計算機科学論文における手法の利点・欠点に着目したデータの構築と分析

白井 穂乃 [1]　井之上 直也 [1,2]　鈴木 潤 [1,2]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所 AIP センター
{hshirai, naoya-i, jun.suzuki, inui}@ecei.tohoku.ac.jp

1 はじめに

論文の出版数が急増している．STM 協会の報告[*1]によると，年間で 300 万を超える論文が出版されている．出版数の増加によって，人手による論文の情報収集には限界が来ている．

この状況を打開するため，学術論文から有用な情報を自動で抽出する取り組みが行われている．例えば，分野に依存しない研究として，文献間の引用関係に基づいて，文献の重要度・技術のトレンドを自動的に解析する Citation Network [5] や，学術論文の各文を「背景」「先行研究」などに分類する Argumentative Zoning [14] がある．一方，分野に依存する研究として，生物医学分野の文献を対象として，タンパク質等の専門用語，タンパク質間の関係，物質とその副作用などを抽出する BioNLP [3] がある．このような，論文から有用な情報を自動で抽出する取り組み，論文解析は盛んに行われており，Semantic Scholar [11] や Dr. Inventor [8] のようなアプリケーションツールとして活用されている．

本研究では，計算機科学の分野の論文に対する情報抽出の問題を考える．そもそも，計算機科学の論文は，ある問題に対する先行研究の解決手法とその利点・欠点を論じ，新しい解決手法の提案を行う文書である．よって，計算機科学の研究者が新しい手法を提案する上で，先行研究の手法とその利点・欠点は全て追うべき情報である．冒頭で述べたように，論文が急増する中で，研究を効率化するために，手法の利点・欠点の情報を自動で取得するツールが望まれると考える．

しかし，上述のような，論文を扱った先行研究では，手法とその利点・欠点の抽出は行われていない．例えば，前述の Citation Network や BioNLP は引用関係やエンティティの抽出を行っているのみで，手法の利点・欠点は抽出の対象として扱っていない．

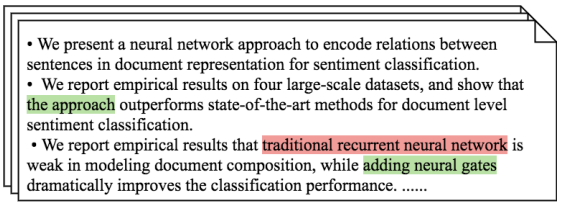

図1: 利点・欠点の抽出イメージ

本研究では図 1 のような，手法とその利点・欠点の自動抽出を試みる．自然言語処理では，利点・欠点と似た概念として評判分析 (Sentiment Analysis) がある．よって，評判分析を論文で述べられている手法に対して適用する．適用するために，手法とその利点・欠点についてアノテーションスキームを定め，実際にアノテーションを行い，一致率を調査した．また，構築したアノテーションデータを用いて，タスクの難しさを検証するため，自動抽出実験を行い，結果に対して分析を行った．構築したアノテーションコーパスは公開している．[*2]

2 データ構築

計算機科学論文における手法の利点・欠点の情報抽出をする先行研究は我々の知る限りないため，自動抽出モデルの構築・評価のためのデータは存在しない．このため，アノテーションスキームを設計した．また，データを構築し，一致度を検証するために，人手によるアノテーションを行った．以降，アノテーションスキームとアノテーション実験について報告する．

2.1 アノテーションスキーム

手法とその利点・欠点をそれぞれ抽出するために，手法 TERM ラベルと手法の評価 Sentiment ラベルを定義した．それぞれのラベルの定義について説明する．

2.1.1 TERM

問題解決の手法に関する記述をアノテーションするために，TERM というラベルを導入する．TERM は，モ

[*1] https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf

[*2] https://github.com/cl-tohoku/scientific-paper-pros-cons

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

デル・アルゴリズムといった仕組み，仕組みの持つ機能，仕組みが動作する方法を表す名詞句に対して付与するラベルである．例えば，例 (1) では，*recursive neural network* と *AdaRNN* はそれぞれニューラルネットワークというモデルとその一種であるため，TERM ラベルを付与する．

(1) *We employ a novel adaptive multi-compositionality layer in recursive neural network, which is named as AdaRNN (Dong et al., 2014).*

また，論文の文書に述べられている全ての利点・欠点について把握するため，一般的な手法の名前だけでなく，限定詞を含む手法についても TERM として扱う．例えば，例 (2) の *Such approaches* は TERM のラベルが付与される．

(2) *Such approaches have a number of disadvantages.*

2.1.2 Sentiment

TERM ラベルが付与された手法に対する評価を捉えるために Sentiment というラベルを導入する．

評判分析の先行研究 [7] に倣い，利点・欠点をポジティブ・ネガティブで表現し，極性のないニュートラルを含めた POSITIVE，NEGATIVE，NEUTRAL の 3 種類を Sentiment とする．ただし，Sentiment は TERM に対する属性として付与する．

Sentiment は文内でのローカルな極性であり，TERM の含まれる文内で Sentiment を判断する．例 (3) において TERM である *the whole-sentence-based classifier* は，*performs the best* というポジティブな評価がされている．よって，この TERM には POSITIVE ラベルを付与する．

(3) *The results indicate that the whole-sentence-based classifier performs the best.*

また，例 (1) の TERM(*recursive neural network*, *AdaRNN*) のように，単に問題解決手法の特徴，性質を述べている場合，NEUTRAL ラベルを付与する．

2.2 データ

2.1 節で定義したアノテーションスキームを適用するデータについて述べる．アノテーション対象とするのは ACL anthology の論文である．ただし，論文全体ではなく，イントロダクションの節のみ扱う．これは，イントロダクションは一般的に既存手法・提案手法について述べられているためである．

アノテーション実験を行なうにあたって，本研究では *coreference resolution* がタイトルまたは本文に含まれている論文を選出した．*coreference resolution*（共参照解析）は自然言語処理の分野において長年研究対象になっているため，広い年代で様々な手法が提案されているためである．Google のカスタム検索を用いて，92 本の論文を選出した．選出した論文は 1999 年から 2017 年に出版された論文である．

2.3 アノテーション実験

自動抽出のデータを構築するため，人手でアノテーションを行った．スキームが人間にとって理解でき，正確にアノテーションできるかを検証するため，複数人でアノテーションすることで，その一致度について調査した．

2.3.1 実験内容

2.2 節で述べた論文に対してアノテーションを行った．自然言語処理を専門とする留学生 3 人にスキームについて説明し，アノテーションを依頼した．

アノテーションの一致率について調査するため，1 つの論文につき 2 人がアノテーションするように割り当てた．また，アノテーションには brat [12] を用いてユーザーインターフェースを作成した．

2.3.2 結果・考察

アノテーションの結果と考察について報告する．TERM のアノテーションについて,完全一致率は 24.0 %，部分一致を含む一致率は 38.2 % であった．

一致率が低い結果となったため，不一致の事例を分析したところ，一方のアノテーターが TERM を付与した箇所について，もう一方のアノテーターが付与しなかった事例が多く存在した．これは，文書中に含まれる単語が TERM かどうか判断が難しかったためと考えられる．例えば，*joint inference* や *a learned cluster ranker* が一方のみのアノテーションとしてみられた．

TERM が部分一致しているアノテーションについては，名詞句でアノテーションすべき箇所が，正しくアノテーションされていなかった．これは，アノテーションスキームが作業者に正確に伝わっていなかったことが原因として考えられる．具体的には，*a simplified semantic role labeling (SRL) framework* のような修飾語を含む TERM はアノテーションの範囲の不一致が見られた．

また，TERM が完全一致したアノテーションについ

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表1: 文書セットごとの Sentiment の混同行列

A,B	Pos	Neu	Neg
Pos	7	2	0
Neu	3	78	10
Neg	0	3	25

(i) κ : 0.7031

A,C	Pos	Neu	Neg
Pos	10	5	1
Neu	1	100	9
Neg	0	4	23

(ii) κ : 0.7044

B,C	Pos	Neu	Neg
Pos	4	2	0
Neu	2	72	13
Neg	0	6	13

(iii) κ : 0.4903

て，Sentiment の混同行列及びアノテーター間の一致度 (κ 統計量) を表 1 に示す．3 つの混同行列は 3 人のアノテーターを A, B, C とした時，各アノテーターがアノテーションした文書セットの共通部分に対する結果である．

Sentiment について，TERM が完全一致している場合は一致率が高かった．Sentiment が一致しない原因として，ドメイン知識が必要な事例が存在することがわかった．具体的には，例 (4) の *a graph representation* が NEUTRAL と POSITIVE でアノテーションが割れた．*a more adequate clusterization phase* を獲得することが利点なのかどうか，ドメイン知識が必要なためと考えられる．

(4) *We argue that a more adequate clusterization phase for coreference resolution can be obtained by using a graph representation.*

3 自動抽出実験

自動抽出タスクとしてどの程度難しいかを検証するため，ベースラインとなるモデルを構築し実験を行った．

3.1 タスク設定

本研究では 1 文を入力として与え，TERM の位置と Sentiment を出力するフレーズ抽出タスクとして定義する．評価指標は，予測ラベルが正解ラベルと位置・ラベルともに一致した時のみ正解とし，F 値で評価する．

3.2 データ

2.3 節の実験で作成したデータを用いる.ただし,2 人のアノテーションを結合したデータを使用する．データの結合は，できる限り多くの手法 (TERM) とその利点・欠点 (POSITIVE, NEGATIVE) を採用する方法を用いた．具体的には，TERM について，1 人がアノテーションしていれば TERM とし，部分一致している場合については，範囲が広い方を TERM として採用した.また,Sentiment が不一致の場合は，POSITIVE ・ NEGATIVE のラベルを優先し，2 人がそれぞれ POSITIVE ・ NEGATIVE をアノテーションしている場合は NEUTRAL ラベルを採用した．

表2: アノテーションデータの詳細

データ	sentence	TERM POSITIVE	TERM NEUTRAL	TERM NEGATIVE
noisy	1,872	254	1,102	116
clean	2,058	255	1,100	116

また，文書のスクリーニングと，アノテーションの修正を人手で行ったデータも作成した．スクリーニングについて，文・単語のトークン化が誤っている場合を修正した．アノテーションついては，TERM の範囲が名詞句の範囲であるように修正を行った．以降，この修正前後の 2 つのデータセットをそれぞれ noisy データ・ clean データとする．データの規模・ラベルの数は表 2 のとおりである．

3.3 実験設定

論文ごとに訓練・開発・テストを 8:1:1 に分割し，10 分割交差検証を行った．また，新しい論文に対して，過去の論文データでも対応できるか検証するため，最新年である 2017 年とそれ以外の年のデータをそれぞれテスト，訓練とする分割でも実験を行った．評価指標である F 値は 10 分割交差検定の平均値で示す．ただし，テストデータの F 値は開発データの F 値が最も高くなったエポックにおける結果を示す．

3.4 モデル

ベースラインのモデルとして，NERtagger[*3]を用いた．このモデルは Lample らの提案した，特徴量や言語に依存せずに固有表現抽出を行う BiLSTM-CRF モデル [6] である．単語埋め込みベクトルとして，ACL Anthology Corpus [1] で学習済みの word2vec を使用した．このモデルを Baseline モデルとする．

また，Baseline モデルに加え，1 Billion Word Benchmark で訓練済みの ELMo [9] ベクトルを使用するモデル (ELMo モデル) も用いた．

3.5 実験結果・考察

実験結果を表 3 に示す．実験結果について，ウィルコクソンの符号順位検定を棄却域 5 % で行った．clean データにおける ELMo モデルについて，他 3 つの結果と比較して精度に有意差があると示された．

最新年以外の clean データについて ELMo モデルで訓練した結果，2017 年データの F 値は 42.69 % であった．

モデルの予測結果から，タスクの難しさについて考察する．以降，最も精度が高い clean データにおける

[*3] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

表3: 自動抽出実験の結果

データ	モデル	dev F1	test F1
noisy	Baseline	44.48	43.69
	ELMo	47.79	48.60
clean	Baseline	50.70	49.79
	ELMo	54.23	52.35

ELMo モデルの予測結果を用いて述べる．ただし，例文において，下線の上付き文字を正解ラベル，下付き文字を予測ラベルとして表記する．

評価が明快な語が含まれている場合はモデルが予測できている．例 (5) では，*This approach* に対して *suitable* という評価をしているため，モデルが TERM-POSITIVE ラベルを予測できたと考える．

(5) $\underline{\textit{This approach}}^{\textsc{term-positive}}_{\textsc{term-positive}}$ *to feature engineering is suitable not only for knowledge-rich but also for knowledge-poor datasets .*

しかし，手法を意味する単語を含まない場合，TERM の予測は難しいため，偽陰性の誤りが生じている．例 (6) では，*concept maps* は TERM-POSITIVE が正解ラベルであるが，モデルでは予測できていない．これは，*concept maps* には明示的に TERM を表す単語が含まれていないためと考えられる．

(6) *Several studies report successful applications of* $\underline{\textit{concept maps}}^{\textsc{term-positive}}$ *in this direction...*

また，暗黙的な評価をしている場合は Sentiment の予測が難しいため，ラベル付与の誤りが生じている．例えば，例 (7) では *these approach* は TERM-NEGATIVE が正解ラベルであるが，モデルは TERM-NEUTRAL を予測している．*require labeled training data* という評価が *these approach* にとって悪い評価であることを予測できないためと考えられる．

(7) *While successful ,* $\underline{\textit{these approaches}}^{\textsc{term-negative}}_{\textsc{term-neutral}}$ *require labeled training data , consisting of mention pairs and the correct decisions for them .*

4 関連研究

Tateisi ら [13] の研究は，情報科学論文に出現する用語間の関係を構造化するためのタグ付けスキーマを提案している．また，自然言語処理分野の評価型ワークショップ SemEval では，論文ドメインでの情報抽出タスクが提案されている．SemEval-2017 の評価タスクである ScienceIE [2] は複数ドメインの論文からフレーズと関係の抽出を行うタスクである．SemEval-2018 Task 7 [4] は ACL Anthology Corpus の Abstract について，エンティティ同士の関係を分類するタスクを提案している．

冒頭でも触れたが，自然言語処理の分野では，レビュー文章のドメインにおける観点付き感情極性分析が行われている．例えば，SemEval-2015 Task 12 [10] は，ホテルやレストランの料理の価格やサービスの質などの定義した観点に基づいて評価分析を行うタスクである．また，論文に対する評判分析では，文献内で引用している文献に対する著者の感情極性を解析する引用評価極性解析 (Citation Sentiment Analysis) [15] が行われている．

5 結論

本研究では，計算機科学論文における，手法の利点・欠点のアノテーションと自動抽出を試みた．今後の課題として，自然言語処理の論文以外で検証を行うこと，ドメイン知識の必要な事例・新しい年の論文に対応した自動抽出モデルの構築に取り組むことが考えられる．

謝辞

本研究は，JST CREST (課題番号: JPMJCR1513) の支援を受けて行った．

参考文献

[1] Akiko Aizawa et al. “Construction of a New ACL Anthology Corpus for Deeper Analysis of Scientific Papers”. In: *SCIDOCA*. 2018.
[2] Isabelle Augenstein et al. “SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications”. In: *SemEval*. ACL, 2017.
[3] Louise Deléger et al. “Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016”. In: *Proc. of the 4th BioNLP Shared Task Workshop*. ACL, 2016.
[4] Kata Gábor et al. “SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers”. In: *SemEval*. ACL, 2018.
[5] Yuya Kajikawa et al. “Creating an academic landscape of sustainability science: an analysis of the citation network”. In: *Sustainability Science* 2.2 (2007).
[6] Guillaume Lample et al. “Neural Architectures for Named Entity Recognition”. In: *Proc. of NAACL*. ACL, 2016.
[7] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.
[8] Diarmuid P O’Donoghue et al. “Towards Dr inventor: a tool for promoting scientific creativity”. In: ICCC. 2014.
[9] Matthew Peters et al. “Deep Contextualized Word Representations”. In: *Proc. of NAACL*. ACL, 2018.
[10] Maria Pontiki et al. “SemEval-2015 Task 12: Aspect Based Sentiment Analysis”. In: *SemEval*. ACL, 2015.
[11] *Semantic Scholar*. https://www.semanticscholar.org.
[12] Pontus Stenetorp et al. “brat: a Web-based Tool for NLP-Assisted Text Annotation”. In: *Proc. of EACL*. ACL, 2012.
[13] Yuka Tateisi et al. “Typed Entity and Relation Annotation on Computer Science Papers”. In: *LREC*. ELRA, 2016.
[14] Simone Teufel et al. “Argumentative zoning: Information extraction from scientific text”. PhD thesis. Citeseer, 1999.
[15] Abdallah Yousif et al. “A survey on sentiment analysis of scientific citations”. In: *Artificial Intelligence Review* (2017).

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.