ラベルの不均衡を考慮した End-to-End 情報抽出モデルの学習

山口 泰弘　進藤 裕之　渡辺 太郎
奈良先端科学技術大学院大学 先端科学技術研究科
{yamaguchi.yasuhiro.yw2, shindo, taro}@is.naist.jp

1 はじめに

情報抽出とは自然言語で書かれた文書から構造化された情報を構築する技術である．一般的な情報抽出は，与えられた文から目的の用語を示すスパンを抽出する固有表現認識と，抽出されたスパンの間の関係を予測する関係抽出，スパンの共参照関係を予測する共参照解決など複数のタスクからなる．これらのタスクはこれまで別々に研究が進められていて，特に関係抽出や共参照解決ではタスクでは，文と対象のスパンが与えられたもとで関係を予測する手法が多く研究されている．情報抽出を行う際は，固有表現認識を行った結果を関係抽出モデルの入力とするパイプラインシステムとして処理される．こうしたパイプライン処理においては上流のタスクでの誤りが下流のタスクに伝播しまうことが問題となる．

そこで，最近ではスパンの抽出とスパンの間の関係予測を同時に 1 つのモデルで行うマルチタスク End-to-End 情報抽出の手法が検討されている [1][2][3]．これらの手法は次のような手順で処理される．まず，与えられた文から可能なすべてのスパンを列挙する．次に，列挙された各スパンについてスパン埋め込みを計算する．そして，固有表現認識の場合はスパン埋め込み，関係抽出や共参照解決の場合はスパンのペアの埋め込み表現に対して分類を行う．

先述した手法の問題点として，ラベルがつけられているスパン・スパンペア (正例) の数とラベルのないもの (負例) の数に大きな差があることが挙げられる．処理する文書のトークン数を N としたとき，可能なすべてのスパンの数は $N(N+1)/2$ となる．また，スパンペアの数は，すべてを評価した場合 $\{N(N-1)/2\}^2$ となる．一方，正例のスパンの数は多くの場合 N 以下であり，正例と負例の数に不均衡が生じる．先行研究では評価するスパンを定数 $L \leq N$ 以下の長さのみに限定していて，この場合スパンの数は $L(2N-L+1)/2$ であるが，依然として正例と負例の数には大きな差が生じる．表 1 に SciERC[1] データセットにおける 1 文中の正例と負例のスパンの統計を示す．スパン数は $L = 10$ とした場合に列挙されるスパンの数を計算したものである．正例の割合は，スパン数に対するラベル付けされたスパンの割合を表す．この統計から，実際のデータセットにおいても正例・負例の数が不均衡であることがわかる．

トークン数 N	正例数	スパン数	正例の割合 (%)
24.4 ± 1.74	3.01 ± 1.74	199 ± 101	3.01 ± 1.74

表 1　SciERC における不均衡

こうしたラベルの不均衡は，一般的に機械学習モデルの学習を妨げることが知られいる [4][5]．そこで本研究では，Under Sampling, Over Sampling, Hard Example Sampling の 3 つのサンプリング手法を用いて学習に利用するデータを選択することで，このラベルの不均衡の問題を解決することを試みた．

これらのサンプリング手法を用いて End-to-End 情報抽出モデルを学習した結果，Hard Example Sampling を用いた場合に最もモデルの予測性能の改善が見られた．

1.1 End-to-End 情報抽出モデル

本研究ではマルチタスク End-to-End 情報集出モデルとして、DyGIE[2][3] をもとにしたモデルをベースラインとして用いた．図 1 に処理の手順を示す．モデルは以下のようなステップで実行される．

スパンの列挙: 与えられた文中から，トークン数が $L(\leq N)$ 以下の可能な全てのスパンを列挙する．L はハイパーパラメータとして設定する．

スパン埋め込み: 列挙されたスパンに対応するベクトル表現を計算する．各トークンに対する埋め込み $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ からスパン (s_i, e_i) の始点と終点の埋め込みを選択し，それらを結合したベクトル $\boldsymbol{g}_i = [\boldsymbol{x}_{s_i}; \boldsymbol{x}_{e_j}]$ をスパン埋め込みとする．

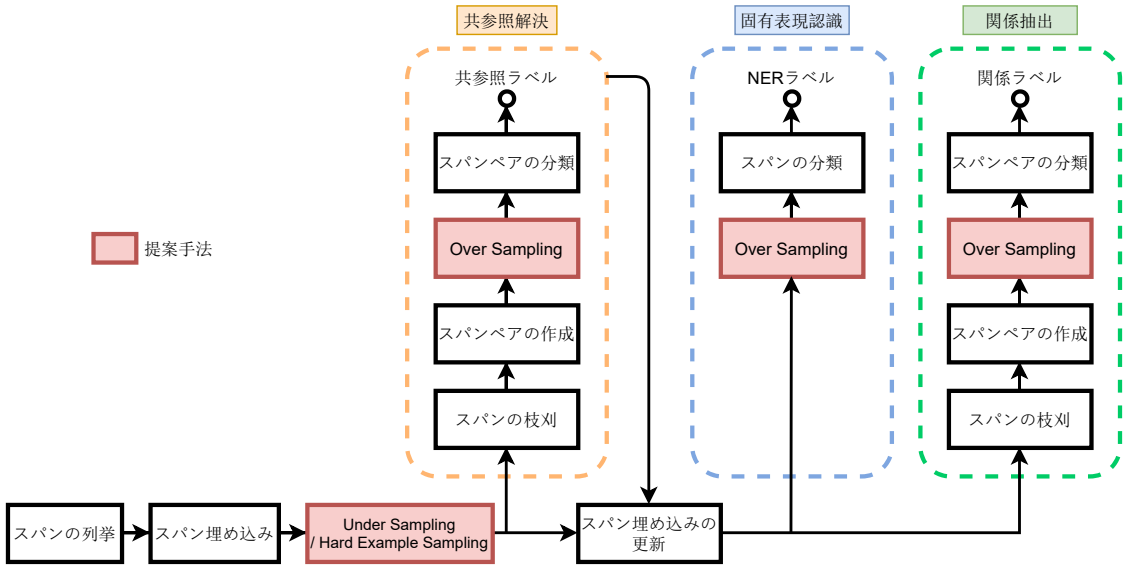

図 1　実験に利用したモデルの構造．ベースラインには赤枠の処理を除いたモデルを利用した．

スパンの枝刈: DyGIE[3][2] の手法に基づき，各スパンに対応するスコアを計算し，スコアの大きいものを選択する．選択するスパンの数は定数 γ を設定し，$\lceil \gamma N \rceil$ とする．

スパンペアの作成: 2 つのスパン埋め込み $\boldsymbol{g}_i, \boldsymbol{g}_j$ を結合し，スパンペアのベクトル $[\boldsymbol{g}_i; \boldsymbol{g}_j]$ を作成する．

スパン・スパンペアの分類: スパン，またはスパンペアのベクトル ($\boldsymbol{g}_i$ または $[\boldsymbol{g}_i; \boldsymbol{g}_j]$) を，順伝播型ニューラルネットワーク (FFN) を用いて各ラベルに対するスコアを計算する．

スパン埋め込みの更新: DyGIE[3][2] と同様に，予測した共参照関係に基づいてスパン埋め込みの更新を行う．

損失: 固有表現認識 (NER)，共参照解決 (CR)，関係抽出 (RE) の各タスクにおける交差エントロピー誤差 $\mathcal{L}_{\text{NER}}, \mathcal{L}_{\text{CR}}, \mathcal{L}_{\text{RE}}$ をそれぞれ計算する．モデルの学習ではこれらの誤差の和 $\mathcal{L}$ を最小化する．

$$\mathcal{L} = \mathcal{L}_{\text{NER}} + \mathcal{L}_{\text{CR}} + \mathcal{L}_{\text{RE}}$$

提案手法ではこのベースラインモデルにサンプリング処理を追加することでラベルの不均衡に対処することを試みた．

2　サンプリング手法

本研究では，先述のラベルの不均衡を考慮するために以下の 3 つのサンプリング手法を検討した．

2.1　Under Sampling

各ラベルの学習事例の数が偏っている場合に，多数派のデータはその一部のみを使って学習することで少数派とのデータサイズをバランスさせる手法である [4]．ここでは列挙されたスパンのうち，NER ラベルが付与されていないスパンからランダムに選択したスパンを学習に使用する．一方，ラベルが付与されたスパンはすべて学習に使用する．

2.2　Over Sampling

Over Sampling は少数派のデータを増加させる手法であり、様々な手法が提案されている。本研究では SMOTE[5] と呼ばれる手法を利用した．

SMOTE アルゴリズムでは，まず，増やしたいラベルの対象データをランダムに 1 つ選択し，その近傍 k 個を求める．そして，この k 個の近傍から 1 つを選んで，対象データとの間に同じラベルをもったデータを生成することでデータを増やす．

固有表現認識，共参照解決，関係抽出の各タスクごとに，スパン（あるいはスパンペア）の分類器に渡す直前の埋め込み表現 ($\boldsymbol{g}_i$ または $[\boldsymbol{g}_i; \boldsymbol{g}_j]$) に対して SMOTE アルゴリズムを用いて事例を増やす．ここではラベルのつけられたスパン・スパンペアに対して Over Sampling を行う．なお，ラベル間のデータ数の不均衡については考慮しないものとする．

2.3　Hard Example Sampling

予測の難易度にばらつきのあるデータを学習する際に用いられる手法の 1 つに Hard Example Mining (HEM)[6][7] がある．HEM では，学習時にモデルの予測スコアに基づいて学習の難しい事例を発見し，より困難な事例を用いてモデルの学習を行うことで

モデルの性能向上が期待できる．本研究では，このHEM の考え方をもとに，学習の難しい事例を選択する Hard Example Sampling を提案する．

可能なすべてのスパンを列挙した場合，実際にラベルがつけられたスパンとオーバーラップするスパンが複数生成される．“This paper reports on two contributions to [large vocabulary continuous speech recognition] .” という例文では．括弧で囲まれた範囲が実際のスパンである．ことのき，列挙されるスパンの中には “continuous speech recognition”, “continuous speech”, “contributions to large vocabulary” といった実際のスパンとオーバーラップしたものが含まれる．こうしたオーバーラップしたスパンは互いに共通する部分文字列を持ち，似通っているため分類が困難になと考えられる．

HES を用いることで，ラベルの不均衡を改善するとともに，こうした予測の困難な事例に対する予測性能の改善が期待できる．

3 実験

提案した各サンプリング手法を用いてモデルを学習し，固有表現認識，共参照解決，関係抽出の各タスクにおける予測性能の比較を行った．

3.1 データセット

学習と評価には SciERC[1] を利用した．このデータセットは 500 本の科学論文のアブストラクに対して固有表現認識，共参照解決，関係抽出のためのアノテーションを行ったデータセットである．このデータセットのうち，学習データと検証データを用いてモデルの学習とハイパーパラメータの決定を行い，テストデータで各手法の性能を比較した．

3.2 モデルの設定

各トークンの埋め込みを得るために，科学論文で事前学習された BERT モデルである SciBERT[8] を利用した．対象の文を SciBERT に入力し，得られたトークンの埋め込みを元にスパン埋め込みを算出した．学習時には SciBERT のパラメータの更新を行った．

スパンの列挙では，$L = 10$ として，トークン数が 10 以下のすべてのスパンを予測の対象とした．また，スパンの枝刈りにおいては共参照解決と関係抽出の両方で $\gamma = 0.5$ を設定した．

固有表現認識，共参照解決，関係抽出の各タスクでは，スパン・スパンペアの埋め込みを 2 層の順伝播型ニューラルネットワークを通して各ラベルの予測スコアを計算した．各層の次元はすべて 150 とした．

	Base	US	OS	HES	HES+OS
固有表現認識					
F1	71.4	71.1	68.9	**71.7**	70.4
Precision	70.3	69.7	62.4	**71.2**	67.2
Recall	72.6	71.7	**76.9**	72.2	73.1
共参照解決					
F1	54.9	52.6	52.7	**55.5**	53.3
Precision	**69.0**	61.9	57.9	67.2	62.3
Recall	45.7	45.5	**47.8**	47.3	45.9
関係抽出					
F1	45.3	46.8	46.2	**48.0**	46.9
Precision	**53.8**	51.0	43.8	52.1	45.3
Recall	39.1	43.3	**48.8**	44.6	47.6

表 2 実験結果

3.3 サンプリングの設定

Under Sampling と Hard Example Sampling では，正例全てに加えて，バッチの繰り返しごとに負例を最大で 50 個ランダムに選択して学習に利用した．また，Over Sampling では $k = 10$ として近傍を計算し，SMOTE アルゴリズムを用いて各文に対して 10 個の正例を増やした．

3.4 評価

各タスクの性能はテストデータを用いて F1, Precision Recall を計算した．関係抽出タスクでは，選択されたスパンの範囲と予測された関係ラベルのみに基づいて各スコアを計算した．したがって関係予測の評価においては選択されたスパンの NER ラベルの予測の正しさについて考慮しないものとした．

4 結果と考察

実験の結果を表 3.4 に示す．手法 Base, US, OS, HES, HES+OS はそれぞれ，ベースラインモデル，Under Sampling, Over Sampling, Hard Example Sampling, HES と OS の併用，を表す．すべてのタスクの F1 スコアにおいて，HES が最も高い精度となった．OS は固有表現認識，共参照解決のタスクにおいてベースラインの F1 スコアを下回った．一方，Recall を見るとすべてのタスクにおいて OS が最も高い結果なった．また，Base と OS, HES と HES+OS をそれぞれ比較すると，OS を用いた場合に Precision が低

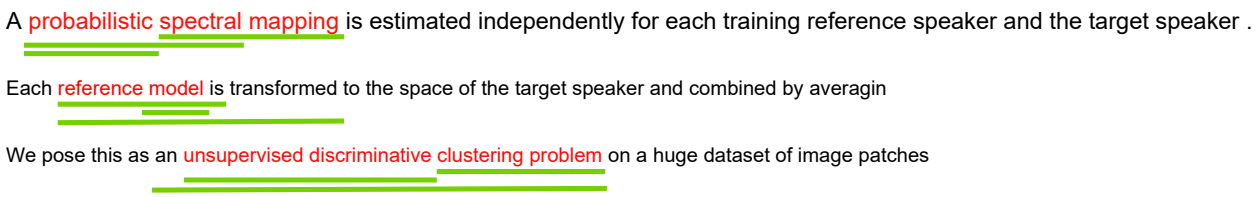

図 2　スパンとその近傍

下し，Recall が上昇する傾向がみられた．

4.1　スパン埋め込みの近傍

Over Sampling では近傍の事例を用いてデータを生成するため，サンプリング対象となる事例の近傍の状態がモデルの学習やその性能に影響を及ぼすと考えられる．そこで，SciBERT から計算したスパン埋め込みを用いて，ラベルのつけられたスパンの埋め込みの近傍となるスパンを分析した．

表 2 に正例とその近傍のスパンの例を示す．赤で示された範囲が正例のスパンであり，下線で示したものが最近傍のスパンの範囲である．近傍のスパンを見ると，その多くが正例とのオーバーラップを含んでいることがわかる．SMOTE アルゴリズムを利用してデータを増やした場合，こうしたまぎらわしい負例との間に正例データが生成されるため，正例と負例の識別が困難になったと考えられる．そして，まぎらわしい事例に対してラベルを付与してしまうことで，Over Sampling を用いた手法では Recall は高く，Precision は低くなるという結果になった．

一方，Hard Example Sampling ではオーバーラップを含むスパンを学習に使用し，正例と負例の境界付近の事例を学習したことで性能の向上に繋がったと考えられる．

4.2　サンプリングサイズの影響

サンプリングサイズの違いによる予測性能への影響を調べるために，異なるサンプリングサイズでモデルを学習し，予測性能の比較を行った．表 4.2 に HES における比較を示す．HES 10, HES 50, HES 100 はそれぞれ負例のサンプリングサイズを 10, 50, 100 として学習を行ったものである．

表 4.2 から，HES 50 が全てのタスクにおいて最も高い F1 スコアとなった．一方，HES 10 ではベースラインと比較して固有表現認識と共参照解決においては F1 スコアが下回る結果となった．この結果から，サンプリングサイズが小さすぎる場合には事例が不足して十分な学習が行われず，HES を用いても予測性能は向上しないと考えられる．また，HES 50 と HES 100 では HES 50 の F1 スコアが高いことから，学習に必要な事例数と正例・負例のバランスを調整する必要があることが示唆される．

	Base	HES 10	HES 50	HES 100
固有表現認識				
F1	71.4	69.9	**71.7**	71.5
Precision	70.3	68.9	**71.2**	70.4
Recall	**72.6**	70.1	72.2	72.4
共参照解決				
F1	54.9	53.2	**55.5**	53.6
Precision	**69.0**	61.1	67.2	65.7
Recall	45.7	44.2	**47.3**	45.3
関係抽出				
F1	45.3	45.8	**48.0**	47.3
Precision	**53.8**	50.2	52.1	52.5
Recall	39.1	43.1	**44.6**	43.3

表 3　サンプリングサイズごとの予測性能

5　おわりに

本研究では，マルチタスク End-to-End モデルの学習におけるラベルの不均衡の問題を解決するために，Under Sampling, Over Sampling, Hard Example Sampling の 3 つのサンプリング手法による学習を検討した．実験の結果，Hard Example Sampling を用いた場合に，固有表現認識，共参照解析，関係抽出の全てのタスクにおいて F1 スコアの改善が見られた．一方，SMOTE による Over Sampling ではまぎらわしい正例が生成されることで Precision の低下と Recall の上昇が確認された．そして，Hard Example Sampling においてはサンプリングサイズが予測性能に影響を与えていて，予測性能を向上させるためには適切なサンプリングサイズを選択する必要があることが示唆された．

また，本研究で用いたようなスパンとその関係を同時に予測するモデルは，情報抽出以外のタスクにおいてもその利用が検討されている [9]．そこで，今後は他のタスクにおいてもこの提案手法による予測性能への影響を検証したい．

参考文献

[1] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction. In *Proc.\ Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2018.

[2] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, Relation, and Event Extraction with Contextualized Span Representations. 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. 5784–5789, Hong Kong, China, nov 2019. Association for Computational Linguistics.

[3] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 3219–3232. Association for Computational Linguistics, 2020.

[4] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186. Morgan Kaufmann, 1997.

[5] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, Vol. 16, pp. 321–357, 06 2002.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 9, pp. 1627–1645, 2010.

[7] Abhinav Shrivastava, A. Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, 2016.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. 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. 3613–3618, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. Generalizing Natural Language Analysis through Span-relation Representations. pp. 2120–2133. Association for Computational Linguistics (ACL), jul 2020.