1E-6

SemCode2: オントロジーに基づくアノテーションとトランスコーディング

長尾　確 †
† 名古屋大学 エコトピア科学研究機構

1　はじめに

コンテンツをより賢くするための常套手段としてアノテーション（あるいはメタデータ）を関連付けるというものがあるが、アノテーションがコンテンツに直接関連付けられた情報である限り、そのアノテーションを他のコンテンツに適用するのは一般に困難である。

作成コストが非常に高く、同時に適用範囲も広い情報の代表例は（領域）オントロジーであろう。本論文では、アノテーションの二次利用を念頭において、コンテンツに対する直接のアノテーション（第一層アノテーション）からオントロジーに相当する部分を抽出してメタアノテーション（第二層以降のアノテーション）として再構造化し、トランスコーディングに応用する。

オントロジーを使ったトランスコーディングの例として、テキスト内の用語の言い換えを紹介する。これは、テキストに含まれる専門用語をより平易の言葉で置き換えるものであり、ユーザーがブラウザ上で語をクリックすることによって実行される。この処理はインタラクティブであると同時にインクリメンタルであり、言い換えられた結果にさらに用語が含まれる場合は、続けて言い換えを行うことができる。

2　セマンティック・アノテーションの問題点

アノテーションとは、コンテンツに対するコンテンツつまりメタコンテンツ一般のことである [2]。高精度の検索・要約・翻訳等のコンテンツの高度利用にアノテーションが有効であることはすでに多くの人が認めているところであるにも関わらず、意味的アノテーションに関する具体的な活動が遅々として進んでいない理由は、その作成コストが大きく、現在の技術では自動化できる部分が少ないためである。

さらに、よく考えてみると、コンテンツに直接関連付ける意味的アノテーションには以下のような問題があると思われる。

1. 第一に、第一層アノテーションはコンテンツの変更に柔軟に対応できないことである。これはアノテーションが個々のコンテンツに固有の意味的内容を顕在化したものであるという性質上無理のないことである。

2. 一つ目の問題と関連するが、第一層のアノテーションは、特定のコンテンツの特徴を明示的に記述したものなので、それ以外のコンテンツに流用することが一般に困難である。同様の意味的内容を持つコンテンツになら流用できたほうが適切であるが、そのためにはコンテンツが似ているということを厳密に定義しなければならない。

3. そして三つ目は、意味的アノテーションというより意味表現一般に関する問題であるが、表現されている内容が適切な抽象度と論理性（あるいは推論可能性）を備えているか、という問題である。これは、コンテンツの意味をどう捉えるかによって問題の複雑さが変わってくる。

やるべきことは、意味的アノテーションの再構造化を行って一般性の高い部分を抽出し、コンテンツと独立に管理可能な形式にして、コンテンツが変更されても利用可能であり、他のコンテンツに対しても適用可能な、メタアノテーションを作成することである。

3　オントロジーに基づくアノテーション

以上の問題を解決するための一つの有力のアプローチが、オントロジーの概念の導入である。オントロジーには今ではいろいろな意味があるが、ここでは、辞書的に用いられる概念体系（つまり、何らかの名前から検索される形式的で論理的な概念記述の集合）とする。

本研究でのオントロジーは、第一層のアノテーションから他のコンテンツでも使えそうな一般化可能な部分を抽出して、さらに必要な属性を考慮して再構造化していくことによって構築される。

ここでのオントロジー構築の手順は以下のようになる。

1. テキストコンテンツの第一層アノテーションを作成する。これは主に言語構造の解析と修正である。

2. 言語構造の末端となる語彙に関して、多義語か専門用語と考えられる場合は、辞書を検索して適切な語義の ID を付与する。

3. つまり多くの場合、オントロジーへのポインタとなる ID を決める必要がある。たとえば、語＋品詞＋通し番号のような ID を自動的に生成して、該当する語にアノテートする。

4. 次に、この ID に対応したオントロジーのエントリーを作成する。これは単なる XML ではなく、RDF (Resource Description Framework) を用いて記述する。RDF の利点は、内部データ構造として有向グラフを扱えることである。オントロジーの特徴はネットワーク構造を用いた推論ができることであるため、RDF を用いる意義がある。

3.1　オントロジーエディタ

オントロジーエディタは、言語的アノテーションのオーサリングツールと連動してオントロジーデータを作成・編集するためのツールである。

ここでのオントロジーの従来システムとの大きな違いは、基本的に概念はすべて語義であると考え、必ず、その解説文に言語的アノテーションを付与したものを用意し、それへのリンクを含むことである。

SemCode2: Ontology-Based Annotation and Transcoding

† NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
EcoTopia Science Institute, Nagoya University (†)
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

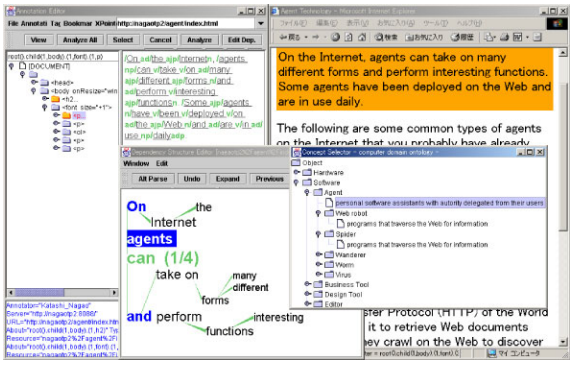

図 1: 言語的アノテーションの編集画面

3.2 オントロジーの作成プロセス

オントロジーエディタを用いたオントロジーデータの作成プロセスは次のようになる。

1. 図 1 のように、テキストコンテンツに対する言語的アノテーションを作成する。語義を付与すべき語のタグを選択し、オントロジー ID を付与する。

2. 選択された語義の解説文に対して言語的アノテーションを付与する。解説文に含まれる用語の検索と語義アノテーションもここで行う。

3. オントロジーの基本フレームが生成され、スロット情報を編集する。この結果は RDF 形式で保存される。具体的には以下のような形式である。

```
<?xml version="1.0" encoding="Shift_JIS"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:sc="http://.../SemCode/semcode-onto-2005-01-01.rdfs#">
 <rdf:Description rdf:about="urn:sense:agent-n-01">
  <sc:is-a>urn:sense:software-n-01</sc:is-a>
  <sc:resource>http://.../text/agent/into.html
  </sc:resouce>
  <sc:definition>http://.../glossary/agent-n-01.html
  </sc:definition>
 </rdf:Description>
</rdf:RDF>
```

4 オントロジーに基づくトランスコーディング

次に、オントロジーを用いたトランスコーディングを考える。ここでは、典型的な例の一つとして、専門用語の言い換えを取り上げる。これは、ユーザーの選択した専門用語をより平易な表現に置き換えるというものである。

まず、オントロジーはコンテンツの第一層アノテーションと用語辞典の解説文に対する言語的アノテーションの二つのリソースを結びつける働きがあるとする。また、オントロジーはその概念を言語化するときに必ず含めるべき属性とそうでない属性の区別を持っているとする。後者の条件は、解説文を使って用語を言い換える場合、どの部分が省略可能でどの部分がそうでないかを決定するのに利用される。

また、ここでの言い換えは、インタラクティブであると同時にインクリメンタルである [1]。つまり、ユーザーにとって理解が困難な用語や、文脈から自分が誤解している感じる用語をオンデマンドに言い換えるのが適切であり、自動的にすべての用語を変換するべきではないだろう。また、用語の説明に別の用語が含まれることはよくあるため、言い換え結果にまだ理解困難な用語が含まれている場合がある。このとき、ユーザーはさらに言い換えを要求することができるようになっているべきである。そのため、言い換え処理はインクリメンタルに実行可能である必要がある。

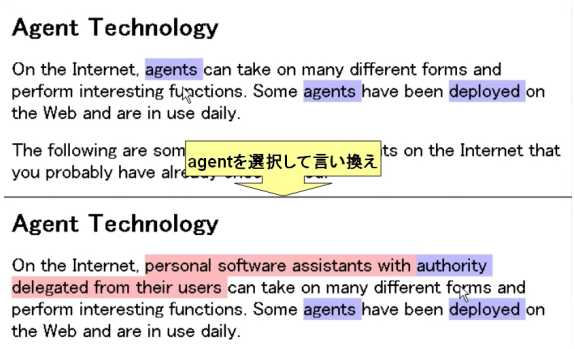

図 2: 言い換えトランスコーディングの画面例（上が変換前、下が変換後）

4.1 言い換えトランスコーディング

図 2 のように、ブラウザ上で専門用語を選択してクリックすると、トランスコーディングによってコンテンツに埋め込まれたオントロジーインタフェースを通して、オントロジーサーバーにリクエストが伝達される。このとき、オントロジーサーバーは、関連するオントロジーデータを呼び出すと同時に、アノテーションサーバーから言語的アノテーションを含めた解説文データを受け取り、それに基づいて言い換えトランスコーディングを実行する。

オントロジーインタフェースは、言い換えだけでなくオントロジーの内容を確認したり、オントロジーに属性を追加する場合にも用いられる。

5 今後の課題

オントロジーの構築は容易ではないが、意味的アノテーションの作成コストを相対的に引き下げるためには、複数のコンテンツに適用可能なより一般的な概念体系を作って、意味的アノテーションの一部とするやり方が妥当である。

今後の課題の一つ目は、複数のコンテンツから別々に派生したオントロジーをまとめあげるための環境作りである。これは、コンテンツや第一層アノテーションを見比べながら文脈の類似性と差異性を考慮して、オントロジーを修正していくためのツール類である。

さらに、ここでのオントロジーは、その概念が言語化された場合の表現と常に密接に関連付けられているから、オントロジーのネットワークを構成すると同時に、言語間の関係も明確になっていくと思われる。これによって、特定分野の用語辞典が自然に出来上がっていく仕組みや、既存の用語辞典を修正していくような仕組みができると思われる。これももちろん今後の課題の一つである。この仕組みは、巨大な Web から知識を構築していくメカニズムの根幹をなすものと考えられるだろう。

参考文献

[1] Ryuichiro Higashinaka and Katashi Nagao, “Interactive Paraphrasing Based on Linguistic Annotation,” Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-02), pp.1218-1222, 2002.

[2] Katashi Nagao, Yoshinari Shirai, and Kevin Squire, “Semantic Annotation and Transcoding: Making Web Content More Accessible,” IEEE MultiMedia, Vol.8, No.2, pp.69-81, 2001.