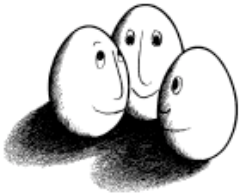
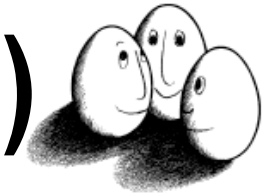


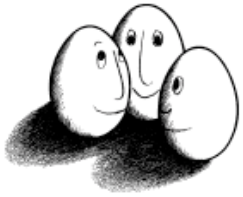
Information Extraction

Information Extraction (IE)



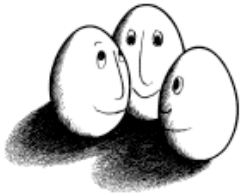
- Extraktion von strukturierten Informationen aus textuellen Dokumenten
- Textart:
  - beliebig, unbekannt
  - natürlichsprachlich
- Suche nach zuvor festgelegten Informationen
- Weiterverarbeitung: direkte Darstellung, Datenbank, Tabellenkalkulation
- **Ziel:** Entwicklung von Systemen, die spezielle Informationen aus freien Texten aufspüren und strukturieren können

Information Retrieval (IR)



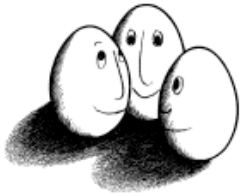
- Dokumente aus Dokumentensammlung auswählen
- Benutzeranfrage durch Stichwörter
- Ergebnis: geordnete Menge relevanter Dokumente

Szenario



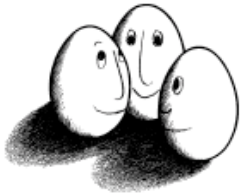
- Benutzer sucht Informationen über Aktienkurse von Firmen mit Besitz in Bolivien
- Daten sollen in einer Tabellenkalkulation weiterverarbeitet werden

IR vs. IE



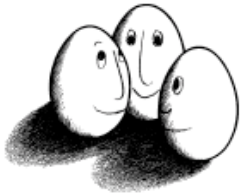
- IR-Systeme finden Texte und präsentieren sie dem Benutzer.
- Vorgehen:
  - Eingabe einer Liste von relevanten Suchbegriffen ins System.
  - Rückgabe: Menge von Dokumenten, die ähnliche Begriffe enthalten.
  - Benutzer entscheidet selbst, welche Texte relevant sind, und übernimmt die Weiterverarbeitung.

IR vs. IE

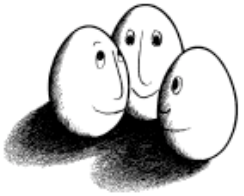


- IE-Systeme analysieren Texte und präsentieren nur die spezifischen Informationen
- Vorgehen:
  - IE-System füllt die Datenbank oder Tabellenkalkulation selbstständig mit den Namen der Firmen und deren Aktienkursen

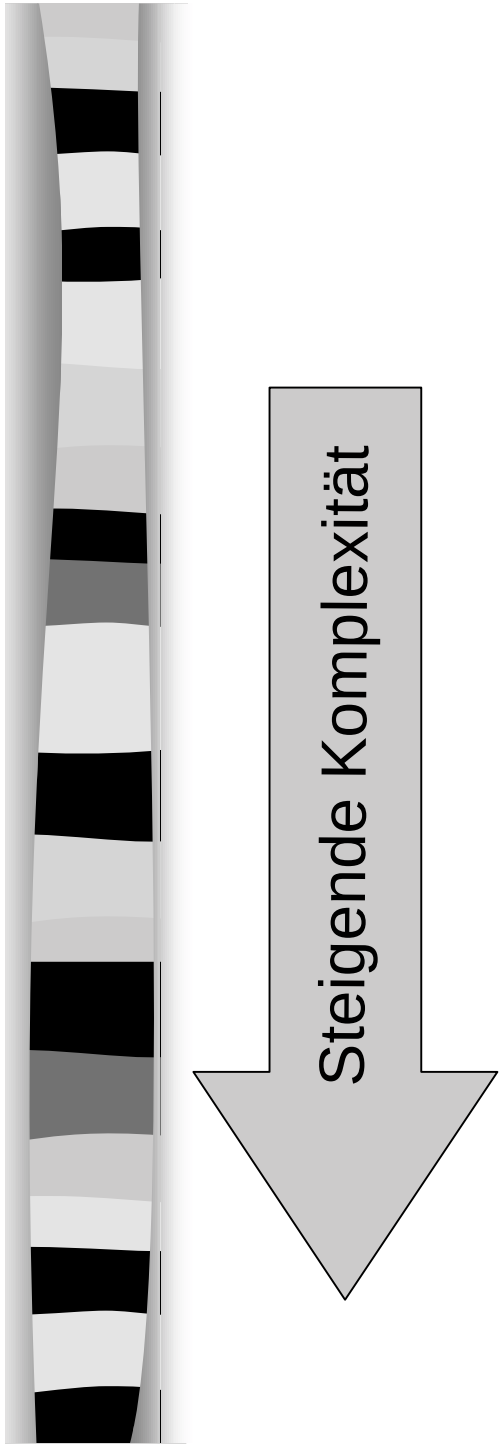
Vor- & Nachteile



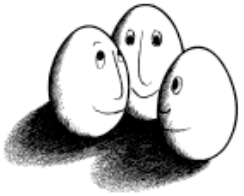
- IE Systeme sind aufwendig und wissensintensiv
- In der Regel muss das System an verschiedene Bereiche und Szenarien angepasst werden
- IE ist rechenintensiver als IR
- Zeitersparnis durch IE, lesen der Ergebnistexte entfällt



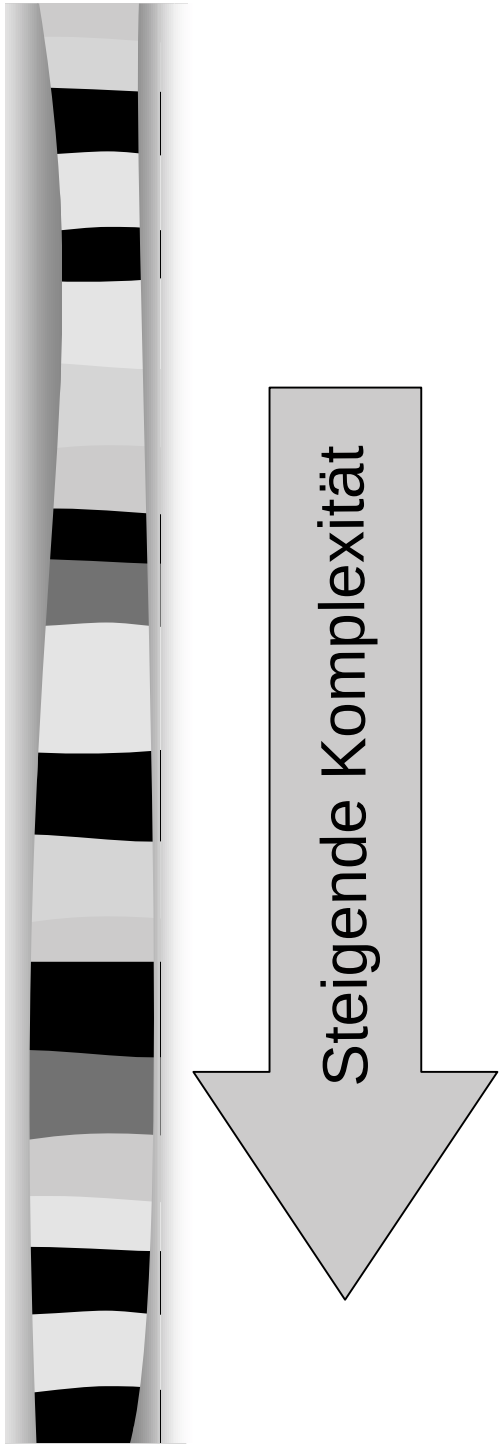
IE Typen



- Named Entity recognition (NE)
  - Finden und Klassifizieren von Namen
  - Heuristik: Name bezieht sich auf etwas Bestimmtes (Einzelding)
- Coreference Resolution (CO)
  - Identifiziert Referenzen zwischen den NE-Objekten
- Template Element construction (TE)
  - Fügt beschreibende Informationen zu NE Ergebnissen hinzu

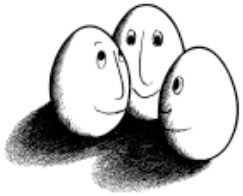


IE Typen



- Template Relation construction (TR)
  - Erkennt Beziehungen zu den NE Ergebnissen
- Senario Template production (ST)
  - Setzt TE und TR Ergebnisse in spezielle Ereignis-Szenario

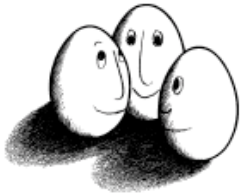
Anwendungsbeispiel



Die glänzend rote Rakete wurde am Dienstag gestartet. Sie ist die Erfindung von Dr. Hans Müller. Dr. Müller ist ein Wissenschaftler bei Raketenwerke Inc.

- NE: z.B.: Rakete, Dienstag, Dr. Müller (Entities)
- CO: „Sie“ bezieht sich auf Rakete

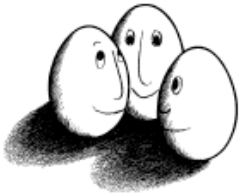
Anwendungsbeispiel



Die glänzend rote Rakete wurde am Dienstag gestartet. Sie ist die Erfindung von Dr. Hans Müller. Dr. Müller ist ein Wissenschaftler bei Raketenwerke Inc.

- TE: Rocket = glänzend rot Erfindung von Dr. Müller
- TR: Müller arbeitet für Raketenwerke Inc.

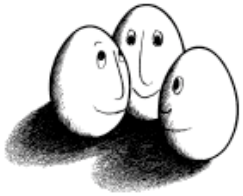
Anwendungsbeispiel



Die glänzend rote Rakete wurde am Dienstag gestartet. Sie ist die Erfindung von Dr. Hans Müller. Dr. Müller ist ein Wissenschaftler bei Raketenwerke Inc.

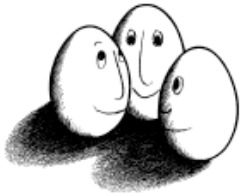
- ST: Raketenstart ist Ereignis, in welches die verschiedenen Entities verwickelt waren

MUC



- **M**essage **U**nderstanding **C**onference (1-7)
- **M**essage **U**nderstanding **C**ompetition
- Gegründet bei DARPA Ende der 80‘iger

Evaluationskriterien für IE

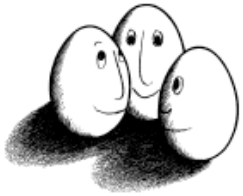


- Evaluation von IE-Systemen auf der MUC
- Präzision P = Anteil der korrekt gewonnenen Wissenseinheiten WE

	relevante Texte	irrelevante Texte
ausgegebene Texte	A	B
nicht ausgegebene Texte	C	D

$$P = \frac{A}{A + B}$$

Evaluationskriterien für IE

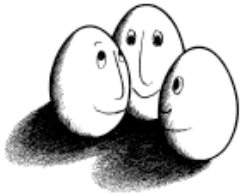


- Pη = fast alle gefundenen WE sind relevant
- Vollständigkeit V = Anteil der korrekt gewonnenen WE gegenüber allen WE

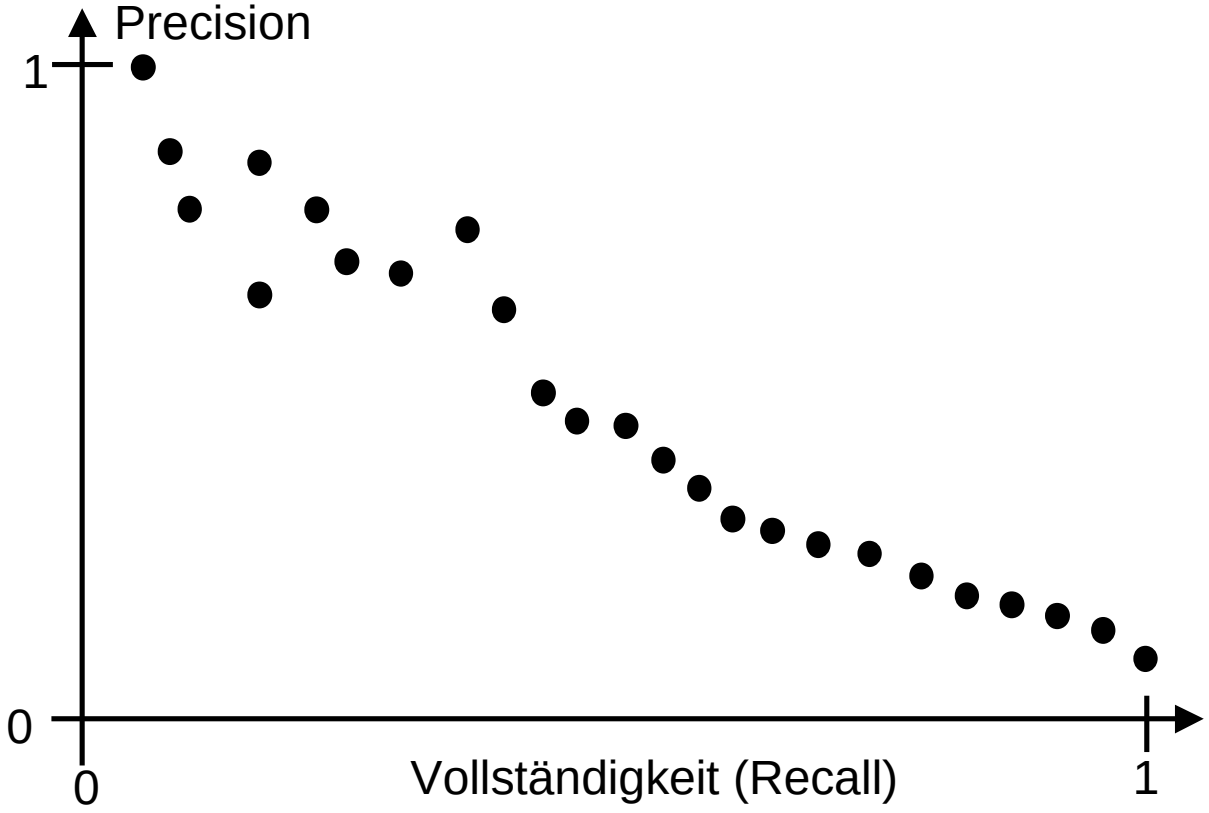
	relevante Texte	irrelevante Texte
ausgegebene Texte	A	B
nicht ausgegebene Texte	C	D

$$V = \frac{A}{A + C}$$

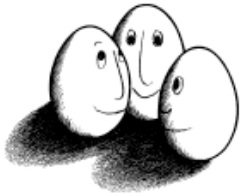
Precision-Recall-Diagramm



- Maße sind gegenläufig ➔ Bei guter Vollständigkeit ergibt sich eine geringe Präzision und umgekehrt



Evaluationskriterien für IE



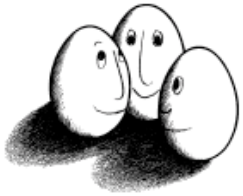
- Optimierung bezüglich hoher Präzision:
  Relevante WE werden möglicherweise nicht erkannt
- Optimierung bezüglich hoher Vollständigkeit:
  Aufnahme von irrelevanten WE
- F-Maß gibt die Güte des IE-Prozesses an

$$F=\frac{(ß^2+1,0)*P*V}{(ß^2*P)+V}$$

In der Regel ß=1

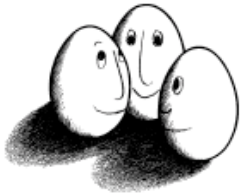
- Abweichung von 1 gibt an ob:
  P oder V stärker gewichtet werden sollte

Dokument Views



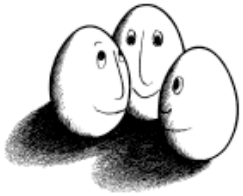
- Terms View
- Mark-Up View
- Layout View
- Typographic View
- Linguistic View

Seminaranmeldung Beispiel



```
<0.21.3.95.14.12.11.ed47+@andrew.cmu.edu.0>
Type:      cmu.andrew.official.cmu-news
Topic:     ECE Seminar
Dates:     30-Mar-95
Time:      4:00 - 5:00 PM
Place:     Scaife Hall Auditorium
PostedBy: Edmund J. Delaney on 21-Mar-95 at 14:12 from andrew.cmu.edu
Abstract:


   COMPUTERIZED TESTING AND SIMULATION OF CONCRETE CONSTRUCTION

                          FARRO F. RADJY, PH.D.

                          President and Founder
                        Digital Site Systems, Inc.
                            Pittsburgh, PA

 DATE:  Thursday, March 30, 1995
 TIME:  4:00 - 5:00 P.M.
 PLACE:  Scaife Hall Auditorium
REFRESHMENTS at 3:45 P.M.

------------------------
```

Terms View

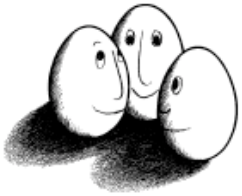


1	1	3	2	2	4	1	1	1	...	...	...	...	term frequency
abstract	and	andrew	at	auditorum	cmu	computerized	concrete	construction	...	...	...	...	

Inverse document frequency:

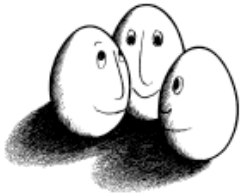
4	n	3	n	2	1	2	1	1	...	...	...	...	n/doc
abstract	and	andrew	at	auditorum	cmu	computerized	concrete	construction	...	...	...	...	

Mark-Up View

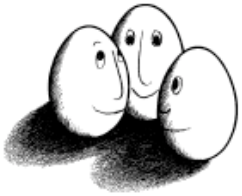


- Betrachtung von Terms und Metaterms
- Metaterms geben „Rolleninformationen“ über verschiedenen Terms

Mark-Up View Beispiel



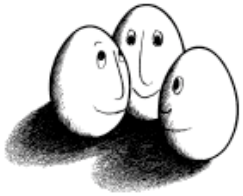
```
<html>
<head>
<title>Dayne Freitag's Home Page</title>
</head>

<body bgcolor="#FFFFFF">

<center><h2>Dayne Freitag</h2>
<hr>
<h3><font face="Helvetica">Contents</font></h3></center>

<center>

<table>

<tr><td>
<font face="Courier">
Introduction.....................................
<a href="intro.html"><i>intro.html</i></a>
</font>
```



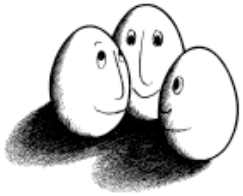
Layout View

- LV betrachtet die 2-dimensional Anordnung und Größe von Terms
- Vorgehen: Alle non-whitespace characters werden durch Sterne (*) ersetzt
- Ziel: Erkennung von wichtigen Text-objekte, z. B. Paragraphen Über-schriften, Mail headers, usw.

Layout View Beispiel

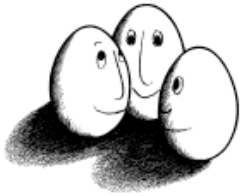


Typographic View



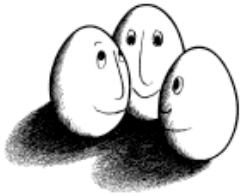
- Zeichen werden in folgende Klassen eingeteilt:
  - Numerische
  - Zeichensetzung
  - Großbuchstaben
- Zahlen ersetzt durch „9“
- Zeichensetzung keine Änderung
- Großbuchstaben ersetzt durch „A“ sonstige Buchstaben durch „a“
- Ziel: Relevante Stellen lokalisieren

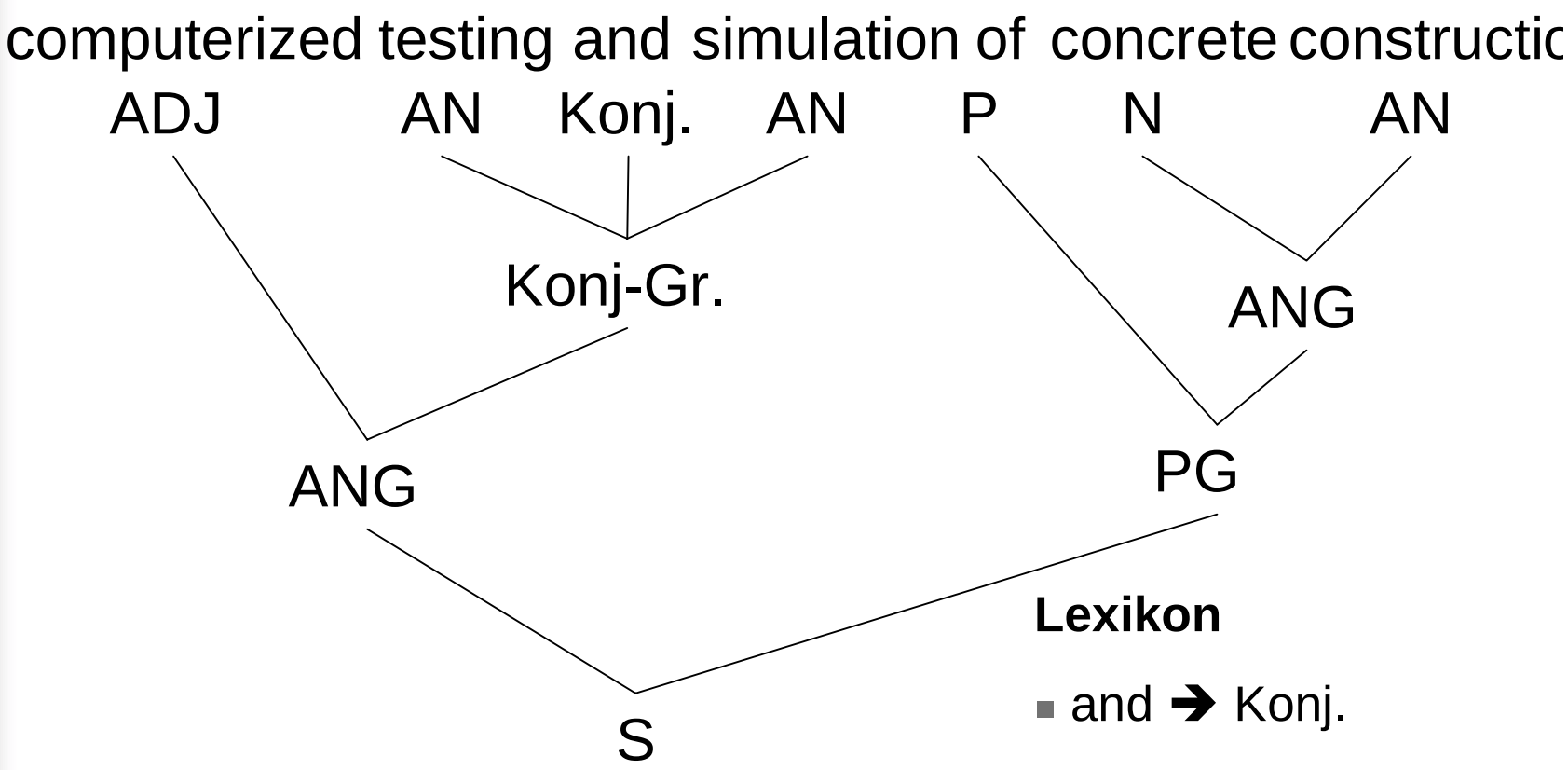
Typographic View Beispiel



```
<9.99.9.99.99.99.99.aa99+@aaaaaa.aaa.aaa.9>
Aaaa:     aaa.aaaaaa.aaaaaaaa.aaa-aaaa
Aaaaa:    AAA Aaaaaaa
Aaaaa:    99-Aaa-99
Aaaa:     9:99 - 9:99 AA
Aaaaa:    Aaaaaa Aaaa Aaaaaaaaaa
AaaaaaAa: Aaaaaa A. Aaaaaaa aa 99-Aaa-99 aa 99:99 aaaa aaaaaa.aaa.aaa
Aaaaaaaa:


   AAAAAAAAAAAA AAAAAAA AAA AAAAAAAAAA AA AAAAAAAA AAAAAAAAAAAA

                          AAAAA A. AAAAA, AA.A.

                          Aaaaaaaaa aaa Aaaaaaa
                        Aaaaaaa Aaaa Aaaaaaa, Aaa.
                            Aaaaaaaaaa, AA

 AAAA:  Aaaaaaaa, Aaaaa 99, 9999
 AAAA:  9:99 - 9:99 A.A.
 AAAAA:  Aaaaaa Aaaa Aaaaaaaaaa
AAAAAAAAAAAA aa 9:99 A.A.
```

Linguistic View Beispiel

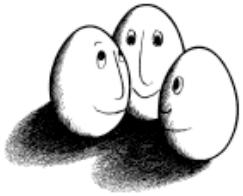




Lexikon

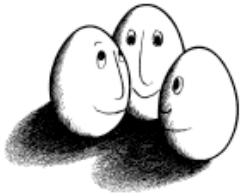
- and ➔ Konj.
- computerized ➔ADJ
- of ➔ P
- testing ➔ AN

Maschinelles Lernverfahren für IE



- Standard IE-Systeme sind sehr domainspezifisch:
  - Anpassung ist sehr aufwendig und muss von Experten durchgeführt werden
- **Ziel:** System, dass auf verschiedenen Domains angewendet werden kann
- **Besser:** Ausgehend von einer Trainingsmenge von bereits mit Ergebnissen versehenen Textdokumenten werden automatisch Regeln zum Füllen von Templates abgeleitet

Maschinelles Lernverfahren für IE Beispiel



- Annotiertes Trainingsbeispiel:

<PNG>Sue Smith</PNG>, 39, of Menlo Park, was appointed

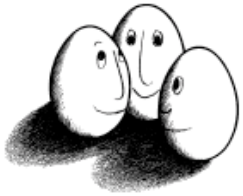
<TNG> president </TNG> of <CNG> Foo Inc. </CNG>

- Abgeleitete Template-Regel:

  Noun-group(PNG, head(isa(person))), noun-group(TNG, head(isa(title))), noun-group(CNG, head(isa(company))), prep, head(of OR at OR by)), verb-group(VG, type(passiv), head(named OR elected OR appointed)), subject(PNG,VG), object(VG,TNG),post-nominal-prep(TNG,PREP), prep-obj(PREP,CNG)

  ➔management-appointment(person(PNG), title(TNG), company(CNG))

Maschinelles Lernverfahren für IE Beispiel



NG

PNG	TNG	CNG	Prep	VG
is a person	is a title	is a company	is a title	type: passiv
subject	object	Prep. object	Object	{named, elected appointed}
				Object

appointment
Person: PNG

Title: TNG

Company: CNG

Grammatik

S➔NG,VG

VG ➔ V, TNG

TNG ➔ T, Prep

Prep ➔ Prep, CNG

Lexikon

... Is a person