

## XMLでマークアップされた 言語資料の検索と加工

千葉庄寿 (麗澤大学)

## 用例検索システムの動向

- システムとコーパスが緊密に結びついたもの  
コーパスに付属する専用検索ソフト  
専用のオンライン検索システム
- データを汎用的に検索するもの  
grep ツール (XMLのデータ構造は行を単位とせず、一般的なツールでは必ずしも検索がうまくいかない)  
特定のannotationに特化した検索ツール  
独自のプログラム、スクリプトを作成・利用
- XML に対応する汎用的な検索ツールは多くない (ほとんどが CUI)  
cf. XAIRA (<http://www.oucs.ox.ac.uk/rts/xaira/>)

## 言語研究とXMLの相性

- タグの有用性  
extracting information, re-usability, multi-functionality (Leech, 1997: 3-6)
- annotation の要件を満たす  
recoverability, extricability, conformity to an standard (*idem*. p. 6-7)
- SGMLから受け継いでいる問題点 (豊島, 2001)  
表現力の制約 (文書構造のhierarchyには必ず従わなければならない)

## XMLを用いた用例検索の要件

- お仕着せ検索の問題 (豊島, 2001:9)  
「XMLのマークアップ校正を繰り返したテキストに対して、tagを剥ぎ取った形での頒布を求められるのは、しばしばある事である。これは、XMLデータの検索技術等、(今の処)誰も信用していないからであろう。」
- ブラックボックス的でない、XMLのマークアップを検索に積極的に利用できる汎用の検索ツールが必要

## 言語研究のための検索ツール開発

- 求められるコンセプト  
多言語対応 (Unicode)  
GUIによる操作性の向上と検索プロセスの視覚化  
XML文書を比較的自由に処理できる汎用性  
一般性の高いXML関連技術の利用 XPath (Clark et. al, 1999)

## XMLの検索加工技術標準

- XPath  
XMLの木構造の一部を取り出す  
- XSLTの基盤技術: 他の技術標準にも取り入れられてきている
- XSLT (eXtensible Style Language Transformations)  
XPathで指定するパターンについて構造を変換  
- 変換パターンをテンプレートxsl:template要素で記述
- XSL-FO (Formatting Object)  
- 文書のスタイル情報を記述  
- 用途を考慮し、XSLの機能を分化

## XMLの検索と加工の実践

- XPath
  - ノード (要素, 属性, ルート, テキストなど, XMLのさまざまな部分をまとめてこう呼ぶ)
  - ロケーションパス location path
    - ロケーションステップを/で連結したもの:  
軸 axis :: ノードテスト[述語]\* (0個以上)
    - 軸: コンテキストノードでの検索先の位置関係, 13種類定義されている。
  - 述語: 関数や演算子を記述
    - ノードに条件を指定し, 検索結果を絞り込む
    - 組み合わせることも, 複数列挙することも可能

2004-12-11

言語資料のXMLによるマークアップ

## より高度な技術標準

- 規格の普及にはまだまだ時間がかかると思われるが...
- XPointer: XPathよりも複雑なパターン検索を可能にするポイント機能, 2003年に W3CのRecommendation 規格に。
- XLink: XML 文書のリンク機能, 2001年に W3CのRecommendation 規格に。
- XQuery, XML Query: XML データベースの検索

2004-12-11

言語資料のXMLによるマークアップ

## XPath, XSLTの限界と利点

- XPath 1.0 で検索できるのは, ノードを始点とする木構造的な階層のみ, 中のデータから必要なものだけを取り出す, ということができないことがある
  - 正規表現など, 他の検索技術の併用でOK?
- XPath 1.0 はDTDに対応していない, あくまで表面的な構造をたどる。
- XSLTは XPathの表現力に依存する, 柔軟な変換を実現するにはかなりの訓練がいる。
- 一方で, これまでは取り出しにくいパターンを拾い出すことができる, 例:
  - ある属性, 要素を含む語を含む文を表示
  - 段落の最初の2文のみを表示

2004-12-11

言語資料のXMLによるマークアップ

## 結論

- 電子テキストの柔軟な検索に XML 関連技術の利用は有効
- XML関連技術の利用を本格化させよう
  - XMLおよび関連技術の普及: XPath, XSLTは実用レベルに
  - XMLによる構造記述を個々の研究者の関心に合わせ利用: 「ブラックボックス」的でない用例検索の実践

2004-12-11

言語資料のXMLによるマークアップ

## コーパス研究の知識インフラ

- XMLを利用した用例検索の教育的価値
  - XMLの基本知識, XPath, 正規表現, Unicode の基礎知識はより高度なテキスト処理, 言語分析への足がかり
    - XMLデータの作成, 加工: より高度なXMLの知識とマークアップのための技術的訓練が必要
  - 標準規格としてのXML, XPath の知識は多方面に利用できる

2004-12-11

言語資料のXMLによるマークアップ

## ディスカッション

2004-12-11

言語資料のXMLによるマークアップ