アノテーションの副次生成とテキスト処理への応用

Generating Annotation Data Secondarily and its Application to Text Processing

伊藤 一成♥ 斎藤 博昭♦

Kazunari ITO Hiroaki SAITO

我々は，汎用アノテーション記述言語 MAML（Multimedia Annotaion Markup Language）を提案している．MAML は，“如何なるデータ，事象，物体，人物や行為に対して同一の仕様でアノテーションできる”ことを目標に掲げている．アノテーションデータは通常，付与自体を目的とする明示的作業によって生成されるが，何らかの別の作業から副次的に生成されるテキストもアノテーションデータとみなすことが可能である．

本論文では MAML の適用領域の一つとして，アノテーションの概念を用いた新しいテキスト処理技法を提案する．またこれを用いたメール及び Web に関するテキストコンテンツの処理について例示する．

We have proposed MAML (Multimedia Annotation Markup Language), a generalized annotation description format for media data. MAML sets a goal that anyone can annotate any data, events, objects, person and acts by a unite specification. In almost all cases, an annotator does the job explicitly. But text which is generated secondarily by other work can also be treated as annotation data.

This paper proposes a new text processing technique based on the concept of annotation, and illustrates a few applications.

1. はじめに

通常アノテーションデータはユーザの明示的作業によって生成される．我々は明示生成されたアノテーションデータを用いた応用事例として，動画像の検索システム及び要約システムを実装してきた[1][2]．しかしながらアノテーションデータ生成に係る時間コストがその生成及び流通の大きな障害となっている．アノテーション関連技術の研究自体は盛んであるが，実際に稼動されているシステムや製品がなかなか出現しない理由もそこにある．ところで，アノテーションの概念を用いた大規模システムの成功事例に Google のイメージ検索[3]が挙げられる．画像がエンベッドされている HTML テキストをアノテーションデータと考えることにより，キーワードによるイメージ検索を実現している．ユーザが自らの意志で作成したホームページのテキスト自体がアノテーションデータになっている．

我々が提唱している汎用アノテーション記述言語 MAML は，このように本来の作業から副次的に生成されるアノテーションデータも対象領域としている．本論文では，ユーザの非意識的な操作及び文章入力によって生成あるいは抽出されるテキストコンテンツに対して，アノテーションの概念に基づく新しい処理技法を提案する．

2. MAML の記述例

本論文では MAML の記述として受信メールの文書を例に挙げる．MAML の詳細については文献[4]及び[5]を参照されたい．“受信メールに返信する”ということは，受信メールのメールタイトル又は本文（の一部）に対してアノテートしているとみなすことが出来る（図 1 参照）．図 1 において，矢印の起点に接続された四角領域の内部に記述されているテキストがアノテーションデータ，終点がその対象である．

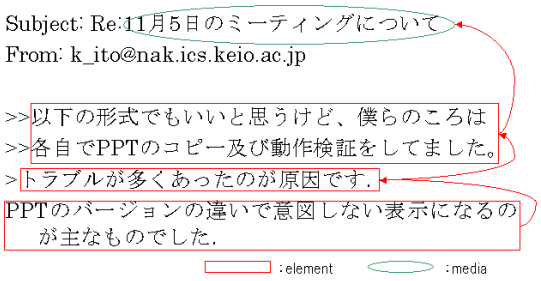

図 1 返信メールの例
Fig.1 An Example of a Reply Mail

図 1 の返信メールを MAML に変換した例を図 2 に示す．

```
<?xml version="1.0" encoding="UTF-8"?>
<maml>
 <media type="text"
  media-word="11 月 5 日のミーティングについて">
  <element id="1">
   <contents>
    <copy>以下の形式でも…</copy>
   </contents>
  </element>
  <element id="2" target="1">
   <contents>
    <copy>トラブルが多く…</copy>
   </contents>
  </element>
  <element id="3"
   annotator="k_ito@nak.ics.keio.ac.jp"
   target="2">
   <contents>
    <comment>PPT のバージョン…</comment>
   </contents>
  </element>
 </media>
</maml>
```

図 2 MAML の記述例（本文は一部省略）
Fig.2 MAML Description of a Reply Mail

本論文において，以後アノテーションの基本単位（図 1 の四角に相当）をエレメント（element），その対象でエレメントでは無いものをメディア（図 1 の楕円に相当）と呼ぶこととする．通常メディアはファイルを指す場合が多いが，人物，事象等 URI で参照できない場合は図 2 の media タグの media-word 属性値に示される様に，メディアを自然言語で表現する．また element タグの id 及び target 属性値によりエ

♥ 学生会員 慶應義塾大学大学院理工学研究科後期博士課程 k_ito@nak.ics.keio.ac.jp

♦ 正会員 慶應義塾大学理工学部情報工学科 hxs@nak.ics.keio.ac.jp

レメント間の参照関係が表現されている．

3. アノテーションデータの処理

本章では，MAML ファイルの処理プロセスについて解説する．

3.1 概要

MAML ファイルに含まれるエレメントは，メディアに対する直接のアノテーションと，他のエレメントに対するアノテーションとに大別される．あるエレメントを子とし，そのアノテーションの対象を親とする親子関係を定義できる．するとエレメント全体はメディアをルートとするツリー構造とみなすことができる(図 3 の矢印)．

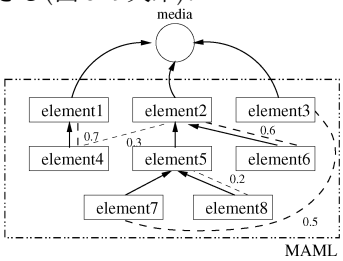

図 3　アノテーションデータの構造化
Fig.3　Structuralization of Annotated Data

はじめに MAML ファイルに対する結合演算子＋，集合演算子∪及び差分演算子－を定義する．ここで 2 つの MAML テキストをそれぞれ識別子を用いて A 及び B と表すとする．A＋B は 2 つの MAML ファイル中のエレメントを融合する．A と B に共通するエレメントが存在する場合，1 つに集約される．共通するエレメントとはテキストが全く同一のものを指す．集合演算∪とは，結合演算＋に類似するものである．∪の場合，2 つのファイルに共通のエレメントが存在しても別のエレメントと扱い融合される．一方 A－B は差分をとる．B において親を持つか，又は親も子もいずれも持たないエレメントが A に存在すれば，そのエレメントを削除し，すべての子エレメントも再帰的に削除する．例を図 4 に示す．

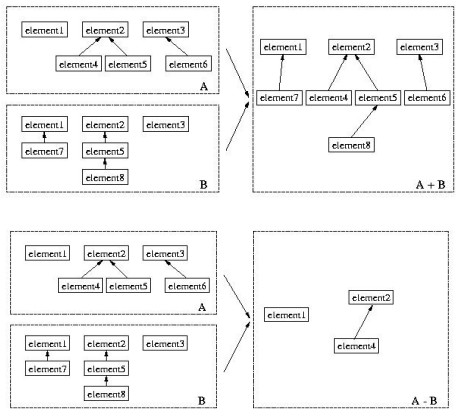

図 4　アノテーションデータの結合及び差分演算
Fig.4 Union and Difference Operation of Annotation Data

また，各エレメントのアノテーションテキスト間で自然言語処理に基づいて関連性を定義する(図 3 の点線)．

識別子を X と定義した MAML テキストに対して，3.2 節で説明するスコア算定処理を施し，アノテーションの対象関係及び言語関連性に基づいて図 3 に示すような半有向グラフとして表現する．それに必要な前処理（chasen[6]による名詞句抽出及び TF（Term Frequency）値の算出等）を行ったデータ形式を以後グラフ表現 G(X)と表現することとする．

以上のデータ構造化により，検索，条件無（大域）要約，条件付要約，関連情報抽出等の処理プロセスを提供する．

はじめに検索処理及び条件付要約処理について概説する．検索処理，条件付要約処理は 2 つのグラフを基に行う．1 つ目がクエリに相当するグラフ（G(Q)とする），2 つ目が処理対象のグラフである（G(A)とする）．以後これら 2 つのグラフによる検索処理を search(G(Q),G(A))，条件付き要約処理を summarize (G(Q),G(A))と記述する．エレメントが 1 つも存在しない MAML ファイルを N と定義すると summarize (G(N),G(A))が条件無（大域）要約に相当する．また A のエレメント群の部分集合を A'とすると search(G(A'),G(A-A'))が A'に対する関連情報抽出処理となる．

検索処理 search(G(Q),G(A))及び要約処理 summarize (G(Q),G(A))を行う場合には，Q∪A に対して次節で述べるスコア算出処理を施す．

3.2 スコア算出処理

個々のエレメントに対するスコア（重要度）の算出方法について説明する．Q∪A に関して，エレメント e_i と e_j の構造に基づく類似度 $sim_{str}(e_i,e_j)$ を以下の式により与える．

$$sim_{str}(e_i,e_j)=\begin{cases}1 & e_i \text{と} e_j \text{が親子関係の場合}\\ 0 & \text{それ以外}\end{cases}$$

次に，すべてのエレメントのテキストから個々の名詞句の単語重要度をそれぞれ TF 値に基づいて算出する．2 つのエレメント e_i,e_j について，e_i に含まれる名詞句を $k'_1,k'_2,\ldots,k'_n$，e_j に含まれる名詞句を $k''_1,k''_2,\ldots,k''_m$，e_i と e_j に共通の名詞句を $k'''_1,k'''_2,\ldots,k'''_l$ とする．e_i と e_j の言語的類似度 $sim_{lex}(e_i,e_j)$ を Jaccard 関数によって与える．

$$sim_{lex}(e_i,e_j)=\frac{\sum_{i=1}^{l} tf(k'''_i)}{\sum_{i=1}^{n} tf(k'_i)+\sum_{i=1}^{m} tf(k''_i)-\sum_{i=1}^{l} tf(k'''_i)} \quad \cdots(1)$$

ただし式(1)の右辺の分母が 0 となる場合，$sim_{lex}(e_i,e_j)=0$ とする．

次に 2 つのエレメント，e_i の e_j 類似度 $sim(e_i,e_j)$ を次式で定義する．

$$sim(e_i,e_j)=\alpha sim_{str}(e_i,e_j)+(1-\alpha)sim_{lex}(e_i,e_j) \quad \cdots(2)$$

ここで，α（0≦α≦1）は構造か言語かどちらの要素を重視するかを決定付けるパラメータである．

自然言語処理分野の研究において，文間の単語による結束性の高い文が重要であるとみなすことにより重要文を抽出する手法がある[7]．本件においても同様に，他のエレメントとの構造的及び言語的結束性の高いものは重要エレメントであると考える．そこで，A に含まれる個々のエレメントに対してそのスコア $w(e_i)$ を以下の式によって求める．

$$w(e_i)=\sum_{\substack{j=1\\ i\neq j}}^{N}\beta_j sim(e_i,e_j)$$

ここで，N は Q∪A のエレメントの総数であり，β_j は重みパラメータである．

$$\beta_j = \begin{cases} 1 & e_j\text{がAに属する場合} \\ \gamma\ (>>1) & e_j\text{がQに属する場合} \end{cases} \quad \cdots(3)$$

検索処理の場合は，A の中で決められた閾値以上のスコアを持つエレメントそれぞれを結果とする．式(3)のγの値を大きくすることによりQと関連性の高いエレメントのスコアがより大きくなる．要約処理は，エレメントの集合を結果とする．エレメントのスコアの高いものから順にまず結果集合に含めていく．その際，親にあたるエレメントも再帰的に結果集合に含める．この処理を繰り返し，結果集合に含まれるエレメントが要約率によって決められる一定の数に達したらそれを最終結果とする．また式(2)のαを大きくし，構造上のつながりを重視することで主題部分及びそれに対するアノテーション群だけが結果集合となるようにする．処理の概念図を図 5 に示す．

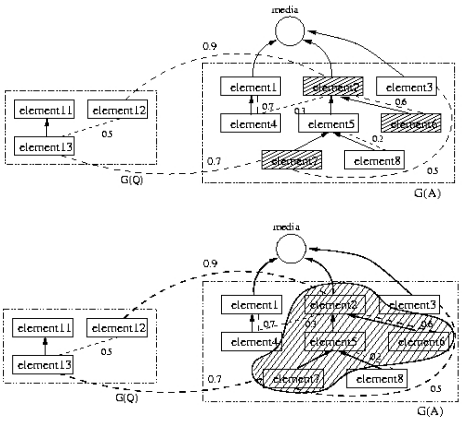

図 5　検索処理・要約処理の概念図（上:検索　下:要約）
Fig.5 Conceptual Diagram of Search and Summarization (Upper:Search Lower: Summarization)

最後にグラフ同士の差分演算子－を定義する．G(A)－G(B)とは，G(A)のスコア算出にあたって，A に含まれる名詞句 *k* が B にも含まれるならば 3.2 節の処理過程の式(1)において *tf*(*k*)を 0 として言語的類似度を決定することを意味する．

4. 応用事例

前章で解説した概念を基にした応用事例について述べる．

4.1 複数メールの要約

我々は Annotation Mailer という名称のメーラを実装している．メーラでは，すべての新規･受信･返信･転送メールに対して図 2 で例示した MAML が内部で生成･保存される．またメーラのアドレス帳からも図 6 に示すような MAML が生成されている．

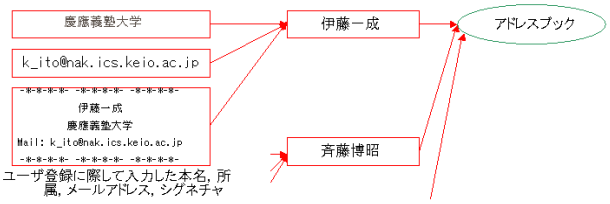

図 6　アドレス帳からのアノテーションデータ生成例
Fig.6　Generating Annotation Data from Address Book

Annotation Mailer は複数のメールを一つに融合し，かつ任意の要約率に圧縮し提示する機能を有する．処理の流れ図を図 7 に示す．

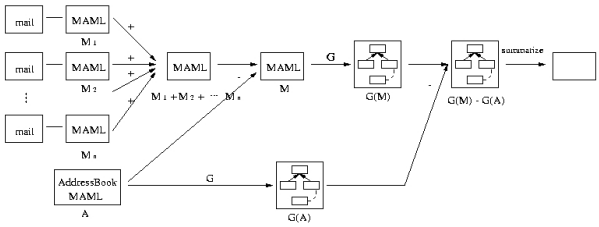

図 7　複数メール要約の処理系統図
Fig.7　Flowchart of Mail Summarization

はじめに，ユーザが指定したメールのサブジェクトと同一のメール群それぞれを対象にした MAML ファイル群（M_1,M_2,…,M_nとする）を結合し，さらにアドレス帳の MAML ファイル（A とする）との差分をとる．生成された単一 MAML ファイル M（$= M_1+M_2+\cdots+M_n-A$）及び A に対し，グラフ構造化を施す．次に要約処理 summarize(G(N),G(M)－G(A))を行う．ここで N はエレメントが一つも存在しない MAML ファイルを意味する．要約結果のエレメント集合を基にメール本文を再合成する． ファイル群の融合の際に A との差分を取るのは，通常メールでは最後にシグネチャが付与されるが，これを可能な限り取り除く処理をしている．また冒頭に“ こんにちは，伊藤です．”等の挨拶文や，返信メールの生成において“ 伊藤一成さんが書きました: ”といったメーラによって自動付与される文が記述されることがあるが，要約対象を G(M)ではなく，G(M)－G(A)とすることより，このような文のスコアを減少させ，要約対象文にならないようにしている．さらにシグネチャが先の処理で取り除くことが出来なかったとしても，それに相当するエレメントのスコアも大きく減じる効果も持たせている．

自然言語処理分野における重要文抽出手法のみで，電子メールのような対話形式のテキストデータの要約を行うと，その対話構造が保持されない．しかしながらこの例では，どの文がどの文に対する返答文か，つまりアノテーションかという構造が記述されているので，どのような要約率を設定しても，対話構造を保持した要約文書を生成できる．

4.2 ユーザ適応型 Web 検索

ユーザはブラウザを使う際に様々な操作や，入力を行う．以下に列挙する．

1) アドレスバーに URL を入力してページにアクセスする．
2) リンクを参照してページにアクセスする．
3) お気に入りリストに登録する．お気に入りリストからページを参照する．
4) 検索バーに検索単語を入力し検索する．

これらの操作･入力もすべてアノテーション行為とみなし，MAML データを生成する Annotation Browser という名称のタブ型ブラウザを実装している．それぞれの操作･入力に対するエレメントの生成例を図 8 に示す．

Annotation Browser は，ユーザ適応型 Web ページ検索機能を有する．検索処理の流れ図を図 9 に示す．図 1，2 で例示した Annotation Mailer における新規作成及び返信したメールの MAML と Annotation Browser のブラウジング履歴の MAML（図 8 の 1,2,3,4-a）をユーザのプロファイルとみなし結合

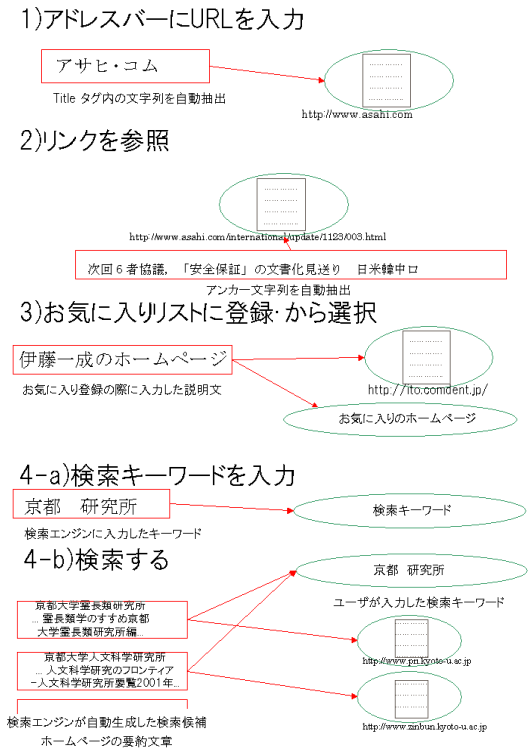

図８　ユーザの操作からのアノテーションの生成例
Fig.8 Annotation Data Generated from User's Operation

する(P とする)．また Annotation Mailer のアドレス帳から生成された MAML（図 6）を A，Annotation Browser のお気に入りリストから生成される MAML（図 8 の 3）を F とする．これら P, F 及び A を融合し，グラフ構造化する(G(P+F+A))．

一方ユーザから入力された検索キーワードから Google API を利用して，検索一覧を入手する．Google API から入手した個々の Web ページのタイトルと要約表現をエレメントとした MAML ファイル（図 8 の 4-b に相当）を生成し同様にグラフ構造化する（G(R)とする），さらに検索キーワードをエレメントとする MAML ファイル（図 8 の 4-a に相当）をグラフ構造化する(G(Q)とする)．グラフ G(P+F+A)をクエリとしてグラフ G(R)－G(Q)を対象とする検索処理 search (G(P+F+A),G(R)－G(Q))を行うことで，R 中の個々のエレメントに対してスコアを算出する．これにより Google の検索一覧から自分に関連性の高い Web ページの情報を選択抽出し，スコア順に Google の検索結果表示画面と同様なフォーマットで提示する．ここで検索対象を G(R)ではなく G(R)－G(Q)としているのは，いうまでもなく R の個々のエレメントはすべて検索キーワードを含んでおり，その単語の重要度は非常に高く，よってその共起によりすべてのエレメントのスコアが一律に大きく上がってしまうのを防ぐ為である．また summarize(G(N),G(P))を定期的に行い不要なエレメントを削除することで，プロファイルの肥大化を防いでいる．

このように，応用システムを構築する上でデータ形式に MAML を採用することで，対象メディアや生成形態が異なる様々な MAML データを容易に流用および併用することが可能である．さらに一元的に処理可能である点が最大の利点であり特徴であると考える．

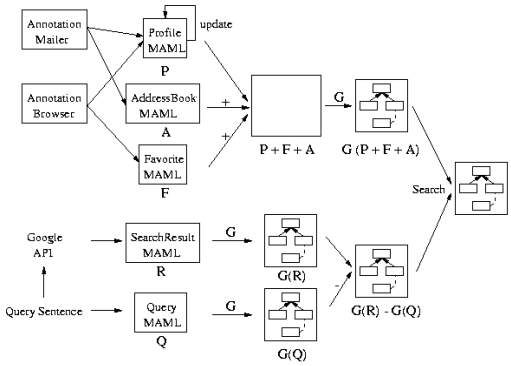

図９ ユーザ適応型 Web 検索の処理系統図
Fig.9　Flowchart of User Oriented Web Search

5. まとめと今後の予定

本論文では，アノテーションの概念を取り入れた新しいテキスト処理技法について概説した．今後は，高度メディア利用環境の実現を目指して，MAML の特徴を生かした様々な応用アプリケーション・システムの構築を行っていく予定である．

[謝辞]

本成果は，平成 15 年度未踏ソフトウェア創造事業の一部である．IPA（情報処理推進機構）及びプロジェクトマネージャーの京都大学　田中克己教授に深く感謝いたします．

[文献]

[1] Google イメージ検索ホームページ
http://images.google.co.jp/
[2] 伊藤一成，斎藤博昭: メタデータ解析に基づくメディア検索システム，情報処理学会研究報告，DBS131－69，pp. 515－520，(2003)．
[3] 伊藤一成，酒井康旭，斉藤博昭: 音声と映像の一貫性を考慮した要約動画の生成，DEWS，(2004)．
[4] 伊藤一成，斎藤博昭: 汎用アノテーション記述言語 MAML，情報処理学会研究報告 FI74－9，pp.63－69，(2004)．
[5] 汎用アノテーション記述言語　MAML ホームページ
http://ito.comdent.jp/maml.html
[6] 日本語構文解析ソフト 茶筌 ホームページ
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html
[7] 吉見毅彦，奥西稔幸，山路孝浩，福持陽士: 表題へのつながりに基づく文の重要度評価，自然言語処理，Vol.6, No.1，pp.43－57，(1999)．

伊藤 一成　Kazunari ITO
現在，慶應義塾大学大学院理工学研究科後期博士課程在学中．自然言語処理及びマルチメディア情報処理に関する研究に従事．日本データベース学会，情報処理学会，電子情報通信学会，電気学会，人工知能学会，各学生会員．

斎藤 博昭　Hiroaki SAITO
慶應義塾大学工学部数理工学科卒業．現在同大理工学部情報工学科専任講師．工学博士．自然言語処理，音声言語理解に興味を持つ．日本データベース学会，情報処理学会，言語処理学会，日本音響学会，電子情報通信学会，ACL 各会員．