

Basic Techniques for the Extraction and Annotation of Machine Understandable Information

Manuela Kunze, Dietmar Rösner



Otto-von-Guericke Universität Magdeburg
Fakultät für Informatik
Institut für Wissens- und Sprachverarbeitung

Semantic Web

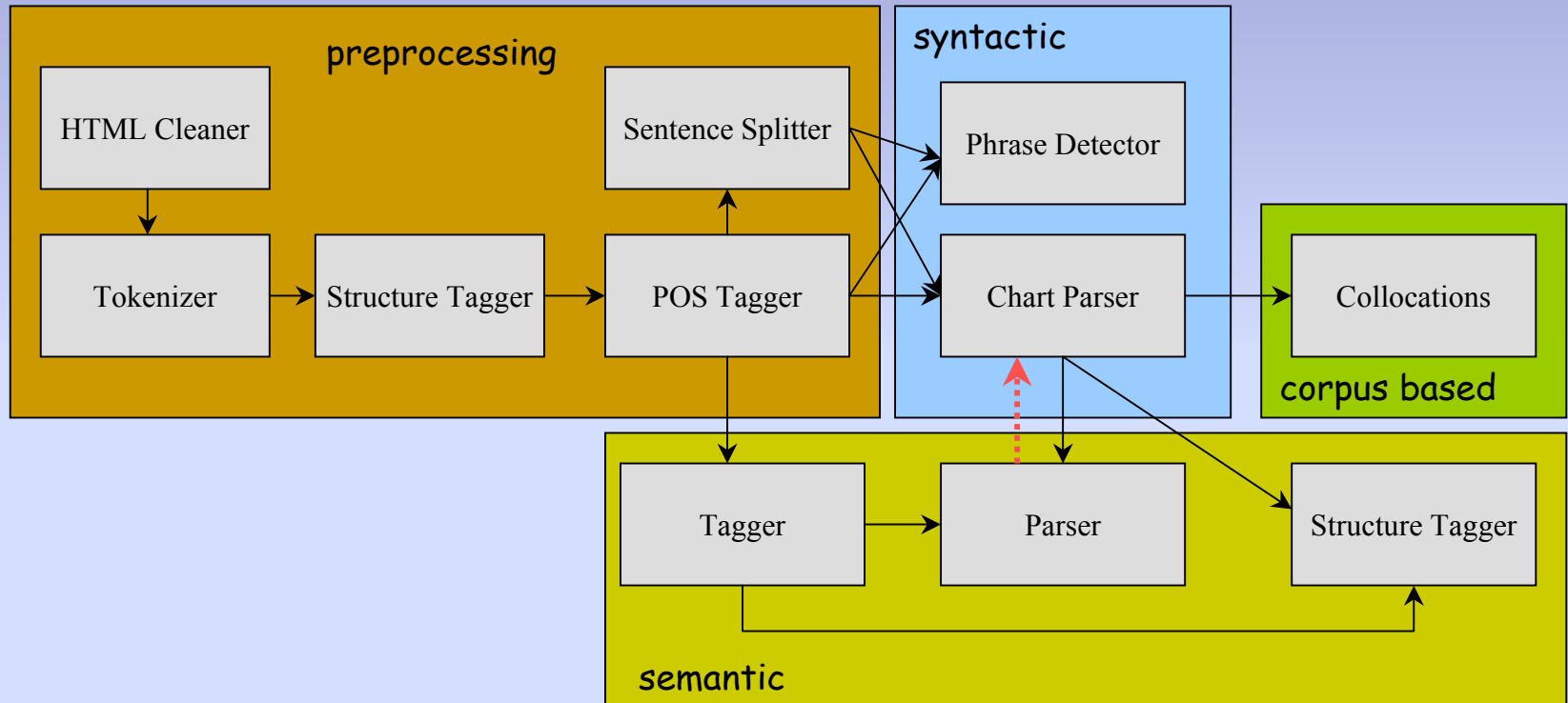
- ▶ core
 - machine understandable information about resources
- ▶ but
 - documents are mostly without any annotation

- ▶ Idea: Natural Language Processing for the Semantic Web
 - annotate a paragraph with additional information

DocSuite XDOC

- ▶ linguistic and semantic analyses
 - German
- ▶ language processing for
 - extraction of metadata
 - automatic markup of documents
- ▶ XML based resources and results
 - XSL transformations

DocSuite XDOC



Semantic Tagger

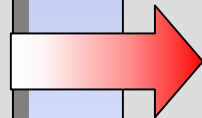
1. Bekleidung : 3 lange graue Unterhosen . Darueber eine dunkelblaue Trainingshose Tuerkisfarbenes Unterhemd , darueber ein blau-gelb-kariertes Herrenoberhemd der Marke Friendship , Groesse 37 , darueber ein pink-grau-gemustertes Sweat-Shirt . Die Oberbekleidung ist vollstaendig durchtrennt (med . Massnahme) . Am linken **Handgelenk** eine Uhr der Marke Casio . Bekleidung und Schmuckgegenstand verbleiben bei der Leiche . 2. Leiche eines 61 Jahre alten Mannes von 1,61 m Koerperlaenge und 58,1 kg Koerpergewicht . Koerperbau regelmaessig , schlankwuechsiger Koerperbautyp . 3. An der Koerperrueckseite mit Aussparung der Aufliegeflaeche sowie an der Koerpervorderseite mit Aussparung der linken Gesichtsseite dunkelblaurote , nicht mehr wegdrueckbare Totenflecke . Totenstarre in den grossen **Gelenken** in Loesung . 4. Die Haare befinden sich nur noch kranzfoermig im Hinterkopfbereich und an den Seiten , Laenge bis zu 4 cm . 5. In der **Kopfhaut** keine **Schwellungen** oder **Durchtrennungen** feststellbar . 6. **Stirnhaut** unversehrt . In der Koerpermittellinie , mit dem Zentrum 13,0 cm ueber der Nasenwurzel , weissliche **Narbe** von 4 : 1 cm . 7. **Augen** geschlossen . Sulzige Durchtraenkung im Bereich des rechten Augenober- und -unterlides . Deutliche Gefaesszeichnung erkennbar . Keine fleckfoermigen **Blutungen** in den Lidumschlagsfalten . Pupillenweite beidseits 4 mm . 8. **Haut** ueber der **Nase** unversehrt . Im rechten Nasenloch etwas roetliche Fluessigkeit . Im linken Nasenloch etwas Schleim . 9. **Ohren** wohlgebildet . In den aeusseren Gehoergaengen kein ungoeheriger Inhalt . 10. **Haut** ueber den Wangen unversehrt . Bis 0,1 cm langer Bartwuchs . 1,2 cm unterhalb der Mitte der Unterlippe waagrecht gestellte weissliche **Narbe** von 0,8 cm . 11. Mund geschlossen . **Mundschleimhaut** ohne **Unterblutungen** . In der Mundhoehle etwas Schleim . Die **Zunge** steht hinter den Kiefern . Deutlich un gepflegter Zustand der Zaehne , nur noch einzelne Zaehne vorhanden , keine frischen Zahnausbrueche . 12. Kopf relativ leicht , jedoch nicht widernatuerlich beweglich . 13. **Halshaut** unversehrt . 14. **Brusthaut** unversehrt . In der **Haut** ueber dem unteren Anteil des **Brustbeines** , in einer Ausdehnung von 5 : 8 cm Textildruckspur . 15. **Bauchhaut** unversehrt . **Gruenverfaerbung** im gesamten Unterbauch . Nabelbucht verschmutzt . 16. Aeusseres Genitale vom maennlichen Typ . Keine Verunreinigungen . 17. Obere **Gliedmassen** muskelschwach . An der Innenseite des rechten **Unterarmes** , praktisch vom **Ellbogengelenk** bis zum **Handgelenk** reichend , in einer Gesamtausdehnung von 22 : 4 cm , weissliches Narbengebiet . Ueber der Handgelenkinnenseite waagrecht gestellte **Narben** von bis zu 1 cm Laenge erkennbar . Fingernaegel verschmutzt . Sie ueberragen die Fingerkuppen deutlich . 18. Weissliche **Narbe** ueber dem linken Ellenbogen von 2,5 : 0,5 cm . In der Mitte des linken **Unterarmes** an der Innenseite weissliche **Narbe** von 5 : 3 cm . An der daumenseitigen Kante des **Handgelenkes** Roetung der **Haut** von 1,0 cm im Durchmesser . Auf Einschnitt mit dem Gewebe verfilzte **Unterblutung** . 19. Untere **Gliedmassen** muskelschwach . Weissliche **Narben** ueber dem rechten Fussgelenk mit Uebergreif auf die **Haut** des Fussrueckens in einer Gesamtausdehnung von 8 : 6 cm . Linkes **Bein** unversehrt . 20. **Haut** des Rueckens unversehrt . In die **Haut** des Rueckens wird ein Laengsschnitt gelegt und die **Haut** zu beiden Seiten flaechenhaft abpraepariert . In der **Rueckenhaut** rechtsseitig , mit dem Zentrum 10 cm oberhalb des Darmbeines und 4 cm rechts der Koerpermittellinie , lackrote **Unterblutung** in der **Muskulatur** von innen nach aussen in einer Ausdehnung von 1,5 : 1,0 cm . **Schulterblaetter** und **Domfortsaetze** der **Wirbelsaeule** intakt . 21. Die **Schaedelschwarte** ist unversehrt . 22. **Schlaefenmuskulatur** beidseits ohne Besonderheiten . 23. **Schaedeldach** intakt . Es misst im Saegesschnitt 0,3 - 0,6 cm . Keine **Blutung** zwischen **Schaedeldach** und harte **Hirnhaut** . Keine **Blutung** unter die harte **Hirnhaut** . 24. Deutliche Blutfuelle der **Gefaesse** am Hirn . Hirngewicht 1350 g . Weiche **Hirnhaeute** ohne Besonderheiten . Furchen abgeflacht , Windungen verstrichen . Druckring am Kleinhirn . Auf den Hirnschnittflaechen kein Anhalt fuer **Blutungen** oder Erweichungen

Semantic Parser

...

Urformen ist
Fertigen fester
Koerper aus
formlosem Stoff *

...



```

<CONCEPT TYPE="PROCESS">
  <WORD>Fertigen</WORD> → production
  <DESC>Schaffung von etwas</DESC>
  <SLOTS>
    <RELATION TYPE="RESULT">
      <ASSIGN_TO>OBJECT</ASSIGN_TO>
      <FORM>N(gen, fak) P(akk, fak, von)</FORM>
      <CONTENT>fester Koerper</CONTENT> → of solid objects
    </RELATION>
    <RELATION TYPE="SOURCE">
      <ASSIGN_TO>MATERIAL</ASSIGN_TO>
      <FORM>P(dat, fak, aus)</FORM>
      <CONTENT>aus formlosem Stoff</CONTENT> → from formless matter
    </RELATION>
  </SLOTS>
</CONCEPT>

```

*primary shaping is the production of solid objects from formless matter

Semantic Structure Tagger

```
<RELATION>
```

```
  <TYPE>has</TYPE>
```

```
  <DESC>has_attribute</DESC>
```

```
  <FILLER>ADJ</FILLER>
```

```
  <PRESENTATION>
```

```
    <ELEMENT>N</ELEMENT>
```

```
    <ELEMENT>ADJ</ELEMENT>
```

```
</PRESENTATION>
```

```
<SYNTACTIC>
```

```
  <STRUCTURE>NP1</STRUCTURE>
```

```
  <STRUCTURE>NP2</STRUCTURE>
```

```
  <STRUCTURE>NP3</STRUCTURE>
```

```
  <STRUCTURE>MA1</STRUCTURE>
```

```
</SYNTACTIC>
```

```
</RELATION>
```

resource for the semantic structure analysis

noun phrase: *darkred liver*

syntactic structure:

NP1: adjective noun

semantics of darkred: **color**

predicate: **has_color**

result:

has_color(liver,darkred)

Semantic Structure Tagger

```

<topicmap>
<topic id="Leber">
<instanceof><topicRef
xlink:href="#organ"></instanceof>
<basename>
<basenamestring>Leber</basenamestring>
</basename>
</topic>
<topic id="dunkelrot">
<instanceof><topicRef
xlink:href="#color"></instanceof>
<basename>
<basenamestring>dunkelrot</basenamestring>
</basename>
</topic>

```

```

<association>
<instanceof><topicRef
xlink:href="#has_color"></instanceof>
<member>
<rolespec><topicRef xlink:href="#organ"></rolespec>
<topicref xlink:href="#Leber">
</member>
<member>
<rolespec><topicRef xlink:href="#color"></rolespec>
<topicref xlink:href="#dunkelrot">
</member>
</association>
</topicmap>

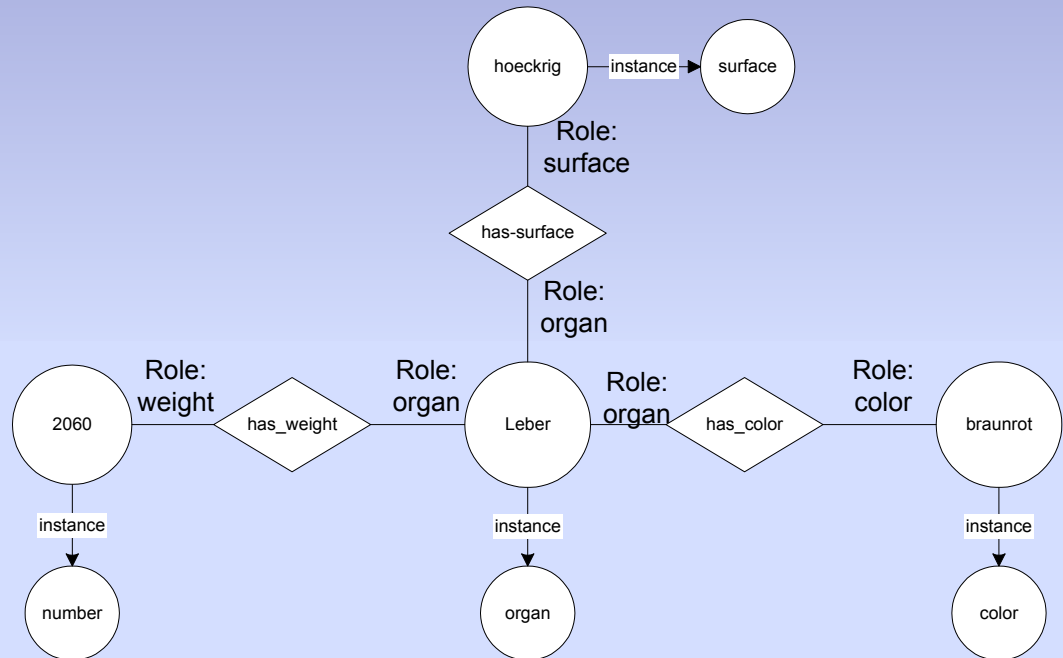
```

Semantic Structure Tagger

...

Gewicht der **Leber** **2060 g**. Auf Einschnitt **braunrote Farbe** des Gewebes. Geringgradige **hoeckrige Oberflaeche**. Keine Verfestigung des Gewebes feststellbar. Von der Schnittflaeche laesst sich reichlich schmutzig-roetliche Fluessigkeit abstreifen.

...



Conclusion

- ▶ interpretation of documents
 - combining linguistic and semantic tools

- ▶ XML based
 - resources and results described by DTD
 - independent of the finally required target markup language
 - RDF(S), Topic Maps, OWL, ...

Conclusion

► current application fields

- semantic analysis of autopsy protocols
- extraction of company profiles (WWW pages)
- information extraction of Medline abstracts

