データ拡張による固有表現抽出の不確実性推定

橋本 航　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hashimoto.wataru.hq3, kamigaito.h, taro}@is.naist.jp

概要

本研究では，固有表現抽出タスクにおけるデータ拡張が不確実性推定に与える影響について調査する．医療や金融のような高い安全性が求められる領域では，予測結果が正しいだけではなく，その信頼度が高い必要がある．しかし，事前学習済みモデルを含む深層学習モデルはしばしば実際の正解率と乖離した確信度を出力するため，それらの領域への適用が妨げられている．また，その問題に対応する既存手法は推論コストが高い．我々は，ジャンル横断および言語横断の設定にて，固有表現抽出におけるデータ拡張が不確実性の推定性能に与える影響を調査した．その結果，データ拡張は多くの場合不確実性の推定性能を向上させ，データ拡張によって生成した文の Perplexity が低い場合はデータ拡張サイズを増やすことで，さらに不確実性の推定性能が改善することが判明した．

1　はじめに

固有表現抽出 (Named Entity Recognition; NER) は，テキストから人名や組織名などの固有表現を抽出する，自然言語処理における基礎的なタスクの一つである．自然言語処理タスクは深層学習（Deep Neural Networks; DNNs）によって大きな成功を得ており，その中でも BERT [1] や DeBERTa [2] を始めとした事前学習済みモデル (Pre-trained Language Models; PLMs) に基づいた手法が，NER も含めて強力なベースラインとなっている．

しかし，一般的に PLMs を含む DNNs は図 1 のように，モデルの出力する確信度が実際の正解率と乖離している傾向にある [3]．この問題により，DNNs は高い確信度での誤った予測の出力をしばしば引き起こすため，医療や金融のような誤りに対するコストが大きい領域での DNNs の適用が制限される．このような問題に対処するため，自然言語処理における様々な領域で不確実性をより正しく推定するための研究が進んでいる [4, 5]．しかし，これらの研究では MC Dropout [6] や Bayesian Neural Networks [7] のような 1 つのデータインスタンスに対して複数回の予測を行う必要がある手法が用いられており，その推論コストは莫大になるため実用的ではないという問題がある．

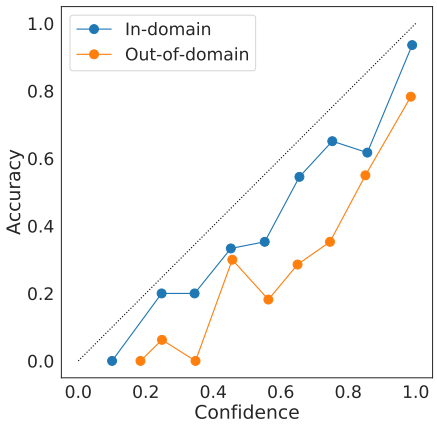

図 1　分布内および分布外のデータに対する，mDeBERTaV3 による固有表現抽出の信頼度プロット[1)]．対角線は，確信度と正解率が完全に一致していることを示す．

そこで，我々はデータ拡張に着目する．コンピュータビジョン領域では，データ拡張はモデルを入力に対してロバストにするため，不確実性推定の改善につながることが報告されている [8, 9]．さらにデータ拡張自体はモデル構造を変えないため，推論時間の増加を伴うことなく不確実性推定が改善されると期待される．データ拡張は NER にも適用されているが [10, 11]，低資源設定における汎化性能の観点に限定されているものが多い．

本研究では，NER におけるデータ拡張が不確実性推定に与える影響を，ジャンル横断および言語横断の設定において調査した．その結果，我々の実験からいくつかの知見が得られた．第一に，NER におけ

1)　分布内および分布外のデータは，それぞれ OntoNotes 5.0 の bn および tc である．

るデータ拡張は多くの場合不確実性の推定性能の改善につながる．特に，文脈に応じてエンティティを新しく生成する Masked Entity Language Modeling (MELM) [11] や，エンティティを学習データ内の同じエンティティタイプのエンティティで置換する Mention Replacement (MR) [10] により不確実性の推定性能が改善した．第二に，データ拡張によって生成された文の Perplexity が低いほど，データ拡張サイズを増やしたときにさらに不確実性の推定性能が改善することが判明した．

2 手法

本節では，実験で用いる既存手法および NER におけるデータ拡張について説明する．

2.1 既存手法

Baseline 最終層の Softmax 関数を通した後の最大確率を用いる．

Temperature Scaling (TS) TS は DNNs が出力する確信度を補正するための後処理方法である．Softmax 関数を適用する前に logit を温度パラメータでスケーリングする．

Label Smoothing (LS) LS は機械学習で一般的な正則化手法であり，確信度を他ラベルにも割り振ることで自信過剰な予測を防ぐ．

Monte-Carlo Dropout (MC Dropout) MC Dropout は DNNs の不確実性推定に利用できる正則化手法であり，複数の確率的推論を必要とする．本研究では $M = 20$ の確率的推論を行い，その平均を出力する．

2.2 固有表現抽出におけるデータ拡張

Label-wise Token Replacement (LwTR) LwTR では，まずトークンを置換するかどうかを決定するために二項分布を用いる．二項分布により選ばれたトークンは，学習データ上のラベルごとのトークン分布に基づいて，同じラベルを持つ別トークンとランダムに置き換えられる．

Mention Replacement (MR) MR は，トークンの代わりに学習データに存在する同じラベルを持つ別のエンティティで文中のエンティティを置き換える．学習データ中のエンティティは様々なトークン数から構成されているため，MR は LwTR と異なり元のラベル列を保持するとは限らない．

Synonym Replacement (SR) SR は LwTR と同様に置換するトークンを二項分布で選び，対象のトークンを WordNet の同義語で置換する．同義語は複数のトークンを持つことがあるため，SR は元のラベル列を保持するとは限らない．

Masked Entity Language Modeling (MELM) MELM では，まずエンティティマーカーによってエンティティ部分がマスクされた文に対して，文脈上適切なエンティティを予測するような言語モデルを学習する．その言語モデルを用いて，指定した比率でマスクされたエンティティ部分に対して元エンティティとは異なるエンティティを出力することでデータ拡張を行う．

3 実験設定

本研究では，系列ラベリングの枠組みで NER を行い，事前学習済みモデルの mDeBERTaV3 (`microsoft/mdeberta-v3-base`) [12] をエンコーダとして採用する．[2] 各実験は異なるシードで 10 回試行し，評価指標は平均値および標準偏差を報告する．また，視認性の向上のため，報告値は 100 倍した．

3.1 データセット

本研究では，ジャンル横断および言語横断設定における不確実性推定性能を測るため，OntoNote 5.0 データセット [13] および MultiCoNER データセット [14] を用いる． OntoNote 5.0 データセットは broadcast conversation (`bc`)，broadcast news (`bn`)，magazine (`mz`)，newswire (`nw`)，telephone conversation (`tc`)， および web data (`wb`) の 6 つのジャンルから構成される． MultiCoNER データセットは 11 の言語から構成される NER データセットであり，本研究では 英語 (`EN`)， ドイツ語 (`DE`)，スペイン語 (`ES`)，およびヒンディ語 (`HI`) の 4 言語を用いる．

3.2 評価尺度

本研究ではエンティティの確信度を評価するため，エンティティを構成するトークンの確信度の積をエンティティの確信度とみなして評価を行う．用いる評価指標を以下に示す．

Expected Calibration Error(ECE) ECE は各ビンごとの正解率と確信度の差を測り，その期待値を取る．本研究ではビンの個数を 10 とする．

Maximum Calibration Error(ECE) MCE は ECE と類似しており，期待値ではなく最大値をとる．

2) MELM におけるエンティティ予測のための言語モデルにも，同様に mDeBERTaV3 を用いる．

表 1 OntoNotes 5.0 を用いたジャンル横断設定における結果．太字は手法間での最良を示す．

OntoNotes 5.0 (bc)

Methods	bc		bn		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	18.87±0.73	23.58±1.01	17.54±0.67	25.90±1.29	18.83±0.89	25.65±1.09	23.52±0.77	34.25±1.41	**26.20±1.23**	**28.76±1.30**	57.47±0.87	62.96±0.56
TS	18.86±0.68	23.22±0.86	17.19±0.81	24.93±1.27	19.42±1.48	26.32±1.97	23.51±1.08	33.68±1.72	26.85±2.11	29.36±2.35	57.66±1.32	62.96±1.15
LS	19.29±1.04	24.11±1.57	17.45±0.96	25.43±1.77	19.38±1.03	26.36±1.56	23.72±1.01	34.23±1.95	26.34±1.78	28.81±2.04	**56.98±1.17**	**62.51±0.91**
MC Dropout	18.69±0.71	23.54±1.31	17.50±0.66	25.77±1.58	19.22±1.21	26.39±1.16	23.67±0.73	34.51±1.59	26.32±1.10	28.66±1.12	57.51±1.29	62.80±0.90
LwTR (DA)	19.15±0.55	23.70±0.77	17.58±0.44	25.45±1.34	19.34±1.34	26.11±1.56	23.65±0.53	33.89±1.13	27.50±1.73	29.70±2.01	58.68±1.51	63.83±1.22
MR (DA)	19.13±0.95	23.17±1.10	17.43±0.62	24.99±1.36	**18.38±1.62**	**24.93±1.73**	**23.28±0.54**	33.35±1.16	26.78±2.19	28.85±2.21	59.01±0.99	64.06±0.76
SR (DA)	**18.16±0.63**	**21.99±0.91**	**17.01±0.39**	**24.45±0.74**	20.01±1.56	26.94±1.72	23.42±0.66	**33.29±1.33**	26.62±1.59	28.81±1.76	58.14±0.79	63.02±0.59
MELM (DA)	18.59±0.60	22.67±0.95	17.22±0.65	24.55±1.41	19.41±0.80	26.01±1.06	23.66±0.85	33.75±1.46	30.11±1.39	32.59±1.71	58.72±1.42	63.71±1.18

OntoNotes 5.0 (bn)

Methods	bc		bn		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	19.30±0.82	24.37±1.47	11.50±0.75	16.14±1.97	20.55±1.59	26.62±2.55	20.05±0.98	28.44±2.25	25.42±0.73	27.56±0.64	**59.02±1.16**	63.61±0.66
TS	19.20±0.88	24.18±1.75	11.25±0.55	15.43±1.41	21.21±1.14	27.20±1.72	20.34±0.73	28.80±2.12	25.33±1.28	27.57±1.27	59.11±1.06	**63.60±0.60**
LS	**18.37±0.60**	**22.52±1.41**	11.42±0.52	15.31±1.24	21.61±0.47	27.04±1.04	19.98±0.41	27.64±1.11	**24.66±0.48**	**26.69±0.44**	59.92±0.75	63.87±0.77
MC Dropout	18.76±0.97	23.34±1.56	11.38±0.71	15.73±1.60	20.91±0.96	26.62±1.82	20.04±0.57	28.25±1.62	25.21±1.27	27.52±1.17	59.09±0.99	63.63±0.54
LwTR (DA)	20.30±0.87	25.42±1.18	11.72±0.42	16.37±1.21	20.71±1.01	27.14±1.16	20.51±0.41	29.04±1.26	26.36±2.08	28.67±2.09	59.32±0.97	64.00±0.55
MR (DA)	19.78±1.26	24.35±1.85	11.59±0.34	15.89±0.92	20.19±0.47	26.08±1.07	20.42±0.60	27.83±1.74	25.69±0.77	27.75±0.81	59.57±0.96	64.13±0.50
SR (DA)	19.61±0.97	24.08±1.64	11.38±0.44	15.44±0.96	**19.79±0.75**	**25.52±1.22**	19.81±0.39	27.18±1.30	26.20±1.56	28.42±1.68	59.86±0.67	63.66±0.40
MELM (DA)	19.93±0.69	23.98±1.09	**10.75±0.46**	**14.11±0.69**	20.40±0.65	25.54±1.19	**19.73±0.65**	**26.80±1.19**	28.47±2.14	30.59±2.15	60.51±0.57	64.44±0.33

OntoNotes 5.0 (tc)

Methods	bc		bn		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	36.70±1.65	44.25±1.66	35.47±2.48	45.75±2.46	37.15±1.77	47.34±1.79	39.08±0.56	52.50±1.41	31.17±1.56	33.81±1.67	**46.38±1.28**	**54.29±1.37**
TS	35.69±2.21	43.34±2.18	34.15±2.65	44.48±2.56	36.38±1.79	46.71±1.43	38.59±1.53	52.58±1.38	27.95±2.51	30.70±2.55	47.20±0.92	55.31±1.10
LS	33.91±1.86	41.50±1.75	31.40±2.35	41.24±2.43	34.14±1.91	44.37±1.42	37.04±2.25	50.00±1.92	26.46±1.36	28.89±1.42	48.48±1.29	56.10±0.89
MC Dropout	35.83±2.02	43.93±1.75	33.87±2.02	44.31±1.92	36.18±2.43	46.31±2.43	38.97±0.83	52.80±1.08	29.01±2.50	31.94±2.81	46.92±2.04	54.95±2.13
LwTR (DA)	34.94±2.42	43.20±1.90	32.61±3.16	43.28±2.55	34.44±1.83	44.98±1.88	37.85±2.13	52.09±1.60	28.78±2.27	31.31±2.14	46.78±1.26	54.94±1.84
MR (DA)	35.18±2.89	42.62±2.30	33.50±3.77	42.66±3.20	34.35±2.78	44.78±2.69	37.97±2.64	50.85±3.46	28.65±3.20	31.23±3.18	48.61±1.70	55.78±1.90
SR (DA)	34.58±2.40	42.51±1.55	32.66±4.13	42.57±3.28	**32.69±3.21**	43.01±2.83	38.50±1.51	52.00±1.56	27.30±4.37	29.85±4.54	46.99±1.27	54.86±1.40
MELM (DA)	**33.05±1.75**	**40.55±2.16**	**29.46±1.55**	**37.81±1.56**	33.46±1.66	**42.78±2.55**	**36.79±1.27**	**49.33±2.26**	**25.71±1.73**	**28.19±1.87**	50.52±1.10	57.27±1.27

Area Under the Precision-Recall Curve (AUPRC) AUPRC は Precision/Recall (PR) 曲線の下の領域の面積である．高いほど，特定の Recall において Precision が高い傾向となる．

4 検証

本節では，ジャンル横断および言語横断の設定における不確実性推定の結果を提示する．

4.1 ジャンル横断における評価

表 1 に，OntoNotes 5.0 でソースジャンルが bc，bn，および tc の場合における，ジャンル横断的な不確実性推定の評価結果を示す．この表から，エンティティごとの不確実性の推定性能の評価において，一般的な分類問題に有効であると考えられてきた TS，LS および MC Dropout よりも，データ拡張の方が性能が高い傾向にあることがわかる．特に，ソースジャンルが tc の場合，MELM をはじめとするデータ補強法は，Baseline と比較して，ECE で最大 6.01 %，MCE で最大 7.94 % 改善し，優れた校正性能を示す．また，MR と SR も MELM に続いて良好な校正性能を示す．一方で，データ増強法では wb における校正性能は改善されない傾向にある．また，AUPRC を表 4 に示す．データ拡張の中では，MR が優れた性能を示す一方で，MELM は ECE や MCE のような校正エラーに基づく評価指標ほど改善されない．

4.2 言語横断における評価

MultiCoNER でソース言語を EN とした場合の言語横断設定における不確実性推定の結果を表 2 に示す．ジャンル横断設定の場合と異なり，MR が分布内および分布外において優れた不確実性推定性能を示した． Zhengping ら [15] により，言語的な距離が大きいほど確信度校正性能が低下する傾向にあることがわかっているが，本実験でも同様の傾向が得られた．一方，ジャンル横断的な校正では優れた不確実性推定性能を示す傾向にある MELM は，言語横断的な設定では良い性能を示さない．

また，本研究で用いた mDeBERTaV3 の事前学習に用いられている CC100 データセット [16] の言語比率を見ると，英語が最も多く次いでドイツ語，スペイン語，ヒンディ語の順に多く，不確実性推定性能の順位と相関している． さらに，Tomasz らによ

表 2　MultiCoNER を用いた言語横断設定における結果.

Methods	MultiCoNER (EN)											
	EN			DE			ES			HI		
	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)
Baseline	28.29±0.30	30.51±0.39	93.04±0.18	31.31±0.52	34.91±0.83	91.97±0.23	31.22±0.28	33.70±0.39	90.87±0.27	46.84±1.64	48.13±1.51	82.04±2.24
TS	28.46±0.43	30.70±0.52	93.13±0.17	31.45±0.70	35.08±1.05	92.02±0.24	31.24±0.41	33.77±0.38	90.92±0.18	46.83±1.38	48.35±1.25	83.01±1.45
LS	28.50±0.57	30.60±0.68	93.12±0.13	31.50±0.64	34.81±0.66	91.93±0.26	31.43±0.58	33.83±0.67	90.82±0.10	46.36±1.23	47.95±1.03	84.00±1.60
MC Dropout	28.57±0.34	30.83±0.54	92.97±0.34	31.64±0.48	35.24±0.68	91.86±0.37	31.47±0.42	33.98±0.40	90.79±0.22	47.42±1.30	48.77±1.23	81.39±3.30
LwTR (DA)	28.17±0.54	30.48±0.77	92.80±0.28	31.13±0.59	34.60±0.78	91.57±0.34	31.10±0.35	33.61±0.51	90.66±0.27	46.70±1.47	47.95±1.30	82.57±1.96
MR (DA)	**28.01±0.42**	**30.08±0.49**	**93.30±0.24**	**31.12±0.74**	34.71±0.81	**92.05±0.20**	**30.75±0.34**	**33.24±0.36**	**91.03±0.15**	46.96±1.20	48.28±1.12	81.75±2.52
SR (DA)	28.15±0.42	30.36±0.48	93.08±0.26	31.17±0.39	**34.42±0.70**	92.02±0.39	31.60±0.55	33.86±0.56	90.65±0.33	**45.85±0.53**	**47.38±0.47**	**84.91±0.91**
MELM (DA)	28.53±0.38	30.68±0.43	92.72±0.22	32.61±0.49	36.14±0.65	91.17±0.29	32.09±0.44	34.38±0.52	90.14±0.30	47.91±1.79	49.18±1.79	81.13±2.41

表 3　データ拡張によって生成した文の Perplexity.

Algorithm	OntoNotes 5.0 (bc)	OntoNotes 5.0 (bn)	OntoNotes 5.0 (tc)	MultiCoNER (EN)
LwTR	7.05	7.59	7.33	6.78
MR	**5.36**	**5.27**	**5.83**	**5.83**
SR	5.91	6.35	6.02	6.35
MELM	5.56	5.65	5.90	6.14
(Train)	5.18	4.84	5.80	5.54

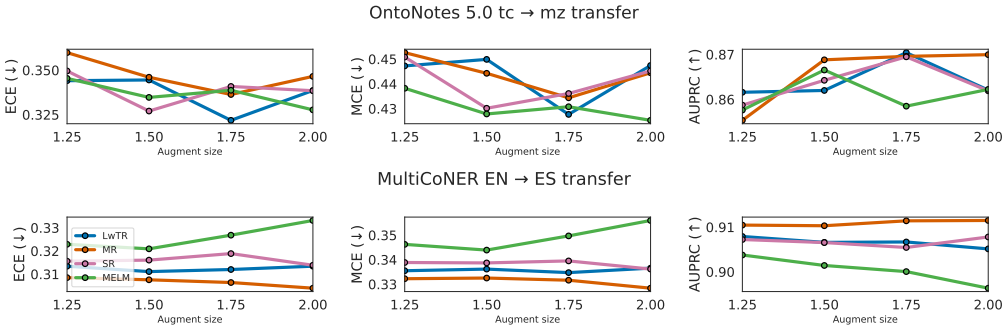

図 2　データ拡張サイズを増やした場合における各評価指標の推移.

り [17]，トークン化において言語間で語彙の重複が生じやすい場合，NER における言語横断設定においてよい予測性能を示す傾向にあることが示されたが，不確実性推定においても同等の傾向を示す.

5　分析

本節では，NER のデータ拡張の不確実性推定性能への理解を深めるため，生成した文の Perplexity およびデータ拡張サイズの影響を調査する.

まず，データ拡張によって生成した文の不確実性推定性能への影響を調査するため，我々は GPT2 [18] を用いてデータ拡張による生成文の Perplexity を測った．表 3 に，各データ拡張手法で生成した文および元の学習データの Perplexity を示す．いずれの場合においても，MR が最も Perplexity が低い一方で，MELM は MR ほど低い Perplexity は示さなかった．MELM において文脈に適合しているが実際には存在しないエンティティが生成されると，文の Perplexity が悪影響を受ける可能性がある.

さらに，図 2 に，OntoNotes 5.0 tc → mz, MultiCoNER EN → ES のような分布外の横断設定において，データ拡張サイズを増やした場合の評価指標の推移を示す（分布内横断設定の結果は 図 3 に示す）．多くの場合，MR がデータ拡張サイズを増やした場合に不確実性推定がさらに改善される．先述の通り MR は複数のデータセットで最も Perplexity が低いため，Perplexity の低さがデータ拡張サイズを大きくした場合の不確実性推定性能にとって重要であることがわかる.

6　まとめ

本研究では，NER におけるデータ拡張が不確実性推定性能に与える影響をジャンル横断および言語横断の観点から調査した．その結果，MELM や MR のようなデータ拡張を用いた場合に優れた不確実性推定性能を示すことが判明した．一方で，TS や MC Dropout のような良い不確実性推定性能を示すとされてきた既存手法は，NER の場合においては良い不確実性推定性能を示さない．また，データ拡張により生成した文の Perplexity が低い MR のような手法において，データ拡張サイズを増やすことでさらに不確実性推定性能が改善されることが判明した.

参考文献

[1] 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, 2019.

[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In **International Conference on Learning Representations**, 2021.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In **Proceedings of the 34th International Conference on Machine Learning**, Vol. 70 of **Proceedings of Machine Learning Research**, pp. 1321–1330. PMLR, 2017.

[4] Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In **Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence**, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[5] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In **International Conference on Learning Representations**, 2021.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In **Proceedings of The 33rd International Conference on Machine Learning**, Vol. 48 of **Proceedings of Machine Learning Research**, pp. 1050–1059, 2016.

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 1613–1622, Lille, France, 2015. PMLR.

[8] Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, and Dustin Tran. Combining ensembles and data augmentation can harm your calibration. In **International Conference on Learning Representations**, 2021.

[9] Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, and Deva Ramanan. Soft augmentation for image classification. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 16241–16250, June 2023.

[10] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3861–3867, 2020.

[11] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2251–2262, 2022.

[12] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing. In **The Eleventh International Conference on Learning Representations**, 2023.

[13] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In **Proceedings of the Seventeenth Conference on Computational Natural Language Learning**, pp. 143–152, 2013.

[14] Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3798–3809, 2022.

[15] Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Calibrating zero-shot cross-lingual (un-)structured predictions. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 2648–2674, 2022.

[16] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, 2020.

[17] Tomasz Limisiewicz, Jiří Balhar, and David Mareček. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 5661–5681, 2023.

[18] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

[20] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, p. 2623–2631, 2019.

表 4　OntoNotes 5.0 を用いたジャンル横断設定における AUPRC.

Methods	OntoNotes 5.0 (bc)					
	bc	bn	mz	nw	tc	wb
Baseline	94.72±0.21	95.13±0.43	96.40±0.40	93.27±0.41	92.69±0.57	93.03±0.56
TS	**94.89±0.59**	95.14±0.35	96.15±0.51	93.26±0.45	92.78±1.01	92.97±0.83
LS	94.74±0.54	95.09±0.37	96.15±0.30	93.15±0.43	92.60±0.79	92.73±0.36
MC Dropout	94.71±0.31	95.09±0.18	96.07±0.24	93.11±0.43	92.76±0.67	92.88±0.33
LwTR (DA)	94.53±0.28	95.02±0.37	96.22±0.33	93.23±0.23	92.76±0.64	92.91±0.52
MR (DA)	94.44±0.29	94.88±0.24	**96.53±0.43**	**93.4±0.29**	**92.82±0.60**	92.74±0.42
SR (DA)	94.44±0.35	95.09±0.32	95.70±0.40	93.21±0.37	93.24±0.43	**93.06±0.39**
MELM (DA)	94.51±0.16	**95.15±0.34**	96.01±0.29	93.09±0.44	92.64±0.52	92.90±0.47

Methods	OntoNotes 5.0 (bn)					
	bc	bn	mz	nw	tc	wb
Baseline	**95.12±0.30**	97.23±0.20	95.83±0.45	95.29±0.27	93.62±0.59	93.13±0.40
TS	95.05±0.39	**97.38±0.17**	95.33±0.31	95.23±0.20	**93.96±0.51**	**93.25±0.29**
LS	94.99±0.22	97.32±0.20	95.60±0.22	95.11±0.37	93.49±0.43	92.90±0.47
MC Dropout	95.03±0.34	97.30±0.18	95.78±0.46	95.29±0.19	93.80±0.44	93.22±0.35
LwTR (DA)	94.36±0.54	97.29±0.14	95.74±0.16	95.15±0.20	93.64±0.51	93.08±0.49
MR (DA)	94.57±0.50	97.20±0.19	**96.27±0.31**	95.11±0.22	93.64±0.55	92.91±0.52
SR (DA)	94.76±0.65	97.28±0.15	95.85±0.33	95.30±0.17	93.78±0.63	93.06±0.24
MELM (DA)	94.34±0.47	97.24±0.21	96.18±0.32	**95.32±0.32**	93.51±0.50	92.97±0.48

Methods	OntoNotes 5.0 (tc)					
	bc	bn	mz	nw	tc	wb
Baseline	87.10±1.25	89.22±0.71	84.94±1.61	81.28±2.58	93.45±0.77	89.62±1.10
TS	**87.74±1.12**	89.45±0.47	85.95±1.65	82.50±1.35	93.11±0.98	89.93±0.88
LS	87.07±1.00	89.57±0.76	**86.67±1.75**	82.79±1.09	92.75±1.06	90.66±0.61
MC Dropout	87.25±0.73	89.02±1.08	85.12±1.62	81.95±2.56	93.36±0.89	90.05±0.84
LwTR (DA)	86.95±0.61	89.74±0.72	86.20±1.67	83.08±1.78	93.70±0.64	90.28±0.55
MR (DA)	86.78±1.12	**90.06±0.61**	86.36±1.64	**83.81±2.79**	**93.69±0.61**	**90.69±1.23**
SR (DA)	86.78±1.49	89.61±0.56	86.42±2.36	81.83±2.85	93.53±0.72	90.04±0.97
MELM (DA)	86.38±1.16	89.05±1.18	86.65±1.37	81.89±2.77	93.30±0.59	89.12±1.47

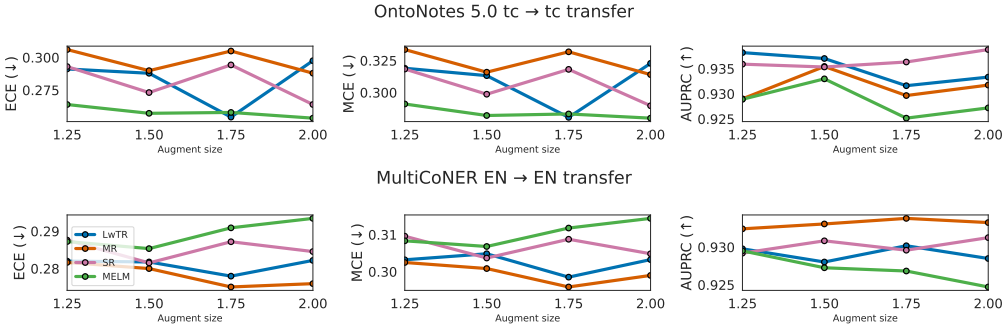

図 3　データ拡張サイズを増やした場合における各評価指標の推移.

A　OntoNotes 5.0 におけるジャンル横断設定の場合の AUPRC

表 4 に，OntoNotes 5.0 におけるジャンル横断設定の場合の AUPRC を示す.

B　分布内設定におけるデータ拡張サイズを増やした場合の各指標の推移

図 3 に，OntoNotes 5.0 tc → tc，MultiCoNER EN → EN の横断設定においてデータ拡張サイズを増やした場合の評価指標の推移を示す.

C　詳細な学習設定

損失関数の最適化に linear scheduler つき AdamW [19] を用いた．ミニバッチサイズは 32 であり，初期の学習率を 1e-5 とした．Early Stopping の停止基準に評価セットの F1 を用い，5 回連続でスコアが改善されない場合に学習を打ち切った.

TS の温度パラメータを調整するため Optuna [20] を用いた． [0.001, 0.002, ..., 5.000] の範囲で探索され，100 回の試行の中で評価セットの損失が最小となる温度パラメータが採用された． LS におけるスムージングパラメータの選択のため，[0.01, 0.05, 0.1, 0.2, 0.3] の範囲内でグリッドサーチを行った．LwTR，MR，および SR における二項分布のハイパーパラメータ選択では，[0.1, 0.2, ..., 0.8] の範囲内でグリッドサーチを行い，最適なものを選択した．MELM においては，エンティティ予測モデルの微調整のための学習データマスク率 η およびエンティティ予測時におけるマスクパラメータ μ の選択が必要なため，それぞれ [0.3, 0.5, 0.7] の範囲でグリッドサーチを行い，最適となる組み合わせを採用した．4 節におけるデータ拡張では，最終的な学習セットのサイズが元の学習セットのサイズの 1.5 倍になるように学習データを増加させた.