科学技術論文からの情報抽出のための ラベルなしドメイン内データの活用

加藤 明彦　　進藤 裕之　　松本 裕治
奈良先端科学技術大学院大学 先端科学技術研究科
{kato.akihiko.ju6, shindo, matsu}@is.naist.jp

1 はじめに

科学技術論文から情報抽出を行うモデルの学習データとして，固有表現抽出（NER），関係抽出（RE），共参照解析（COREF）の注釈（ 図 1）を付与したコーパス [2] は重要である．しかし科学ドメインにおいて，上記 3 種全ての注釈を付与した既存のデータセットは比較的小規模である [7]．

そこで本稿では，科学ドメインの大規模ラベルなしデータから言語モデルを通じて，上記の解析タスク群（NER/RE/COREF）にとって有用な特徴量を得る手法について検討する．具体的には，(1) 事前学習された ELMo [8] の fine-tuning，(2) Entity を考慮した言語モデル（Entity-aware LM）の学習 の 2 手法を検討対象とする（ 4 章）．

実験の結果，entity mention のタイプ推定，関係抽出，共参照解析のマルチタスク学習 [7] で，ラベルなしデータから得た特徴量が有用であることを確認した．

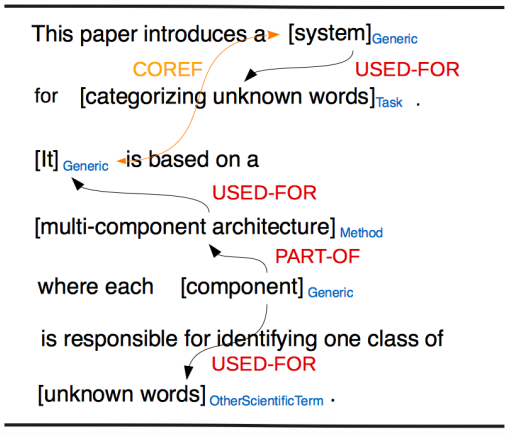

図 1: 科学論文からの情報抽出タスクのアノテーションの例．

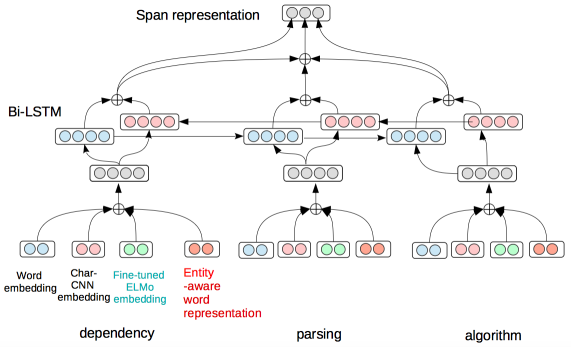

図 2: 各 entity mention の表現ベクトルの算出に関するモデルアーキテクチャー．双方向 LSTM への各時刻の入力には．言語モデル由来の特徴量（ 4.3）が含まれる．図には 表 1 の内，Fine-tuned ELMo + Entity-aware LM の場合を示している．

2 タスク定義

本稿では，Singh ら [9] のタスク設定に基づき，科学論文からの情報抽出タスクを以下の様に定式化する．

入力 文書および文書中の entity mention の範囲

出力
- 各 entity mention のタイプ（entity type tagging）
- 各文内の entity mention 間の関係の有無および relation type（関係抽出）
- 文書中の各 entity mention の先行詞（共参照解析）

3 モデル

本稿では，Luan ら [7] のマルチタスク学習に基づくモデル（SciIE）を用いる（ 図 2）．SciIE では，入力文書中の各文が別々の双方向 LSTM に入力され，各時刻の出力をもとに，各 entity mention の表現ベク

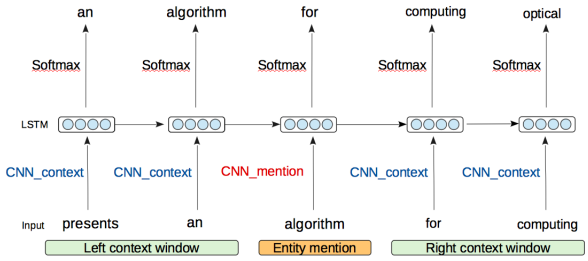

図 3: Entity を考慮した言語モデル．図では順方向 LSTM のみ示している．逆方向 LSTM も同様のアーキテクチャーである．

モデル	入力特徴量
Pre-trained ELMo（ベースライン）	(a), (b), (c)
Fine-tuned ELMo	(a), (b), (d)
Pre-trained ELMo + Entity-aware LM	(a), (b), (c), (e)
Fine-tuned ELMo + Entity-aware LM	(a), (b), (d), (e)

表 1: モデル一覧と入力特徴量（ 4.3）

トル（スパン表現）が Lee ら [6] の手法に基づいて算出される．ここで entity mention とは，entity に関する文書中の個別の言及である[1]．各 entity mention のスパン表現は，各タスク専用の多層パーセプトロン（MLP）群に入力され，これらの MLP からの出力を用いて，各タスクの正規化前のモデルスコア $\Phi_E(e, s_i), \Phi_R(r, s_i, s_j), \Phi_C(s_i, s_j)$ が計算される．ここで $\Phi_E(e, s_i)$ はスパン s_i が entity type e を持つ事象，$\Phi_R(r, s_i, s_j)$ は，スパンの対（s_i,s_j）が relation type r を持つ事象，$\Phi_C(s_i, s_j)$ は，スパンの対（s_i,s_j）が共参照関係にある事象のスコアである．各タスクのソフトマックス層はこれらのスコアを受け取り，正規化したモデルスコアを出力する．目的関数は，各タスクの負の対数尤度の重み付き和である [7]．

4 言語モデル

Luan らの手法 [7] には学習データが比較的小規模であるという問題点がある．事前学習した言語モデルの下流タスクでの有用性 [8, 3] を考慮して，Luan らは入力特徴量を生成するために ELMo [8] の事前学習済モデルを利用しているが，科学ドメインのデータは事前学習に用いていない．そこで本章では科学ドメインの大規模ラベルなしデータを利用した 2 種類の言語モデルの学習と，科学技術論文からの情報抽出における言語モデル由来の特徴量の利用方法について述べる．

4.1 事前学習された ELMo の fine-tuning

第一に事前学習された ELMo [8] の fine-tuning について述べる．まず，注釈付きコーパスから entity mention の辞書を作成する．次に，科学ドメインの大規模ラベルなしデータから辞書マッチングで少なくとも 1 つ entity mention を含む文を収集する．そして，収集したドメイン内ラベルなしデータで ELMo の事前学習済モデルを fine-tuning する．

4.2 Entity を考慮した言語モデルの学習

第二に entity を考慮した言語モデルについて述べる．近年，下流タスクでの有用性が報告されている言語モデル [8, 3] のほとんどは，ELMo（ 4.1）も含め，entity を考慮していない．すなわち，これらの言語モデルにはトークンの系列のみが入力され，文中の entity mention の範囲に関する情報は与えられない．

我々は，前者に加えて後者も言語モデルに入力として与えることで，関係抽出や共参照解析に有用な特徴量を言語モデルから得られると期待し，entity を考慮した文脈依存の単語表現を得ることのできる言語モデル（Entity-aware LM）を提案する．

Entity-aware LM（ 図 3）は，ELMo をベースとした双方向 LSTM であるが，entity mention と周辺文脈がそれぞれ異なるパラメーターで符号化されるという点で ELMo とは異なる．両者の違いを詳しく述べると，ELMo では，トークンの表現ベクトルは文字レベルの畳み込みニューラルネットワーク（Char-CNN）[5] によって算出されるのに対し，Entity-aware LM では，トークンが entity mention に属するかどうかによって，別々の Char-CNN を用いた符号化を行う．

Entity-aware LM の学習データは以下の手順で作成する．まず，ラベルなしデータから辞書マッチングで少なくとも 1 つ entity mention を含む文を収集する．次に各文の各 entity mention を中心とする左右 k 単語（k はハイパーパラメーター）を切り出して訓練事例とする（ 図 3）．

4.3 言語モデル由来の特徴量の利用

Luan ら [7] は，双方向 LSTM の各時刻での入力特徴量（ 図 2）として，(a) 単語分散表現，(b) 文字レベルの CNN から求めたトークン表現，(c) 事前学習

[1] 図 1 では，“system” などが entity mention に相当する．

したELMoから求めたトークン表現 の3者を用いている．一方，ラベルなしデータから学習した言語モデル由来の特徴量としては，(d) fine-tuningしたELMoから求めたトークン表現 および (e) entityを考慮した言語モデルから求めたトークン表現 が利用可能である．そこで本稿では，これらの特徴量の利用方法として， 表1に示した4パターンを検討する．

5 実験

5.1 実験設定

5.1.1 データセット

ラベル付きデータセットとしては，SciERC [7] を用いる．SciERCは人工知能分野の500本の論文のアブストラクトに対して固有表現抽出，関係抽出，共参照解析の注釈を付与したコーパスである．データセットの訓練/開発/テストへの分割はSciERCのものに従った．言語モデルの学習に利用するラベルなしデータとしてはSemantic Scholar コーパス [1] を用いる．このコーパスは，計算機科学，神経科学，医学生物学分野の約3900万本の論文から構成されている．

5.1.2 ハイパーパラメーターとモデル選択

バッチサイズは10文とし，最適化にはAdamを用いた．学習率は0.001とした．入力特徴量（ 4.3）については，(a) 単語分散表現 を300次元，(b) 文字レベルのCNNから求めたトークン表現 を150次元とした．言語モデル由来の特徴量 (c)-(e) は各1024次元とした．文をエンコードする双方向LSTMは1層200次元，各タスクの多層パーセプトロンは2層150次元とした．また，目的関数において各タスクの負の対数尤度を足し合わせる際の重みについて，以下の2つの設定で実験を行った．

Best_RE $(\lambda_{ETT}, \lambda_{RE}, \lambda_{COREF}) = (0.05, 1.0, 0.3)$

Best_COREF $(\lambda_{ETT}, \lambda_{RE}, \lambda_{COREF}) = (0.33, 0.33, 0.33)$

Best_RE設定では，関係抽出について，開発セットで最大のF値をもたらすモデルを用いてテストセットでの評価を行った．Best_COREF では，共参照解析について同様のモデル選択を行った．

5.2 事前学習されたELMoのfine-tuning

Entity mentionの辞書は，SciERCコーパス [7] のアノテーションから作成した．Semantic Scholar コーパス [1] に対する辞書マッチングによって，少なくとも1つentity mentionを含む450万文を収集し，これらを用いて，Petersら [8] が配布しているELMoの事前学習済モデルを10エポック，fine-tuningした．この結果，得られた言語モデルのパープレキシティは，SciERCコーパスの開発セットで30.19となった．

5.3 Entityを考慮した言語モデルの学習

ELMoのfine-tuningに用いたものと同一の，少なくとも1つentity mentionを含む450万文から作成した訓練事例を用いて，Entity-aware LMの学習を10エポック行った．Entity mentionの左右の文脈窓はk=7単語とした．ネットワークパラメーターの初期値としては，ELMoの事前学習済モデルの対応するパラメーターの値を利用した．この結果，得られた言語モデルのパープレキシティは，SciERCコーパスの開発セットで56.18となった．

5.4 実験結果

まずBest_RE設定での関係抽出のF値（ 表2）に着目すると，各提案手法は開発/テストの双方でPre-trained ELMoを上回った．また，Fine-tuned ELMoはPre-trained ELMo + Entity-aware LMよりも開発セットで+1.76ポイント，テストセットで+1.52ポイント高い精度をもたらした．Fine-tuned ELMoをEntity-aware LMと組み合わせた結果，開発セットで+0.48ポイント，テストセットで+0.74ポイント，さらに精度が向上した．

Best_COREF 設定での共参照解析のF値（ 表3）についても，各提案手法は開発/テストの双方でPre-trained ELMoを上回った．Fine-tuned ELMoはPre-trained ELMo + Entity-aware LMよりも開発セットで+2.15ポイント，テストセットで+0.84ポイント高いF値となった．Fine-tuned ELMo + Entity-aware LMをFine-tuned ELMoと比較すると，開発セットでは1.95ポイント低いF値となったが，テストセットでは0.43ポイントの向上が見られた．

Pre-trained ELMo + Entity-aware LMの精度がFine-tuned ELMoよりも低い一因として，Entity-aware LMの学習データ量がFine-tuned ELMoよりも少ない点が挙げられる．これはEntity-aware LMの訓練事例を作成する際に，entity mentionを中心とする左右k単語を文から切り出しているためである．学習データ量を揃えての比較は今後の課題とする．

	Dev F1			Test F1		
Model	**RE**	**COREF**	**ETT**	**RE**	**COREF**	**ETT**
Pre-trained ELMo（ベースライン）	57.66	62.35	84.17	59.64	60.95	81.16
Fine-tuned ELMo	60.35	**63.34**	86.13	62.32	61.42	83.37
Pre-trained ELMo + Entity-aware LM	58.59	62.54	85.30	60.80	59.44	82.43
Fine-tuned ELMo + Entity-aware LM	**60.83**	62.74	**87.12**	**63.06**	**61.46**	**83.89**

表 2: Best_RE 設定での実験結果．各数値は 4 回の独立した試行の平均である．
RE, COREF, ETT はそれぞれ関係抽出，共参照解析，entity type tagging を示す．

	Dev F1			Test F1		
Model	**RE**	**COREF**	**ETT**	**RE**	**COREF**	**ETT**
Pre-trained ELMo（ベースライン）	54.01	66.62	83.65	58.57	60.62	80.72
Fine-tuned ELMo	**58.38**	**69.31**	86.41	**62.38**	61.84	83.55
Pre-trained ELMo + Entity-aware LM	55.03	67.16	85.64	60.56	61.00	81.82
Fine-tuned ELMo + Entity-aware LM	56.22	67.36	**86.53**	62.21	**62.27**	**83.75**

表 3: Best_COREF 設定での実験結果．各数値は 4 回の独立した試行の平均である．

6 関連研究

Peters ら [8] は ELMo を提案している．ELMo は事前学習した双方向言語モデルの全ての層の重み付き和として計算され，この重みは下流タスクのモデルと同時に学習される．ELMo の双方向言語モデルに各時刻で入力される，文脈に依存しないトークン表現は，文字レベルの CNN によって算出されているため，未知語に対してより堅牢なモデルとなっている．Singh ら [9] は，entity mention のタイプ推定，関係抽出，共参照解析の同時推論に取り組んでいる．彼らはタスク間の依存関係を考慮するために，複数のタスクの潜在変数の同時分布を因子グラフとしてモデル化している．彼らは entity mention の範囲がモデルに入力として与えられる問題設定を扱っており，本稿のタスク定義はこれに基づいている．また，Ji ら [4] が提案している，entity を陽にモデル化した言語モデルでは，単語系列の確率分布をモデル化するだけでなく，入力文の各トークンに対し，当該のトークンが指し示す entity のインデックスなどの付加的な確率変数の値の生成も行っている．

7 おわりに

本稿では，科学論文からの情報抽出にとって有用な特徴量を科学ドメインの大規模ラベルなしデータから得る課題に取り組んだ．具体的には，事前学習された ELMo の fine-tuning （Fine-tuned ELMo) および Entity を考慮した言語モデル（Entity-aware LM）を検討した．実験の結果，entity mention のタイプ推定，関係抽出，共参照解析のマルチタスク学習の枠組みで，言語モデルに基づく特徴量が有効に働くことを確認した．また，Fine-tuned ELMo に Entity-aware LM を組み合わせることで関係抽出の精度がさらに向上することを確認した．

8 謝辞

本研究は JST CREST（課題番号: JPMJCR1513）および JSPS 科研費（課題番号: 18K18109）の支援を受けて行った。

参考文献

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *NAACL*, 2018.

[2] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *SemEval@ACL*, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[4] Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. Dynamic entity representations in neural language models. In *EMNLP*, 2017.

[5] Yoon Kim, Yacine Jernite, David A Sontag, and Alexander M. Rush. Character-aware neural language models. In *AAAI*, 2016.

[6] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *EMNLP*, pages 188–197. Association for Computational Linguistics, 2017.

[7] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*, 2018.

[8] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, 2018.

[9] Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. Joint inference of entities, relations, and coreference. In *AKBC*, AKBC '13, pages 1–6, New York, NY, USA, 2013. ACM.