学会発表ポスターのコンテンツとデザインを分離したアノテーションコーパスの構築

吉田 奈央† 宮尾 祐介† 阿辺川 武‡ 相澤 彰子‡†
†東京大学大学院情報理工学系研究科 ‡国立情報学研究所
{naou.yoshida, yusuke}@is.s.u-tokyo.ac.jp, {abekawa, aizawa}@nii.ac.jp

1 はじめに

情報伝達においては、情報の受け手に対し多くの情報を短時間で効率的に把握させるため、視覚情報を充実させた手段をとることは重要である。駅で見られる路線図や、サッカーや野球などのスポーツ観戦におけるチーム選手の配置図など、日常生活においても情報量の多さを視覚的情報に訴えることで効率化した事例は枚挙にいとまがない。

学術分野においても、研究内容や成果を短時間で効率的に伝えるために、言語情報だけでなく視覚的情報を利用することは広く行われている。学会等のオーラルプレゼンテーションに用いられるスライドやポスタープレゼンテーションに用いられるポスター、論文誌で提出を求められるグラフィカルアブストラクトはその好例である。

著者らは、これらのうち特に学会発表のポスターに着目し、アノテーションコーパスを構築した。特に、ポスターのコンテンツ（主に言語情報）とデザイン（視覚的情報）を分離して構造化データとすることを目指す。学会発表ポスターは一枚の掲示物で研究内容全体を伝達することが求められるため、論文の要約であるという面（コンテンツ）と、限られたスペースで効率的に情報を伝達するためのデザインが要求されるという二面性があるメディアである。これらの側面を個別に研究すること、さらにはコンテンツとデザインの関係を分析することで、言語情報と非言語情報のインタラクションを明らかにしていくことを狙っている。本コーパスを用いることで、例えば、ポスターのコンテンツと論文を紐づけることで論文の要約としてのコンテンツの分析を行ったり、要約と配置の同時最適化といった両側面の関係性の研究を進めることができると期待される。

本研究の概略を図 1 に示す。ポスター（PDF ファイル）に対し、論文内容の要約である言語情報（図や表を含む）と、それらをポスター上に配置するデザインをそれぞれ抽出する。前者は XHTML、後者は CSS で記述し、これらを組み合わせると元のポスターを再現できるようにする。

本研究では、計算言語学と自然言語処理の国際会議論文のアーカイブである ACL Anthology[1]から 107 のポスターの PDF ファイルを収集し、そのうち 30 のポスターに対して、XHTML と CSS によるアノテーションを行った。以下では、アノテーション方法の詳細と、構築したコーパスの概要について述べる。

2 関連研究

論文テキストやポスターを入力情報とし、その内容をビジュアライズしたものを出力する研究について、そのデータセットの存在や種類に着目しながら関連研究として挙げる。

Qiang ら [3] は、読みやすくかつ見やすいデザイン配置になるポスターのレイアウトを生成し、それに基づいてポスターを自動生成する手法を提案した。ウェブから収集した 600 のポスターの中から選択した ‘良いデザイン’ の 25 ポスターを対象に、それらのレイアウトと各パネル（長方形にわけられたブロックにテキストと図表がはいったもの）の情報をベイジアンネットワークを利用して学習した。彼らはポスターの読みやすさ、美しさに主眼をおいた分析と実験をおこなっており、テキストの長さやパネル内での当該部分の面積の割合などを PDF ファイルにタグ付けしたデータを作成し、学習データとして使用している。ポスターのコンテンツは、論文テキストに対して TextRank[1] を適用して要約文を抽出することで生成している。これに続く Qiang ら [2] の研究では、[3] より対象データ数をふやし、600 のポスターの中から ‘良いデザイン’ の 85 のポスターを選択、それらにレイアウト情報をタグ付けした XML ファイルのコーパスを作成している（未公開）。彼らはデザイン性を重視したコンテンツの配置を目指しているが、コンテンツの生成には既存の要約手法をそのまま適用しており、コンテンツの構造化や分析は行っていない。

安村ら [4] は、学術論文を入力とし、プレゼンテーション用スライドを生成するシステムを提案した。彼らは論文テキスト中の出現単語の重要度を TFIDF 法で計算し、そこから編集単位（パラグラフやセクションなど）の重要度を算出し、スライド生成に利用した。

[1] https://aclanthology.info/

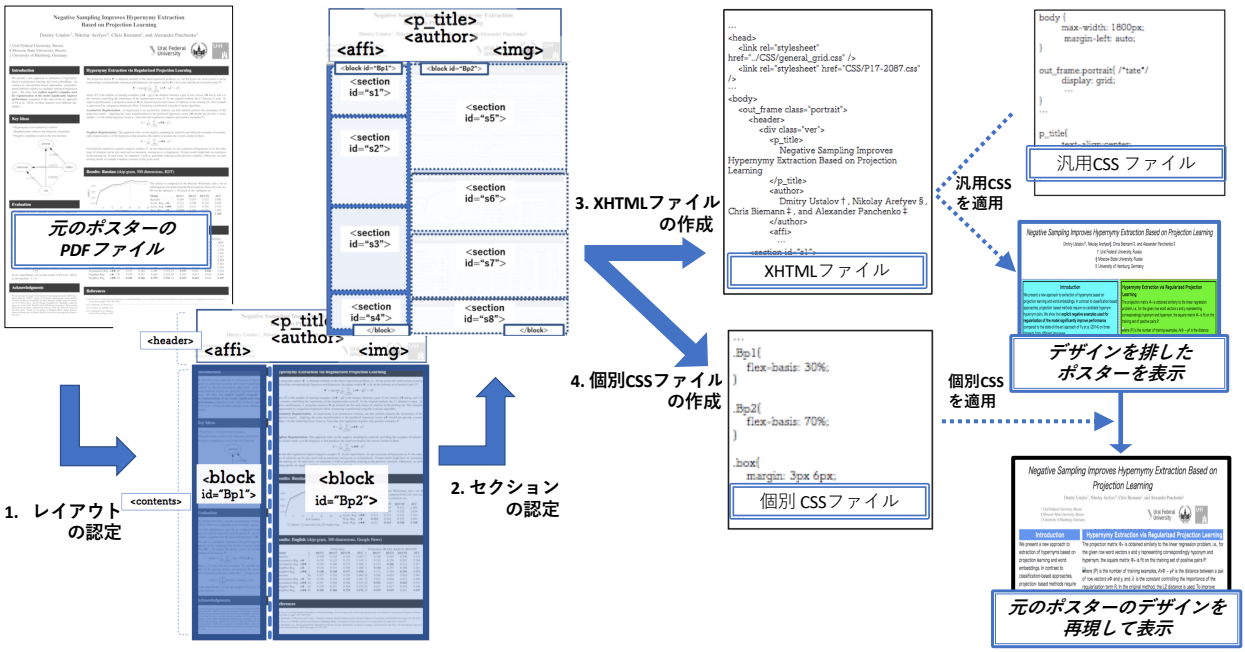

図 1: ポスター PDF ファイルから XHTML ファイルと CSS ファイルを作成する手順

各編集単位のスライドへの割当枚数をユーザーに入力させ、重要度や編集単位間の依存関係をもとに、論文内容の要約の集合を作成し、それらを XHTML で記述されたスライドレイアウトに当てはめたもの出力する。これはスライドの自動生成や作成支援システムに重きが置かれた研究であり、データセットの整備や公開は行われていない。

柴田ら [5] は、談話構造解析をスライド自動生成に応用する手法を提案した。構造化されたテキストはコスト高であるとして、加工を加えていない生テキストを入力としている。テキスト情報のビジュアライズという点については、箇条書きやインデントのみを対象としている。この研究も、データセットの整備や公開は行っていない。

このように、本稿で述べるようなコンテンツとデザインの両方を構造化したデータセットの構築はこれまでに例がないといえる。

3 アノテーション方法

3.1 基本方針

ポスターは論文内容を端的に要約してまとめた言語情報と、掲示物として見やすさを考慮し体裁を整えるという視覚的情報の二つが複合的に体現されたものと言える。それらを分離してアノテーションコーパスとする目的で、言語情報の構造化に XHTML、ポスターのレイアウトやビジュアルデザインには CSS を利用してこれらを記述する。

本研究では、XHTML5 [2]で定義されるタグに加え、ポスター内のメタ情報や論理構造を明示できるように新たにタグを定義する。これは、ポスターの言語情報は論文の要約であるとの仮定に基づき、論文の論理構造と対応づけられるように設計している。アノテーションに使用したタグセットを表 1 に、これらのタグを用いた XHTML の例を図 2 に示す。タグの使用方法については、3.2 節で詳述する。

ポスターのデザインについては、2 段階で CSS を定義・適用する。まず、全ポスター共通の CSS ファイルを定義する（以下、汎用 CSS）。汎用 CSS は、すべてのポスターに適用可能な必要最低限のレイアウト（ブロックの並び方）のみを与える。これを適用すると、元のポスターから個々のビジュアル的特色やデザイン性を取り除いた状態で XHTML ファイルを表示することができる。次に、ポスターごとに個別 CSS を作成する。個別 CSS ファイルを汎用 CSS に追加して適用することで、現物のポスターの外観をブラウザ上で再現できるようにする。

ポスターのレイアウトは、いくつかの例外はあるものの、[3] でも言及されている通り、本データでも矩形で区切ることができる事例が 9 割以上だった（本研究で収集したポスター 107 のうち 102 が該当）。そこで、レイアウト部分は柔軟に矩形の変形レイアウトを組むことができる CSS Grid のフレームワークを使用し、その他のデザイン（フォントなど）については単純な CSS のみを使用する。

[2]https://www.w3.org/TR/html/

表 1: 本研究で定義したタグセット

ポスターのレイアウトのタグ	
poster	ポスター全体を表すタグ。
header	ポスター中でタイトルや著者などを表示する場所を表すタグ。
p_title	ポスターのタイトル。
author	ポスターの著者名。各人の名前をこのタグで括る。
affi	著者の所属。各所属をこのタグで括る。表記のない場合は記述しない。
mailadd	著者の連絡先メールアドレス。表記のない場合は記述しない。
contents	ポスター本体（論文の内容を提示する部分）を表すタグ。
p_block	ブロック部分の範囲を指定する。

セクションのタグ	
section	セクションの範囲を表すタグ。
section_title	セクションにタイトルがある場合は、その該当する部分に付与する。タイトルが存在しない場合は記述しない。

セクション内の構成要素のタグ	
text	箇条書きでなく、平文が書かれている部分を表す。
img	図、表、数式等、テキストで表現できない部分を表す。
ul / ol	箇条書きからなるテキスト部分を表す。

```
<poster class="portrait">
  <header>
    <p_title>学会発表ポスターのコンテンツとデザインを分離したアノテーションコーパスの構築</p_title>
    <author>吉田奈央</author>
    ...
  </header>
  <contents class="tt2">
    <p_block id="Bp1">
      <section id="s1">
       <section_title>研究目的</section_title>
        ...
      </section>
      ...
    </p_block>
    ...
  </contents>
</poster>
```

図 2: ポスター XHTML の例

3.2 アノテーション手順

以下では、図 1 を参照する形で手順を説明する。

1. レイアウトの認定 まず最初に、ポスター全体の掲示方向と、タイトル部分、ポスター本体、ブロックのレイアウトを認定する。ポスターの掲示方向は、タイトル位置と可読方向から、縦長方向 (portrait) か横長方向 (landscape) のどちらの掲示を想定しているかを判断する。これは poster タグの class 属性に記述する。タイトルや著者情報など論文のメタ情報が記述されている部分は header タグで記述し、その下に各メタ情報を表 1 左のタグを用いて記述する。

ポスター本体部分（contents タグで記述）においては、多くのポスターは複数の矩形領域に区切ってコンテンツが配置されているため、この矩形領域をブロックと定義し、p_block タグを用いて記述する。ブロックは、必ずしも論文・ポスターの論理構造とは一致せず、主に視覚的情報を担う単位である。この際、論文内容から推測されるコンテンツの読み順は考慮せず、あくまでポスター上で視認される区切りに沿ってブロックの区切りおよび並び順を認定する。

アノテーション作業の効率化のため、この段階でポスターレイアウトのパターンを分類し、各パターンについて XHTML テンプレートを利用する。具体的には、ポスターの掲示方向と、ブロックの数と配置の向きの組み合わせを一つのレイアウトパターンとする。例えば図 1 では、ポスターは縦向きで、ブロックが横に 2 つ配置されているレイアウトパターンと認定する。レイアウトパターン名は contents タグの class 属性に指定し、各パターンについて p_block の並び方を規定するスタイルを汎用 CSS に記述する。すなわち、同じレイアウトパターンのポスターについては、汎用 CSS ではまったく同じスタイルが適用される。また、このパターンごとに XHTML ファイルのテンプレートを用意し、以降の作業はこのテンプレートに具体的な内容を追加することで行う。

ただし、コンテンツを矩形のブロックで区切ることが難しいポスターも散見された。この場合、最も近しいと思われるパターンに分類しそのテンプレートを利用した。

2. セクションの認定 次に、各ブロック内部のコンテンツの論理構造を示すタグ付けを行う。一般的に、各ブロックは複数の論理的かたまりから構成されることが多く、これをセクションと認定する。将来的には、各セクションを論文のセクションやパラグラフと対応付けることで、論文とポスターとの関係を明示化することを狙っている。セクションの認定には、視覚的情報（区切り線や色分けなど）のみならず、コンテンツの意味的内容も用いるが、作業の簡単化のため論文は参照しない。

セクションにはタイトルが付けられることが多く、その場合は section_title タグを用いて記述する。セクションの中にさらに小さい論理的かたまりが認定される場合は、section タグを入れ子にして記述する。

3. XHTML ファイルの作成 各セクションについて、その中のテキストや図表を XHTML 化する。テキスト部分については、箇条書きか否かによって、text, ul, ol タグを使い分ける。図、表、数式など、自然言語テキストでは記述できない要素は、すべて画像ファイルとしてポスター PDF ファイルから切り出し、img タグを用いて XHTML ファイルに挿入する。

表 2: ポスターのパターン分類とその該当数

	2 ブロック	3 ブロック	4 ブロック
縦×平行	37	2	—
縦×垂直	3(2)	7 (2)	1
横×平行	4	—	—
横×垂直	11(1)	34	4

表 3: 構築したコーパスにおけるタグの統計

	p_block	section	sec_title	text	img	ul/ol
最小	2	4	4	0	0	0
最大	4	17	17	20	31	18
平均	2.4	8.9	8.6	6.6	11.9	5.6

4. 個別 CSS ファイルの作成　現物のポスターの外観を再現するために、各 XHTML タグに対してスタイルを定義する個別 CSS ファイルを作成する。個別 CSS では、フォントの文字色・エフェクトや、ブロックやセクションのサイズや背景色など、ポスターの詳細なデザインを規定する。

3.3 例外処理

上記のアノテーション手順で、ブロック構造と論理構造が XHTML の記述ルールでは成立しない、つまり、ブロックの中で section が閉じない場合が散見された。この事例はタグ付けを行った 30 ファイル中 5 例確認された。ブロック構造をまたいで文章が続く場合が 4 例、セクションに対応する図表だけが次のブロックに現れる場合、同一セクション内の要素が次のブロックに現れる場合が各 1 例あった。

この場合、ブロックの終点で一度 section タグを閉じ、次のブロックの開始の直後に section を再開させるようにアノテーションを行う。そして、後者の section に特別な属性 goeswith を設けて前者の section の id を指し示すようにし、両方の section が一つながりであることを記述する。

4 コーパスの概要

本研究では、ACL Anthology に収録されたポスターのうち、2013 − 2018 年までの ACL, EACL, NAACL, EMNLP で発表された 107 ポスターを収集した（全て PDF 形式）。そのうち 30 ポスターを無作為に抽出し、3 節で説明したアノテーション作業を行った。

今回収集した 107 のポスターは、9 種類のレイアウトパターンに分類された。表 2 に、各パターンに分類されたポスターの数を示す。縦列は「掲示方向 (縦：portrait か横:landscape）×ブロックの掲示方向に対しての平行 or 垂直の別」、横列は各ポスターの「ブロックの数」として、それらの組み合わせとそれに該当するポスターの数を示したものである。() 内の数値は、例外的なレイアウトを当該パターンとして扱った数を示す。

表 3 に、本論文で定義したタグの統計を示す。アノテーション作業を行った 30 ポスターのうち、各タグの数の最小値、最大値、平均値を示した。ブロック数は 2–4 で抑えられるが、section や text, img 等のタグの数はポスターによって大きく異なり、ポスターにおける情報提示方法にさまざまなバリエーションがあることが見てとれる。

今後、構築したコーパスを利用して、ポスターのコンテンツとデザインの構造についてより詳細な分析を行う予定である。

5 今後の課題

本稿では、学会発表ポスターを対象として、テキスト等の言語情報と、それらのレイアウトやデザイン等の視覚的情報をそれぞれ XHTML と CSS に分離して構造化したコーパスの構築について述べた。現在までに、ACL Anthology から収集した 107 ポスターのうち、30 ポスターについてアノテーション作業を完了した。本コーパスは、オープンデータセットとして公開することを予定している。

今後は、論文内容とポスターの内容の対応や、図表や数式の内部の構造化を行い、より詳細な分析や応用に活用できるようコーパスの整備を進めていく計画である。また、本コーパスを利用して、ポスターの自動生成やデザインの最適化等の研究を行う予定である。

謝辞　本研究は、JST、CREST、JP-MJCR1513 の支援を受けたものです。

参考文献

[1] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *EMNLP 2004*, 2004.

[2] Yu-ting Qiang, Yan-wei Fu, Xiao Yu, Yan-wen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers by probabilistic graphical models. *Journal of Computer Science and Technology*, Vol. 34, No. 1, pp. 155–169, 2019.

[3] Yuting Qiang, Yanwei Fu, Yanwen Guo, Zhi-Hua Zhou, and Leonid Sigal. Learning to generate posters of scientific papers. In *AAAI 2016*, 2016.

[4] 安村禎明, 武市雅司, 新田克己. 論文からのプレゼンテーション資料の作成支援. 人工知能学会論文誌, Vol. 18, pp. 212–220, 2003.

[5] 柴田知秀, 黒橋禎夫. 談話構造解析に基づくスライドの自動生成. 自然言語処理, Vol. 13, No. 3, pp. 91–111, 2006.