

学術論文からの実験情報抽出とその可視化

平井 久貴^{1,a)} 新妻 弘崇^{1,b)} 太田 学^{1,c)} 高須 淳宏^{2,d)}

概要: 論文のサーベイにおいて、どのような実験結果が得られたのか等、実験に関する情報を収集、整理することは重要である。そこで我々はこれまでに、機械学習手法の一つである Conditional Random Field (CRF) を用いて実験に関する図表や、段落を抽出する手法を提案した。しかし実験結果には、その論文の価値を決めるような重要なものとそうでないものがある。また、実験結果を示す表は、多くの場合、数値が記載されているが、どの数値が優れているか等、実験情報の重要性は一見しただけでは判別しにくい。そこで本稿では、CRF を用いて実験に関する段落と表を抽出し、段落の強調表示や、表のグラフ化によって実験情報を可視化する手法を提案する。

キーワード: 学術論文, 情報抽出, 可視化

Experimental information extraction from academic articles and its visualization

HISATAKA HIRAI^{1,a)} HIROTAKA NIITSUMA^{1,b)} MANABU OHTA^{1,c)} ATSUHIRO TAKASU^{2,d)}

Keywords: academic article, information extraction, visualization

1. はじめに

近年、CiNii^{*1}等の学術論文データベースの充実により、膨大な数の論文を手軽に入手できるようになった。しかし、論文は機械が処理できるデータと異なり、人が読まなければ有効に活用できない。これまでに、自動要約やキーワード抽出をはじめ、文章を効率良く読むことを支援する研究 [1][2][3][4] が行われている。また、研究者にとって論文の実験の手法や結果をまとめることは、研究内容の整理や比較に必要不可欠である。

そこで本研究は、論文中の実験に関連のある論文構成要

素を実験情報と呼び、これらの情報を論文から自動抽出し、可視化することを目的とする。

これまでに、論文の概要から重要文を抽出する研究 [5] が行われている。しかし、概要には実験の内容理解に必要な具体的な情報が記述されていない場合も多い。例えば、概要には論文中で報告された実験の中で最も効果のあった結果のみが記載されることはあっても、通常実験の詳細については記載されない。また、実験内容は図表を用いて説明されることが多く、これまでに図表や図表説明文を抽出する研究 [6][7] も行われてきた。しかし、図表には実験と関連のないものもある。本研究では、これまでに提案した方法により抽出した、実験に関する表や段落を可視化する手法を提案する。

本稿の構成は次の通りである。まず、2節で学術情報の抽出や可視化に関する研究を紹介し、3節で論文からの論文構成要素抽出と、論文構成要素抽出結果を利用した実験情報抽出について説明する。続く4節で実験情報の抽出実験について述べ、5節で抽出した実験情報の可視化につい

¹ 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology,
Okayama University

² 国立情報学研究所
National Institute of Informatics

a) hirai@de.cs.okayama-u.ac.jp

b) niitsuma@suri.cs.okayama-u.ac.jp

c) ohta@de.cs.okayama-u.ac.jp

d) takasu@nii.ac.jp

*1 <http://ci.nii.ac.jp/>

て述べる。6節で可視化に関する評価実験を行い、最後に7節で本稿をまとめる。

2. 関連研究

2.1 情報抽出

学術情報の抽出に関する研究として檜本ら [8][9] は、学術論文から図、表、脚注、参考文献領域といった論文構成要素を、ルールと Support Vector Machine(SVM) により自動抽出する手法を提案した。彼らは、学術論文の XML テキストにこれらの論文構成要素タグを付与することでこれらの要素を抽出し、DEIM2013^{*2} の学術論文 50 件に対して論文構成要素抽出実験を行った。ルールにより個々の論文構成要素を抽出し、抽出精度として平均の F 値 0.897 が報告されている。一方、SVM による論文構成要素の抽出は、二段階で行った。まず各論文構成要素の開始位置と終了位置を SVM により抽出し、抽出精度として平均の F 値 0.742 が得られている。次に、SVM の出力結果を素性として使い、再び SVM で論文構成要素を抽出した。その結果、平均の F 値が 0.749 と向上し、一段階の抽出よりも有効であると報告されている。また、彼らは SVM と同様の素性を用いて CRF により論文構成要素を抽出し、平均の F 値 0.850 を確認している。

竹島ら [10] は、論文読解を支援するために、図表説明文の抽出法を提案した。図や表は論文の重要な内容を表していると考えられるが、図や表のみでは論文の内容を理解することは難しい。そこで、竹島らは図や表を参照する文を起点として文に付加した重みを他の文へ伝播させることで、図表説明文を抽出した。4 件の論文から 24 の図表と図表説明文を抽出し、その結果、平均の適合率 0.768 が報告されている。

2.2 可視化

福田ら [11][12] は、論文と特許から技術動向に関する情報を抽出し、抽出した情報を利用して技術動向マップを自動作成して可視化する手法を提案した。この技術動向マップに要素技術として提示されている用語をユーザがクリックすることで、その要素技術がどのような分野で利用されているのかを、年代順に表示することができる。年代順に表示することで、要素技術の変遷が分かる。例えば「半導体レーザ」は 2002 年までは、主に画像系の分野において使われていたが、2004 年に入ると論理回路の分野でも利用されるようになったことが技術動向マップにより判明した。さらに福田らは、技術動向マップの作成に、特定分野において使用された基礎的な要素技術とその効果を利用した。このような要素技術とその効果の変遷を知ることは、その分野における技術動向を把握するために重要である。

そこで、要素技術とその効果に関する表現を自動的に抽出するための手法を提案した。福田らは、国立情報学研究所主催の第 8 回 NTCIR ワークショップ (NTCIR-8)^{*3} 特許マイニングタスクで提供されたデータと SVM を用いて実験し、表題及び概要から、要素技術やその効果を示す用語を抽出した。その結果、論文解析における用語抽出は再現率 0.254、適合率 0.496、F 値 0.336、特許解析における用語抽出は再現率 0.441、適合率 0.537、F 値 0.484 となったことが報告されている。

村田ら [13] は自然言語処理に関わる論文の概要から重要な情報を抽出する手法を提案した。彼らは重要な情報を「精度表現」、「自然言語処理における分野」、「言語名」、「組織・人名」の 4 つの分野として定めたが、これらの情報は様々な目的で役立つ。例えば、関連する論文を検索するためのキーワードとしての利用や、自然言語処理分野の論文のサーベイの自動構築等の目的での利用ができる。彼らは、SVM を用いたテキストチャンカーである YamCha^{*4} を利用してこの 4 分野の単語を抽出し、抽出精度の平均の F 値は 0.8 だった。また、村田らは抽出した各分野の単語を表やグラフで可視化する手法も提案した。表の列毎に抽出した重要情報をまとめることで、論文の特徴や状況を一度に把握できるようになった。さらに、抽出した重要な情報をグラフにより可視化することで、論文数の分布を知ることができ、自然言語処理に関わる分野の論文のサーベイや、各言語を扱う論文の傾向の理解に役立つと報告されている。しかし、村田らは論文の概要のみを抽出対象としているため、詳細な情報は抽出できない。本研究では、論文全体から実験に関する情報を抽出し、実験に関する詳細な情報の可視化を試みる。

3. 実験情報抽出

論文中の実験についての記述は様々である。例えば、データセットに関する記述、評価指標に関する記述、実験結果に関する記述がある。本稿では、それらの記述をまとめて実験情報と呼び、抽出対象の論文として、NTCIR-9^{*5} の Spoken Document タスクに投稿された論文を扱う。

本研究で実験情報抽出は、2 段階で行う。まず論文からルールにより論文構成要素を抽出する。次に抽出した論文構成要素を用いて、CRF により実験情報を抽出する。本稿では、抽出後の可視化を目的とするため、論文構成要素として論文に記載されている表キャプション、表、段落の 3 種類を抽出目標とする。

我々は [14] でこの抽出手法を提案したが、本研究では、この手法で抽出した実験情報を可視化するため、手法の概

^{*3} <http://research.nii.ac.jp/ntcir/ntcir-ws8/meeting/index-ja.html>

^{*4} <http://chasen.org/taku/software/yamcha/>

^{*5} <http://research.nii.ac.jp/ntcir/ntcir-9/index-ja.html>

^{*2} <http://db-event.jpn.org/deim2013/>

表 1: 論文構成要素タグ一覧

論文構成要素	開始タグ	終了タグ
表キャプション	<TBLCAP>	</TBLCAP>
表	<TBL>	</TBL>
段落	<PAR>	</PAR>

表 2: 論文構成要素タグの付与ルール

論文構成要素タグ	付与ルール
<TBLCAP>	1 単語目が表を意味する単語, 2 単語目が数字, 3 単語目が大文字から始まる TEXT に付与
</TBLCAP>	<TBLCAP> が付与された TEXT, もしくは <TBLCAP> を付与した TEXT と x 座標が同じ TEXT に付与
<TBL>	前の TEXT に </TBLCAP> が付与されている TEXT に付与
</TBL>	1 つ以上前の TEXT に <TBL> が付与されており, 次の TEXT と y 座標の差が閾値より大きい TEXT に付与
<PAR>	TEXT の幅が閾値内であり TEXT の最初の文字が字下げされている TEXT に付与
</PAR>	次の TEXT の文字が字下げされているか, 次の TEXT との y 座標の差が閾値以上ある TEXT に付与

要を以下に述べる.

3.1 ルールによる論文構成要素抽出

本研究では, 我々が提案した論文構成要素抽出プログラム [14] を利用する. この論文構成要素抽出プログラムは, 論文の PDF を pdf2xml^{*6} で変換した XML の TEXT タグに, 論文構成要素タグを付与することで抽出する. TEXT タグは XML で記述された文章の行等に付与される. 他に表や段落等に付与される BLOCK タグ, 単語等に付与される TOKEN タグなどがある. 本研究で付与する論文構成要素タグは, 表キャプション, 表, 段落のみである. これらの論文構成要素タグの一覧を表 1 にまとめる.

一方これらの論文構成要素抽出ルールを表 2 にまとめる. 表 2 のルールは, 表キャプションと表は <TBLCAP>, </TBLCAP>, <TBL>, </TBL>, 段落は <PAR>, </PAR> の順にルールを適用する. またここで表を意味する単語とは “Table”, “Tbl”, のことである.

3.2 CRF による実験情報抽出

本研究では, 自然言語処理等の様々な分野で利用されている識別モデルである CRF を利用して実験情報を抽出する. 抽出には 3.1 節で述べた論文構成要素を CRF の素性として利用する. また, 本研究では, [14] で我々が提

表 3: 素性テンプレート

種類	素性	内容
Unigram (レイアウト素性)	<text.y(0)>	TEXT の y 座標
	<text.w(0)>	TEXT の幅
Unigram (言語的素性)	<component(0)>	ルールにより付与された論文構成要素タグ
	<table(0)>	表を意味する単語の有無
	<point(0)>	表に関連の深い単語の有無
	<measure(0)>	実験の評価指標を表す単語の有無
	<effect(0)>	評価指標の効果を表す単語の有無
	<chapter(0)>	実験情報を示唆する節題の有無
Bigram	<y(-1), y(0)>	タグの遷移

表 4: 段落の手がかり語とその種類

種類	手がかり語
TABLE	Table, Tbl
POINT	show, list, draw, illustrate
MEASURE	F-measure, recall, precision, score, performance, experiment, result, dry-run, formal-run, method
EFFECT	improve, maximum, compare, average, point, degrade, best, better, submit, evaluation

案した手法と同様に, CRF++0.58^{*7} を用いる. ここで, CRF++0.58 で利用する素性テンプレートを表 3 に示す. 表 3 の素性は, レイアウト素性と言語的素性で構成されている. 表 3 で, ルールにより付与された論文構成要素タグは, 表キャプション, 表, 段落のそれぞれの開始, 終了タグ 6 種類であり, これらの論文構成要素タグは, 3.1 節で述べたルールを利用して付与されたものである. 言語的素性で, 表を意味する単語とは表 4 の TABLE, 表と関連の深い単語は表 4 の POINT, 実験の評価指標を表す単語は表 4 の MEASURE, 評価指標の効果を表す単語は表 4 の EFFECT に含まれる手がかり語である. また, 実験情報を示唆する節題とは, 表 5 のいずれかの単語を含む節題である. これらの単語は抽出対象の論文から人手で収集したものである. 表 3 の素性を用いて, 実験情報の表キャプションのタグ, 実験情報の表のタグ, 実験情報の段落のタグを XML の TEXT に付与する. また, 付与される実験情報タグの接続に関する情報は Bigram 素性によって考慮される.

^{*6} <http://souceforge.net/projects/pdf2xml>

^{*7} <https://taku910.github.io/crfpp/>

表 5: 節題の手掛かり語リスト

ABSTRACT, ANALYSIS, CONCLUSION, DESCRIPTION, DISCUSSION, EVALATION, EXPERIMENT, METHODOLOGY, RESULT, RETRIEVAL, TRAINING, TEST, SUMMARY, SYSTEM

表 6: 論文構成要素タグ付与結果

論文構成要素タグ	recall	precision	F 値
<TBLCAP>	1.000	1.000	1.000
</TBLCAP>	1.000	1.000	1.000
<TBL>	1.000	1.000	1.000
</TBL>	0.897	0.897	0.897
<PAR>	0.926	0.895	0.911
</PAR>	0.878	0.847	0.862
平均	0.950	0.939	0.945

4. 実験情報の抽出実験

4.1 論文構成要素へのタグ付与実験

データセットには, NTCIR-9 の Spoken Document タスクに投稿された論文 10 件を用い, 評価指標には, 以下に示す recall, precision, F 値を用いた.

$$\text{recall} = \frac{\text{正しく付与したタグの数}}{\text{正解データのタグの数}} \quad (1)$$

$$\text{precision} = \frac{\text{正しく付与したタグの数}}{\text{提案手法が付与したタグの数}} \quad (2)$$

$$\text{F 値} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

データセットに対し, 各論文構成要素タグを付与し, F 値で評価した. 結果を表 6 に示す. 表 6 より, 平均の recall, precision, F 値それぞれが 0.94 程度となった. このように, ルールで付与した論文構成要素タグを, 実験情報抽出において CRF の素性として利用する.

4.2 CRF による実験情報抽出実験

テストデータを抽出対象の NTCIR-9 の Spoken Document タスクの論文 10 件, 学習データを NTCIR-9 で投稿されたその他のタスクの論文 81 件とした. 評価指標は 4.1 節と同様に recall, precision, F 値とし, 実験情報として抽出した表キャプション, 表, 段落毎に算出した. 以下に, これらの評価指標の定義を示す.

$$\text{recall} = \frac{\text{正しく抽出した実験情報の数}}{\text{正解データの実験情報の数}} \quad (4)$$

$$\text{precision} = \frac{\text{正しく抽出した実験情報の数}}{\text{抽出した実験情報の数}} \quad (5)$$

$$\text{F 値} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

表 3 の素性テンプレートを利用し, 実験情報の抽出を行っ

表 7: CRF による実験情報抽出

実験情報	recall	precision	F 値
表キャプション	0.810	0.857	0.833
表	0.870	0.620	0.724
段落	0.528	0.682	0.589
平均	0.736	0.720	0.715

た. 結果を表 7 にまとめる. 表 7 の抽出結果を 5 節で述べる実験情報の可視化に利用する.

5. 実験情報の可視化

4 節の実験で抽出した実験情報を可視化する手法について説明する. 本研究では, 論文の PDF を変換した HTML を用いて可視化する. まず, 可視化対象の論文を pdf2htmlEX^{*8}により HTML に変換する. 次に, 4 節の方法で XML ファイルから抽出した実験情報の表キャプション, 表, 段落の文字列情報と論文の HTML の文字列情報を照合して, HTML において実験情報の位置を特定する. その結果を利用して段落の強調表示や表のグラフ化を行う. 抽出した実験情報の段落を強調表示することで, 実験に関する記述箇所が一目で分かる. また, 表をグラフ表示することで表の視認性が向上するため, 論文を効率良く読むことができる. 以下で, 実験情報の段落の可視化と実験情報の表の可視化について述べる.

5.1 実験情報の段落の可視化

図 1 に実験情報の段落を強調表示した論文の HTML を示す. これは [16] の論文を HTML 表示したものである. 図 1 では実験情報の段落を薄い青で強調表示した. このように, 強調表示された論文の HTML を見ることで, 実験に関する内容の記載箇所を容易に把握できる.

しかし本稿では実験情報を, データセット, 評価指標, 実験結果などに分類していない. そのため, 強調表示された段落がどのような実験情報について書かれた記述か分からない. よって, 実験情報をこのような種類に分けて抽出し, 抽出した実験情報の種類に合わせて表示の色を変えることなどを検討している.

5.2 実験情報の表の可視化

表を用いて実験結果をまとめる場合は多い. しかし, 指標や数値が多い場合, 視認性に問題があり, 本文などをよく読まないで表の内容の理解が困難な場合がある. そこで本研究では, 抽出した実験情報の表のグラフ化を提案する. これにより, 視認性を向上させ, どの数値が優れているかなどの判断が容易に行えるようにする. 以下で表の解析手法, グラフ表示方法について述べる.

^{*8} <https://github.com/coolwanglu/pdf2htmlEX>

5.2.1 表の解析手法

XML ファイルから抽出した表を一行ずつ解析し、グラフを構成する要素であるグラフ名、凡例、数値、手法名を判別する。以下で、各要素の判別手法について述べる。

- グラフ名
論文から抽出した実験情報の表キャプションをグラフ名として用いる。
- 凡例
実験情報の表から、数値の数が5個以下の行の各単語を凡例として用いる。
- 数値
各数値を数値表現として用いる。この時、数値欄に“-”が記載される場合があるが、本研究では“-”は値なしとする。
- 手法名
数値表現と判別されなかった行の文字列を手法名とする。

5.2.2 表のグラフ表示

5.2.1 節で判別した各グラフ要素を Javascript に変換し、Google Chart API*⁹ により HTML としてグラフ表示する。Google Chart API は数値データを自動的にグラフ変換し、HTML として取得できるサービスである。また、変換したグラフを論文 HTML と一緒に表示するため、論文の XML ファイルから抽出した実験情報の表キャプションの文字列と論文の HTML ファイルの文字列情報を照合して、論文 HTML 中の表キャプションのタグ内にイベント属性である“onclick”を含める。これにより、クリックで論文上にグラフを表示できる。図2は[16]とその表をグラフ化したものを HTML 表示した例である。図2では、グラフ名、手法名、凡例と数値が適切に抽出され、可視化できている。また、グラフを表示するウィンドウは任意の位置に自由に動かすことができる。

6. 評価実験

5.2 節で、表のグラフによる可視化について述べた。しかし、表の書き方は様々なため、この方法でどの様な表が適切にグラフ化できるかを実験により評価する。本研究の評価実験では、実験情報を抽出した Spoken Document タスクの10件の論文から著者が一つずつ表を選定し、それらの表をグラフ変換した。グラフ変換の評価は以下の式で定義した可視化率で行う。

$$\text{可視化率} = \frac{\text{グラフ化に成功した表の数}}{\text{選定した表の数 (10)}} \quad (7)$$

表のグラフ化に成功しかどうかの正解判定は著者が行い、表の凡例、数値、手法名が過不足なく表示されているグラフを正解と判定した。

*⁹ <https://developers.google.com/chart/>

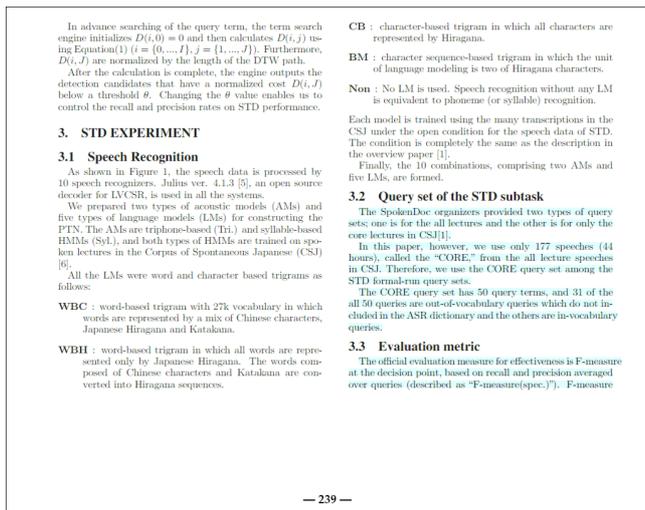


図 1: 強調表示した論文 HTML[16]

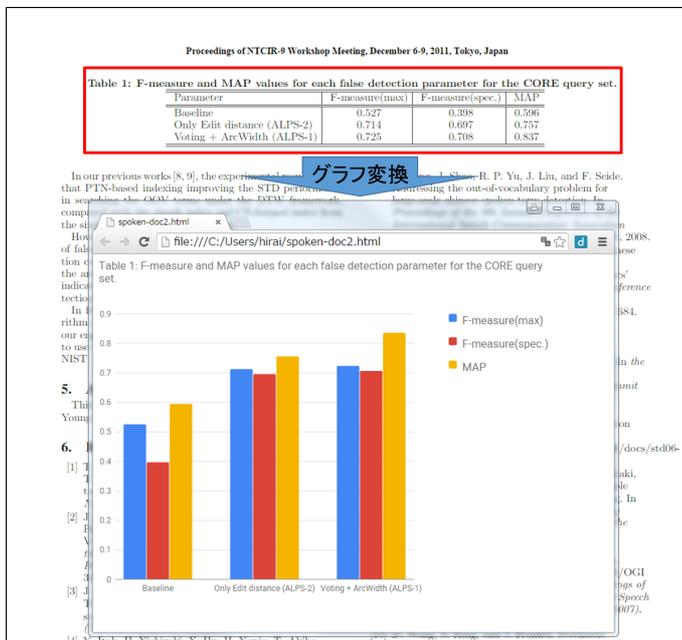


図 2: グラフの HTML 表示例 [16]

表 8: 実験情報の表のグラフ変換結果

提案手法	グラフ化に成功した表の数	グラフ化に失敗した表の数	可視化率
	7	3	0.7

グラフ変換の結果を表8に示す。表8より、グラフ化に成功した表の数は7個で、可視化率は0.7となった。図3にグラフ化に成功した表[17]とそのグラフを載せる。図3より、変換前の表を一見しただけではどの数値が最も優れているか等の判別が困難であるが、グラフ化によりその問題を改善できた。しかし、グラフ化には成功したが、グラフ表示する上で問題のある表もあった。図4にその表[18]とグラフの例を載せたが、このグラフでは一部の数値が識別できない。これは、指標の種類などによって数値の尺度が異なるため、一つの目盛で対応できないことが原因である。そのため、数値の尺度を指標等ごとに判別し、その種

Transcript type	Segmentation type	uMAP	pwMAP	fMAP
BASELINE		0.0670	0.0520	0.0536
manual	tt	0.0859	0.0429	0.0500
manual	C99	0.0713	0.0209	0.0168
ASR	tt	0.0490	0.0329	0.0308
ASR	C99	0.0469	0.0166	0.0123
ASR_nsw	tt	0.0312	0.0141	0.0174
ASR_nsw	C99	0.0316	0.0138	0.0120

グラフ変換

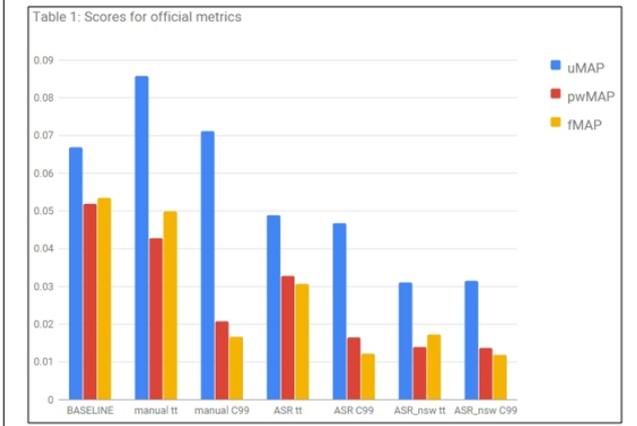


図 3: グラフ化に成功した表の例 [17]

類に応じて縦軸を増やすなどして、適切なグラフの表示方法を検討していきたい。また、グラフ化に失敗した表 [19] の例を図 5 に示す。この表は、図 5 の表中の手法名にあたる文字列が複数行にわたって記載されているため、5.2.1 節の方法で行を解析するだけでは手法名を取り出すことが出来なかった。そのため、表の解析手法についてさらに検討し、多様な記述方法にも対応できるようにしていきたい。

7. まとめ

本研究では、我々が [14] で提案した手法を用いて学術論文から表や段落などの実験情報を抽出し、抽出した実験情報の可視化を試みた。実験情報について書かれた段落の強調表示により、実験情報の段落を明示するインタフェースを示した。また、実験情報の表をグラフ化することで、実験情報の視認性が向上することを確認した。

今後の課題として、段落の強調表示については実験情報を細分類できるように抽出手法を改善し、その結果を可視化に反映させたいと考えている。また表のグラフ化については、いくつかの標準的な形式をもつ表に対して、適切なグラフ化の手法を検討していきたい。さらに、可視化したグラフと本文の記述を対応づけるなど、論文を効率良く読むための機能の実装も検討している。

謝辞

本研究の一部は、科学研究費補助金基盤研究 (C)(課題番号 25330384, 15H02789) および国立情報学研究所公募型共同研究の援助による。ここに記して深謝する。

Table 4: Result of IV retrieval (LVCSR+n-gram)

	LVCSR	n-gram(5-best)	LVCSR+n-gram	DTW	LVCSR+DTW
Detect	103	96	147	88	141
Correct	98	84	130	83	131
Recall	0.51	0.44	0.68	0.43	0.69
Precision	0.95	0.86	0.88	0.94	0.93
F-measure	0.67	0.59	0.77	0.59	0.79

グラフ変換

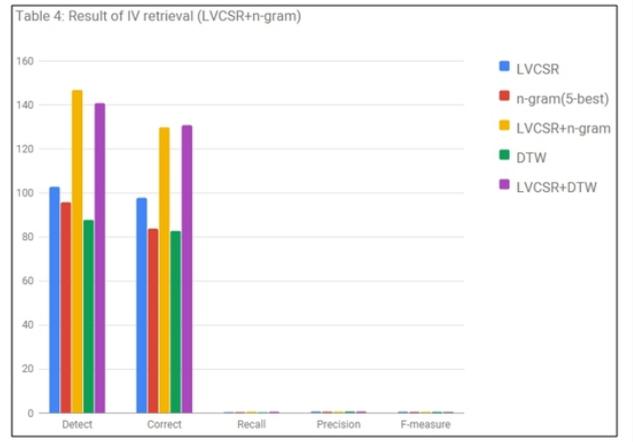


図 4: 適切なグラフ表示とはいえない表の例 [18]

Table 5: The best performance using four results and introducing new local distances

MODEL	F-measure(%)	MAP(%)
Demiphone + SPS + Intensive triphone + Triphone by organizer	65.81	81.43

図 5: グラフ化に失敗した表の例 [19]

参考文献

- [1] 奥村学, 難波英嗣: テキスト自動要約に関する研究動向, 自然言語処理, vol. 6, no. 6, pp. 1-26, 1999.
- [2] H.P.Luhn: The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, 1958.
- [3] 松尾豊, 石塚満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, vol. 17, no. 3, pp. 213-227, 2002.
- [4] 鉢木稔浩, 太田学, 高須淳宏: Web 資源を利用した学術論文閲覧支援システム, 情報処理学会研究報告, vol. 2009-DBS-149, no. 14, pp. 16, 2009.
- [5] 徳永康次, 延澤志保, 太原育夫: テキスト構造に着目した学術論文の要旨自動生成のための重要文抽出, 第 6 回情報科学フォーラム, E-032, pp. 215-216, 2007.
- [6] 市野順子, 箕牧数成, 山口和泰, 垣智, 東郁雄, 古田重信: 図表検索のための図表情報自動抽出の試み, 情報処理学会研究報告, vol. 2002, no. 28, pp. 143-150, 2002.
- [7] D.Pinto, A.McCallum, X.Wei and W.B.Croft: Table extraction using conditional random fields, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 235-242, 2003.
- [8] 榎本達矢, 太田学, 高須淳宏: 学術論文からの構成要素抽出の一手法, DEIM Forum 2014, C5-2, 2014.

- [9] 榎本達矢, 太田学, 高須淳宏 : 学術論文からの構成要素抽出法の改良, DEIM Forum 2015, B4-6, 2015.
- [10] 竹島亮, 渡邊豊英 : 文と単語の相互依存性に注目した図表説明文の抽出, 電子情報通信学会技術研究報告, vol. 110, no. 85, pp. 43-48, 2010.
- [11] 福田悟志, 難波英嗣, 竹澤寿幸 : 論文と特許からの技術動向情報の抽出と可視化, 情報処理学会論文誌, vol. 6, pp. 16-29, 2013.
- [12] 福田悟志, 難波英嗣, 竹澤寿幸 : 技術文書からの動向情報抽出と可視化, 言語処理学会第 17 回年次大会, pp. 276-279, 2011.
- [13] 村田真樹, Stijn De Saeger, 橋本力, 風間淳一, 山田一郎, 黒田航, 馬青, 相澤彰子, 鳥澤健太郎 : 論文データからの重要情報の抽出と可視化, 第 23 回人工知能学会全国大会, 3F2-NFC3-9, 2009.
- [14] 平井久貴, 新妻弘崇, 太田学, 高須淳宏 : 学術論文からの実験情報抽出の一手法, DEIM Forum 2015, F3-1, 2015.
- [15] J.Lafferty, A.McCallum and F.Pereira : Conditional Random Fields : Probabilistic Models for Segmenting and labeling Sequence Data, In Proc. of 18th International Conference on Machine Learning, pp. 282-289, 2001.
- [16] H. Nishizaki, H. Furuya, S. Natori, and Y. Sekiguchi : Spoken term detection using multiple speechrecognizers' outputs at NTCIR-9 SpokenDoc STD subtask, In Proceedings of the Ninth NTCIR Workshop Meeting, 2011.
- [17] M. Eskevich and G. J. F. Jones : DCU at the NTCIR-9 SpokenDoc passage retrieval task, In Proceedings of the Ninth NTCIR Workshop Meeting, 2011.
- [18] K. Iwami and S. Nakagawa : High speed spoken term detection by combination of n-gram array of a syllable lattice and LVCSR result for NTCIR-SpokenDoc, In Proceedings of the Ninth NTCIR Workshop Meeting, 2011.
- [19] H. Saito, T. Nakano, S. Narumi, T. Chiba, K. Kon' No and Y. Itoh : An STD system for OOV query terms using various subword units, In Proceedings of the Ninth NTCIR Workshop Meeting, 2011.