B7IM2032

修士論文

計算機科学論文からの技術の利点・欠点のマイニング

白井穂乃

2019年2月5日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (情報科学) 授与の要件として提出した修士論文である。

白井穂乃

審査委員：

乾 健太郎 教授　（主指導教員）

鈴木 潤 准教授　（副指導教員）

計算機科学論文からの技術の利点・欠点のマイニング*

白井穂乃

内容梗概

近年，論文の出版数が急増し，人手による論文からの情報収集に限界が来ている．本研究では，計算機科学の論文における重要な情報を獲得することを目的として，論文中に述べられている技術とその利点･欠点の自動抽出を提案する．本稿では，自動抽出を行うための注釈付きコーパスを構築し，ベースラインモデルによるタスクの性質の検証を行った結果について報告する．具体的には，論文への利点･欠点に関するアノテーションスキームを定義し，人手によるアノテーション実験を行った．また，自動抽出タスクとしての難しさを検証するため，アノテーション実験で構築したデータを用いて，利点・欠点の自動抽出モデルを構築した．アノテーション実験によって，約 100 本の論文に対してアノテーションラベルが付与されたデータを構築した．ベースラインによる自動抽出実験によって，タスクを解くためにドメインに関する知識・推論が必要なことが分かった．

キーワード

論文解析, テキストマイニング

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B7IM2032, 2019 年 2 月 5 日.

Mining pros and cons of technique in computer science papers*

Hono Shirai

Abstract

In recent years, the number of scholarly papers published has increased rapidly, and there is a limit to collect information from the papers by hand. In this paper, we propose automatic extraction of the technique and its pros and cons described in the papers of computer science. We report on the annotated corpus we constructed and the automatic extraction experiment to verify the task properties by using the baseline model. Specifically, we define the annotation scheme of pros and cons to the papers and report on the result of the manual annotation experiment. In order to verify the difficulty of an automatic extraction task, we construct an automatic extraction model of pros and cons using data constructed in annotation experiments and report the experiment results. We create data that is given annotation label about 100 paper. The automatic extraction experiment results show that our task is difficult and a better solution may require domain-specific knowledge and inference.

Keywords:

scientific literature analysis, text mining

*Master's Thesis, System Information Sciences, Graduate School of Information Sciences, Tohoku University, B7IM2032, February 5, 2019.

目 次

図 目 次

表 目 次

1　はじめに・研究背景

近年，学術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．我々研究者は，こうした膨大な学術文献の中から関連分野の情報を適切に取捨選択し，把握することが年々難しくなってきている．

このような状況を打開すべく，学術論文から有用な情報を自動抽出するための様々な研究が盛んに行われている．例えば，分野に依存しない研究として，Teufuls [1] は，学術論文の各文を「背景に関する記述」「先行研究に関する記述」などに分類する Argumentative Zoning というタスクに取り組んでいる．また，文献間の引用関係に基いて，文献の重要度や，技術のトレンドなどを自動的に解析する取り組み (Citation Network Analysis) も盛んに行われている [2]．ScienceIE [3] では，物理学，材料科学，計算機科学を対象として，フレーズの抽出や同義語・下位語の関係抽出などの研究が行われている．引用評価極性解析 (Citation Sentiment Analysis) の分野では，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研究が盛んに行われている [4]．

一方，分野に依存する研究として，BioNLP [5] では，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の認識タスク，たんぱく質間の関係や，物質とその副作用などを抽出する関係抽出タスクの研究が行われている．

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 [6] や Dr. Inventor [7] 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，計算機科学分野の学術論文に対する自動解析の研究に取り組む．

そもそも，計算機科学の論文は，ある問題 (タスク) に対する先行研究の解決する技術とその利点・欠点を論じ，新しい技術の提案を行う文書である．例えば，次のような一文が論文内にあったとする．

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

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

例 (1) では， *the whole-sentence-based classifier* という技術に対して， *performs the best*，すなわち「性能という観点」において,「良い」という評価がなされている．このような技術の利点・欠点の情報は，計算機科学の研究者が新しい技術を提案する上で，全て追うべき情報である．よって，論文が急増する中で，研究を効率化するために，技術の利点・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．また，個々の論文における問題解決のための技術とその利点・欠点を自動認識し，その結果を集約することは，分野全体の技術の特徴を手早く俯瞰できるようになると期待できる．

しかし，上述のような，論文を扱った先行研究では，技術とその利点・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，利点・欠点は抽出の対象として扱っていない．また，自然言語処理の分野では，このような評価の抽出問題は，評価分析 (Sentiment Analysis)，または意見マイニング (Opinion Mining) として，古くから取り組まれてきた [8]．しかしながら，こうした解析を学術論文のドメインに適用する試みはこれまでになく，タスクの具体的な設計方針（何を評価対象とみなすか，何を観点とみなすかなど），コーパス等の言語資源も整備されていない．

本研究では技術とその利点・欠点の自動抽出を試みる．利点・欠点と似た概念としてある評判分析 (Sentiment Analysis) を論文で述べられている技術に対して適用する．適用するために，技術とその利点・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 (3 節)．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，ベースラインモデルによる自動抽出実験を行い，結果に対して分析を行った (4 節)．最後に，構築したアノテーションデータとベースラインモデルの予測結果を用いた検索インタフェースの作成について述べる (5 節)．

2　関連研究

本節では，本研究と関連する研究である論文解析と評判分析について説明する．

2.1　論文解析

建石ら [9] 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．これは論文中に存在する用語すべてに対して意味クラスを付与し，用語間にも関係のクラスを付与することで論文内容の構造化を行うことを目的としている．

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 [3] は物理学，材料科学，計算機科学の論文からフレーズと関係の抽出を行うタスクである．フレーズは Task, Process, Material の 3 つのクラスに対応している．フレーズのクラスを分類するタスク，フレーズ同士が類義語・下位語の抽出を試みるタスクが提案されている．SemEval-2018 Task 7 [10] は ACL Anthology Corpus [11] 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．エンティティは概念を表現する名詞句と規定されており，比較・結果などの 5 つの関係を抽出・分類するタスクである．

上記の先行研究は，論の章構造・エンティティ同士の関係について扱う研究である．また，技術の評価について扱う研究はわれていない．

2.2　評判分析

評判分析 (Sentiment Analysis) は文・文章の感情極性を扱う研究である．具体的には，商品レビューやソーシャルメディアの投稿をポジティブ・ネガティブ・ニュートラルのいずれかの極性に分類することを目的としている．細かい粒度で感情極性を捉える研究として，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 (Aspect-Based Sentiment Analysis) が行われている．例えば，SemEval-2015 Task 12 [8] 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 (Aspect) に基づいて評価分析を行うタスクである．また，特定のエンティティに関する評判分析を行う Targeted Sentiment Analysis [12] も行われている．Targeted Sentiment Analysis は人名・企業名・製品名などの対象 (Target) が文中に含まれる文において，対象に対する評判を判定する．対象とする文は主に Twitter のツイートであるが，Student Comment を対象にした先行研究も存在する [13].

しかし，我々の知る限り，論文ドメインにおいては観点に基づいた評価分析は行われていない．

3　データ構築

計算機科学論文における技術の利点・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．

3.1　アノテーションスキーム

技術とその利点・欠点をそれぞれ抽出するために，技術 TERM ラベルと技術の評価 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．

3.1.1　TERM

問題解決の技術・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．具体的には，モデル・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を TERM として付与する．

例えば，例 (2) 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．

(2) *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，一般的な技術の名前だけでなく，限定詞を含む技術についても TERM として扱う．具体的には，例 (3a) の *Such approaches* ， 例 (3b) の *they* はそれぞれ TERM のラベルが付与される．

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

b. *First, they require additional resources, such as lists of polarity shifters or discourse connectives which signal specific relations.*

3.1.2 Sentiment

TERM ラベルが付与された技術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 [14] 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．例えば，例 (4) 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．

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

また，例 (2) の TERM(*recursive neural network*， *AdaRNN*) のように， 単に技術の特徴， 性質を述べている場合は NEUTRAL ラベルを付与する．

以降，POSITIVE, NEGATIVE, NEUTRAL が付与された TERM をそれぞれ TERM-POSITIVE, TERM-NEGATIVE, TERM-NEUTRAL と表記する．

TERM に Sentiment が付与される事例について，つまり TERM-POSITIVE, TERM-NEGATIVE の事例について，(1) ドメイン非依存な評価と (2) ドメイン依存な評価表現が文中に含まれていることが考えられる．

ドメイン非依存な例として以下の事例がある．

(5) *... our system can generate high-quality labeled data.*

例 (5) は *our system* に TERM-POSITIVE のラベルが付与されている．これは *can* や *high-quality labeled data* といった，ドメイン非依存でポジティブな単語が含まれている．

一方で，ドメイン依存な例として以下の事例がある．

(6) *This approach requires no manually-specified information about the meaning of the connectors, just the connectors themselves*

例 (6) は *This approach* が TERM-POSITIVE のラベルが付与されている．*This approach* に関する評価表現 *requires no manually-specified information* にはポジティブな単語が含まれていないが,「タスクを解決する上で必要とする情報が少ない」という観点において TERM にとってポジティブであると推論できる．

このような TERM に対する評価表現として考えられる観点は (1) 性能，(2) 前提条件，(3) 機能がある．具体的な例を列挙する．

例 (7) は TERM-POSITIVE である *it* に対して性能に関する評価表現が記述されている事例である．*a strong and robust performer* という評価表現は性能に関してポジティブな評価を表明している．

(7) *... it is a strong and robust performer*

また，前述の例 (6) における評価表現 *requires no manually-specified information* は TERM が動作するための前提条件に関する評価表現である．

最後に，例 (8) は機能に関する評価表現を含む事例である．この事例は *The model* が TERM-POSITIVE である．TERM に関する評価表現 *can capture* は TERM の持つ機能という観点を想起させ，*complex semantic information* という難しい情報を捉える，と表明することで機能がポジティブであると評価している．

(8) *The model can also avoid overfitting to features derived from neutral or objective sentences.*

以上のように，Sentiment の判断は利点・欠点にあたる評価表現によって行う．そのため，当初は評価表現のフレーズもアノテーションしてもらうことを検討したが，利点・欠点は TERM の周辺に記述されているため，アノテーションする必要がないと判断した．よって，本研究では利点・欠点の評価表現はアノテーションせず，TERM に対してアノテーションするのみとしている．

3.2 データ

3.1 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．

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アノテーションする論文として選んだ．*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，広い年代で様々な技術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り，workshop の論文も含んでいる．

3.3 アノテーション実験

3.1 節のアノテーションスキームに基づいて，データ構築のためのアノテーションデータを構築する．

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．

3.3.1 実験内容

3.2 節で述べた ACL anthology の論文をアノテーションする．

92 本の文書について，自然言語処理を専門とする留学生の学生 3 人にアノテーションしてもらった．

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．アノテーションツールには brat [15] を用いてアノテーションインタフェースを作成した． 実際のアノテーションインタフェースを図 1 に示す．

3.3.2 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．

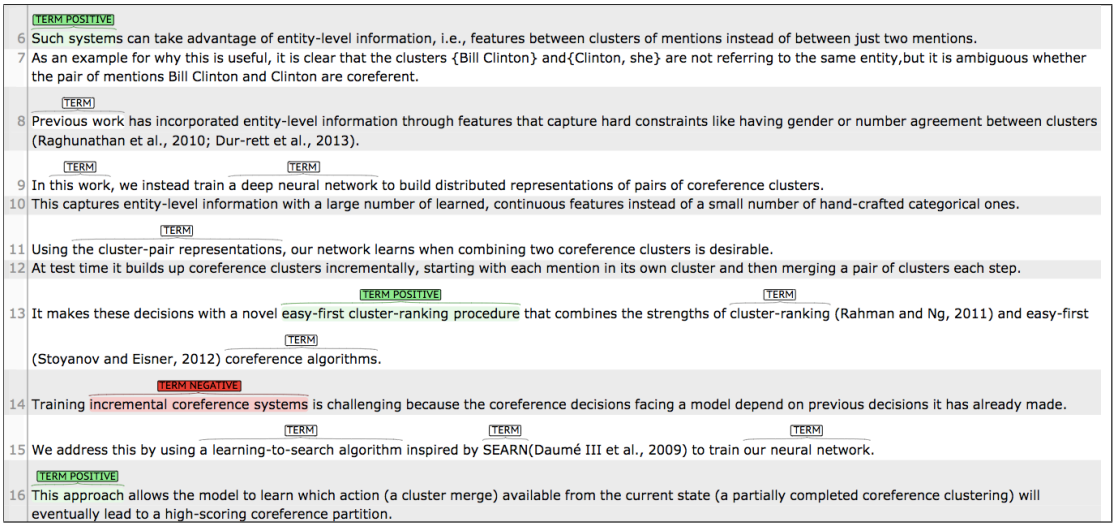

図 1: brat によるアノテーション

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，*the* や *a* などの助詞が含むか含まないかでアノテーションの範囲が不一致している場合は部分一致が起こっている．また，*a simplified semantic role labeling (SRL) framework* のような修飾語を含む TERM もアノテーションの範囲の不一致が見られた．

次に，TERM が完全一致したアノテーションについて，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

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

(a) κ : 0.7031

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

(b) κ : 0.7044

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

(c) κ : 0.4903

トの共通部分に対する結果をそれぞれ示している．

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 (9) 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．

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

また，学術論文の文章では暗黙的な評価を用いることも，Sentiment の判断が難しい原因として考えられる．一般的に，学術論文では明示的に既存手法を批判することは避けられるため，暗黙的な評価表現を用いる．例 (10) では *the cascades approach* が TERM-NEGATIVE と TERM-NEUTRAL でアノテーションが割れた．*only when the first level decision-making is done* という評価が *the cascades approach* を暗にネガティブ に評価していると推論できるため，一方のアノテーターが TERM-NEGATIVE のラベルを付与したと考える．

(10) *Importantly, our pruning and scoring functions operate sequentially ateach greedy search step, whereas in the cascades approach, the second level function makes its prediction only when the first level decision-making is done*

アノテーション実験全体についての考察は以下に挙げる．まず，TERM のアノテーションの難しさがある．アノテーションの一致率を上げるために，アノテー

ションスキームの細かい定義に関する説明が必要だと考える．具体的には冠詞を含む，修飾語を含んだ名詞句の範囲を取るなどの対策が考えられる．Sentiment のアノテーションには論文のトピックに関する知識が必要であることがわかった．また，論文の文章は暗黙的な評価を用いることが多いため，Sentiment のアノテーションに揺れがおこることがわかった．

4　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．

4.1　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．

4.2　データ

3.3.1 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 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．

また，文書のクリーニングと，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．これは，アノテーションデータとして用いたテキストの文・単語のトークン化が誤っていたためである．具体的には，2 文が繋がっているなど文区切りが間違っている場合，句読点が単語と正しく区切られていない場合について，正しい区切りとなるよう修正した．アノテーションの修正ついては，TERM の範囲が名詞句の範囲であるように修正を行った．具体的には，限定詞が TERM に含まれていないラベル，TERM のアノテーション範囲が名詞句でないラベルを正しい範囲に修正した．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．

表 2: アノテーションデータの詳細

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

4.3 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 2017 年より前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．

4.4 モデル

ベースラインのモデルとして，NERtagger[2]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 [16] である．単語埋め込みベクトルとして，ACL Anthology Corpus [17] 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 [18] ベクトルを使用するモデル (ELMo モデル[3]) も実験に用いた．

4.5 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 データについて，Baseline，ELMo のいずれのモデル

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

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

でも noisy データと比較して精度に有意差があると示された．また，ELMo モデルについて，noisy．clean いずれのデータでも Baseline モデルと比較して精度に有意差があると示された．

最新年である 2017 年より前の clean データについて ELMo モデルで訓練した結果，テストデータである 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 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 (11) 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．

(11) *This approach*$^{\textsf{TERM-POSITIVE}}_{\textsf{TERM-POSITIVE}}$ *to feature engineering is suitable not only for knowledge-rich but also for knowledge-poor datasets .*

しかし，技術・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 (12) 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．

(12) *Several studies report successful applications of concept maps*$^{\textsf{TERM-POSITIVE}}$ *in this direction...*

同様に TERM の予測が難しい場合として，共参照が必要な場合が考えられる．

例 (13) では，*they* の TERM-NEGATIVE をモデルは予測できていない．モデルは共参照を無視しているため *they* が TERM かどうかの判断ができないためと考える．

(13) *Second , they*$^{\textit{TERM-NEGATIVE}}$ *have limitations in their expressiveness : the information extracted from the two mentions alone may not be sufficient for making an informed coreference decision ,* ……

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 (14) では，*This model* 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．音声統合だけでなく非言語モダリティに適用できるという評価が POSITIVE であると予測できないためと考える．

Sentiment の予測は 3.3 節のアノテーション実験における Sentiment の不一致と同様の傾向がみられた．つまり，評価が明快でない場合，ドメインに関する知識や推論が必要な場合はモデルも予測が難しいと考えられる．

(14) *This model*$^{\textit{TERM-POSITIVE}}_{\textit{TERM-NEUTRAL}}$ *is not specifically tailored to gesture-speech integration, and may also be applicable to other non-verbal modalities .*

次に，最新年である 2017 年をテストデータとし，2017 年より前の年のデータで学習した実験設定について考察する．F 値 42.69 %は前述の実験設定の F 値よりも大きく精度が下がった．原因として，学習データに存在しない未知語がテストデータに含まれていたことが挙げられる．

例 (15) では *memory network* が TERM であるが，モデルは予測できていない．*memory network* は新しく登場した技術であるため，学習データには存在しない TERM である．よってモデルの予測は難しいと考える．

(15) *We compare the prediction accuracy of memory network with an existing state-of-the-art coreference resolution system* …

また，例 (16) の *the WD classifier* と *the CD classifier* が TERM であるがモデルが予測できていない．*WD* と *CD* は同じ論文中で *within a document (WD) and*

across multiple documents (CD) と定義している単語であり，学習データには存在しない単語である．

(16) *Specifically, the WD classifier uses features based on event mentions and their arguments while the CD classifier relies on* ···

自動抽出実験によって，本タスクがドメインに関する知識・推論が必要なタスクであることが分かった．また，共参照や未知語に対応したモデルを構築することもタスクを解くために必要であると考えられる．

5　検索インタフェース

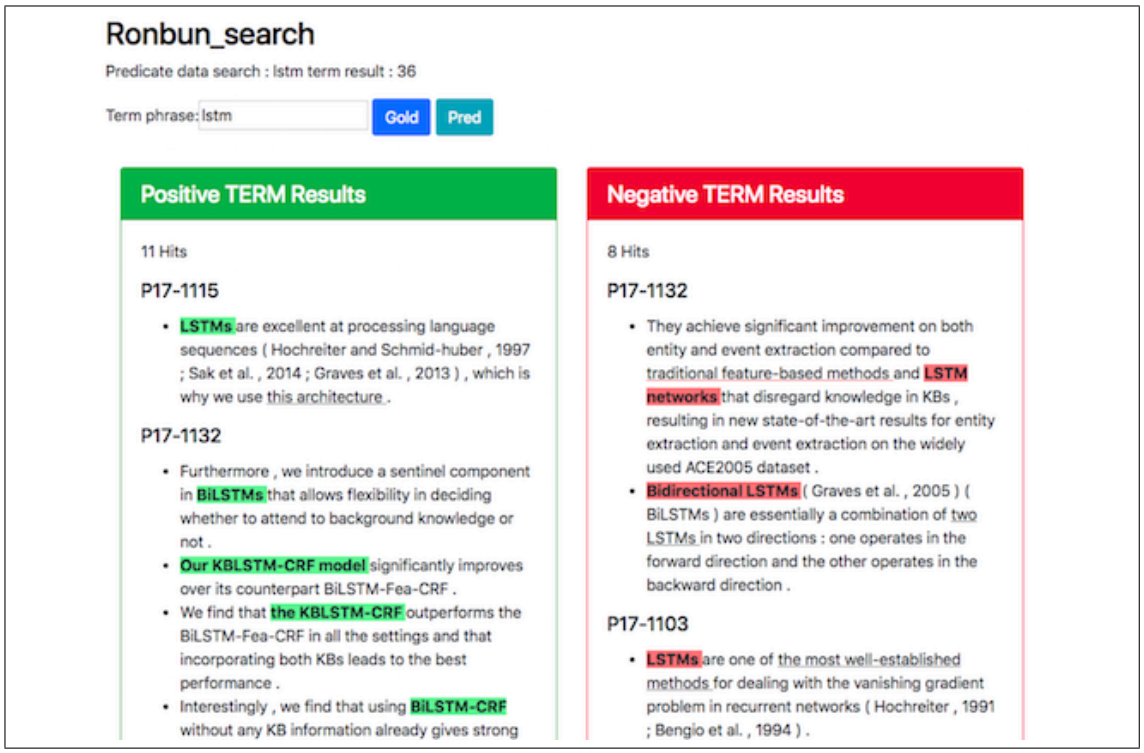

図 2: 検索インタフェース画面

3 節で作成したデータ，及び 4 節でモデルが予測したデータを利用し，検索インタフェースを作成した (図 2).

3 節で作成したデータは，4.2 節で作成したクリーニング済みデータ (clean データ) を用いた．モデルの予測したデータは，4 節の実験において最も精度が高かったモデルを用いて 2017 年の論文を予測した結果をデータとして用いている．

検索インタフェースでは，*Term phrase* に検索したい TERM を入力すると，左側に TERM-POSITIVE，右側に TERM-NEGATIVE が一覧で表示される．図 2 は *LSTM* が TERM の検索結果を表示している．例えば，図 2 の POSITIVE の検索結果として *Furthermore , we introduce a sentinel component in BiLSTMs that allows flexibility in deciding whether to attend to background knowledge or not .* が表示されている．この文から *BiLSTMs* は *a sentinel component* が *allows*

flexibility in deciding whether to attend to background knowledge or not という点において優れているという情報を得ることができる．このように，検索結果を利用して，ある技術の利点・欠点を俯瞰することが可能である．

6　結論

本研究では，計算機科学論文における，技術の利点・欠点のアノテーションと自動抽出を試みた．

アノテーションスキームを定義し，複数人によるアノテーション実験を行った．また，アノテーション実験で構築したデータを用いて，ベースラインモデルによる自動抽出実験を行った．

今後の課題として，まず，データの大規模化が挙げられる．そのために，アノテーション作業の効率化・半自動化が望まれる．また，今回定義したアノテーションスキームは自然言語処理の論文以外においても利用可能である．そのため，様々な学術領域においてアノテーションデータを作成し，自動抽出が可能か検証を行うことが考えられる．最後に，抽出精度の向上のために，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

本研究を進めるにあたり，多くの皆様のご協力，ご助言をいただきました．心より感謝申し上げます.

乾健太郎教授には，研究室配属前から，進路・研究に関する様々なご指導・ご助言をいただきました．鈴木潤准教授にも，同じく研究について多くのご指導をいただきました．深く感謝申し上げます．研究方法や論文執筆に関しまして，直接のご指導をいただきました井之上直也助教に心より感謝申し上げます．また，本論文の審査をお受けしていただきました北村喜文教授及び木下哲男教授に深く感謝申し上げます.

最後になりましたが，研究生活を支えてくださった乾・鈴木研究室のスタッフの皆様をはじめ，2 年間の大学院生活を支えてくださったすべての方に厚く御礼申し上げます.

参考文献

[1] Simone Teufel, et al. *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh, 2000.

[2] Yuya Kajikawa, Junko Ohno, Yoshiyuki Takeda, Katsumori Matsushima, and Hiroshi Komiyama. Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, Vol. 2, No. 2, p. 221, Jul 2007.

[3] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 546–555, 2017.

[4] Myriam Hernández-Alvarez and José M. Gomez. Survey about citation context analysis: Tasks, techniques, and resources. *Nat. Lang. Eng.*, Vol. 22, No. 3, pp. 327–349, 2016.

[5] Louise Delėger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferrė, Philippe Bessieres, and Claire Nėdellec. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pp. 12–22, 2016.

[6] Semantic scholar. `https://www.semanticscholar.org`.

[7] Diarmuid P O'Donoghue, Horacio Saggion, Feng Dong, Donny Hurley, Y Abgaz, X Zheng, O Corcho, Jian J Zhang, J-M Careil, Babak Mahdian, et al. Towards dr inventor: a tool for promoting scientific creativity. ICCC, 2014.

[8] Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, and Michal Konkol. UWB at SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proc. 10th*

Int. Work. Semant. Eval., pp. 342–349, 2016.

[9] 建石由佳, 仕田原容, 宮尾祐介, 相澤彰子. 情報科学論文からの意味関係抽出に向けたタグ付けスキーマ. 言語処理学会第 19 回年次大会, 2013.

[10] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[11] Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark T Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. *Proc. Sixth Int. Conf. Lang. Resour. Eval. (LREC 2008)*, pp. 1755–1759, 2008.

[12] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, June 2011.

[13] Charles Welch and Rada Mihalcea. Targeted sentiment to understand student comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2471–2481, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[14] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press, 2015.

[15] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference*

of the European Chapter of the Association for Computational Linguistics, pp. 102–107, Avignon, France, April 2012. Association for Computational Linguistics.

[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270. Association for Computational Linguistics, 2016.

[17] Akiko Aizawa, Takeshi Sagara, Kenichi Iwatsuki, and Goran Topic. Construction of a new acl anthology corpus for deeper analysis of scientific papers. In *SCIDOCA*.

[18] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

発表文献一覧

受賞一覧

- NLP 若手の会 (YANS) 第 13 回シンポジウム奨励賞 (2018).

国内会議・研究会論文

- 白井穂乃, 井之上直也, 鈴木潤, 乾健太郎. 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. 言語処理学会第 25 回年次大会, March, 2019.
- 白井穂乃, 井之上直也, 乾健太郎. 情報科学論文からの技術の利点・欠点のマイニングに向けて. NLP 若手の会 (YANS) 第 13 回シンポジウム, August, 2018.
- 白井穂乃, 井之上直也, 乾健太郎. 情報科学論文における問題解決手法と評価表現の付与仕様の検討. 人工知能学会全国大会 (第 32 回), June, 2018.
- 白井穂乃, 田然, 松田耕史, 乾健太郎. コノテーションに基づいた名詞の感情極性の予測. NLP 若手の会 (YANS) 第 12 回シンポジウム, September, 2017.