Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An application of Data-Oriented Computional Linguistics

Tylman Ule

and

Jorn Veenstra

{ule,veenstra}@sfs.uni-tuebingen.de

December 3, 2002

& January 14, 2003

One day

I was asked by another team at Tübingen University to solve their real-world NLP problem

- Prometheus
  `www.prometheus.uni-tuebingen.de`
- BMBF-funded: *Neue Medien in der Bildung*

One day

I was asked by another team at Tübingen University to solve their real-world NLP problem

- Prometheus
  `www.prometheus.uni-tuebingen.de`
- BMBF-funded: *Neue Medien in der Bildung*
- their goal:
  training prospective doctors,
  simulating a hospital

One day

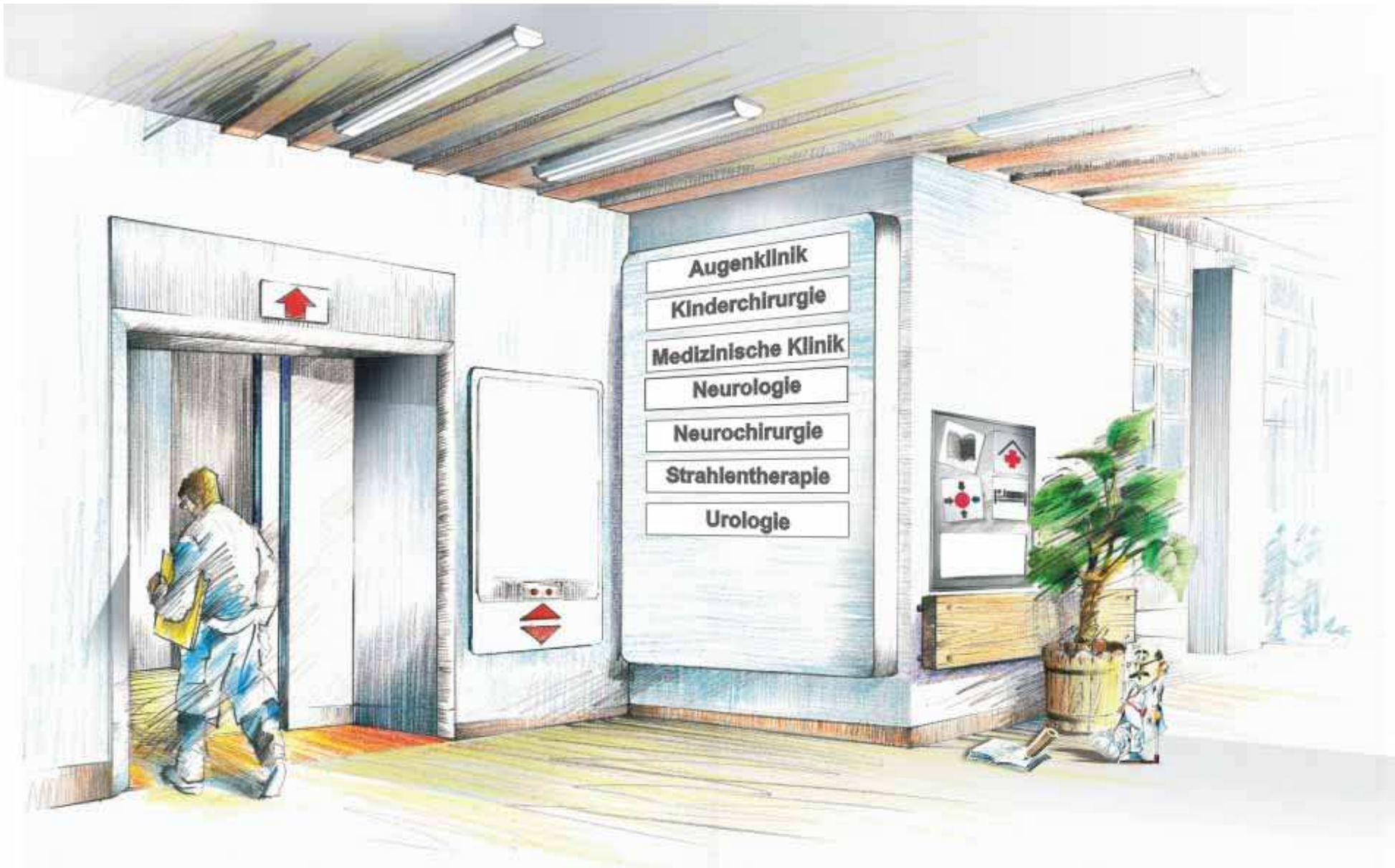
I was asked by another team at Tübingen University to solve their real-world NLP problem

- Prometheus
  `www.prometheus.uni-tuebingen.de`
- BMBF-funded: *Neue Medien in der Bildung*
- their goal:
  training prospective doctors,
  simulating a hospital
- our goal:
  see what we can achieve with our tools and (linguistic) knowledge

The Hospital: Entrance

The Hospital: Information Desk

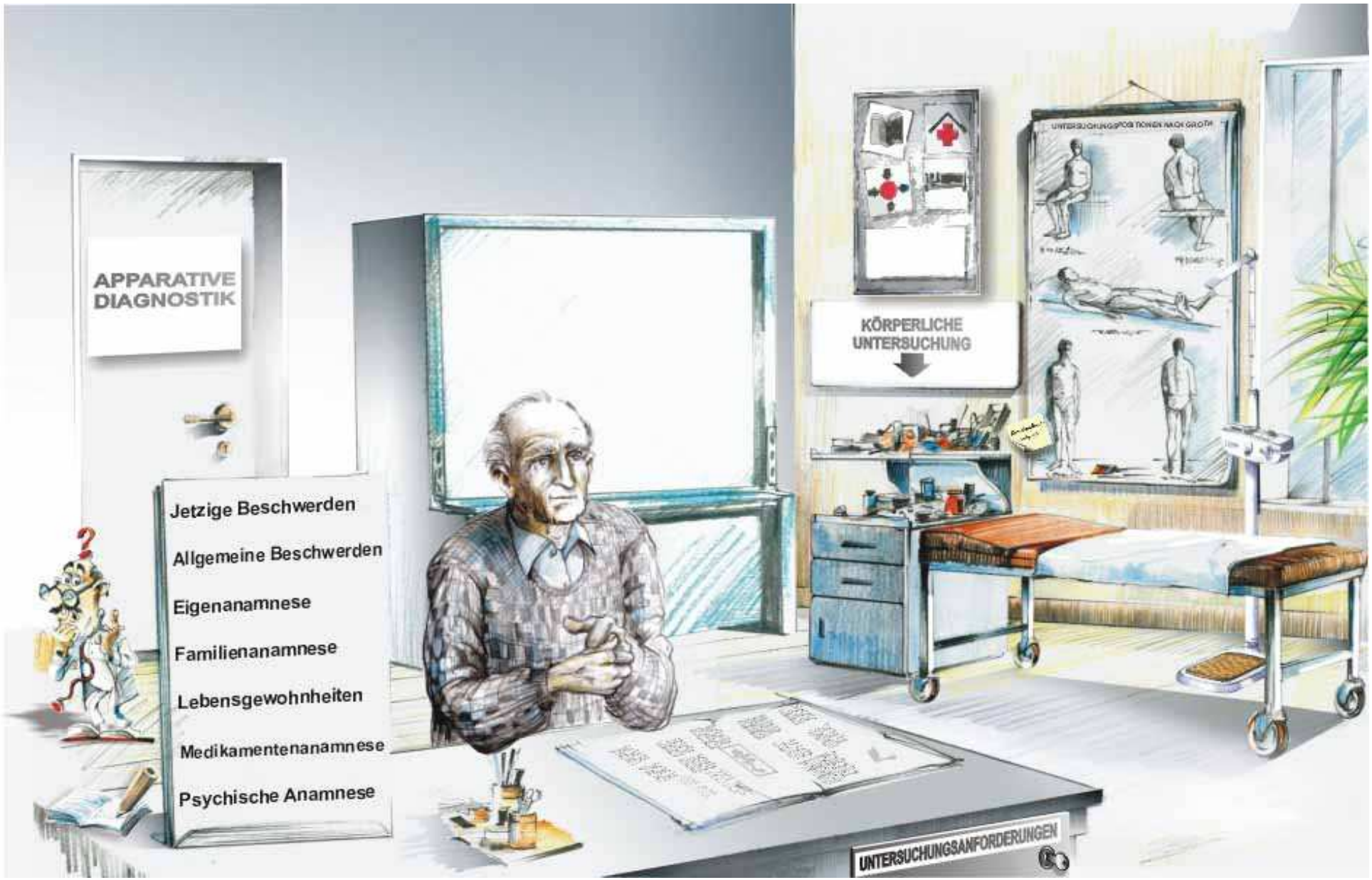
The Hospital: Elevator



The Hospital: Neurology

The Hospital: The Patients

The Hospital: Patient



The Hospital: Other Examinations

The Hospital: Other Examinations

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Hospital: Interview

The Hospital: Anamnesis

Dialogue between doctor and patient in a hospital

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Hospital: Anamnesis

Dialogue between doctor and patient in a hospital

- Hello, Mr. White. How are you?
- I have this strange feeling in the chest for a few weeks now. My general practitioner sent me here.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Hospital: Anamnesis

Dialogue between doctor and patient in a hospital

- Hello, Mr. White. How are you?
- I have this strange feeling in the chest for a few weeks now. My general practitioner sent me here.
- Can you tell me where exactly it hurts?
- It hurts right in the middle of my chest. It is always the same area where I feel the pain.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Hospital: Anamnesis

Dialogue between doctor and patient in a hospital

- Hello, Mr. White. How are you?
- I have this strange feeling in the chest for a few weeks now. My general practitioner sent me here.
- Can you tell me where exactly it hurts?
- It hurts right in the middle of my chest. It is always the same area where I feel the pain.
- Is it a pungent pain, or rather a constant pain?
- The pain usually stays for several ten minutes, and then goes away. It is quite constant.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Outline of the talk

Our goal is two-fold, main theme is
NLP serving a purpose

- How to reach our goal

  Given a certain task, what linguistic and statistical methods should we employ

- What do the numbers mean

  Do 70% mean that the application is sufficiently implemented?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Application

Setting: Teaching medicine students

- e-learning
- only part of it is anamnesis, i.e.
  a doctor is querying a patient about his problems
- setting should be as natural as possible

our problem should be invisible

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Anamnesis: How it should look like

- Wo tut's denn sonst noch weh?
- Ansonsten fühle ich mich auch seit einiger Zeit recht schlapp.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Anamnesis: How it should look like

- Wo tut's denn sonst noch weh?
- Ansonsten fühle ich mich auch seit einiger Zeit recht schlapp.
- Sind Sie denn so schlapp, weil Sie zu wenig schlafen?
- Ich schlafe eigentlich so viel wie immer, und trotzdem bin ich in letzter Zeit immer müde.

Anamnesis: How it should look like

- Wo tut's denn sonst noch weh?
- Ansonsten fühle ich mich auch seit einiger Zeit recht schlapp.
- Sind Sie denn so schlapp, weil Sie zu wenig schlafen?
- Ich schlafe eigentlich so viel wie immer, und trotzdem bin ich in letzter Zeit immer müde.
- Hatten Sie in letzter Zeit auch Kreislaufprobleme?
- Immer, wenn ich schnell aufstehe, wird mir schwindelig.

Anamnesis: what makes it feasible

How do you teach a computer to talk like this?

Short answer:

We have a database

- with sample questions
- and the virtual patients' answers

Longer answer . . .

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

- Are you tired because you don't sleep enough?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

- Are you tired because you don't sleep enough? Is your job very tiring?

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

- Are you tired because you don't sleep enough?
  Is your job very tiring?
  Are you chronically short of sleep?

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

- Are you tired because you don't sleep enough?
  Is your job very tiring?
  Are you chronically short of sleep?
  Are you tired more often now than in the past?

Anamnesis: what makes it feasible

We have several sample questions that would trigger the same answer from the patient:

- Are you tired because you don't sleep enough?
  Is your job very tiring?
  Are you chronically short of sleep?
  Are you tired more often now than in the past?

- I sleep as much as I always did, but still I am tired all the time lately.

Anamnesis: what makes it feasible

All sample questions are labelled by their topic, e.g.

- Are you tired because you don't sleep enough? → fatigue / tiredness
- Does the pain stop when you lie down? → improvement / decline

Anamnesis: what makes it feasible

All sample questions are labelled by their topic, e.g.

- Are you tired because you don't sleep enough? → fatigue / tiredness
- Does the pain stop when you lie down? → improvement / decline

This topic links to a single answer from the patient:

- fatigue / tiredness → I sleep as much as I always did, but still I am tired all the time lately.
- improvement / decline → The pain is less strong when I take a rest and lie down.

Anamnesis: what makes it feasible

We have:

- topics
  - define all different areas that any input question may possibly target
  - contain (topics | sample questions)

Anamnesis: what makes it feasible

We have:

- topics
  - define all different areas that any input question may possibly target
  - contain (topics | sample questions)
- some of the topics are leaf topics
  - contain no other topics, only sample questions

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

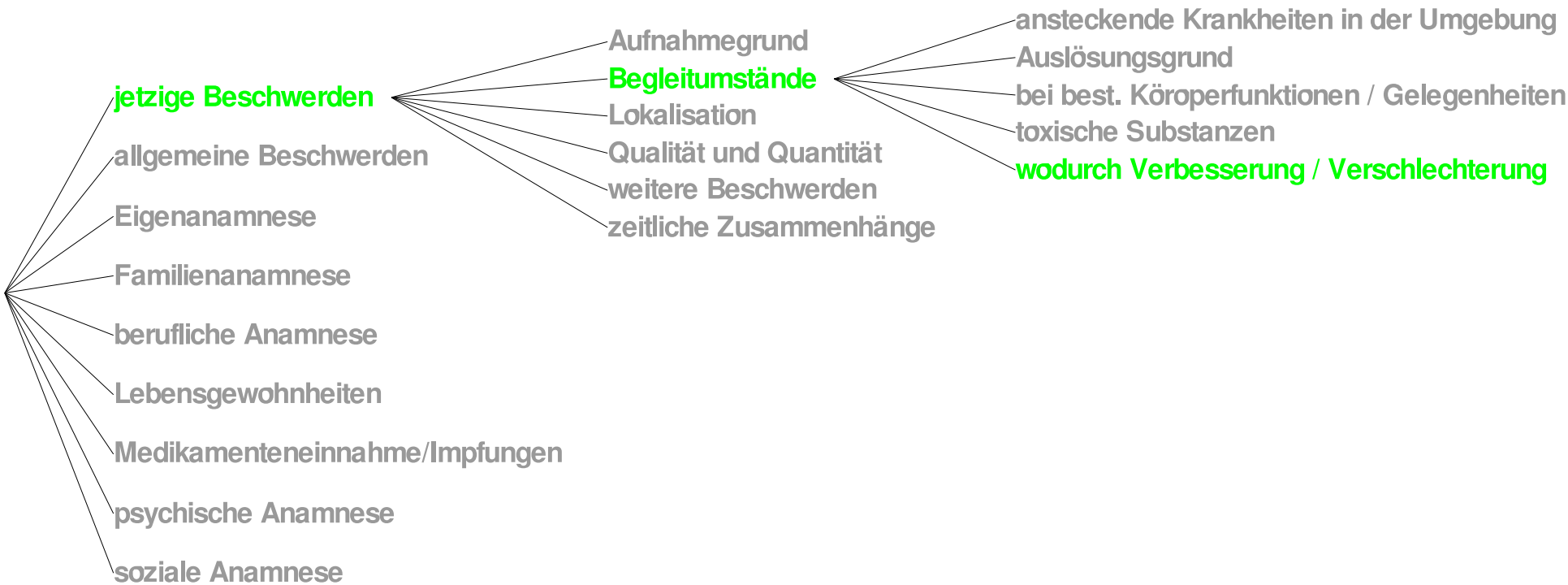
Anamnesis: what makes it feasible

We have:

- topics
  - define all different areas that any input question may possibly target
  - contain (topics | sample questions)
- some of the topics are leaf topics
  - contain no other topics, only sample questions
- sample questions
  - ideally all questions that a doctor in training may ever ask

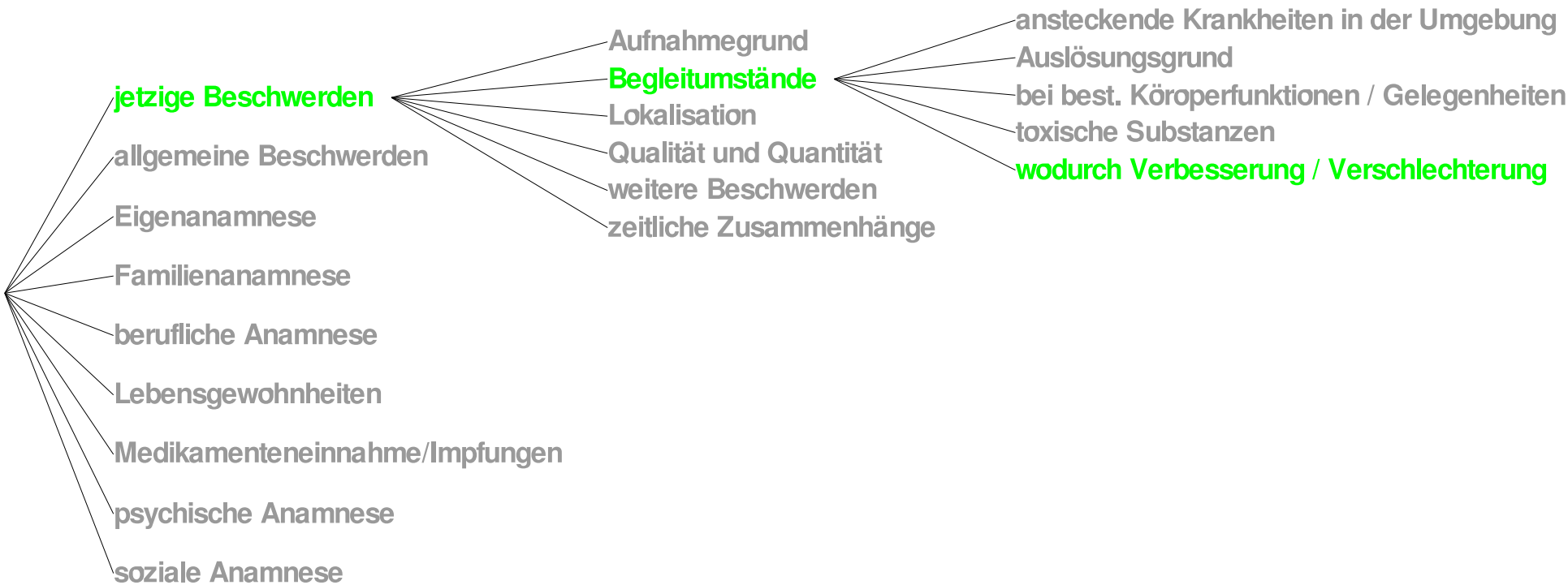
Anamnesis: what makes it feasible

The topics are arranged in tree order:



Anamnesis: what makes it feasible

The topics are arranged in tree order:

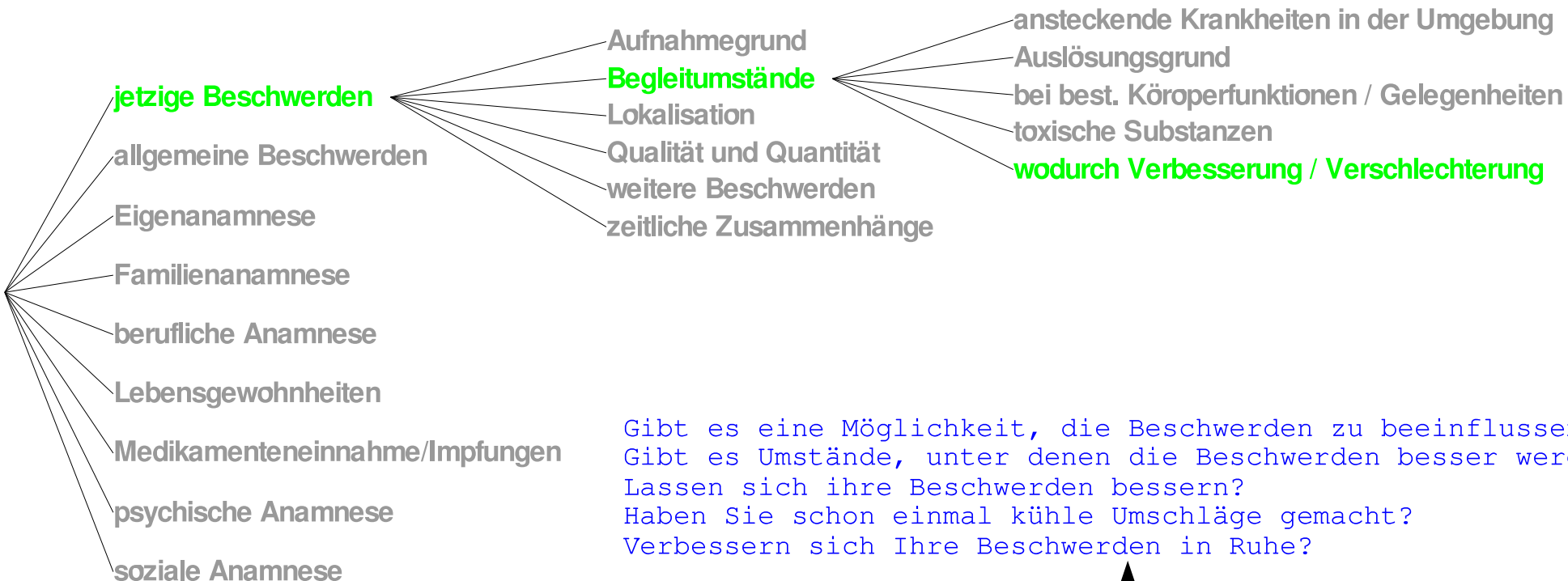


`Sind die Schmerzen besser, wenn Sie ruhen?`

- find sample question most similar to input question

Anamnesis: what makes it feasible

The topics are arranged in tree order:

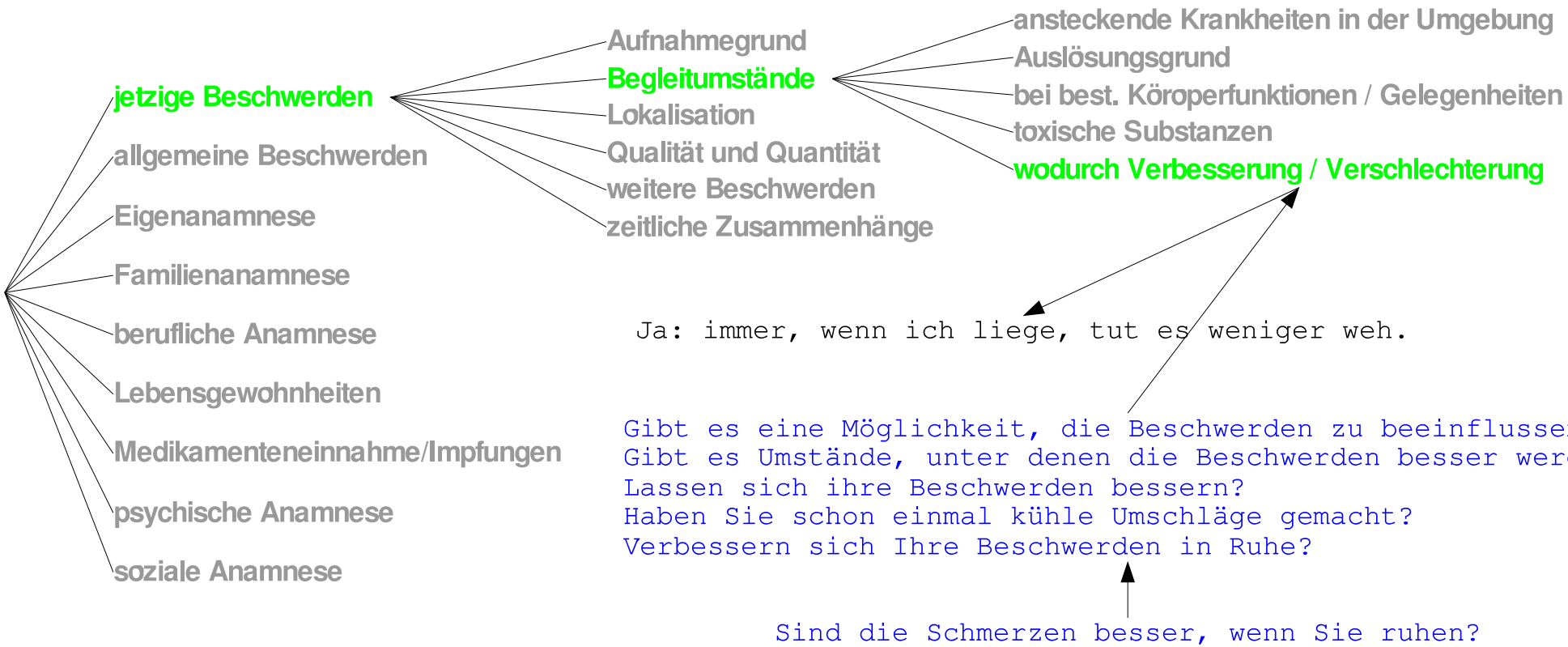



```
Gibt es eine Möglichkeit, die Beschwerden zu beeinflussen?
Gibt es Umstände, unter denen die Beschwerden besser werden
Lassen sich ihre Beschwerden bessern?
Haben Sie schon einmal kühle Umschläge gemacht?
Verbessern sich Ihre Beschwerden in Ruhe?
                          ↑
Sind die Schmerzen besser, wenn Sie ruhen?
```

- find sample question most similar to input question

Anamnesis: what makes it feasible

The topics are arranged in tree order:



- find sample question most similar to input question
- look up the patient's record for his answer

The Data: Some Numbers

A tree structure, i.e.

- 90 topics
- 77 leaf topics
- 3126 sample questions

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The Data: Some Numbers

A tree structure, i.e.

- 90 topics
- 77 leaf topics
- 3126 sample questions

- at least 30 questions per category
- on average 40 questions per leaf topic
- on average 105 questions per topic

The Data: Some Numbers

A tree structure, i.e.

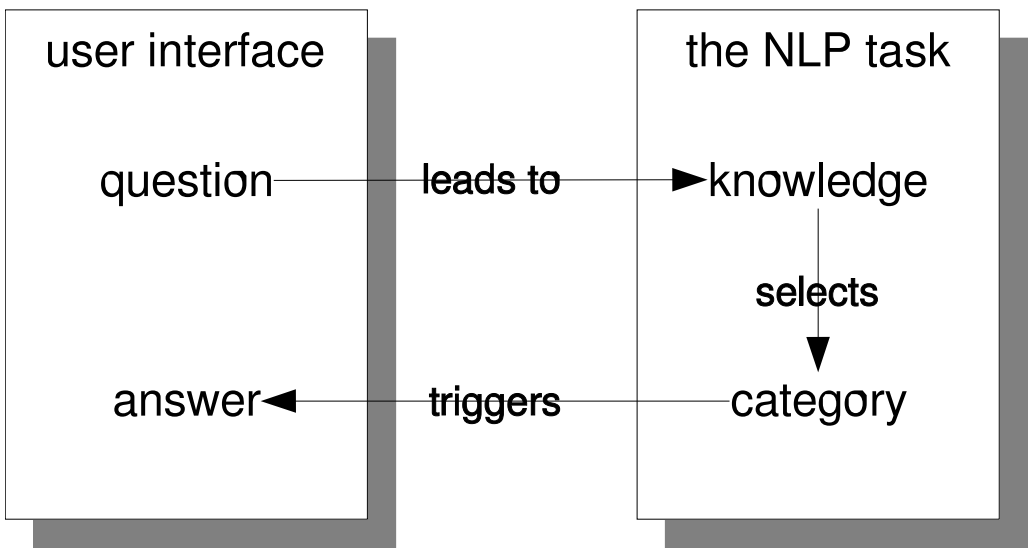
- 90 topics
- 77 leaf topics
- 3126 sample questions

- at least 30 questions per category
- on average 40 questions per leaf topic
- on average 105 questions per topic

\+ for each patient a collection of answers

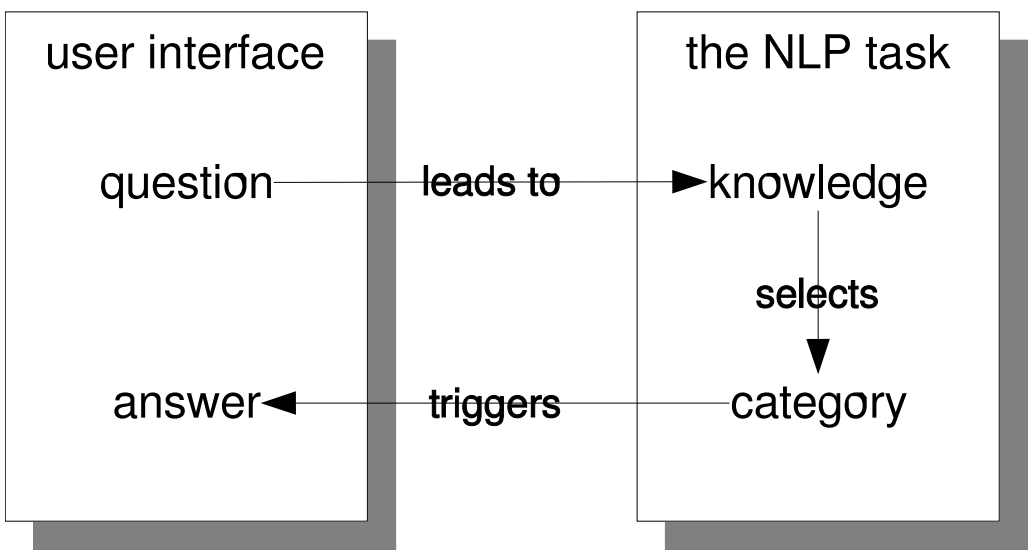
Anamnesis: Definition of the NLP task

- given
  - question: specified by user
  - 1-to-1 mapping: category $\longrightarrow$ answer



Anamnesis: Definition of the NLP task

- given
  - question: specified by user
  - 1-to-1 mapping: category $\longrightarrow$ answer



- our problem
  - question $\longrightarrow$ knowledge: *feature generation*
  - knowledge $\longrightarrow$ category: *classification*

Classification and features

- Confronted with a real-life NLP task we usually can see the task as a classification task: a certain input has got to be mapped to a certain output.
- In our case: an open set of questions has to be mapped onto one question which is mapped deterministically onto an answer. This is that *classification* task. The *features* are the bits of information that are used to perform this task·
- So we have the stage of feature generation, which can be done linguistically and statistally; and the stage of feature usage, which we call classification.

What are features?

1. Have you been in coma lately?
   (a) Question
   (b) “you” is subject.
   (c) “in coma” is prepositional object.
   (d) “coma” is an important word.
2. Did you have carcinome in your lungs?
   (a) “carcinome” is cancer tissue.
   (b) “lungs” is “lung” in plural.

Classification and features

1. Feature generation.
2. Feature Usage (Classification).
3. Evaluation and Feature selection.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Feature generation

Extract information from the data that can be used for classification. We will discuss:

1. Linguistic feature generation:
   (a) Morphological analysis.
   (b) Shallow parsing.
   (c) Semantic knowledge as Germanet.
2. Statistical feature generation (strings and counting):
   (a) String matching techniques for morphology.
   (b) Automatic keyword extraction.

Classification

Once we have the features we need a method to do the actual mapping from input (free questions) to output (given question and their coupled answers). We could do this in several ways, e.g.:

1. Rule-based:
   (a) Decision tree.
   (b) Finite-state automaton.
   (c) Harvard Virtual Patient
2. Machine Learning (statistically):
   (a) Memory-based learning.
   (b) Support vector machines.

Feature selection

Once we have chosen the method for classification we can consider which features are relevant, and which features can be left out. This is the topic of next week.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Break

Those who smoke can do so now. Those who don't can just have a break. After the break we will go deeper into feature generation.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The remaining talk

- feature generation
  - linguistic analysis
    - robust partial parsing
    - morpho-syntax, morphology, lemma
    - semantic knowledge
  - counting and string operations
    - $tf * idf$
    - m_1, m_2, m_3
    - edit distance
    - “morphstrings”
- classifier setup, feature selection $\longrightarrow$ next time

Feature Generation

We need features that allow our patient to react

- accurately
- quickly

Let's go through feature generation by example:

Zeigt der Fieberverlauf starke Schwankungen?

(*Are there considerable fluctuations in the fever curve?*)

Allgemeine Beschwerden – Fieber/Schüttelfrost

Linguistic Features: Robust Partial Analysis

```
[VCLVF
     .VVFIN    Zeigt ]
{MF
     [NC
          .ART      der
          .NN       Fieberverlauf ]
     [NC
          [AJAC
               .ADJA     starke ]
          .NN       Schwankungen ] }
.\$.        ?
```

- lexical verb: zeigen
- NC with head nouns: Schwankung

Linguistic Features: Robust Partial Analysis

Steigt die Temperatur rasch an?

- lexical verb
- NC with head nouns
- separable verb prefix: an#steigen
- adjective/adverb phrase: rasch

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linguistic Features: Robust Partial Analysis

Steigt die Temperatur rasch an?

- lexical verb
- NC with head nouns
- separable verb prefix: an#steigen
- adjective/adverb phrase: rasch
- PC with preposition/head noun: unter

Leiden Sie unter Fieberschüben?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linguistic Features: Robust Partial Analysis

Steigt die Temperatur rasch an?

- lexical verb
- NC with head nouns
- separable verb prefix: an#steigen
- adjective/adverb phrase: rasch
- PC with preposition/head noun: unter
- wh pronoun: Wann

Leiden Sie unter Fieberschüben?
Wann findet der Fieberanstieg statt?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linguistic Features: Robust Partial Analysis

Steigt die Temperatur rasch an?

- lexical verb
- NC with head nouns
- separable verb prefix: an#steigen
- adjective/adverb phrase: rasch
- PC with preposition/head noun: unter
- wh pronoun: Wann
- sentence structure

Leiden Sie unter Fieberschüben?
Wann findet der Fieberanstieg statt?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Linguistic Features: Lemma

- Temperatur, rasch, unter
  $\rightarrow$ Temperatur, rasch, unter

- Wann $\rightarrow$ *wann
- Zeigt $\rightarrow$ *zeigen
  findet $\rightarrow$ finden
- Schwankungen $\rightarrow$ Schwankung
- Fieberanstieg $\rightarrow$ $zero$

normalise all to lowercase

Linguistic Features

Steigt die Temperatur rasch an?
Wann findet der Fieberanstieg statt?
Leiden Sie unter Fieberschüben?
Zeigt der Fieberverlauf starke Schwankungen?

We have, for each clause

- wh pronoun
- adverb
- head noun
- preposition
- head noun of PC
- seperable verb prefix
- lexical verb

Linguistic Features

Steigt die Temperatur rasch an?
Wann findet der Fieberanstieg statt?
Leiden Sie unter Fieberschüben?
Zeigt der Fieberverlauf starke Schwankungen?

We have, for each clause

```
-,rasch,temperatur,-,-,an,steigen
wann,-,fieberanstieg,-,-,statt,finden
-,-,unter,fieberschub,-,leiden
-,-,fieberverlauf,-,-,-,zeigen
-,-,schwankung,-,-,-,zeigen
```

Linguistic Features: Semantics

GermaNet, of course

```
fever, febricity, pyrexia, feverishness
-- (a rise in the temperature of the body; frequently a symptom of infection)
       => symptom
          -- ((medical) any sensation or change in bodily function that is
              experienced by a patient and is associated with a particular
              disease)
           => evidence, grounds
              -- (your basis for belief or disbelief; knowledge on which to
                  base belief; "the evidence that smoking causes lung cancer
                  is very compelling")
               => information
                  -- (knowledge acquired through study or experience or
                      instruction)
                   => cognition, knowledge, noesis
                      -- (the psychological result of perception and
                          learning and reasoning)
                       => psychological feature
                          -- (a feature of the mental life of a living
                              organism)
```

Linguistic Features: Semantics

```
fever, febricity, pyrexia, feverishness
-- (a rise in the temperature of the body; frequently a symptom of infection)
       => symptom
          -- ((medical) any sensation or change in bodily function that is
              experienced by a patient and is associated with a particular
              disease)
           => evidence, grounds
              -- (your basis for belief or disbelief; knowledge on which to
                  base belief; "the evidence that smoking causes lung cancer
                  is very compelling")
```

- Fieberschub? Fieberanstieg?
- rheumatisches Fieber? Pfeiffersches Drüsenfieber?
- we focus on subpart of medical domain
- even distribution of distances?

International Classification of Diseases (ICD-10)

- specialised
- free of charge

Kapitel XX

Äußere Ursachen von Morbidität und Mortalität (V01-Y98)

Folgeerscheinungen äußerer Ursachen von Morbidität und Mortalität sind in den Schlüsselnummern Y85-Y89 enthalten.

Dieses Kapitel gliedert sich in folgende Gruppen:

V01-X59	Unfälle	
	V01-V99	Transportmittelunfälle
		V01-V09 Fußgänger bei Transportmittelunfall verletzt
		V10-V19 Benutzer eines Fahrrades bei Transportmittelunfall verletzt
		V20-V29 Benutzer eines Motorrades bei Transportmittelunfall verletzt
		V30-V39 Benutzer eines dreirädrigen Kraftfahrzeuges bei Transportmittelunfall verletzt

Linguistic Features: Semantics, ICD-10

- too detailed
- mapping to our list of topics?
- instead of using the semantic relations directly, use as a source for relevant morphemes

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Automatic feature generation

1. Keyword generation:
   (a) $M3$.
   (b) $tf * idf$.
2. Morphological generation:
   (a) String edit distance.
   (b) Morphstring.

Automatic feature generation: keywords

The doctor wants to know whether the patient has had cancer before, she can phrase this in several ways:

1. Do you suffer from long cancer?
2. Have you suffered from cancer?
3. Did you suffer carcinome?
4. Suffered from similar complaints?

How can we find the relevant keywords from these phrases?

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Automatic feature generation: keywords II

How can we find the relevant keywords from these phrases?

1. Look for words that occur more with one class than with the other classes.
2. Ng and Lee proposed a method to do this: $M3$.
3. tf*idf is a method from the information retrieval world.

Keyword extraction (Ng and Lee 1996)

$M1$: The word occurs in more than $M1$ of the cases with one class.

$M2$: The word occurs at least $M2$ times in the training set.

$M3$: Only the $M3$ most frequently occurring keywords are extracted.

tf*idf

tf*idf is a way to find words that are characteristic in a certain context. it stands for term frequency * inverse document frequency: you take the frequency of a term in a certain document and divide that by the frequency over the total data.

tf: Count the number of occurrences of a term in the relevant data, in our case this is the number of occurrences of a term in a certain class.

idf: Count the number of occurrences of a term in the total data and invert this (power -1).

$tf * idf$: Multiply tf*idf. This gives a list of keywords per class.

Edit distance

A way to automatically determine the similarity between words. Two words that have the same stem tend to have a short edit distance.

- spelling errors: long cancer
- spelling variants: behavio*ur*
- morphological variance: suffer, suffered

Edit distance

But . . .

- dissimilar but small distance:
  ablösung, lösung
- compounds:
  schimmelpilzallergie, allergie, schimmelpilz

MorphString

Generate new word forms from a list of words:

1. Make a list of all words in the list.
2. Look for words which contain another word in this list.
3. Suppose the rest of a contained word is a suffix or affix.
4. Make a list of these suffixes and affixes and go through the list of words again, looking for words which contain these.
5. repeat this procedure until no more words are found.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

MorphString

Original:

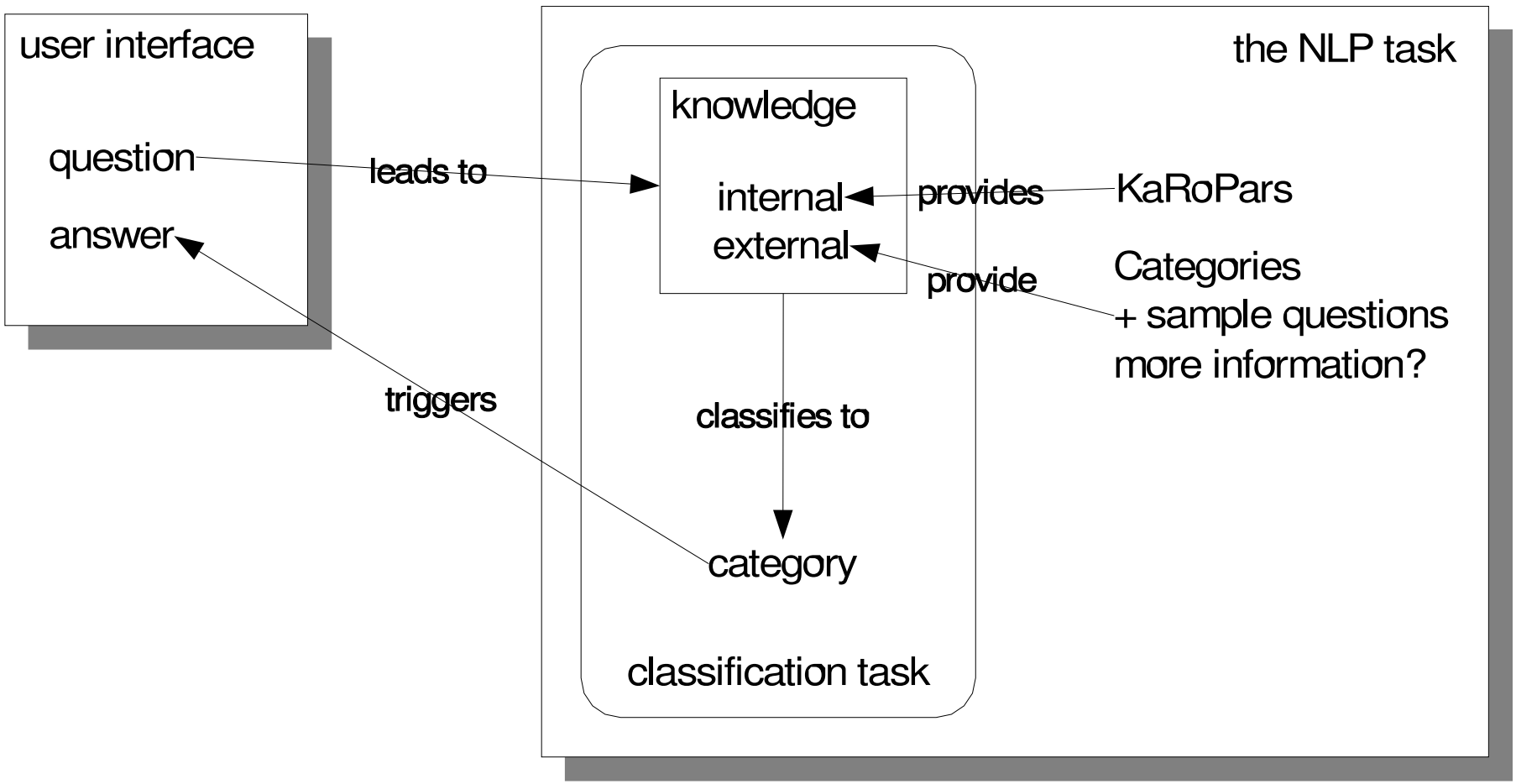
```
abführmittel
abführmitteln
mittel
mitteln
```

Yield:

```
abführ * -mittel(0) -mitteln(0)
adenom * -ektomie(1) pankreas-(0)
         prostata-(0) hypophysen-(1)
         physen-(4)
allergie \dots nickel-(1) schimmelpilz-(1)
```

Conclusion

- feature generation



next time: classifier setup & feature selection

Conclusion

- linguistic tools are just a means to generate features
- the usefulness of these features has still to be assesed → next time

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN