j})$$

品詞タグ付けモデルは，交差エントロピー誤差を用いた確率的勾配降下法により文脈ベクトルから品詞を予測する．一方，依存構造解析モデルは，遷移ベース解析 (transition-based parsing, [17]) の一種である arc-hybrid 型 [18] 解析により，文脈ベクトルから依存構造を特定する．品詞タグ付けの性能は精度 (accuracy)，依存構造解析の性能は LAS (label attachment score) で計算され，それぞれ 3 回の試行の平均値を示す．なお LAS は，予測された依存関係を表す弧の両端とラベルがすべて正しいものの割合を表す．

4.2 データ

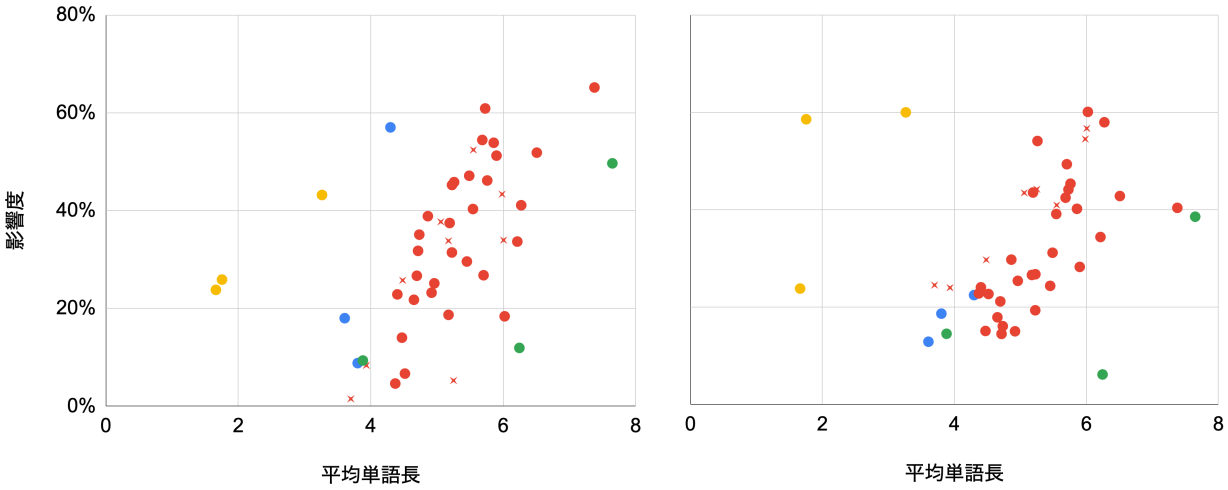
本実験では [2] にならい，大規模コーパス Universal Dependencies (UD, [19]) [2] のうち v2.0 を使用する．対象言語は，[2] で使用された 47 の言語に加え，ノルウェー語の Bokmål (no_bokmaal) の計 48 言語とする．各言語の平均単語長は UD v2.0 に含まれるすべてのデータ（学習，検証，評価データ）から計算される．表 1 に使用する言語の ISO 639 における言語コード，およびデータ数と平均単語長を示す．言語名，各言語の文字体系とデータ数については付録（表 2）に示す．なお，手法の (iii) で言及した具体的な「学習時に出現しなかった文字」は，対象のタスクと言語でそれに当てはまる割合が最も高かった，ASCII コード 92 で表される逆斜線 \ を基本とする．この文字が学習時に出現している場合，すなわち ar, hr, id, ru での依存構造解析では，ASCII コード 125 で表される終わり波括弧 } とする．そして「学習時に出現しなかった文字のみから構成される文字系列情報」は，入力文の空白以外の文字をすべて上記の文字に置き換えたものとする．例えば en（英語）において，*Yuri is so cute* の (iii) の評価時の文字系列情報は \\\\␣\\␣\\␣\\\\ となる．

4.3 結果

図 2a と 2b に品詞タグ付けと依存構造解析における実験の結果をそれぞれ示す．いずれも，横軸は各言語の平均単語長，縦軸は各言語の文字系列情報の

2) https://universaldependencies.org/

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



(a) 品詞タグ付け　　(b) 依存構造解析

図 2　それぞれのタスクでの結果の散布図．赤色の丸印はラテン文字をアルファベットとして用いる言語を，赤色のバツ印はラテン文字ではないアルファベットを用いる言語を，青色はアブジャドを用いる言語を，緑色はアブギダを用いる言語を，黄色は表意文字を用いる言語をそれぞれ表している．

影響度でプロットされた散布図である．48 言語すべてのデータの Pearson の相関係数は，品詞タグ付けと依存構造解析においてそれぞれ 0.50 と 0.26 と計算される．依存構造解析においては平均単語長と文字系列情報による影響度には強い相関が見られないが，品詞タグ付けにおいては若干の強い相関が認められる．

ここでは，言語の文字体系と使用される文字の種類に着目し，特に 48 言語の中で最も多い，ラテン文字を用いる言語（30 言語）に着目する．ラテン文字を使用する言語のデータの Pearson の相関係数は，品詞タグ付けと依存構造解析においてそれぞれ 0.70 と 0.68 と計算される．この値は十分に大きく，ラテン文字を使用する言語において，平均単語長と文字系列情報による影響度に強い相関があることがわかる．なお，各言語でのそれぞれの影響度の値は付録（表 2）に示す．また，言語数やデータ数が十分ではないが，ラテン文字ではないアルファベットを用いる言語についても，同様の傾向が示唆される．

検定　ここでは，相関係数が 0.5 より大きい場合に「平均単語長が長い言語ほど，文字系列情報による性能への影響度が大きい」という仮説が示せるとする．ラテン文字を用いる言語について得られた 2 つの（標本）相関係数を，「母相関係数が 0.5 である」を帰無仮説，「母相関係数が 0.5 より大きい」を対立仮説とし，有意水準 5%の下で検定する．その結果，品詞タグ付け，依存構造解析において p 値はそれぞれ 0.0494，0.0785 と計算される．このことから，品詞タグ付けにおいては帰無仮説を棄却でき，母相関係数が 0.5 よりも大きいことが言える．一方，依存構造解析においては帰無仮説を棄却できず，母相関係数が 0.5 よりも大きいことが言えない．

5　おわりに

本研究では，文字系列情報を利用するニューラルモデルにおいて，学習時に出現しなかった文字を評価時に与えることにより，モデルにおける文字系列情報による性能の影響度を定義した．この指標を利用し，「平均単語長が長い言語ほど，文字系列情報による性能への影響度が大きい」という仮説を判定するために，二つの基礎的な言語処理タスクで実験を行なった．その結果，すべての言語に関しては仮説が支持されなかった一方，ラテン文字を用いる言語においては仮説を支持する傾向が得られた．一方で，依存構造解析での結果に関しては仮説検定の結果から仮説が支持されなかった．今後の展望としては，応用的なタスクを含む様々な言語処理タスクにおいてこの仮説が正しいか確かめる予定である．また，ラテン文字以外の言語についても確かめる余地がある．

謝辞

本研究は JST さきがけ JPMJPR21C8 の支援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] Barbara Plank, Anders Søgaard, and Yoav Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 412–418, Berlin, Germany, August 2016. Association for Computational Linguistics.

[2] Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. From raw text to Universal Dependencies - look, no tags! In Jan Hajič and Dan Zeman, editors, **Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 207–217, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1054–1063, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Tomoya Kurosawa and Hitomi Yanaka. Does character-level information always improve DRS-based semantic parsing? In Alexis Palmer and Jose Camacho-collados, editors, **Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)**, pp. 249–258, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. **Computational Linguistics**, Vol. 45, No. 3, pp. 559–601, September 2019.

[6] Yuan Zhang and Regina Barzilay. Hierarchical low-rank tensors for multilingual transfer parsing. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1857–1867, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[7] Karolina Stańczak, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein. A latent-variable model for intrinsic probing. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, No. 11, pp. 13591–13599, June 2023.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, November 1997.

[9] Ilya Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. In **Proceedings of the 28th International Conference on International Conference on Machine Learning**, ICML'11, p. 1017–1024, Bellevue, Washington, USA, 2011. Omnipress.

[10] Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. Finding function in form: Compositional character models for open vocabulary word representation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1520–1530, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Proceedings of the Thirty-first Conference on Neural Information Processing Systems**, Long Beach, California, December 2017. Curran Associates, Inc.

[12] Kris Cao. What is the best recipe for character-level encoder-only modelling? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5924–5938, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Clara Vania, Andreas Grivas, and Adam Lopez. What do character-level models learn about morphology? the case of dependency parsing. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2573–2583, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[14] Rik van Noord, Antonio Toral, and Johan Bos. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4587–4603, Online, November 2020. Association for Computational Linguistics.

[15] Matthew S. Dryer and Martin Haspelmath, editors. **WALS Online**. Zenodo, 2013. Available online at https://wals.info, Accessed on January 12, 2024.

[16] Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. Languages Through the Looking Glass of BPE Compression. **Computational Linguistics**, pp. 1–59, 10 2023.

[17] Joakim Nivre. Algorithms for deterministic incremental dependency parsing. **Computational Linguistics**, Vol. 34, No. 4, pp. 513–553, 2008.

[18] Marco Kuhlmann, Carlos Gómez-Rodríguez, and Giorgio Satta. Dynamic programming algorithms for transition-based dependency parsers. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 673–682, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[19] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

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

A　言語の詳細と実験結果の数値

表 2 に，本研究で使用した言語の詳細と実験結果の数値を示す。

B　ハイパーパラメータの詳細

表 3 に各種ハイパーパラメータを示す。基本的に先行研究 [1, 2] で提示された値を使用する。

表 3　各種ハイパーパラメータ。POS は品詞タグ付けでの，DEP は依存構造解析での値を示す。

	POS	DEP
エポック数	20	30
学習率	0.1	0.001
単語埋め込みの次元数	128	100
事前学習済みの単語埋め込みの次元数	-	50
文字埋め込みの次元数	100	500
文字 bi-LSTM の次元数	100	100
隠れ層の数	100	100
ガウス雑音の標準偏差 σ	0.2	-
単語のドロップアウト率	0.25	0.25
文字のドロップアウト率	0.25	0.33

表 2　言語名，ISO 639 言語コード，文字の体系と種類，品詞タグ付け (POS) と依存構造解析 (DEP) の影響度。

言語	コード	文字体系	文字の種類	POS 影響度	DEP 影響度
アラビア語	ar	アブジャド	アラビア文字	57.02%	22.48%
ベラルーシ語	be	アルファベット	キリル文字	33.96%	56.69%
ブルガリア語	bg	アルファベット	ラテン文字	45.24%	26.75%
カタルーニャ語	ca	アルファベット	ラテン文字	26.67%	21.19%
コプト語	cop	アルファベット	コプト文字	1.50%	24.53%
チェコ語	cs	アルファベット	ラテン文字	47.13%	31.17%
古代教会スラヴ語	cu	アルファベット	グラゴル文字	25.74%	29.72%
デンマーク語	da	アルファベット	ラテン文字	25.13%	25.41%
ドイツ語	de	アルファベット	ラテン文字	46.16%	45.35%
ギリシャ語	el	アルファベット	ギリシャ文字	52.41%	40.94%
英語	en	アルファベット	ラテン文字	14.01%	15.11%
スペイン語	es	アルファベット	ラテン文字	35.08%	16.10%
エストニア語	et	アルファベット	ラテン文字	60.89%	44.15%
バスク語	eu	アルファベット	ラテン文字	51.85%	42.81%
ペルシア語	fa	アルファベット	ペルシア文字	8.34%	23.99%
フィンランド語	fi	アルファベット	ラテン文字	65.18%	40.39%
フランス語	fr	アルファベット	ラテン文字	21.76%	17.92%
アイルランド語	ga	アルファベット	ラテン文字	6.66%	22.70%
ガリシア語	gl	アルファベット	ラテン文字	38.86%	29.76%
ゴート語	got	アルファベット	ゴート文字	33.81%	26.94%
古代ギリシア語	grc	アルファベット	ギリシャ文字	37.70%	43.46%
ヘブライ語	he	アブジャド	ヘブライ文字	8.81%	18.67%
ヒンディー語	hi	アブギダ	デーヴァナーガリー	9.34%	14.54%
クロアチア語	hr	アルファベット	ラテン文字	40.32%	39.13%
ハンガリー語	hu	アルファベット	ラテン文字	41.10%	57.96%
インドネシア語	id	アルファベット	ラテン文字	51.24%	28.26%
イタリア語	it	アルファベット	ラテン文字	31.77%	14.51%
日本語	ja	表意文字	漢字，仮名	25.89%	58.57%
韓国語	ko	表意文字	ハングル	43.20%	59.99%
ラテン語	la	アルファベット	ラテン文字	26.78%	49.34%
リトアニア語	lt	アルファベット	ラテン文字	18.38%	60.09%
ラトビア語	lv	アルファベット	ラテン文字	54.45%	42.47%
オランダ語	nl	アルファベット	ラテン文字	18.68%	26.61%
ノルウェー語	no	アルファベット	ラテン文字	23.21%	15.05%
ポーランド語	pl	アルファベット	ラテン文字	53.87%	40.19%
ポルトガル語	pt	アルファベット	ラテン文字	22.87%	24.08%
ルーマニア語	ro	アルファベット	ラテン文字	31.43%	19.36%
ロシア語	ru	アルファベット	キリル文字	43.36%	54.51%
サンスクリット語	sa	アブギダ	デーヴァナーガリー	11.92%	6.18%
スロバキア語	sk	アルファベット	ラテン文字	45.82%	54.11%
スロベニア語	sl	アルファベット	ラテン文字	37.46%	43.53%
スウェーデン語	sv	アルファベット	ラテン文字	29.59%	24.37%
タミル語	ta	アブギダ	タミル文字	49.66%	38.57%
トルコ語	tr	アルファベット	ラテン文字	33.67%	34.40%
ウクライナ語	uk	アルファベット	キリル文字	5.25%	44.21%
ウルドゥー語	ur	アブジャド	ウルドゥー文字	18.00%	12.90%
ベトナム語	vi	アルファベット	ラテン文字	4.63%	22.82%
中国語	zh	表意文字	漢字	23.79%	23.82%

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