アノテーションに基づく大域メディア利用環境の実現

―アノテーションの概念によりデータを構造化して処理―

１．背景

近年，メタデータ技術が注目を集めている．これは，コンテンツの意味･内容に関する特徴を別に記述しておき，その記述データを直接の計算機処理対象とすることで処理を代替しようというものである．ここで，記述データはメタデータと呼ばれる．メタデータに類似する表現にアノテーションデータがあげられる．アノテーションデータとは，メタデータと同様にコンテンツに関する内容を別に記述したデータのことをいう．人間が生成，理解及び活用することを主目的とするものであり，計算機による意味理解を視野に入れたデータ構造にする必要はない．そういう意味で，計算機の意味理解を一義として付与するメタデータとは性質を異にする．アノテーション技術に関する研究は，主に動画等の非テキストコンテンツを対象にしているものが主であるが，何も非テキストデータに限定するものではなく，本件では，テキストデータ，さらには実世界上での物体･人物･事柄などにまで対象を広げる．以下では，これらすべての対象を総称してメディアと呼ぶ．アノテーションに基づく応用システムの実現を考える際に，定められた記述仕様に即したアノテーション生成のプロセスをどう克服するかという困難な問題に直面するため，近年まで基礎研究レベルではいくつか成果が出始めているが，実用に結びつく研究や製品は皆無に近い状態であった．つまり，1)ユーザが容易に理解し，記述できる統一的なアノテーション記述仕様が無い．2)ユーザのアノテーション作業は，相当の時間コストを必要とする．　3）たとえアノテーション作業を行ったとしても，即時にその恩恵を受けることが出来ない．ということが問題である．

２．目的

通常アノテーションデータの生成にはアノテータの明示的な作業を前提としており，そのコストが技術普及の障害となっている．そこで，今回はアノテーションの副次生成に焦点を当てる．具体的には，アプリケーションの非意識的な操作や文章入力から自然発生的に画一的かつ有用なアノテーションデータを生成するプロセスの提案及びその有効性を示す各種アプリケーションを実装した．

３．開発の内容

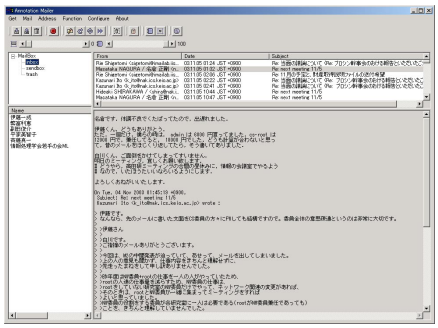
アノテーションデータの記述方式にはすでに我々が提案している MAML（Multimedia Annotation Markup Language）[1]を採用する．MAML データを生成･処理するための汎用ライブラリの構築と,非意識的な操作や文章入力から自然発生的に MAML フォーマットの有用なアノテーションデータを生成するアプリケーション群を実装する．今回はメーラ，ブラウザ，BBS クライアントの 3 点に焦点をあてた．

3.1 汎用生成･処理モジュール

MAMLファイルに含まれる情報は，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（アノテーションの基本単位）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．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．以上よりMAMLデータはアノテーションをノードとする，半有向グラフとみなすことができる．MAMLでは，タグ構造による機械の意味理解は想定しない記述方式となっている．よって表層的な自然言語処理技術を駆使したデータ構造化により，検索，大域要約，条件付要約，関連情報抽出，エレメント群の自動クラスタリング等の処理プロセスを提供する．

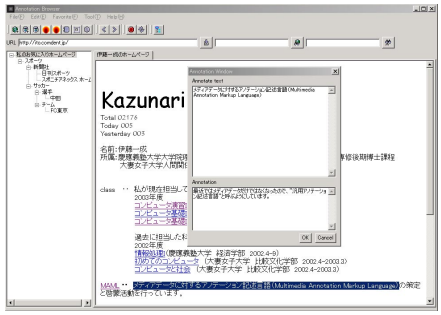
3.2 各種アプリケーション

3.2.1 Annotation Mailer



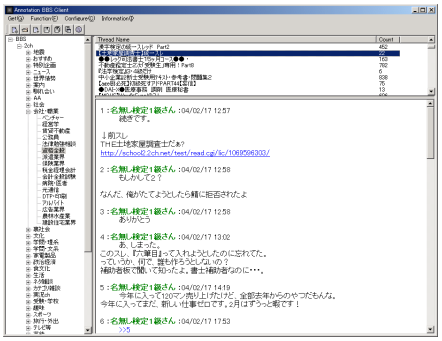
メーラを使う際の操作，入力について，これら一連の操作をすべてアノテーション行為とみなし,MAML データを生成する機能を有するメーラ Annotation Mailer を実装した．
JavaMail API， SWT をベースに実装した．

3.2.2 Annotation Browser



ブラウザを使う際の操作，入力について，これら一連の操作をすべてアノテーション行為とみなし,MAML データを生成する機能を有するブラウザ Annotation Browser を実装した．
HTML 表示に関しては及び IE のコンポーネントを利用し，その他について Java，SWT をベースに実装した．

3.2.3 Annotation BBSClient



BBSClient を使う際の操作，入力について，これら一連の操作をすべてアノテーション行為とみなし,MAML データを生成する機能を有する BBS クライアントツール Annotation BBSClient 実装した．Java，SWT をベースに開発を行った．

3.3 応用

アノテーションの概念を用いた複数メールの要約提示機能，ブラウジングにおけるユーザ適応型検索機能，BBSクライアントにおけるユーザ適応型情報提示機能を実装した．

3.3.1　複数メールの要約提示機能

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できる機能を有する． Annotation Mailerでは，すべての受信メール及び送信メールに対してMAML形式でデータが保存されている．初めに，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ファイル群を結合する．メール本文を対象とするMAMLの親子関係は，親が本文の一部であり，子がそれに対しての返答文に相当する．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，MAMLでは文書データに対するアノテーションを考えた場合，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．

3.3.2　ユーザ適応型Web検索機能

Annotation　Browserは，ユーザ適応型Webページ検索機能を有する．
Annotation　Mailerにおいて新規作成及び返信したメールのMAML群，
Annotation　Browserのお気に入りリスト，Annotation Browserのブラウジング履歴のMAMLをユーザのプロファイルとみなしグラフ構造化する（以後グラフAと呼ぶ）．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それぞれエレメントとした MAMLファイルを生成し同様にグラフ構造化する（以後グラフBと呼ぶ）．グラフAをクエリーとしてグラフBに対して検索処理を行うことで，グラフB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，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．

3.3.3　ユーザ適応型BBS要約機能

Annotation BBSClientは,ユーザ適応型要約機能を有する.書き込み一覧をMAML化し，グラグ構造化したものに対して，前項のMAMLによるユーザのプロファイル（グラフA）をクエリーとして要約処理を行うことで，個々のエレメントに対してスコアを算出する．これにより掲示板の書き込み一覧を要約する．

4．従来の技術(または機能)との相違

本プロジェクトの新規性は大きく２つあげられる．

第１点は，MAML アノテーションを用いた各種アプリケーションを実装する上で汎用ライブラリを利用することによりデータの対象に依存せず短期間でその実装が実現できる．
第２点は，本フレームワークを用いて，今までにない，新しいメディア利用環境を実現できることである．実際にメーラ，ブラウザ，BBS クライアントを実装し，有用な各種機能をアノテーションの概念を取り入れることで実現した．MAML ライブラリを用いることにより非常に単純な工程で実装することが可能となる．
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．

５．期待される効果

任意のメディアを対象とした，機械翻訳，情報検索，自動要約，質問応答，知識発見システムなどの実用化や，より高度なデータの加工提示や情報共有が可能となり，その有用性は計り知れない．
また，このような半構造化データが大量に生成されることは，自然言語処理及びデータベース研究分野において非常に有用あると考えられる．

６．普及(または活用)の見通し

MAML汎用ライブラリは近日公開予定であり，本技術を用いた各種アプリケーションの開発が推進されるものと期待できる．また実装した各種アプリケーションは機能拡張後公開予定である．

７．開発者名(所属，e －mail アドレス)

(＊伊藤一成(慶應義塾大学 大学院後期博士課程 k_ito@nak.ics.keio.ac.jp))

参考文献

[1]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，(2004)．