web 上のテキストからの表形式を出力とする情報抽出

曽我部 泰正 (110067)

北陸先端科学技術大学院大学 情報科学研究科

2003 年 2 月 14 日

キーワード: 情報抽出, 固有表現抽出, ontology, www, 表形式, .

本論文では,web 上のテキストから表形式を出力とする情報抽出の手法を提案する. 表は文章の簡潔な要約とみなすことができ、我々の情報抽出手法は一般のテキストを要約する一手法であると考えることができる。近年 internet の普及により,web 上では様々な情報が公開されつつあり, その量は増加の一途をたどっている. ある事柄について web 上から情報を調べる際には, 主にサーチエンジンを用いた情報検索に頼っている. しかしながら、通常のサーチエンジンは単にあるキーワードを含むサイトをリストアップするに過ぎず, その後にユーザー自身が検索結果のリスト中のサイトのそれぞれにアクセスして文書を読み, 情報を選別するという手間のかかる作業が必要となる. 現在, その手間を減らす一つの手法として自動要約システムが研究されているが, この手法を用いても結局は文章を読むことにはなんら変わりない. また, 上述の手間を減らすもう一つの手法として膨大な情報源の中から必要な情報のみを抜き出す「情報抽出」という技術について研究が行われている. しかしながら, 既存の情報抽出手法は抽出のパターンを人手で生成しなくてはならないという問題点がある. さらに手作業でのパターンの生成には時間がかかる上, 限られたトピックにしか対応できない.

この問題点を解決するため, 本研究では手作業によるパターンを必要としない情報抽出手法を提案する. 我々の手法は web 上に存在する多数の表及び web 上に存在するテキストをもとに教師無し学習を用いることにより自動的に抽出パターンを獲得する. 一般にある対象を表形式で表現するときには, その対象にとって重要な情報のみが,「属性」とその「属性値」の対により, 簡潔な形で表現される. 例えば自己紹介に関する表があると仮定すると「属性」とは,「名前」,「血液型」,「趣味」などで, 属性値とは「太郎」,「A 型」,「アイスホッケー」などである。本研究で提案する手法は，このような「属性」と「属性値」の組を従来の手作業による抽出パターンと置き換え, 通常のテキスト中に存在する重要な情報を表形式にまとめる手法である. 我々の手法を適用するにあたり，1) ある 1 つの種類のオブジェクトに関する情報を記述してある多数の表を収集すること， 2) 表の論理的構造を認識すること，つまり，表のどの部分が属性でどの部分が属性値であるかを認識することの 2 つが前処理として必要となる. この処理には既存の手法 [1] を用いた.

我々の情報抽出の手法ではこのようにして統合された表から得られる ontology、すなわちある対象について記述された表に存在する属性や属性値の集合から得られる性質を用いて，通常のテキストから表の属性や属性値となりえる語を抽出し，テキスト中に存在する主要な情報を表形式で表現する．今後，表の属性や属性値になり得る語のことを *table elements* と呼ぶこととする。本手法の具体的アルゴリズムは概ね次の通りである．まず，前処理として，本研究で入力として用いる統合された表に存在する全ての table elements に対し形態素解析を行い, 表の構成要素を形態素単位に分割する．次にそれぞれの table elements に対し tagging を行う．それぞれの table elements に対して，属性には $\langle ATTRn \rangle$, 属性値には $\langle VALn \rangle$（n はクラス名）を付与する。次に web 上に存在する多数のテキストを収集し，そのテキスト群に対し，形態素解析を行う。そして，テキスト中に出現した table elements に対し，同様のタグを付与する．本研究では，このタグを付与されたテキストを訓練データとみなし，既存の教師あり学習の手法を用いて表の構成要素の出現パターンの学習を行った．

次のステップ、すなわち表の構成要素の出現パターンの学習は自然言語処理における最も基本的かつ重要な処理である tagging タスクの一種であるとみなすことができる。本研究では Yamcha [2] を用いて解析を行った．YamCha は Support Vector Machines を学習アルゴリズムとする，汎用的な tagger であり，すでに様々な自然言語処理のタスクに適用されており，高い解析精度を示している．

YamCha を用いた学習を行う際の素性として，語彙, 読み, 標準形, 品詞細分類, 単語の先頭 4 バイト, 単語の末尾 4 バイトを用いた。さらに，我々は単語の意味クラスを素性として用いた．実験では,「PC のスペック」及び「自己紹介」の 2 つのドメインに対して抽出実験を行った。以下に示す評価基準に基づいて評価を行った。

- “\” や “:” などの特殊記号及び数字のみから構成される形態素は評価の対象から除いた．
- 数字は複数の属性に対する属性値となる可能性があるため, 数字のみから構成される形態素が含まれる表の構成要素に関しては評価の対象から除いた．

以下に抽出実験の結果を示す．

- 自己紹介のドメインに対する実験において，素性として単語の意味クラスを導入しなかった場合, その精度 ($F_{\beta=1}measure$) は 52.7 であった．
- 自己紹介のドメインに対する実験では ($F_{\beta=1}measure$) は 53.8 であった．
- PC のスペックに関する実験における精度 ($F_{\beta=1}measure$) は 52.6 であった．

参考文献

[1] Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii. (2001). A method to integrate tables of the World Wide Web. In the Proceedings of the first International Workshop on Web Document Analysis (WDA 2001). pp. 31-34.

[2] Taku Kudou. YamCha: Yet Another Multipurpose CHunk Annotator . http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/