XDOC - Extraktion, Repräsentation und Auswertung von Informationen

Manuela Kunze and Dietmar Rösner

Otto-von-Guericke-Universität Magdeburg, Institut für Wissens- und Sprachverarbeitung, P.O.box 4120,
D-39016 Magdeburg, Germany
{makunze|roesner@iws.cs.uni-magdeburg.de}

Zusammenfassung Um der wachsenden Informationsflut gerecht zu werden, versuchen wir mit dem XDOC-Ansatz, die für den Benutzer relevanten Informationen aus vorliegenden Dokumenten zu analysieren und dem Benutzer in geeigneter Form zu repräsentieren. Dabei werden die ursprünglichen Texte um XML-Tags erweitert, mit denen die Ergebnisse der einzelnen Analysen abgebildet werden. Für die Interaktion mit dem Benutzer wurde ein auf einen Browser basierendes Interface gewählt. Mit dieser Schnittstelle ist es möglich, die Ergebnisse aus den verschiedenen Modulen von XDOC entspechend der Anforderungen des Benutzers für die Präsentation aufzubereiten.

Einleitung

In der heutigen Zeit wird der Benutzer mit einer ständig wachsenden Zahl an Dokumenten konfrontiert, die er erfassen und verarbeiten muß. Ziel unseres Projektes ist es den Benutzer in der Form zu unterstützen, daß für den Benutzer relevante Informationen durch das XDOC-System in den Texten markiert und extrahiert werden. Da einige Module sich noch in Bearbeitung finden, wurde ein Interface aufgesetzt, welches sowohl für den End-Nutzer als auch für den Entwickler dienlich ist.

System XDOC

In der Abbildung 1 ist eine graphische Darstellung des XDOC-Systems zu sehen. Innerhalb von XDOC werden verschiedene Module miteinander kombiniert, um somit semantische Informationen in den Dokumenten zu selektieren und zu extrahieren. Nachdem eine Vorverarbeitung durch einen Tagger erfolgte, beginnt die syntaktische Analyse, in der auch das Parsen von domain-bezogenen Bezeichnern (z.B. DIN-Angaben) erfolgt. Das Ergebnis der syntaktischen Analyse liefert unter Umständen mehrere Lesarten bezüglich eines Satzes bzw. auch einer Phrase. Diese Ambiguitäten können teilweise über die semantische Analyse

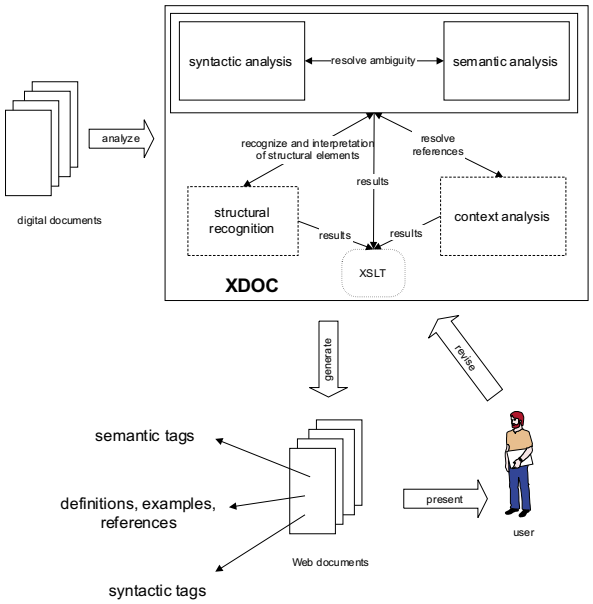

Abbildung1. XDOC-System

aufgelöst werden. Neben der Auflösung von Ambiguitäten liefert uns die semantische Analyse auch erste einfache und komplexere Konzepte [1].

Als Ausgangsdaten für die verschiedenen Analysen innerhalb von XDOC werden syntaktische sowie semantische Lexika und domain-bezogene Ontologien verwendet.
Die Ergebnisse der einzelnen Analysen werden in XML-Notation(siehe [3]) ausgezeichnet (Beispiel aus der syntaktischen Analyse):

```
<PP CAS="AKK">
  <PRP CAS="AKK">durch</PRP>
    <NP TYPE="COMPLEX" RULE="NPC1" GEN="NTR" NUM="SG" CAS="AKK">
      <NP TYPE="FULL" RULE="NP1" CAS="AKK" NUM="SG" GEN="NTR">
        <N>Schaffen</N>
      </NP>
      <NP TYPE="FULL" RULE="NP2" CAS="GEN" NUM="SG" GEN="MAS">
        <DETD>des</DETD>
        <N>Zusammenhalts</N>
      </NP>
    </NP>
</PP>
```

[1] Die Ergebnisse der semantischen Analyse sind primär angedacht für die Erstellung einer Wissensbasis und der Weiterverarbeitung in der Form, daß strukturelle Informationen (Positionen) von z.B. Definitionen, Beispiele erhalten bleiben.

Diese Ergebnisse können innerhalb des XDOC-Systems zum einen als einfacher XML-Text angezeigt werden, bzw. werden mit XSL-Stylesheet aufbereitet, um somit zum Beispiel die Strukturen von gefundenen syntaktischen Lesarten hervorzuheben (siehe Abbildung 2).

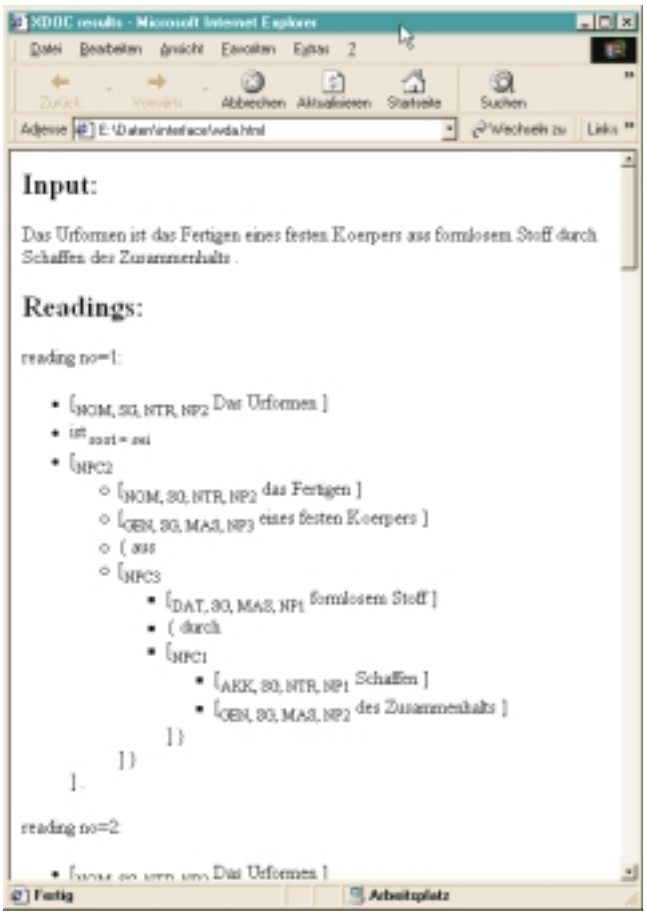

Abbildung2. Darstellung von Lesarten über XSL Transformationen

Die Interaction mit dem Benutzer erfolgt über einen Browser (siehe Abbildung 3). Dabei stehen dem Benutzer verschiedene Funktionen der Module zur Verfügung. Da die Ergebnisse alle im XML-Format vorliegen, können über XSL Transformationen die wichtigsten Resultate der Analysen hervorgehoben werden. Der Benutzer hat über das Interface dann die Möglichkeit gefundene Ergebnisse zu bestätigen oder zum Beispiel die richtige Lesart zu markieren.

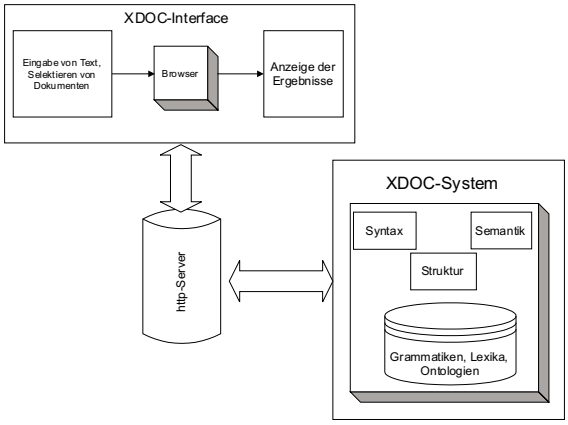

Abbildung3. Interface

Schlussbemerkungen

Die hier beschriebene Workbench vereint verschiedene linguistische Methoden, um durch eine geeignete Kombination dieser Module die Informationen aus Dokumenten zu extrahieren. Durch den modularen Aufbau (ähnlich dem GATE-System [1], [2]) und die Verwendung von XML zur Beschreibung der Ergebnisse wurde die Interoperabilität der Module und eine flexible Kombinierbarkeit der Module sichergestellt. Somit ist es möglich, z.B. statistische Ansätze zur Lösung von Problemen wie z.B. das PP-Attachment (siehe [4]) in XDOC zu integrieren. Durch die Nutzung von XML können zur weiteren Auswertung der Ergebnisse alle derzeit zur Verfügung stehenden Tools für XML genutzt werden. Innerhalb des hier beschriebenen Systems wurde zunächst nur XSL in Kombination mit dem xt-Tool von J.Clark[5] genutzt. Durch die Einbindung verschiedener XSL-Dateien können verschiedene Auswertungen über die teilweise komplexen Ergebnisse der Analysen realisiert werden. Mit XSL erfolgt somit eine weitere Selektierung bzw. eine Repräsentation von für den Benutzer relevanten Informationen.

Literatur

1. H. Cunningham and Y. Wilks, *GATE - a General Architecture for Text Engineering*, Proccedings of COLING-96 (1988).
2. GATE-Site, *http://gate.ac.uk*.
3. M. Kunze and D. Roesner, *Eine XML-basierte Werkbank fuer das Document Mining.*, Proceedings der GLDV-Fruehjahrstagung 2001 (2001), 131–140.
4. M. Volk, *Scaling up. Using the WWW to resolve PP Attachment Ambiguities.*, KONVENS 2000: 5. Konferenz zur Verarbeitung natuerlicher Sprache (2000), 151–155.
5. xt Site, *http://www.w3.org/style/xsl*.