Exploiting Intra- and Inter-Document Relations in Corpus-based Natural Language Processing

Dietmar Rösner, Manuela Kunze, Jörg Kapfer

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

Overview

- background
- feasibility study
- learning from the corpus
- phrases and paraphrases
- future work

Background

- autopsy protocols as valuable sources for research in forensic medicine
- example: injury patterns in car accidents
- up to now:
  - manual search and evaluation
  - no computer support
- project proposal: CL and NLP to support research in forensic medicine
- proposal under evaluation

Feasability study

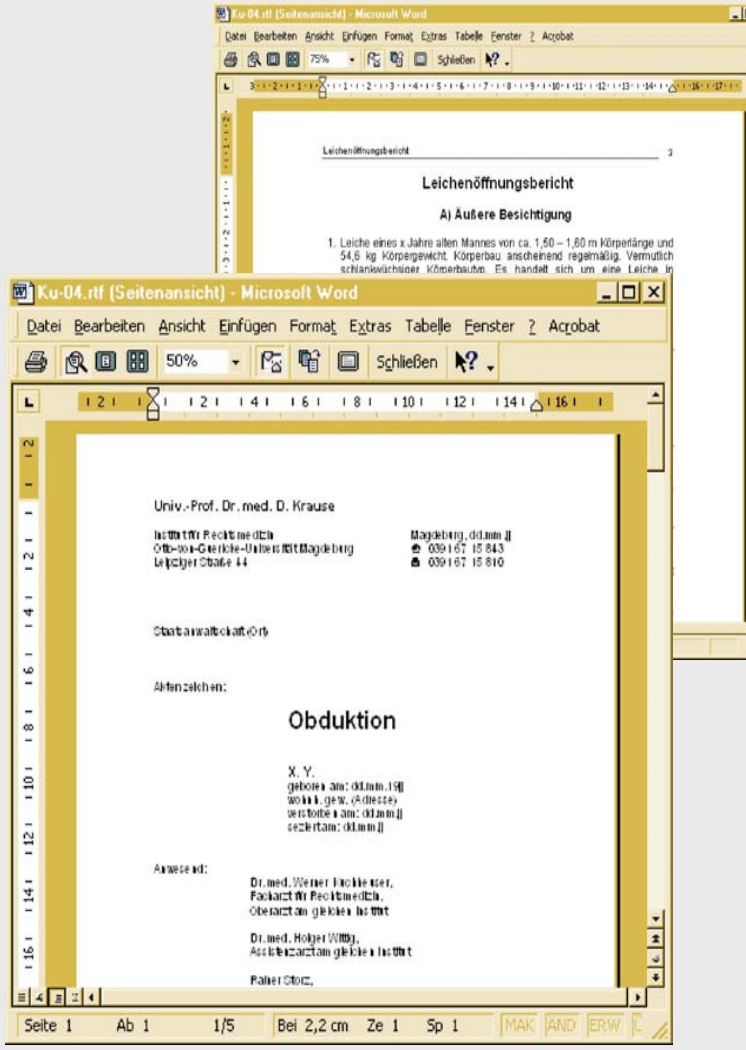
- approx. 600 documents
- > 1.000.000 word forms
- in German
- from time period 1997 till 2002
- Institute for Forensic Medicine, Magdeburg

Aspects of the Corpus

- documents highly structured
- but: structure only implicit
- subdocuments with different sublanguages
- intended audience: nonmedical experts
  - use mundane terms
- plan: creation of a nation wide corpus

Major Subdocuments

- findings (`Besichtigung´)
  - objective recording of results
  - telegrammatic style
- history (`Vorgeschichte´)
  - narrative
  - event and situation oriented
- discussion (`Diskussion´)
  - argumentation
  - interpretation and conclusion



General Aspects

- autopsy protocols as representatives of document class `expert opinion´
  - findings and observations
  - background
  - argumentation and conclusion
- `expert opinion´ documents relevant in many domains
  - medicine
  - engineering
  - jurisdiction
  - ...

A central issue

- How can information from the corpus be exploited for resource creation and maintenance ?
  - lexical resources
  - grammatical resources
  - conceptual resources
- support users as much as possible

Learning tasks

examples:

- detect possible typing errors
- deduce lemmata from word form occurrences
- case frame abstraction
- uncover semantic relations

Detection of possible typos

- compare tokens with other tokens that are `close´
- check for:
  - omissions
  - insertions
  - switch of characters
- if a near match is not interpretable as a possibly inflected or derived form take it as a possible typo
- lexicon free approach
- `ìn the corpus lies the truth´: interdocument relations

Typo detection

- examples from discussion subdocuments:

Todeseintritt 187 : Todeseinritt 1 Todesteintritt 1

Gewalteinwirkung 136 : Gewalteiunwirkung 1 Gewaltenwirkung 1

Gewalteinwirkungen 147 : Gerwalteinwirkungen 1 Gewalteinwirkugnen 1

Blutpfropfbildungen 6 : Blutpropfbildungen 1 Blutpfopfbildungen 1

Weichteilprellungen 14 : Weichteilprellugnen 1 Wichteilprellungen 1

Leichenoeffnungsbefund 410 : Leichenloeffnungsbefund 1
Leichenoeffnungsbefung 1 Leichenoeffnunsgbefund 1

Typo detection

- examples from discussion subdocuments:

Todeseintritt 187 : Todeseinritt 1 Todesteintritt 1

Gewalteinwirkung 136 : Gewalteiunwirkung 1 Gewaltenwirkung 1

Gewalteinwirkungen 147 : Gerwalteinwirkungen 1 Gewalteinwirkugnen 1

Blutpfropfbildungen 6 : Blutpropfbildungen 1 Blutpfopfbildungen 1

Weichteilprellungen 14 : Weichteilprellugnen 1 Wichteilprellungen 1

Leichenoeffnungsbefund 410 : Leichenloeffnungsbefund 1
Leichenoeffnungsbefung 1 Leichenoeffnunsgbefund 1

Typo detection

- examples from `examination´

Zustand 788 : Zusand 2

Bohrloch 13 : Borhloch 1

Bruechen 12 : Brechen 1

Drainage 18 : Dainage 1

Drittels 11 : Drittes 1

Gelenken 326 : Glenken 2

Harnwege 126 : Hanrwege 1

Hohlvene 288 : Hohvene 1

Typo detection

- examples from `history´

Oberkoerper 24 : Oberkoeroper 1

Verletzungen 29 : Verkletzungen 1

Strassengraben 14 : Strassengaben 2

Ermittlungsakte 270 : Ermittlungakte 2

Kollisionsstelle 6 : Kollisionstelle 1

Hauabschuerfungen 1 : Hautabschuerfungen 1

Nachmittagsstunden 6 : Nachmittagstunden 1

Ermittlungsbehoerde 6 : Ermittlungsbeoerde 1

Typo detection

- to minimize `false alarms´:
  - encode knowledge about possibly inflected or derived word forms
  - examples:
    - feminine form: Hausbewohnerin 3 : Hausbewohnern 1
    - declination:

    Schluesselbeine 35 : Schluesselbeines 107 Schluesselbeinen 6 Schluesselbeins 1

    Verkalkungsbeet 5 : Verkalkungsbeete 8 Verkalkungsbeetes 2 Verkalkungsbeeten 1

Lemma deduction

- examples from `examination´:

Schluesselbeine 35 : Schluesselbeines 107 Schluesselbeinen 6 Schluesselbeins 1

Verkalkungsbeet 5 : Verkalkungsbeete 8 Verkalkungsbeetes 2 Verkalkungsbeeten 1

Ellenbogengelenk 37 : Ellenbogengelenkes 73 Ellenbogengelenken 35 Ellenbogengelenke 3

Aufliegegeschwuer 13 : Aufliegegeschwueres 2 Aufliegegeschwuere 2 Aufliegegeschwueren 1

Bekleidungsstueck 2 : Bekleidungsstuecke 3 Bekleidungsstueckes 2 Bekleidungsstuecken 1

Fettgewebsanteile 15 : Fettgewebsanteilen 3 Fettgewebsanteiles 1 Fettgewebsanteils 1

Lemma deduction

- examples from `history´:

Jugendliche 4 : Jugendlichen 3 Jugendlicher 1

Medikamente 20 : Medikamentes 2 Medikamenten 1

Mitbewohner 13 : Mitbewohners 1 Mitbewohnern 1

Polizeibeamte 7 : Polizeibeamten 21 Polizeibeamter 4

Unterschenkel 6 : Unterschenkels 3 Unterschenkeln 1

Lemma deduction

- examples from `discussion´:

Wirkstoff 16 : Wirkstoffe 3 Wirkstoffen 2 Wirkstoffes 1
Kniegelenk 1 : Kniegelenke 4 Kniegelenkes 2 Kniegelenken 1
Medikament 4 : Medikamente 7 Medikamenten 5 Medikamentes 3
Herzinfarkt 16 : Herzinfarktes 5 Herzinfarkte 3 Herzinfarkten 1
Oberschenkel 15 : Oberschenkels 31 Oberschenkeln 3 Oberschenkel- 1
Gegenstossherd 8 : Gegenstossherdes 2 Gegenstossherden 2 Gegenstossherde 1
Sicherheitsgurt 6 : Sicherheitsgurtes 5 Sicherheitsgurten 3 Sicherheitsgurte 2
Schienenfahrzeug 7 : Schienenfahrzeuge 2 Schienenfahrzeuges 2 Schienenfahrzeugen 1

Discussion: Learning from Corpus

- numbers still (too) low
- feasability study only
- larger corpus essential
- corpus under construction

Paraphrases in the corpus

- paraphrases within a document (intra-document paraphrases)
- paraphrases in different documents (inter-document paraphrases)

Inter-Document Paraphrases

- Die Vorsteherdruese ist altersentsprechend.
  {The prostate is corresponding to age.}
- Vorsteherdruese ohne Besonderheiten.
  {Prostrate without anomalies.}
- Vorsteherdruese regelrecht.
  {Prostrate rule conforming}.

Inter-Document Paraphrases

Dimensions:

- Verbosity vs. Ellipsis
- Quasi-synonyms
- Redundancy
- Noun compounds vs. complex noun phrases
- Contextual paraphrases
- Coordination

Inter-Document Paraphrases

Dimensions:

- Verbosity vs. Ellipsis

`Mund geoeffnet.' (mouth open)

`Mund ist geoeffnet.'

or

`Der Mund ist geoeffnet.'

Inter-Document Paraphrases

- Examples: Quasi-synonyms

  `Vorsteherdruese altersentsprechend.'
  (Prostate corresponding to age.)

  or

  `Vorsteherdruese regelrecht.,

  (Prostate rule conforming.)

Inter-Document Paraphrases

- Examples: Redundancy

Lungen intakt.

{Lungs intact.}

Lungen beidseits intakt.

{Lungs on both sides intact.}

Inter-Document Paraphrases

- Examples:

  Noun compounds vs. complex noun phrases

concept: `weight of the liver'

`Lebergewicht' more likely than

`Gewicht der Leber'.

for weight of other organs: `<Organ>gewicht'

more likely than `Gewicht des/der <Organ>'

Inter-Document Paraphrases

- Examples: Coordination

  Vorsteherdruese und Samenblaeschen unauffaellig.

  {Prostrate and seminal vesicles without findings.}

Intra-Document Paraphrases

- to be processed when single documents are interpreted.
- types:
  - Syntactical Variation
  - Generalisation
  - Aggregation
  - Summarisation
  - Coreference
  - Inference via coreference

Paraphrases and the phrasal lexicon

- Becker75:
  - phrasal patterns of a varying degree of variability
  - from completely fixed ('canned')
  - to purely compositional

Paraphrases and the phrasal lexicon

- canned phrases:

Durch die anwesenden Kriminalbeamten wurde versichert, dass die oben genannte Leiche zur Sektion vorliegt.

{By the involved criminal officers was confirmed, that the above named corpse is intended for autopsy.}

Paraphrases and the phrasal lexicon

- patterns with variables:

  Der Tod ist durch X eingetreten.

  {The death has been by X caused.}

X is a nominal phrase of nearly unbounded complexity

Paraphrases and the phrasal lexicon

- patterns with variables: possible fillers for X
  - Mehrfachverletzungen {multiple injuries}
  - einen Schädelbasisbruch {a head base fraction}
  - offene Schädelhirnverletzung infolge Kopfdurchschuss {open head brain injury caused by head shooting}
  - zentrales Herz- und Kreislaufversagen infolge Schädelbruch und Blutung über die harte Hirnhaut mit Druckerhöhung im Schädelinneren {central heart and circulation failure due to head fraction and bleeding above the hard brain skin with pressure increase in the interior of the head}

Paraphrases and the phrasal lexicon

- patterns for arguments:
  - emphasis:
    - Auffällig war auch X {Unusual was also X}
    - Erwähnenswert ist auch X {Worth mentioning is also X}
  - cause-effect:
    - X lässt/lassen sich zwanglos auf Y zurückführen.

    {X can be forcelessly from Y inferred.}

    [Diese Verletzungen] lassen sich zwanglos auf [eine Kollision der rechten Körperseite mit Fahrzeugteilen] zurückführen.

    {[These injuries] can be forcelessly

    from [a collision of the right body side with car parts] inferred.}

Intra-Document Paraphrases

- aggregation: diagnosis
  - discussion: Hirnoedem {brain edema}
  - findings:
    - Hirnwindungen abgeflacht (gyri flattened),
    - Furchen verstrichen (sulci narrowed)
    - ...
- aggregation: procedure
  - Im Bereich der Weichteile über der linken und rechten Brustseite fanden sich trotz schichtweiser Präparation keine bandartigen Unterblutungen.
  - {In the area of the soft tissues above the left and right chest side were found despite layered preparation no band-like haematomata.}

Intra-Document Paraphrases

- summarisation:
  - findings:
    ribs 1--11 on the right, ribs 5--12 on the left ...
  - results:
    many ribs

Intra-Document Paraphrases

Coreference

- Die [Schnittflächen des Gehirns] sind feucht und glänzend. Es finden sich einzelne nicht wegwischbare Blutpunkte.
- {The [dissection areas of the brain] are wet and shiny. There are discrete not wipable bloodpoints.}
- Einzelne nicht wegwischbare Blutpunkte auf den [Hirnschnittflächen].
- {Discrete not wipable bloodpoints in the [brain dissection area]}.

Intra-Document Paraphrases

Coreference cont.

- Es fanden sich Blutungen auf den Hirnschnittflächen.
- {There were bleedings in the brain dissection area.}

Intra-Document Paraphrases

- Im Pkw befand sich eine männliche Person auf dem Beifahrersitz [...]

  {In the car was situated a male person on the passenger seat [...]}
- Die Position von [name of the deceased] auf dem Beifahrersitz [...]
  {The position of [name of the deceased] on the passenger seat [...]}
- note: `In ... befand sich X auf Y´

  vs. `Die Position von X auf Y´

Exploiting cooccurrence data

- Harte Hirnhaut grauweiss.

  {Hard brain skin greywhite.}
- <Concept><Property>
- all property values:
  Harte Hirnhaut: glaenzend, grauweiss, perlmuttergrau,weisslich-gelblich-verfaerbt, intakt, grauroetlich, blaeulich-durchscheinend}
- one very generic property `intakt' (engl. `intact'), all others are characterising the visual appearance

Concept grouping

- `spiegelnd':
  `Herzueberzug', `Lungenueberzug', ...
- `unversehrt':
  `Haut des Rueckens', `Stirnhaut', ...
- `frei':
  `Gehoergange',`Ausfuehrungsgang‘,
  `Kehlkopfeingang', ...

Query Interface for Forensic Research

- similarity of documents?
  - similar findings?
  - similar history?
  - similar discussion?
- tools for efficient `browsing´
- statistical evaluations

Planned work

- representative corpus
- adapted document schema
- broad scale corpus studies
- resource expansion
- semantic representation for subdocuments
- query interface
- evaluation with users

Possible feedback

- project work will probably change authoring process of autopsy protocols
  - XML based creation of protocols
  - improved spelling correction
  - collaboration may further enforce standardisation within forensic medicine

Acknowledgements

We have to thank our cooperation partners from the Institute of Forensic Medicine of our university for their documents and their patience in explaining their contents to medical and forensic ignorants.

Finally

- Thank you
- Aligatoh gozaimas.
- Danke schön!