技術文書中の図表と本文の自動対応付け

Automatic alignment of figures and tables with texts in technical documents

広島市立大学大学院　情報科学研究科准教授
難波　英嗣

2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。博士（情報科学）。東京工業大学精密工学研究所助手等を経て、2010 年より広島市立大学大学院情報科学研究科准教授。自然言語処理、テキストマイニングの研究に従事。

nanba@hiroshima-cu.ac.jp　TEL 082-830-1584

1　はじめに

近年、技術分野の専門分化と共に特許や論文等の技術文書数が爆発的に増加している。その一方で、技術者がそのすべてに目を通すことが困難になっており、技術文書の内容を効率的に把握するための読解支援システムが求められている。

これまでに、様々な観点から技術文書の読解を支援するシステムが提案されてきたが、本稿では、技術文書内の図表に着目したシステムを構築する。図表は、技術文書中の説明文の要点を表したものと捉えることができ、被験者に説明文だけを提示するよりも、図表とともに説明文を提示したほうが、内容の理解に有用であるという調査報告もある[岩槻 98]。本研究では、技術文書中の図表と、個々の図表に対応する文を自動的に対応付け、ユーザに表示することで読解の支援を行う。また、図表と文の対応付けを「系列ラベリング問題」として、機械学習を用いて解く。これにより、技術者が文献調査する際の労力・手間が改善し、円滑な研究の支援ができるようになると考えられる。

本稿の構成は以下のとおりである、2 章では関連研究を紹介する。3 章では図表に関連する文の自動対応付けについて述べる。4 章では評価実験とその結果、考察を述べる。5 章で本稿をまとめる。

2　関連研究

本章では、論文読解支援、技術文書マイニング、文脈情報を考慮した情報抽出に関する関連研究について述べる。

2.1　論文読解支援

Abekawa らは、論文のデジタル化の中で、PDF ファイル内に、文書を理解する上で必要な論理構造を示す情報が含んでいないことに着目し、PDF から必要な情報を抽出するシステムである SideNoter について紹介している[Abekawa 16]。SideNoter は三つの機能を与えている。一つ目は図表検索機能、二つ目は、関連部門検索機能、三つ目は論文 1 ページごとの情報推薦機能である。本研究でも、論文をテーマとした研究で共通しているが、論文内の図表を用いた文脈情報を考慮した文の抽出を行う点で異なる。

2.2　技術文書マイニング

難波らはある分野の特許と論文から、技術動向マップを自動的に作成することを最終目標としている特許マイニングタスクについて紹介している[難波 09]。このタスクでは、「学術論文分類サブタスク」、「技術動向マップ作成サブタスク」の 2 つのサブタスクを設定している。このうちの学術論文分類サブタスクの成果として学術論文自動分類システムを構築、紹介している。このシステムは論文の概要を入力すると、内容語（名詞、動詞、形容詞）を自動抽出し、それらの内容語を検索キーワードとして特許検索システムを用いて関連特許を検索する。この際、検索結果上位 170 件の各特許に付与された IPC コードを抽出し、コード別にスコアを計算し、

スコアの高い順に出力している。

評価は、分類システムが出力した上位 n 件の分類コードが、人手で付与した IPC コードをどの程度正しく抽出できているかによって調査し、再現率は検索結果上位 1 件で約 20%、上位 10 件で約 60% となった。この結果を受けて、再現率のさらなる向上が必要ではあるが、特許の検索初心者にとってはある程度有用であると述べている。本稿においても部分テキストの自動抽出を行うが、図表に関連しているテキストのみを対象に抽出するという点で異なる。

技術動向情報の抽出、可視化を行っている研究に福田らのものがある[福田 13]。福田らは論文および特許の表題と概要を解析することで技術動向情報の可視化を目指している。一般に、論文と特許では、表現や形式が記述スタイルの面で大きく異なっており、ルールに対応付けて解析することは難しい。そこで福田らはこの問題を「要素技術とその効果を示すタグを付与」するという系列ラベリング問題として考え、機械学習を用いたタグの自動付与を行っている。ここでの要素技術とは、研究において使用されたアルゴリズムやツール、技術的手法のことを指し、その要素技術から得られる知見を効果とする。機械学習の素性として用いる要素技術や効果を示す手掛かり語のリスト作成する必要がある。そのため、福田らは手掛かり語のリストを網羅的に作成するために、係り受け関係や上位下位関係による人手での収集、さらに分布類似度を用いた語句の自動収集を行う手法を提案している。実験では、NTCIR-8 特許マイニングタスクのデータを実験データとして使用し、評価尺度には精度・再現率・F 値を用いた。実験の結果、上記の素性を用いた提案手法がすべてのベースライン手法を上回っており、提案手法の有用性を示した。本稿でも同様に、系列ラベリングによる文脈情報を考慮した文の抽出を行うが、文脈情報だけでなく技術文書内の図表の情報も用いる点で異なる。

2.3　文脈情報を考慮した情報抽出

Willot らは論文概要の各文に対し、「背景」、「目的」、「手法」、「結果」のラベルを自動付与する手法を提案している[Willot 15]。Willot らは、論文概要中の各文に付与されるラベルには、例えば「目的」→「手法」→「結果」のようにパターンがあると考え、このタスクを系列ラベリング問題と捉え、リカレントニューラルネットワークの一種である Long short-term memory（LSTM）を用いてラベルの自動付与を実現している。

本稿では、ある図表に関連する一連の文を対応付ける、というタスクであり、系列ラベリング問題として解くことができる。ただし、Willot らの研究では、概要中の文集合のみが与えられるのに対し、本稿では、技術文書中の文集合の他に図表そのものと図表のキャプションも同時に与えられる点が異なる。

3　図表に関連する文の自動対応付け

3.1　タスクの概要

本研究では結果を示す図表を入力として、その図表と対応関係にある文章を自動的に抽出し、表示するシステムを構築する。例を図 1 に示す。

	Train	Dev.a	Dev.b
# of sent.	2,032,679	1,000	1,000
# of En words	48,322,058	31,890	31,935
Enju suc. rate	99.3%	98.9%	98.7%
parse time (sec./sent.)	0.30	0.38	0.48
# of Jp words	53,865,629	37,066	35,921

Table 3: Statistics of the experiment sets.

The statistics of the filtered training set, dev.a, and dev.b are shown in Table 3. The success parsing rate ranges from 98.7% to 99.3% by using Enju2.3.1. The averaged parsing time for each English sentence ranges from 0.30 to 0.48 seconds.

図 1　表とその対応文（[Wu 11] より引用）

この例の場合、図 1 の上部にある表は Table 3 であるため、Table 3 の記載のある文と関連があると考えられる。さらに、キャプションだけでなく、表内の語句や数値が含まれている文も表と関連していると考えられる。特に表内の数値は結果に関するものが記載されていることがあるため、重要であるといえる。

次に、図の場合の例を図 2 に示す。図においても表と同様に、Figure の記載のある文が対応関係にあると考えられる。また、図内の語句が用いられている文についても、対応関係があると考えられる。

このように、技術文書中には図表に対応するテキストが必ず存在する。また、そのテキストは図表に対する説明等の言及をしている場合が多いため、対応文抽出は技術文書の内容を把握するのに重要だと考えられる。本研究ではこの図表に対応文の自動抽出を目的とする。

3.2　システム概要

本節では、本研究で構築するシステムの概要について説明を行う。本システムの流れは以下のようになる。

（1）技術文書中の図表番号の入力

（2）図表画像の解析モジュール

（3）図表関連テキスト抽出モジュール

（4）抽出結果の出力

本研究では、技術文書でも特に、英語で記述された自然言語処理分野の学術論文を対象にする[1]。本システムの入力には Abekawa ら[Abekawa 16] の研究によって XHTML 形式で出力された ACL Anthology の論文データを用いる。3.3 節では、図表画像解析モジュールについて、3.4 節では、図表関連テキスト抽出モジュールについて説明を行う。

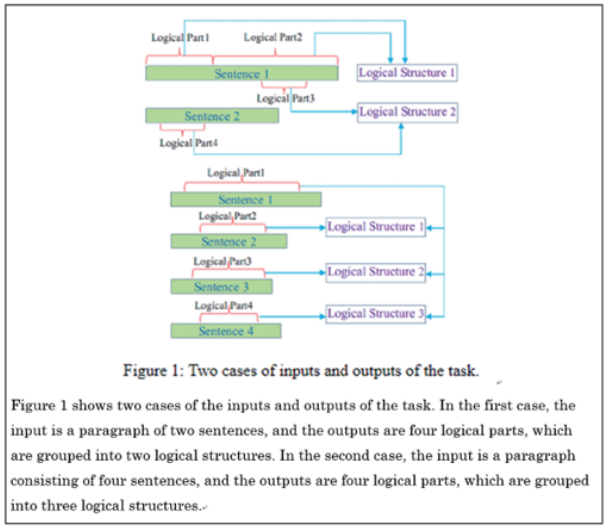

Figure 1: Two cases of inputs and outputs of the task.

Figure 1 shows two cases of the inputs and outputs of the task. In the first case, the input is a paragraph of two sentences, and the outputs are four logical parts, which are grouped into two logical structures. In the second case, the input is a paragraph consisting of four sentences, and the outputs are four logical parts, which are grouped into three logical structures.

図２　論文中の図とその対応文（[Bach 11] より引用）

3.3　図表画像の解析モジュール

本節では、図表中画像を解析し、図表中の文字・数値情報を抽出するモジュールについて説明を行う。Abekawa らは PDF 形式の論文ファイルから、論文表題、概要、著者名、節情報、参考文献等の論文の構造を抽出し、XHTML 形式で出力しており、図表はすべて png 形式で統一されている[Abekawa 16]。本研究ではこの XHTML ファイルを用いて画像解析を行う。画像解析には Google Cloud Vision API を用いる。図３

Chunking	F1	CCG tagging	Accuracy
Our model	93.21	Our model	92.41
Suzuki and Isozaki (2008)	93.88	Xu et al. (2015)	93.00

Table 4: Baseline model, comparison to existing systems

図３　論文中の表の例（[Plank 16] より引用）

Chunking
Our model
Suzuki and Isozaki (2008) 93.88Xu et al. (2015) 93.00
Fl CCG tagging Accuracy
93.21 Our model
92.41

図４　図３の文字認識結果

に論文中の表の例を、図４にその解析結果を示す。

図４から、図３の Table 4 内の文字列がテキスト形式に変換されていることが分かる。また、この解析ではテキスト内の単語の位置情報を獲得することができる。この位置情報を用いることで、例えば、「数字が規則正しく並んでいたら結果を示している図表」というような推測が可能になる。本研究では図表の対応文抽出を行う際に、この画像の解析結果を用いることで、図表内の情報を考慮したシステムの構築を目指す。

3.4　図表関連テキスト抽出モジュール

本節では、学術論文内の図表に関連のあるテキストを抽出するモジュールについての説明を行う。ここでの、図表に関連のあるテキストとは、その図表に対しての言及、または図表内の用語についての説明がある文章（文の集合）のことである。図表とテキストの最も単純な対応関係は、テキスト内に図表のキャプション名（Figure 1、Table 2 等）が含まれているかどうかとなる。しかし、実際にはキャプション名が含まれている文のみが対応関係にあるというケースは少なく、キャプション名の含まれているテキストの前後にわたって対応関係にあるテキストが連なっていることが多い。図５に図表と対応文の例を示す。

図５の例の場合、Table 4 の記載のあるテキストが対応文となるが、この対応文に続く文が “Our chunking” で始まっている。“Our” で始める文はその研究における提案手法を指している場合が多く、結果を示す図表と対応していることが考えられる。この他に、“The result ～” のように、“The” で始まる文が対応文

1　図表とテキストの対応付けに、言語依存およびジャンル依存の情報を用いていないため、本稿で提案する手法は、日本語で記述された文書や特許などにも用意に適用できると考えられる。

Chunking	F1	CCG tagging	Accuracy
Our model	93.21	Our model	92.41
Suzuki and Isozaki (2008)	93.88	Xu et al. (2015)	93.00

Table 4: Baseline model, comparison to existing systems

[Plank 2016]
Baseline model Both or baseline models are comparable to prior work, while being simpler. The results are summarized in Table 4. Our chunking baseline achieves an F 1 of 93.21 on CoNLL,compared to the F 1 of 93.88 of Suzuki and Isozaki (2008), who use a CRF and gold POS tags.

図５　論文中の表とその対応文（[Plank 16] より引用）

に続いている場合は、それらの文も対応づいている場合が多いことが分かっている。また、Table ４のキャプション名に “Baseline model” という語句が入っている。そのため、“Baseline model” から始まっている文も図表との関連が考えられる。さらに、この表に 3.3 節で述べた画像解析を行うと、Table ４内に Suzuki and Isozaki という語句があることが分かる。このことから、Suzuki and Isozaki という語句が含まれる文も Table ４に関連しているのではないかと考えられる。同様に、図表内の数値が含まれている文もその図表との関連があると判断できる。このように、単純な Table、Figure の有無だけでなく、3.3 節の画像解析や前後の文脈を考慮することで、より多くの対応文を抽出することができると考えられる。

そこで、本研究では最初に出現したキャプション名の入っている文を「基準文」とし、基準文の前後の文脈を考慮した対応文抽出を行う。人手で対応付けがされた 290 件の論文データにおいて、基準文と対応文の分布を調査したグラフを図６に示す。

図６の横軸の０が基準文となっており、そこから正の方向が基準文の後ろ、負の方向が基準文の前の対応文となっており、縦軸はその数である。図６から、基準文の前後、特に基準文の後ろの文が対応文となっている場合が多く、対応文抽出においては、文脈情報を考慮す

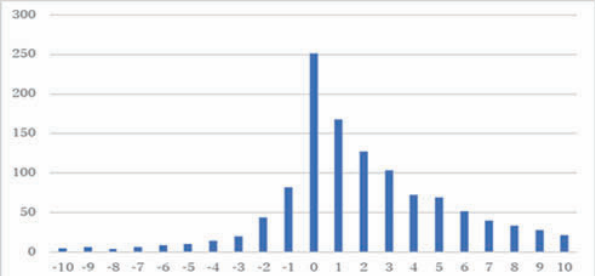

図６　基準文と対応文の分布

ることが重要だといえる。

そこで、本研究では文脈情報を考慮した対応文抽出を自動で行うために、対応文抽出を系列ラベリング問題として解き、リカレントニューラルネットワークの一種である Long short-term memory（LSTM）を用いて解決する手法を提案する。

4　実験

本章では、提案手法の有効性を確認するために行った実験と結果について述べる。4.1 節では実験設定、4.2 節では実験結果及び考察について述べる。

4.1　実験設定

実験データ

本研究では、実験データとして ACL Anthology を用いた。ACL Anthology は自然言語処理分野および計算言語学分野の論文を収集したもので、年々データ量は増加している。本研究では、そのうち自然言語処理分野の論文 31,812 件を取り扱う。

ACL Anthology の論文群は PDF 形式で出力されているが、このままでは形式の違いから図表の解析に用いることができない。そこで、Abekawa らは、ACL Anthology の PDF ファイルから論文の構造を抽出し、XHTML 形式で出力することに成功した[Abekawa 16]。これにより、図表は png 形式、テキスト部分が XML 形式で統一されているため、解析に用いることができるようになっている。図表は全部で 417,237 件ある。本研究ではこの XHTML 形式に変換されたデータセットを用いて実験を行う。

正解データには、人手で図表とテキストの対応付けを行ったデータを用いる。論文中の各図表について、その図表について述べていると判断された文をすべて選択してもらい、リストとした。結果として 290 件の図表と、それに対応している文のリストを用意した。

実験条件

本実験では、データ数が少ないことから、５分割交差検定により実験を行った。最初にキャプション名（Table 1、Figure 2 等）が出現した文にタグを付与し、このデータに LSTM を適用することにより学習を行った。具体的には、これらのタグを付与したデータを、まず 100

次元の実数ベクトルに変換し、次に畳み込み関数を適用して 300 次元のベクトルにする。このベクトルに対し、各次元の最大値を計算した結果として得られる 300 次元のベクトルを LSTM 関数に入力する。LSTM 関数の出力に線形関数を適用し、最も値の大きい次元に対応するラベルを予測ラベルとする[石野 17]。

本実験では、ベースライン手法として条件付き確率場（CRF）を用いて同様の実験を行う。窓枠は 4、素性は前後 4 語の uni、bi グラムとした。

評価尺度

CRF 及び LSTM を用いた実験では、評価尺度として精度、再現率、F 値を使用する。

4.2 実験結果

4.1 節で述べた図表と本文の自動対応付け実験の結果を表 1 に示す。表 1 より、LSTM を用いた手法は CRF と比べ、再現率が 0.1 程度低くなっているが、精度は 0.26 高くなっており、提案手法の有効性を示すことができたと考えられる。

	精度	再現率	F 値
LSTM	0.65	0.13	0.33
CRF (baseline)	0.39	0.23	0.29

表 1　図表と本文の自動対応付け実験結果

4.3 考察

本節では、4.2 節の実験結果に対する考察を行う。4.2 節の表 1 より、ベースライン手法よりも提案手法の精度が高くなっているが、再現率に関しては低い値となってしまっている。原因の一つとして、データ数が少なく、LSTM による学習が十分でないということが考えられる。これは、人手による正解データを増やすことで改善することができる。そして、もう 1 つの原因として、タグの数が少ないことが考えられる。本研究では、最初に出現したキャプション名の入っている文にのみタグを付与している。しかし、キャプション内の語句や図表内の数値の有無は対応文抽出において重要な手掛かりになると考えられる。“The” や “Our” などの図表の対応文に多く用いられている語句も多く存在すると考えられる。これらの要素についてもタグを付与することで精度、再現率を向上させることができるのではないかと考えられる。

5 おわりに

本稿では、論文中の図表に対応づいている文を自動的に抽出する手法を提案した。4.2 節の実験により、提案手法では精度 0.65 を得た。比較手法の精度 0.39 と比べて、0.26 ポイント向上させることができ、提案手法の有効性を確認することができた。

参考文献

[Abekawa 16] Abekawa, Takeshi, and Akiko Aizawa : SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation, the 26th International Conference on Computational Linguistics, (2016).

[Bach 11] Ngo Xuan Bach, Nguyen Le Minh, Tran Thi Oanh, and Akira Shimazu: Learning Logical Structures of Paragraphs in Legal Articles, Proceedings of the 5th International Joint Conference on Natural Language Processing, pp.20-28, (2011).

[福田 13] 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　データベース　Vol. 6, No. 2, pp.16-29, (2013).

[石野 17] 石野亜耶，難波英嗣，竹澤寿幸 : Twitter を利用した旅行計画者の行動分析、観光情報学会 第 16 回研究発表会，(2017).

[岩槻 98] 岩槻恵子 : 説明文理解における要点を表す図表の役割、教育心理学研究．Vol. 46, No. 2, pp.142-152, (1998).

[難波 09] 難波英嗣，藤井敦，岩山真，橋本泰一 : 論文と特許を対象にした技術動向分析、第 7 回、第 8 回　NTCIR ワークショップ　特許マイニングタスク，情報管理 Vol. 52, No. 6. pp. 334-342, (2009).

[Plank 16] Barbara Plank: Keystroke dynamics as signal for shallow syntactic parsing, the 26th International Conference on Computational Linguistics, (2016).

[Willot 15] Paul Willot, Kazuhiro Hattori, and Akiko Aizawa: Extracting Structure from Scientific Abstracts Using Neural Networks, In Proceedings of the 17th Asian Digital Library Conference (ICADL 2015), pp. 329-330, (2015).

[Wu 11] Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata: Extracting Pre-ordering Rules from Predicate-Argument Structures, Proceedings of the 5th International Joint Conference on Natural Language Processing, pp.29-37, (2011).