アノテーション漏れ推定を用いたエンティティ抽出

伊藤雅弘　　山崎智弘
株式会社東芝　研究開発センター
知能化システム研究所　アナリティクスＡＩラボラトリー
{masahiro20.ito, tomohiro2.yamasaki}@toshiba.co.jp

1 はじめに

近年の団塊世代の退職と少子化による労働人口の減少に伴い，産業界では業務に関するノウハウの喪失や継承が関心事となっている．なぜなら，電力・化学プラントなどの建設・運用・保守には，ベテランの長年の経験が重要となるからである．我々は，過去に発生したトラブルの報告書に記録されているベテランのノウハウを，現状の業務や新人の教育に用いる取り組みを進めている．特に，過去の大量の報告書からトラブルに関係するイベント(「配管に亀裂」「水位が低下」など）を抽出し，因果関係として構造化することによって,「特定の原因により引き起こされるリスク」や「トラブルを引き起こした原因」を分析する技術開発に注力している．

我々の構造化プロセスは，まず報告書からトラブルの原因や結果になりうるイベントを系列ラベリング [1] にて抽出し，次に抽出されたイベント間に因果関係があるかを推定する．この 2 段階のプロセスでは，前段でイベントの抽出漏れがあると，後段で正しい因果関係を推定することができない．そのため，前段のイベント抽出はある程度の正しさは担保しつつもなるべく多くのイベントを後段に提供できること，つまり再現率が高いことが望まれる．

再現率を高めるためには系列ラベリングにおける「学習時の目的関数を工夫する」「イベントを示すラベルかどうかの推定値に対する閾値を変更する」なども考えられる．これら方法では，学習データ自体はそのまま用いている．一方で，本研究では学習の情報源となるアノテーションデータのエンティティ付与漏れ（アノテーション漏れ）に着目する．ここでエンティティとは，系列ラベリングでの抽出対象となるイベントや固有表現などが該当する．一般的に，学習データに用いることができる規模で完全に正しいアノテーションデータを作成することは困難である．既存研究 [2] でも，固有表現抽出のデータセットである CoNLL2003 [3] におけるアノテーション誤りが指摘されている．我々もイベント抽出の再現率低下の要因分析をした結果，トラブル報告書内のイベントにアノテーション漏れが存在することが分かっている．エンティティに対するアノテーション漏れがあると，一部のエンティティの特徴を学習できず抽出漏れの要因になると考えられる．

そこで本研究では，アノテーション漏れを推定しそのエンティティを学習データに追加することによって，系列ラベリングを用いたエンティティ抽出の再現率を向上させる方法を提案する．

2 関連研究

これまでの因果関係抽出の研究としては [4, 5] などがあるが，その中でのイベント抽出手法には，主にルールベースと系列ラベリングによる機械学習ベースがある． [4] はルールベースの手法で，手がかり表現を含む文節を基点に，係り受け解析を用いて前後の一定の範囲を原因・結果として抽出する．一方 [5] は系列ラベリングを用いた手法である．文を形態素などのトークンに分割し，1 文中に存在する原因・結果それぞれのイベント範囲の各トークンに対し Cause, Effect のラベルを付与した学習データを用意する．そして，未知の文における原因イベント・結果イベント部分のラベルを推定するモデルを学習する．本研究では，この系列ラベリングを用いてイベントを抽出する．

系列ラベリングの手法としては近年 Bi-LSTM ベースの手法が盛んに提案され，高い性能を達成している [1, 6]. これらの手法は，文字や単語の埋め込みベクトルを Bi-LSTM に与え，最終層で CRF を用いてラベルを推定している．一方 [2] は，アノテーション誤りを含んだ不完全なデータセットにおいて，誤りを含む可能性が高い文を推定し，その文の重みを低くして学習することで，系列ラベリングの性能を向上させる CrossWeigh という手法を提案している． [7] も同様に不完全なデータセットでの学習性能の向上を取り扱っている．しかし，この研究はラ

ベル無しデータに対してルールや辞書に基づき機械的にラベリングしたものを不完全（ノイジー）なデータとして扱っているため，アノテーションデータに誤りが含まれるという [2] や本研究の問題設定とは異なる．

3 提案手法

本研究では，前節で述べた CrossWeigh [2] と同様にアノテーション誤りを含んだ学習データを取り扱う．先に述べたように，機械学習用のデータセットに対してアノテーション作業を実施したとしても，完全に正しいデータを作成することは困難である．これは，特に定義が明確にできないエンティティに対しては顕著であると考えられる．例えば，本研究で取り扱う「トラブル報告書におけるトラブルの原因・結果となるイベント」は，「人名」に比べて明確に定義することが困難で系列も長い．その場合，エンティティの過不足やテキスト範囲の揺れが大きくなる．

[2] は学習データからアノテーション誤りを含む可能性があるデータ（文を構成するトークンの系列とそのラベル）を推定し，文単位で学習時の重みを低くする．つまり当該文の学習への影響を小さくしている． [2] は誤った情報を取り除くため，主に適合率を向上させる効果が高いと考えられる一方，再現率の向上はあまり見込めない．また予備実験において，エンティティのテキスト範囲（エンティティ範囲）の揺れが比較的大きな”イベント”という対象では学習データの多くの文がアノテーション誤りと推定されてしまい，逆に性能低下を引き起こすことが分かっている．

先に述べたように，エンティティ抽出結果を後段の関係抽出処理で用いる場合は，そもそも正しいエンティティが抽出されていなければ正しい関係を生成することができない．そのため，適合率より再現率を重視した性能向上手法が必要となる．そこで本研究では，アノテーション誤りのうちアノテーション漏れと考えられるエンティティを推定し，そのエンティティを学習データに追加するという再現率向上を重視した手法を提案する．これは，誤った情報を取り除く [2] とは逆のアプローチである．

学習データにエンティティを追加することで，当然再現率の向上が想定される．しかし，追加するエンティティを注意深く選定しなければ，ノイズとなるエンティティが増え適合率が大きく下がると考えられる．提案手法では，以下の 3 つの工夫を導入することにより，適合率の低下をなるべく抑制しつつ，再現率の向上を目指す．

(A) 学習データの異なるサブセットを用いて学習した複数モデルによる投票によって，推定エラーではなくアノテーション漏れのみを学習データに追加できる確率を高める．これは，ノイズとなるエンティティが追加されることを抑制する．

(B) (A)において投票数に応じて追加するエンティティを選定する際，範囲が類似するエンティティを「エンティティグループ」としてまとめて扱うことにより，エンティティ範囲の厳密性にこだわらず，追加すべきエンティティ周辺を浮かび上がらせる．

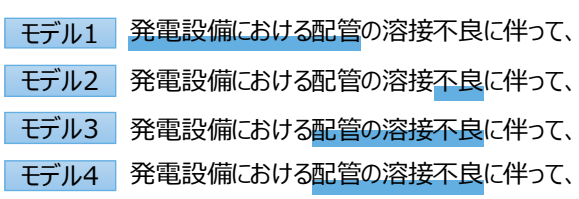
(C) (A)での投票結果を用いて，エンティティグループに属する推定範囲の異なるエンティティから，より正しいものを選択できる確率を高める．これは，追加されるエンティティ範囲のノイズ除去として機能することが期待される．

以下に提案手法のアルゴリズムを示す．

① 学習に用いるアノテーションデータをランダムに並べ替え，k 分割交差検証にて上記アノテーションデータのすべての文からエンティティを抽出する．

② 抽出されたエンティティのうち，アノテーションされていなかったエンティティを記録する．

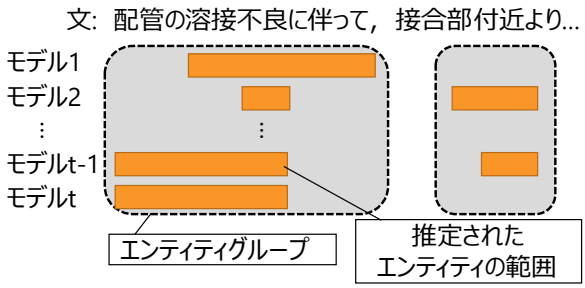
③ ①～②を **t** 回繰り返す．
例えば，t=4 の場合，下図のように 4 つの異なるモデルがそれぞれエンティティ範囲を推定する．



④ 推定されたエンティティ範囲が同一のものの個数を，各エンティティ範囲の「投票数」としてカウントする．

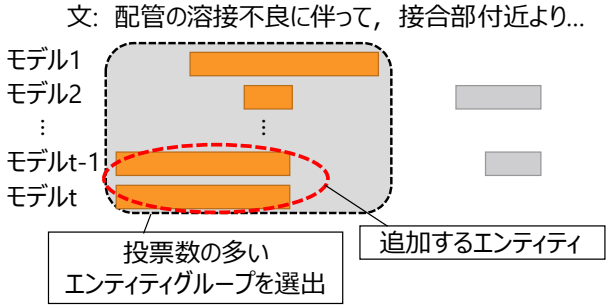
⑤ 範囲が類似するエンティティを「エンティティグループ」としてまとめ，エンティティグループに属するエンティティの投票数を合計する．範囲が類似するエンティティの判定方法は，「1

形態素でも重複する」や「末尾 n 形態素のいずれかが重複する」などが考えられる．例えば，下図のように類似する範囲をまとめ上げたエンティティグループが作成される．



⑥ 投票数が $t \cdot \alpha$ $(0 \leq \alpha < 1)$ 以上のエンティティグループを選出する．α は，同じ文を t 回異なるモデルで推定したうち，どのぐらいの割合の投票数があれば追加するエンティティグループとして選出するかの基準である．

⑦ 選出されたエンティティグループそれぞれに対し，属する複数のエンティティの中で最も投票数の多いエンティティを学習データに追加する．



4 評価実験

提案手法の効果を検証のために実施した実験について述べる．パラメータは以下の通りとした．

k（交差検証における分割数）：10

t（交差検証の繰り返し回数）：20

α（エンティティグループの選出基準）：

0.8，0.6，0.4，0.2，0.0

α が 0.0 の時にすべてのエンティティグループが学習データに追加され，α が 1.0 の時にいずれのエンティティも追加されない．

提案手法における範囲が類似するエンティティの判定方法は「エンティティ末尾 5 形態素のいずれかが重複する」とした．例えば「発電/設備/の/配管/に/亀裂」と「配管/に/亀裂」は類似と判定される．

表 2　データ数（電力プラントトラブル報告書）

	文書数	文数	イベント数
学習データ	715	7,939	8,887
評価データ	40	515	498

これは，日本語のトラブルに関するイベントにおいて「水漏れ」「亀裂」など主要な現象（動作）が末尾付近に出現することが多いため，その部分の抽出を重視し，それ以外のトラブルの対象・条件などの記述の抽出範囲の揺れを許容するためである．なおこの方法では，5 形態素以下のエンティティ同士では「1 形態素でも重複する」と等しくなる．

系列ラベリングモデルは Bi-LSTM-CRF [1] をベースとし，学習済み BERT モデル [8, 9] から得られた単語ベクトルを Bi-LSTM への入力として与えた．オリジナルの学習データおよび提案手法によりエンティティが追加された学習データを用いて 3 回学習し，それぞれの抽出性能の平均値を算出した．

上記設定で実施した以下 2 種類の抽出タスクによる評価について述べる．

4.1 トラブルイベント抽出による評価

電力プラントに関するトラブル報告書（日本語）を用いたトラブルイベント抽出による評価実験を行った．この報告書には，トラブルに関わるイベントとその因果関係がアノテーションされており，データ数は表 1 の通りである．また，性能値算出時の「エンティティが抽出できたか」の判定はエンティティ範囲の完全一致が通常であるが，本実験では先述したイベントの特性から「エンティティ末尾 5 形態素のいずれかが重複する」とした．

実験結果を表 2 に示す．提案手法によりエンティティが追加された学習データを用いた結果，オリジナルのアノテーションデータを用いた場合と比べ，再現率が向上した．一方，エンティティを追加することによって想定通り適合率は下がった．しか

表 1　評価結果（電力プラントトラブル報告書）

		適合率	再現率	F 値
オリジナル		0.861	0.875	0.867
提案手法	α=0.8	0.852	0.885	0.868
	α=0.6	0.842	**0.907**	**0.873**
	α=0.4	0.832	0.909	0.869
	α=0.2	0.815	0.913	0.862
	α=0.0	0.802	0.923	0.859

表 3　データ数（CoNLL2003）

	文数	エンティティ数
学習データ	18,453	29,441
評価データ	3,684	5,698

し，パラメータ α を高くすることによってその程度が緩和される傾向であることから，(A)～(C)の工夫がノイズ除去として機能していることが確認できた．この再現率向上効果とノイズ除去効果のバランスは，データセットやモデルにより変化すると考えられる．本実験では α=0.6 で最も高い F 値となっており，提案手法により F 値を維持もしくは向上させつつ，再現率を向上できることを確認した．

4.2 固有表現抽出による評価

提案手法は，本研究におけるイベントのようにアノテーション揺れが大きい対象に特に有効であると考えられる．一方，一般的な固有表現抽出タスクへの有効性は明らかでないため，広く用いられている CoNLL2003 コーパス（英語）による評価を実施した．当該コーパスには，人名，地名，組織名，その他の 4 種の固有表現がタグ付けされており，データ数は表 3 の通りである．エンティティが抽出できたかの判定はエンティティ範囲の完全一致とした．

実験において，アノテーション漏れを疑似的に再現するため，学習データから個々のエンティティを一定確率（削除確率）で削除した．例えば，削除確率 0.1 だとおおよそ 10%のエンティティが学習データから削除される．削除確率 0.1～0.5 の学習データで実験した結果を表 4 に示す．削除確率の増加に伴って再現率が低下していることが分かる．

まず，削除確率 0.1 のデータを用いて提案手法を適用した結果を表 5 に示す．パラメータ α を下げるに伴って適合率が低下し再現率が向上する傾向は，先述のトラブル報告書における評価と同様である．そして，α=0.4 で F 値が最大となっており，エンティティを削除していないオリジナルの学習データを用いた結果と比べても同等以上となっている．

表 4　エンティティ削除データでの性能

削除確率	適合率	再現率	F 値
削除なし	0.908	0.890	0.899
0.1	0.913	0.881	0.897
0.2	0.922	0.856	0.888
0.3	0.935	0.812	0.869
0.4	0.951	0.727	0.824
0.5	0.965	0.490	0.650

表 5 評価結果（CoNLL2003, 削除確率 0.1）

		適合率	再現率	F 値
削除確率 0.1 データ		0.913	0.881	0.897
提案手法	α=0.8	0.908	0.889	0.898
	α=0.6	0.907	0.890	0.898
	α=0.4	0.909	**0.894**	**0.901**
	α=0.2	0.905	0.895	0.900
	α=0.0	0.895	0.893	0.894

次に，削除確率 0.3 のデータを用いて，提案手法を適用した結果を表 6 に示す．ここでもパラメータ α に対する適合率・再現率の傾向は同様であるが，適合率低下より再現率の向上効果が高いため，推定されたすべてのエンティティグループを追加対象とする α=0.0 で F 値が最大となった．つまり，削除確率 0.3 のデータにおいては，提案手法で述べたエンティティグループを用いたノイズ除去の工夫は必要ではないことを示している．

5 おわりに

本研究では，エンティティのアノテーション漏れを含む不完全なデータセットにおいて，アノテーション漏れのエンティティを推定しデータセットに追加して学習させるという，再現率を重視したエンティティ抽出の性能改善手法を提案した．評価実験により，提案手法がエンティティ追加によるノイズを抑えつつ, 再現率の向上を実現できることを示した．

一方，パラメータ α はデータセットの条件により最適値が異なるため，今後提案手法が効果を発揮するようなデータセットの品質とパラメータ α の組み合わせ条件を調査する．また，アノテーション誤りの可能性のある文の重みを下げる手法 [2] と組み合わせ，適合率・再現率の両面から性能向上させるアプローチも検討する．

表 6 評価結果（CoNLL2003, 削除確率 0.3）

		適合率	再現率	F 値
削除確率 0.3 データ		0.935	0.812	0.869
提案手法	α=0.8	0.924	0.845	0.883
	α=0.6	0.920	0.857	0.888
	α=0.4	0.917	0.866	0.891
	α=0.2	0.911	0.871	0.890
	α=0.0	0.903	**0.880**	**0.892**

参考文献

1. **Lample, Guillaume, et al.** *Neural Architectures for Named Entity Recognition.* : Proceedings of NAACL 2016, 2016.

2. **Wang, Zihan, et al.** *CrossWeigh: Training Named Entity Tagger from Imperfect Annotations.* : Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

3. **Tjong Kim Sang, Erik F. and De Meulder, Fien.** *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.* : Proceedings of CoNLL-2003, 2003.

4. **坂地泰紀, 酒井浩之 , 増山繁.** 決算短信 PDF からの原因・結果表現の抽出. : 電子情報通信学会論文誌 D, 2015.

5. **Dasgupta, Tirthankar, et al.** *Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks.* : Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018.

6. **Akbik, Alan, Blythe, Duncan and Vollgraf, Roland.** *Contextual String Embeddings for Sequence Labeling.* : COLING 2018, 27th International Conference on Computational Linguistics, 2018.

7. **Hedderich, Michael A. and Klakow, Dietrich.** *Training a Neural Network in a Low-Resource Setting on Automatically.* : Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, 2018.

8. (オンライン) 2021 年 1 月 14 日. https://github.com/cl-tohoku/bert-japanese.

9. (オンライン) 2021 年 1 月 14 日. https://huggingface.co/bert-base-uncased.