基本語順に基づく単語並び替えの多言語固有表現抽出

塚本諒　寺岡丈博
拓殖大学工学部情報工学科
{228479@st,tteraoka@cs}.takushoku-u.ac.jp

概要

固有表現抽出は英語などの高リソースの言語で高い精度に達している．一方で，低リソースの言語では精度の向上が課題である．また一般的に広く使われる言語の文法は特徴に偏りがある．本研究では，低リソースの言語の課題を解決しつつ，どの言語であっても精度向上が目指せるよう，言語間の特徴の差を無くすことが重要と考え，どの言語にも見られる基本語順に着目した，多言語固有表現抽出の手法を提案する．

1　背景と目的

近年，固有表現抽出の分野では，高リソースの英語やフランス語などの一部の言語は高い精度に達している．一方で，言語ごとにデータ量は大きく異なり偏りがある．そのため低リソースの言語では精度向上は難しい．したがって，低リソースの言語でも効率的に学習するため，高リソースの言語から低リソースの言語への言語間転移技術が盛んに研究されている．しかし，研究対象の高リソースとされる言語はヨーロッパで使用されることが多く，これらと地理的に近いものや人口の多いものが低リソースの言語として対象となっているため，限定的と言える．世界には約 7000 語あり，そのうち約 4 割が消滅の危機に瀕していると言われている．自然言語処理の研究で対象とされる言語はほんの一部であり，言語ごとにデータ数を増やすことは現実的ではなく，さらに多くの言語で自然言語処理モデルを発展させるためには，言語固有の知識に頼らない普遍的知識を使用した手法で，言語の組み合わせに縛られずに精度を上げる必要があると考える．

本研究は，固有表現抽出のタスクにおいて，どの言語にも応用可能な手法を提案し，低リソースの言語の精度を高リソースの言語の精度と同程度まで向上させることが目標である．

2　関連研究

多言語での学習において，性能を向上させるには，高品質なソース言語の学習データ，言語間の特徴の類似度，意味の類似度が重要であることが知られている．

例えば，Shaffer ら [1] の埋め込みによる分類と言語族分類によって，性能を向上させられる言語の組み合わせを選択する研究や，Garca-Ferrero ら [2] の高リソース言語のラベル付きデータからターゲット言語にラベルを投影することで，低リソースの言語資源を増やす研究がある．

Ceolin ら [3] の研究では，理論言語学の知見を用いて名詞句における形態・統計的性質に関する 94 のパラメータに基づき 58 の言語を分類した．分類によって得られた言語樹は名詞句の構造の類似を示し，言語のグループと関係を見ることができる．

Imai ら [4] の研究では，Ceolin らと同様の理論言語学の知見に基づいた言語の分類で実験を行った．Ceolin らによる言語樹の一部である印欧語族の 25 言語に対してクラスタ分類をした．埋め込みによるクラスタリングに比べて，系統分類によるクラスタリングの方が精度が向上することを報告しているが，一部の言語は系統分類の方が精度が低い．Imai らの研究で対象としている印欧語族の言語は，他の研究でもよく研究される高リソースの言語である．また，全て同じ語族に属することから，より特徴の遠い言語がある場合でも有効な手法であるかは分からない．

Ji ら [5] の研究では，ターゲット言語の語順をソース言語の語順に擬似的に並び替えることで 2 言語間の差を減らす手法を提案した．構造予測タスクにおいて学習データがないターゲット言語に対し，ゼロショットで言語間転送することによって，精度

を向上させることを目的とした．結果として，ソース言語と特徴が大きく異なるターゲット言語への転送で精度が大きく向上することを報告した．それまで多言語モデルは単語レベルで共通化させていたが，Ji らの研究は文構造レベルで共通化をすることで言語間の互換性を持たせることができることを示した．

Politov ら [6] の研究では，低リソース言語への最適化問題を使用したアライメント投影の手法を提案した．アライメント投影によって作成されたデータを使用し，モデルに学習させる手法について，見直しと改良を加えることで固有表現抽出の精度向上を目的とした．結果として，多言語モデルを使用したアプローチよりも精度が上回ったことを報告した．より良いモデルを作ることは重要であるが，より良いデータを作ることも重要であることを示している．ただし，課題として，翻訳やアライメントの精度が低いことがエラーとなり，固有表現抽出の精度低下に繋がっている．

3 提案手法

3.1 概要

本研究では言語の特徴の差を基本語順とし，ソース言語の語順をターゲット言語の語順に揃えてモデルに学習させることを提案する．多言語学習における精度はどの言語と学習させるかが一つの重要な指標であり，Ji ら [5] の研究から言語の基本語順を揃えて学習させることが精度の向上に繋がる．どの言語の組み合わせでも高い精度を実現するには言語の特徴の差を無くすことで解決できると考えた．

3.2 並び替えのための前処理

始めに固有表現抽出のデータセットに対して Google Cloud Translation API[1] を使用して英語に翻訳した．なぜならデータセットには品詞情報が含まれておらず，高品質な英語が利用できるためである．次に翻訳した英文に対して spaCy を使用して動詞を抽出した．学習データ 223200 行のうち，77419 行（34.69 %）のデータで動詞が抽出された．最後に学習データと抽出された動詞との間の単語アライメントを計算した．本研究では並列コーパスを必要としない SimAlign[7] を使用する．SimAlign に使用する多言語モデルは XLM-RoBERTa[8] を使用し，アライメントの計算には二部グラフにおける重みの合計が最大になるようなペアを探す mwmf を使用する．計算に使用する動詞は英文から最初に出現する単語とし，投影先の単語が固有表現であった場合は投影しない．

1） https://cloud.google.com/translate?hl=ja

3.3 主語，動詞，目的語の並び替え

固有表現，動詞，基本語順の情報から学習データの並び替えを行う．ただし，動詞が投影されなかったデータは並び替えを行わない．並び替えをする前に主語と目的語の特定をする．基本語順が SVO の言語の場合，図 1 のように，動詞の単語から左側にある単語はすべて主語とし，動詞の単語から右側にある単語は全て目的語とする．基本語順が「SOV, VSO, VOS」の言語の場合，図 2 のように，それぞれ動詞である単語から「右側，左側，左側」は全て動詞とし，「主語と目的語，主語と目的語，目的語と主語」の区切りは固有表現がある位置の前後をカウントし，2 で割った位置で区切る．主語，動詞，目的語に分類後，ターゲット言語の語順と同じになるようトークン列を並べる．

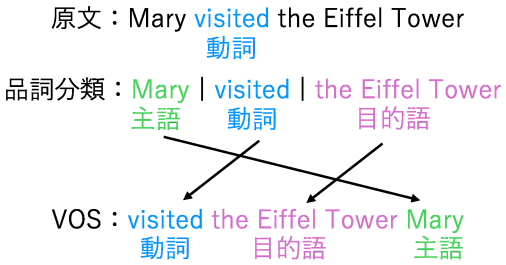

図 1 SVO 言語の並び替え

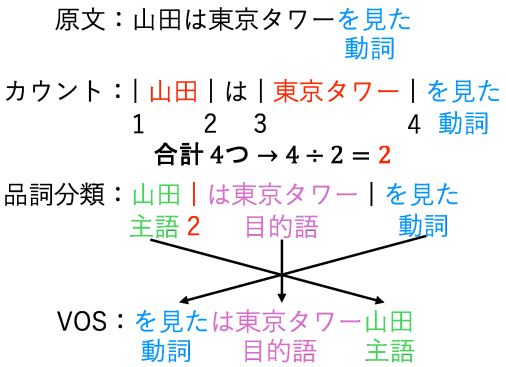

図 2 SOV・VSO・VOS 言語の並び替え

3.4 学習モデルとデータセット

学習モデルは 100 言語，2.5TB のフィルタリングされた CommonCrawl データで学習された大規模な

多言語モデルの XLM-RoBERTa[8] を使用する．多言語モデルを使用することで，個別の言語に依存しない，普遍的知識から学習できる可能性がある．

また，データセットは IOB2 形式の LOC（場所），PER（人），ORG（組織）タグで注釈された 176 言語をサポートする多言語固有表現抽出データセットの WikiAnn[9] を使用する．本研究では基本語順に多様性を持たせるために Ceolin ら [3] による言語の系統樹のうち，Imai ら [4] の研究で使われていない且つ WikiAnn のデータセットがサポートする 17 言語を対象とする．また本研究での低リソースの言語の定義は WikiAnn の各言語の学習データ数の平均である 6006 以下の言語とする．本研究では，マダガスカル語（mg），ウズベク語（uz），カザフ語（kk），キルギス語（ky），テルグ語（te）が低リソースの言語に該当する．基本語順の分類は言語学の書籍 [10] に基づき，17 言語を表 1 のように分類した．

表 1　対象言語の基本語順の分類と先行研究との比較

本研究（17 言語）				先行研究（25 言語）
SVO	SOV	VSO	VOS	
zh-yue zh et fi	eu, ja ko, hu uz, tr kk, ky ta, te	he ar	mg	ro, fr, es, pt, it, scn, af, nl, de, is, en, da, no, fo, el, bg, pl, ru, sl, hr, ps, mr, hi, cy, ga

4　評価実験

4.1　評価方法

学習の設定は学習率を 5e-5，バッチサイズを 32，エポック数を 3，正則化を 0.01 とした．各言語ごとに適合率，再現率，F 値を 10 回の平均で求め，前処理をしない 17 言語のデータをそのまま学習させる手法（ベースライン）と Imai らの言語の系統樹を使用したクラスタリングによる手法（Imai らの手法）と比較した．

4.2　系統樹のクラスタリング

Imai らの手法について，本研究で対象とする 17 言語で言語の系統樹を作成した場合，図 3 のようになる．語族に基づいてクラスタリングするとクラスタ数は最大で 9 に分類することができ，エルボー法で近い部分から合わせていくことで最小でクラスタ数は 2 になる．比較には最も精度が高かった表 2 のクラスタ数 2 を使用する．

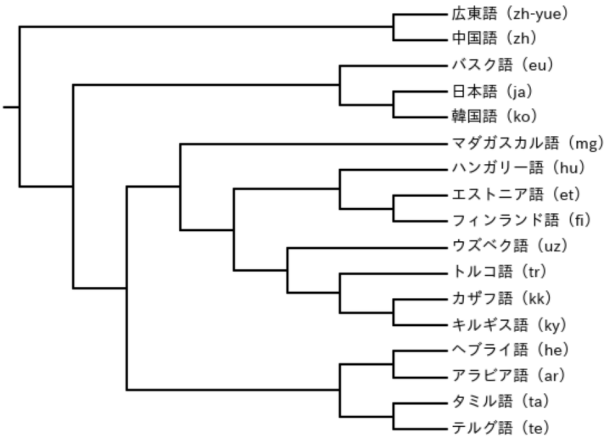

図 3　本研究で使用する言語の系統樹

表 2　Imai らの言語の系統樹を使ったクラスタリング（クラスタ数 2）

クラスタ 1	クラスタ 2
zh-yue, zh	eu, ja, mg, hu, et, uz, he, ta, ko, fi, tr, kk, ky, ar, te

4.3　結果

表 3 からベースラインと比べ，Imai らの手法は 17 言語中 8 言語で精度が向上し，提案手法は 17 言語中 12 言語で精度がわずかに向上した．提案手法は Imai らの手法に比べ，より多くの言語で精度が向上する手法であった．

しかし，Imai らの手法と比べ，提案手法は 17 言語中 8 言語のみ精度が向上した．Imai らの手法で最も良い精度は特徴の近い言語が多く，マダガスカル語からキルギス語までは言語の系統樹で特に特徴の近い言語であり，これら 8 言語中 5 言語で Imai らの手法の方が高い精度であった．一方，提案手法は特徴の近い言語が少ない場合，精度が高い傾向にある．言語の系統樹の広東語（zh-yue）から韓国語（ko），ヘブライ語（he）からテルグ語は特徴の近い言語が少ない言語であり，これら 9 言語中 7 言語で提案手法の方が高い精度であった．

また，低リソースの言語は，提案手法が最も高い精度となったのは，3 手法の中で 5 言語中 1 言語のみであった．

5　考察

5.1　他手法との比較

一般的にトークン列を物理的に並び替えることは文法を破壊する行為であるため，精度を下げる原因

表 3 実験結果（太字は Imai らの手法と比べて高い精度）

	データ数	ベースライン	Imai ら	提案手法
zh-yue	20000	82.24	79.91	**82.23**
zh	20000	78.63	75.52	**78.97**
eu	10000	90.53	90.57	**90.74**
ja	20000	68.77	67.43	**68.79**
ko	20000	85.72	85.63	**85.68**
mg	100	88.90	92.72	90.50
hu	20000	91.88	91.97	91.96
et	15000	90.94	91.10	91.00
fi	20000	90.29	90.26	90.13
uz	1000	91.08	91.32	91.25
tr	20000	91.87	91.95	91.92
kk	1000	86.37	85.86	**86.49**
ky	100	66.79	65.49	65.18
he	20000	83.72	83.68	**83.74**
ar	20000	87.60	87.96	87.78
ta	15000	83.86	84.08	**84.09**
te	1000	80.34	80.32	80.02

である．提案手法はベースラインと比べて良い結果であり，この程度の並び替えは問題がないことを示した．しかし，わずかにしか精度が上がらなかった原因としては並び替えの条件が厳しかったと考えられる．単語並び替えの前後のトークン列の一致率を計算したところ，ターゲット言語が属する語順を含めたトークン列が完全に一致するデータは，語順を SVO に揃えた場合，73.52 %，同様に SOV の場合が最も高く，80.03 %，VSO の場合，59.27 %，VOS の場合が最も低く，51.57 %であった．提案手法は文法を破壊しないよう慎重に並び替えた．また，複数の処理を踏んでいるため，エラーが頻発しないよう処理の結果が不明確なものは並び替えをしなかった．並び替えにより失われた文脈の効果よりも語順を揃えたことによる効果の方がやや上回った可能性がある．一方で，提案手法の並び替えのルールは単純過ぎる可能性がある．提案手法は主語と目的語の位置を動詞の位置からヒューリスティックに特定している．また文法は主語，動詞，目的語以外の要素もある．より厳密な手法で品詞の特定ができれば，精度はより向上する可能性がある．

5.2 各言語について

17 言語のうち，バスク語（eu），マダガスカル語の F 値は 90 %を超えるかなり高い精度であった．この 2 言語はそれぞれ他の言語と比べ，地理的に孤立しており，バスク語はどの語族にも属さない独立語，マダガスカル語は基本語順 VOS，オーストロネシア語族という唯一の特徴を持っていた．あまりに固有な特徴を持つ言語では多言語モデルの言語に依存しない共通の知識からではなく，その言語だけの知識だけで学習している可能性がある．

一方で，特徴の似ている言語がある言語の場合，バイアスの影響がある可能性がある．ウズベク語，トルコ語（tr），カザフ語，キルギス語は似た言語であり，テュルク語族に属す．共通点が多いが，F 値には差があり，ウズベク語，トルコ語は 91 %と高いのに対し，カザフ語，キルギス語は，86 %，65 %であった．ここには大きな差があり，その一つとして文字が異なる．ウズベク語，トルコ語がラテン文字なのに対し，カザフ語，キルギス語がキリル文字である．この 4 言語は共通する知識を使って学習しようとするが，カザフ語，キルギス語はウズベク語，トルコ語に比べ，学習データが少ないため，精度に差が出ている可能性がある．仮にキルギス語がモデルの事前学習の知識も含め，文字以外にも固有の特徴を持っていれば，他の言語に左右されず，高い精度となった可能性がある．

また，分かち書きをしない言語は文字単位のトークンとなっているため，データ数に対して精度が低い．さらに低リソースの言語において提案手法は，5 言語中 4 言語で精度が低下している．ターゲット言語とソース言語の差をなくすこと以上にターゲット言語の質と量を上げることが重要である．

6 まとめ

本研究では，非言語依存な手法で多言語固有表現抽出における，低リソースの言語の精度向上を目的とし，基本語順に着目してソース言語の学習データに単語の並び替えをして学習させる手法を提案した．先行研究と比較した結果，提案手法はより多くの言語で精度が上がり，系統樹や言語の特徴に縛られない手法であることを示した．

今後の課題として，提案手法は翻訳時に英語への依存がある．英語の精度が良いことに期待して英語への翻訳をしたが，英語との特徴が異なるほど翻訳の精度が低くなると考えられ，英語に非依存な手法を考える必要がある．言語の特徴は基本語順だけではなく，音に基づく特徴や名詞格の特徴など様々な分類がある．ゆえに並び替えの条件や分類の基準を変えることで精度の向上に期待できる．また，低リソースの言語の精度を向上させるため，ターゲット言語への対策が必要である．

謝辞

本研究は JSPS 科研費 JP22K00646 と JP25K04218 の助成を受けたものです.

参考文献

[1] Kyle Shaffer. Language clustering for multilingual named entity recognition. In **Findings of the Association for Computational Linguistics: EMNLP**, 2021.

[2] Iker García-Ferrero, Rodrigo Agerri, and German Rigau. T-projection: High quality annotation projection for sequence labeling tasks. In **Findings of the Association for Computational Linguistics: EMNLP**, 2023.

[3] Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, Monica Alexandrina Irimia, Luca Bortolussi, and Andrea Sgarro. At the boundaries of syntactic prehistory. **Philosophical Transactions of the Royal Society B**, Vol. 376, No. 1824, p. 20200197, 2021.

[4] Sakura Imai, Daisuke Kawahara, Naho Orita, and Hiromune Oda. Theoretical linguistics rivals embeddings in language clustering for multilingual named entity recognition. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop**, 2023.

[5] Tao Ji, Yong Jiang, Tao Wang, Zhongqiang Huang, Fei Huang, Yuanbin Wu, and Xiaoling Wang. Word reordering for zero-shot cross-lingual structured prediction. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, 2021.

[6] Andrei Politov, Oleh Shkalikov, Rene Jakel, and Michael Farber. Revisiting projection-based data transfer for cross-lingual named entity recognition in low-resource languages. In **Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies**, 2025.

[7] Masoud Jalili Sabet, Philipp Dufter, Fran¸cois Yvon, HinrichSch ¨ utze. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In **Findings of the Association for Computational Linguistics: EMNLP**, 2020.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, FranciscoGuzm ´ an, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[9] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, 2017.

[10] 山本秀樹. 世界諸言語の地理的・系統的語順分布とその変遷. 溪水社, 2003.

[11] 今井咲良, 河原大輔, 折田奈甫, 小田博宗. 理論言語学の知見を応用した多言語クラスタリング. 言語処理

A ISO コード

表 4 ISO コードとの対応

ISO 639	言語名
本研究（17 言語）	
zh-yue	広東語
zh	中国語
eu	バスク語
ja	日本語
ko	韓国語
mg	マダガスカル語
hu	ハンガリー語
et	エストニア語
fi	フィンランド語
uz	ウズベク語
tr	トルコ語
kk	カザフ語
ky	キルギス語
he	ヘブライ語
ar	アラビア語
ta	タミル語
te	テルグ語
Imai らの研究（25 言語）	
cv	ウェールズ語
ga	アイルランド語
ps	パシュトー語
mr	マラーティー語
hi	ヒンディー語
ro	ルーマニア語
fr	フランス語
es	スペイン語
pt	ポルトガル語
it	イタリア語
son	シチリア語
el	ギリシア語
bg	ブルガリア語
pl	ポーランド語
ru	ロシア語
sl	スロベニア語
hr	セルビア・クロアチア語
af	アフリカーンス語
nl	オランダ語
de	ドイツ語
is	アイスランド語
en	英語
da	デンマーク語
no	ノルウェー語
fo	フェロー語