Vergleiche und Transformationen für XML-Dokumente - Teil 2

Ein Ansatz zur hierarchischen, adaptiven Kollationierung

Varianz in der Gleichheit

- nicht nur der Manuskripte, sondern auch ihrer Teile
- die Elemente sind nicht nur gleich **oder** verschieden, sondern auch gleich **und** verschieden (ähnlich)
- Problem, nach Divergenz wieder zusammenzufinden (Aufsatzpunkte)

Vereinfachende Lösungen

- diff-Tools (für Programmiersprachen) unterscheiden nur gleiche und verschiedene Zeilen (Basis für Patches)
- XML-Vergleichswerkzeuge betrachten Textknoten auch bei kleinen Abweichungen als verschieden.
- Beides nicht tauglich für Kollation von Texten

Ein Maß für Ähnlichkeit

- Die eben vorgestellten Ansätze (Wenger, Hemmerich) überwinden diese Schwierigkeiten, indem sie ein Ähnlichkeitsmaß (nach Myer) einführen.
  - Die Editierdistanz (edit distance) ist die kleinste Zahl von Schritten, die eine Zeichenkette in eine andere überführt.
  - Sie ist ziemlich “teuer”.
  - Basiseinheit sind Zeichen.

Zwei Ideen

- (die negative:) Alle relevanten Unterschiede finden sich im Fließtext, nicht in den Tags.
- (die positive:) Hierarchische Strukturen und zugehörige Lokalitäten verbessern Klassifikation von Unterschieden und können Komplexität reduzieren.
- (vorläufige Moral:) Kollationierung wird durch XML unterstützt, aber nicht durch jedes XML.

Vergleichsstrukturen

- Nur grundlegende Strukturen werden benutzt, in typischen Fällen: Absätze, Sätze, Wörter.
- Die verschiedenen Ebenen werden **unterschiedlich** behandelt. Operationen haben verschiedenen Sinn. Jeweilige Charakteristika können **sinn**gemäß benutzt werden.
- Basis des Vergleichs sind Wörter. Als Ähnlichkeitsmaß eines Satzes kann die Anzahl der übereinstimmenden Wörter dienen.

Vorverarbeitung

- Überzählige Tags müssen ausgefiltert, abweichende Auszeichnungsweisen evtl. angeglichen werden. Für die Vorverarbeitung benützen wir die im ersten Teil vorgestellten und erprobten Werkzeuge.

- Fehlende Auszeichnung der Wortebene und ggf. der Satzebene wird durch eine gleichfalls erprobte Zerlegung von Sätzen in Wortlisten kompensiert.

(Optimistische) Vergleichsstrategie

- Die eigentliche Verarbeitung arbeitet auf Wortlisten. Basis ist der wortweise Vergleich von Sätzen in einem Absatz.

- Zunächst wird im Umfeld sequentiell nach gleichen Sätzen gesucht, im zweiten Schritt dann nach ähnlichen, d.h. solchen mit einem definierten Mindestanteil gleicher Wörter.

- Sätze ohne Entsprechung kommen auf einen Stapel - für spätere Vergleiche

Satzvergleich

- Die einander zugeordneten Sätze werden, ähnlich wie mit MyersDiff, verglichen, allerdings nicht auf Zeichen-, sondern auf Wortebene.
- Wörter werden nur auf Gleichheit, nicht auf Ähnlichkeit untersucht. Bei Unterschieden werden die gesamten Wörter als unterschiedlich ausgegeben (”Keep it simple”).

Hase-Igel-Sätze

```
<p n="40">
  <s n="1">
    Der Igel ist vor dem Hasen.
  </s>
  <s n="2">
    Der Hase isst schneller
    als der Igel.
  </s>
  <s n="3">
    Der Igel gewinnt dennoch.
  </s>
  <s n="4">
    Sehr seltsam.
  </s>
</p>
```

```
<p n="50">
  <s n="1">
    Der Hase ist schneller
    als der Igel.
  </s>
  <s n="2">
    Trotzdem ist der Igel
    vor dem Hasen.
  </s>
</p>
<p n="51">
  <s n="1">
    Der Igel gewinnt dennoch.
  </s>
</p>
```

Ausgabe der Differenz

- Gleiche Zuordnungen und Ausgabe von Strukturänderungen wie bei Wenger bzw. MyersDiff
- aber Veränderung bei Textdifferenz

zeichenbasiert
Der Igel ist vor dem Hasen
Trotzdem ist der Igel vor dem Hasen
bzw.
Der Hase isst schneller als der Igel
Der Hase ist schneller als der Igel

wortbasiert
Der Igel ist vor dem Hasen
Trotzdem ist der Igel vor dem Hasen
bzw.
Der Hase isst schneller als der Igel
Der Hase ist schneller als der Igel

Adaptivität und Erweiterbarkeit

- Anpassbarkeit der Ähnlichkeitsschranke
- Einbezug von anderen relevanten Auszeichnungen: Verse statt Sätze, Überschriften (Reihenfolge)
- Berücksichtigung von Schreibvarianten auf Wortebene
- generell Einbeziehbarkeit von phonematischen, morpologischen und syntaktischen Regeln und Werkzeugen