Projektgruppe

IDSE

Information-Driven Software Engineering

Henning Wachsmuth

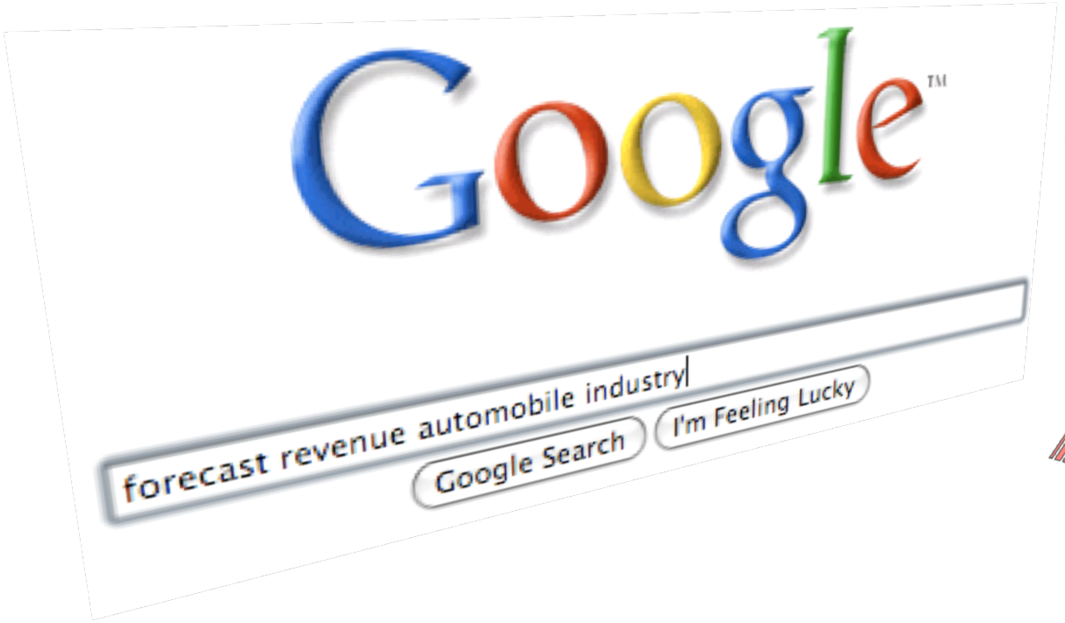
Information Extraction
An Incomplete Overview

12. Mai 2010

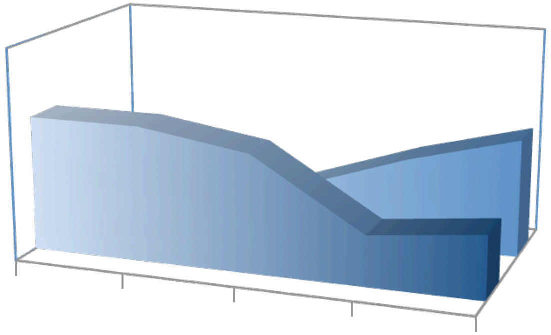
Einführungsvorträge

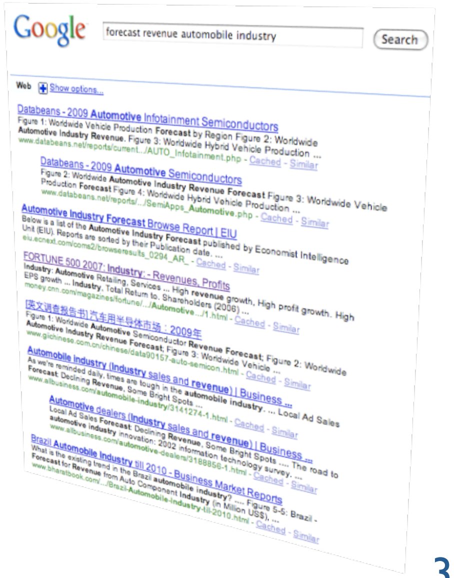
- **Verfassen von Seminarvortrag und –paper**
  Prof. Dr. Gregor Engels, Donnerstag 15.4., 16h-18h
- **Interactive Knowledge – Semantische CMS-Systeme**
  Fabian Christ, Donnerstag 29.4., 16h-18h
- **Maschinelles Lernen**
  Dr. Theodor Lettmann, Freitag 7.5., 15h-17h
- **Information Extraction**
  Henning Wachsmuth, Mittwoch 12.5., 16h-18h
- **Aktuelle Probleme des Software Engineering**
  Benjamin Nagel, **Donnerstag 20.5., 16h-18h**
- **Plagiatanalyse**
  Prof. Dr. Benno Stein, Bauhaus-Universität Weimar, **Freitag 28.5., 14h-16h**
- **Projektmanagement**
  Stefan Sauer, Geschäftsführer s-lab

Motivation: Automatic market forecasting



forecast revenue automobile industry





forecast revenue automobile industry

Databeans - 2009 Automotive Infotainment Semiconductors
Figure 1: Worldwide Vehicle Production Forecast by Region Figure 2: Worldwide Automotive Industry Revenue. Figure 3: Worldwide Hybrid Vehicle Production ...
www.databeans.net/reports/current.../AUTO_Infotainment.php - Cached - Similar

Databeans - 2009 Automotive Semiconductors
Figure 2: Worldwide Automotive Industry Revenue Forecast Figure 3: Worldwide Vehicle Production Forecast Figure 4: Worldwide Hybrid Vehicle Production ...
www.databeans.net/reports/.../SemiApps_Automotive.php - Cached - Similar

Automotive Industry Forecast Browse Report | EIU
Below is a list of the Automotive Industry Forecast published by Economist Intelligence Unit (EIU). Reports are sorted by their Publication date. ...
eiu.ecnext.com/coms2/browseresults_0294_AR_ - Cached - Similar

FORTUNE 500 2007: Industry: - Revenues, Profits
Industry: Automotive Retailing, Services ... High revenue growth, High profit growth, High EPS growth ... Industry, Total Return to. Shareholders (2006) ...
money.cnn.com/magazines/fortune/.../Automotive.../1.html - Cached - Similar

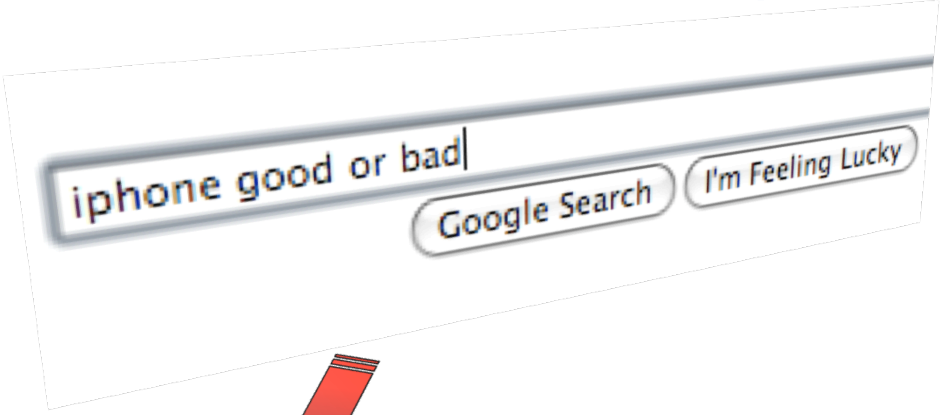
[英文调查报告书]汽车用半导体市场：2009年
Figure 1: Worldwide Automotive Semiconductor Revenue Forecast; Figure 2: Worldwide Automotive Industry Revenue Forecast; Figure 3: Worldwide Vehicle ...
www.gichinese.com.cn/chinese/data90157-auto-semicon.html - Cached - Similar

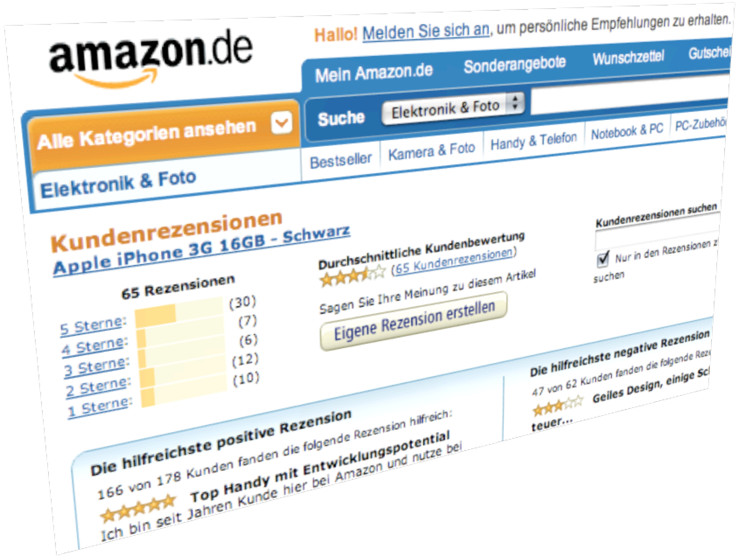
Automobile Industry (Industry sales and revenue) | Business ...
As we're reminded daily, times are tough in the automobile industry. ... Local Ad Sales Forecast: Declining Revenue, Some Bright Spots ...
www.allbusiness.com/automobile-industry/3141274-1.html - Cached - Similar

Automotive dealers (Industry sales and revenue) | Business ...
Local Ad Sales Forecast: Declining Revenue, Some Bright Spots .... The road to automotive industry innovation: 2002 information technology survey. ...
www.allbusiness.com/automotive-dealers/3188856-1.html - Cached - Similar

Brazil Automobile Industry till 2010 - Business Market Reports
What is the existing trend in the Brazil automobile industry? .... Figure 5-5: Brazil - Forecast for Revenue from Auto Component Industry (in Million US$), ...
www.bharatbook.com/.../Brazil-Automobile-Industry-till-2010.html - Cached - Similar

Motivation: Trends and opinions


iphone good or bad



Hallo! Melden Sie sich an, um persönliche Empfehlungen zu erhalten.

Kundenrezensionen

Apple iPhone 3G 16GB - Schwarz

65 Rezensionen

5 Sterne:	(30)
4 Sterne:	(7)
3 Sterne:	(6)
2 Sterne:	(12)
1 Sterne:	(10)

Durchschnittliche Kundenbewertung
★★★☆☆ (65 Kundenrezensionen)

Sagen Sie Ihre Meinung zu diesem Artikel

Kundenrezensionen suchen

☑ Nur in den Rezensionen z

suchen

Die hilfreichste positive Rezension

166 von 178 Kunden fanden die folgende Rezension hilfreich:

★★★★★ Top Handy mit Entwicklungspotential

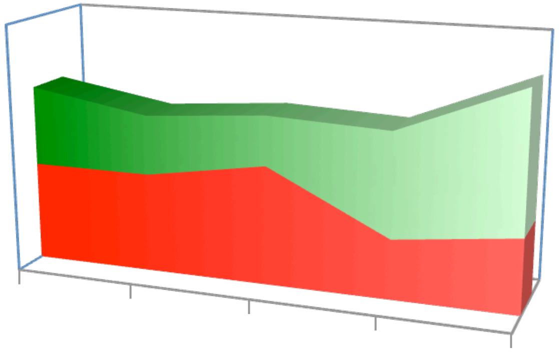
Ich bin seit Jahren Kunde hier bei Amazon und nutze bei

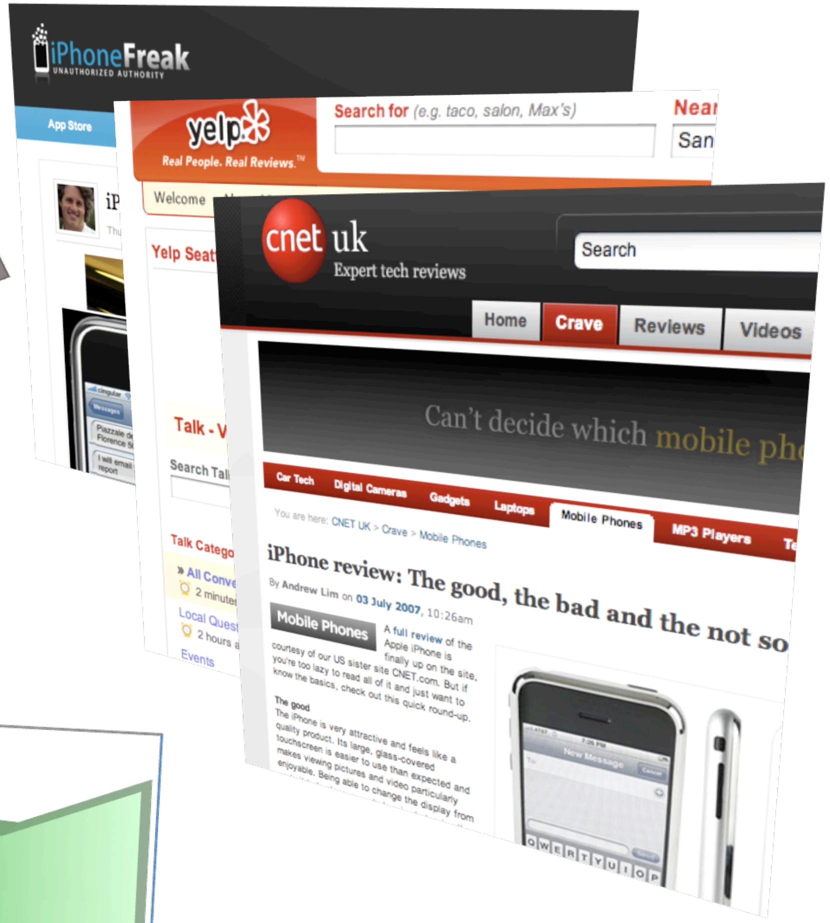
Die hilfreichste negative Rezension

47 von 62 Kunden fanden die folgende Reze

★★★☆☆ Geiles Design, einige Sc

teuer...





You are here: CNET UK > Crave > Mobile Phones

iPhone review: The good, the bad and the not so

By Andrew Lim on 03 July 2007, 10:26am

A full review of the Apple iPhone is finally up on the site, courtesy of our US sister site CNET.com. But if you're too lazy to read all of it and just want to know the basics, check out this quick round-up.

The good

The iPhone is very attractive and feels like a quality product. Its large, glass-covered touchscreen is easier to use than expected and makes viewing pictures and video particularly enjoyable. Being able to change the display from

Outline

What is IE? [1] Introduction, definitions and applications	**[2] Techniques** A common IE pipeline and its algorithms
Evaluation [3] Why, what and how to evaluate	**[4] Outlook** Problems, current state and conclusion

What is Information Extraction (IE)?

„Information Extraction is a technology that is **futuristic from the user's point of view** in the current information-driven world.“

(from the website of the National Institute of Standards and Technology)

NIST National Institute of Standards and Technology

Some marketing from Open Calais...

A little more precise...

„Information Extraction is a technology **based on analysing natural language** in order to extract snippets of information.“

(Hamish Cunningham, University of Sheffield)

„Recovering structured data from formatted text, i.e., **identifying fields**, **understanding relations**, **normalization** and deduplication.“

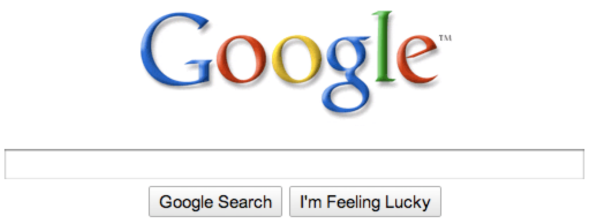
(Kamal Nigam, Google Pittsburgh)

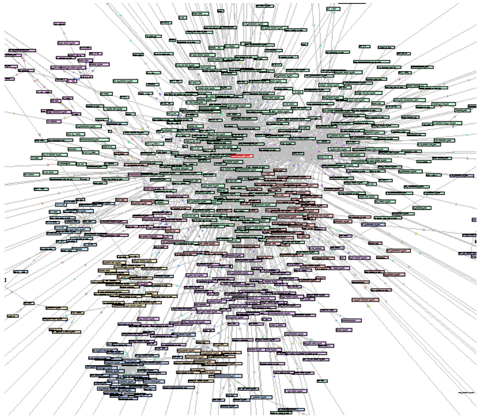
„**Filling slots in a database from sub-segments of text**. IE = segmentation + classification + association + clustering.“

(William W. Cohen, Carnegie Mellon University)

What isn‘t?

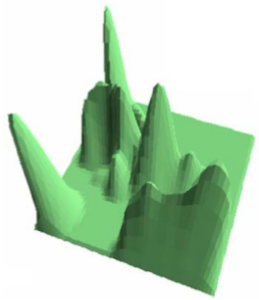
- Google (search) is not IE
  - Search is **Information Retrieval!** Next slide…





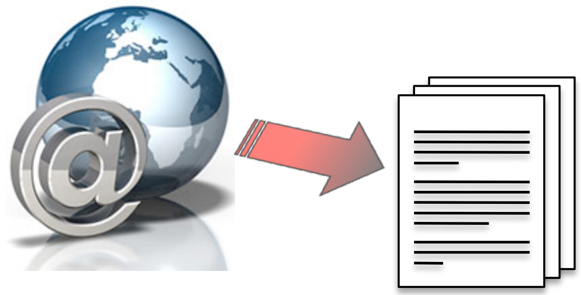
- **Text understanding** is not IE
  - IE is only interested in specific parts of the text
  - IE is usually much more efficient and realistic at present

- Inferring patterns and information from databases isn‘t.
  - That‘s **Data Mining**.

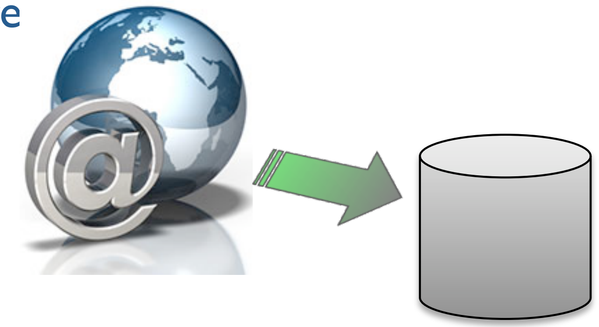


Retrieval vs. Extraction

- **Information Retrieval (IR):** Index and search
  - Returns relevant documents to a query based on similarity measures and statistics
  - YOU analyze the documents!
  - Advantage: widely applicable, problem: sometimes still too much information

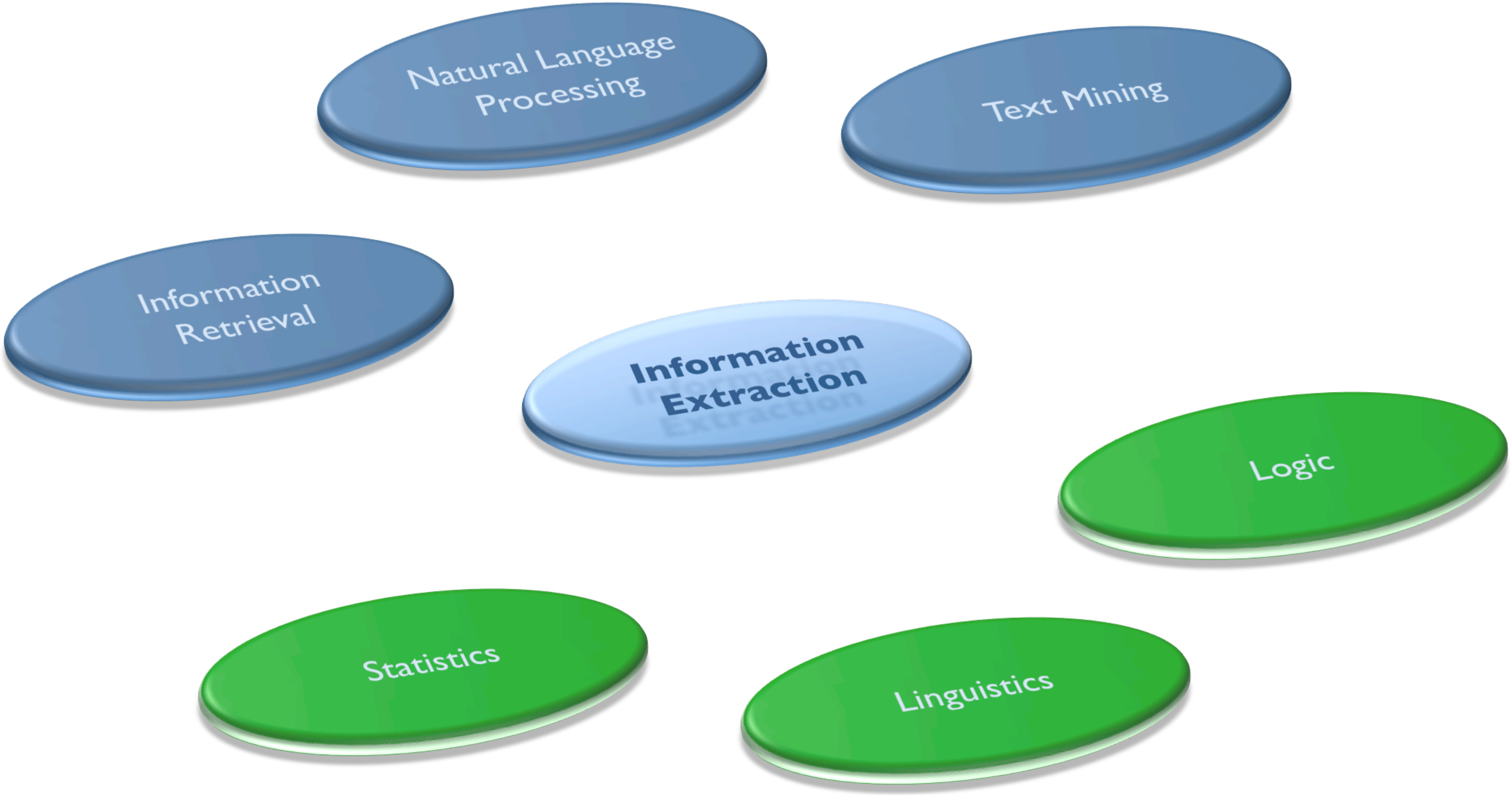


- **Information Extraction (IE):** Find and structure
  - Returns facts and specific information of predefined types
  - YOU analyze the facts!
  - Advantage: allows for automatic aggregation, problem: often domain-specific



IDSE Information-Driven Software Engineering

Related fields and sciences



Applications

- **Entity search**
- **Faceted search**
- Media monitoring
- Tourism offers
- Job Descriptions
- **Trend analysis and opinion mining**
- **Market forecasts**
- Health care delivery
- Monitor terrorist activities
- …

Outline revisited

What is IE? [1]
Introduction, definitions and applications

[2] Techniques
A common IE pipeline and its algorithms

Evaluation [3]
Why, what and how to evaluate

[4] Outlook
Problems, current state and conclusion

Extraction: example

LOEWE AG:VORLÄUFIGE NEUN-MONATS-ZAHLEN 2007

Kronach, 6. November 2007 – Auf vorläufiger Basis konnte das Ergebnis des Loewe Konzerns in den ersten neun Monaten deutlich verbessert werden. (...)

Beim Umsatz strebt der Chef des Konzerns, Rainer Hecker, für das Jahr 2007 weiterhin ein Wachstum von 10 % auf ca. 380 Mio. Euro an.

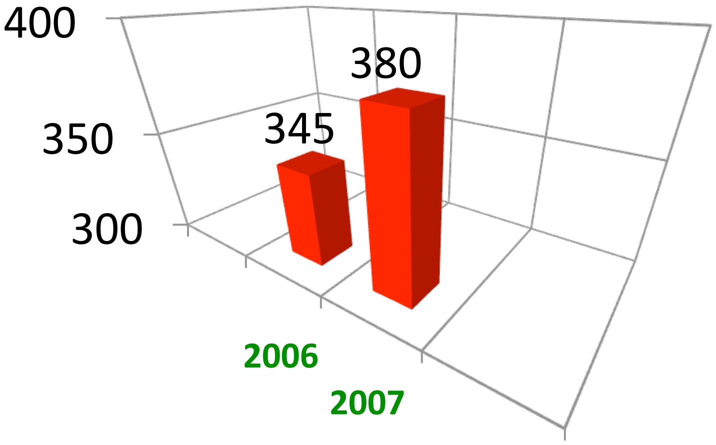
Loewe AG:

T = (01.01.2007, 31.12.2007)

M = (380.0, 10.0)

t = (06.11.2007)

A = (“Rainer Hecker”)



Loewe AG: Revenues in million €

Rule-based vs. statistical approaches

Logic

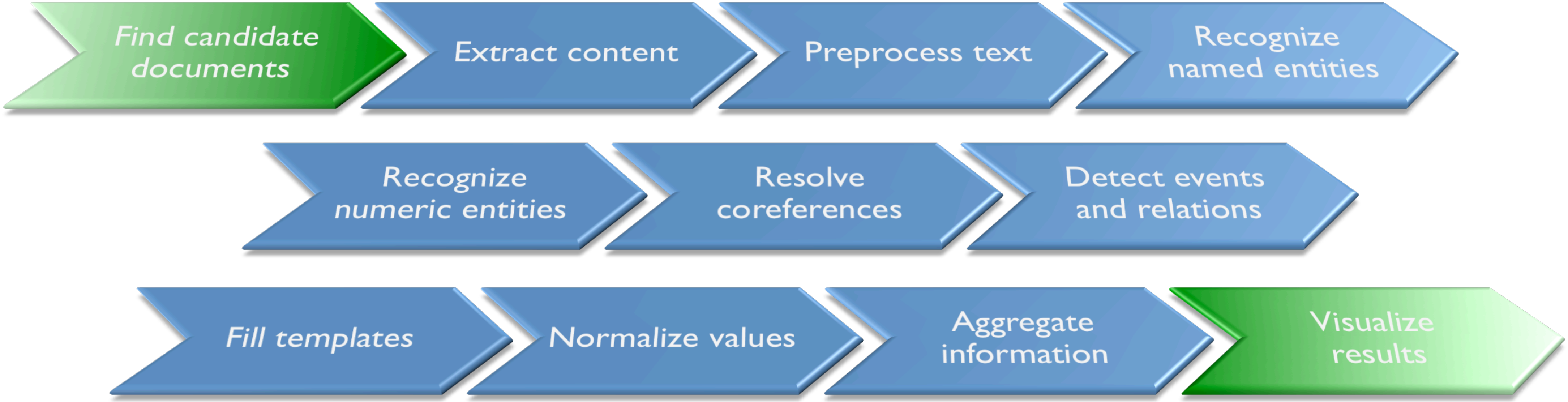
- **Knowledge engineering** based on language experience
- Make use of human intuition
- Only **few training data** needed
- Often very time consuming **hand-crafted** development
- Changes may be hard to accomodate

Statistics

- Use **Machine Learning** techniques
- Results may be counterintuitive
- Large amounts of **annotated training data** needed
- Main work is **feature engineering**
- Changes often easy (but sometimes require reannotation)

High-level IE pipeline

- **Normal IE tasks** may consist of several steps, which are sequentially executed
- An IE pipeline can look like the following:



Find candidate documents

Extract content

Preprocess text

Recognize named entities

Recognize numeric entities

Resolve coreferences

Detect events and relations

Fill templates

Normalize values

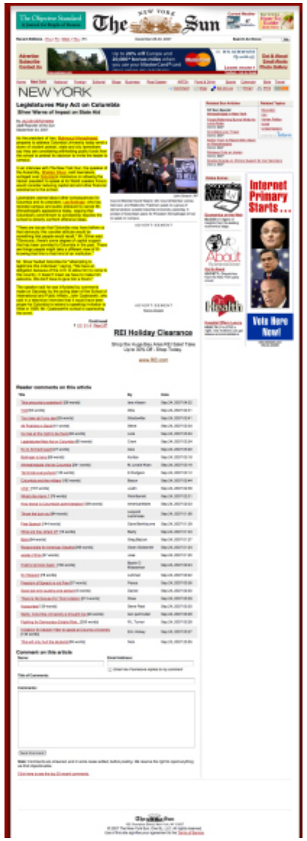
Aggregate information

Visualize results

(Note that the segmentation and naming of steps differ in literature, so take the given pipeline as an example)

Content Extraction

- **Content Extraction** is the process of determining the parts of a document which contain the main textual content
    - Common example: finding the main text of an HTML document
- For semi-structured documents, approaches to this task mostly rely on **analyzing the underlying markup structure**
    - Instead of syntactic trees, often only heuristics on the **density** are used
    - Passages with content often consists of very few tags



from spinn3r.com

Text preprocessing

- Many Information Extraction algorithms are based on the results of **syntactic and linguistic preprocessing**, namely:
  - Tokens and sentences
  - Stems, lemmata and part-of-speech of tokens
  - Phrase structure of sentences

More on this in the talk of Michael Meier

Named Entity Recognition (NER)

- **Named entities** are nameable objects in the world

  - Traditionally, IE concentrates on **persons, organizations**, and **locations,** but many categories are possible (e.g. markets, etc.)
  - NER seeks to find all entities in a text that belong to **predefined categories**
  - NER is the **most popular IE task** and has many applications

- Most state-of-the-art systems use **sequential classifiers** based on Machine Learning techniques:

Die Loewe AG erzielt zum 3. Mal in Folge mehr Umsatz .

O B-ORG I-ORG O O O ...

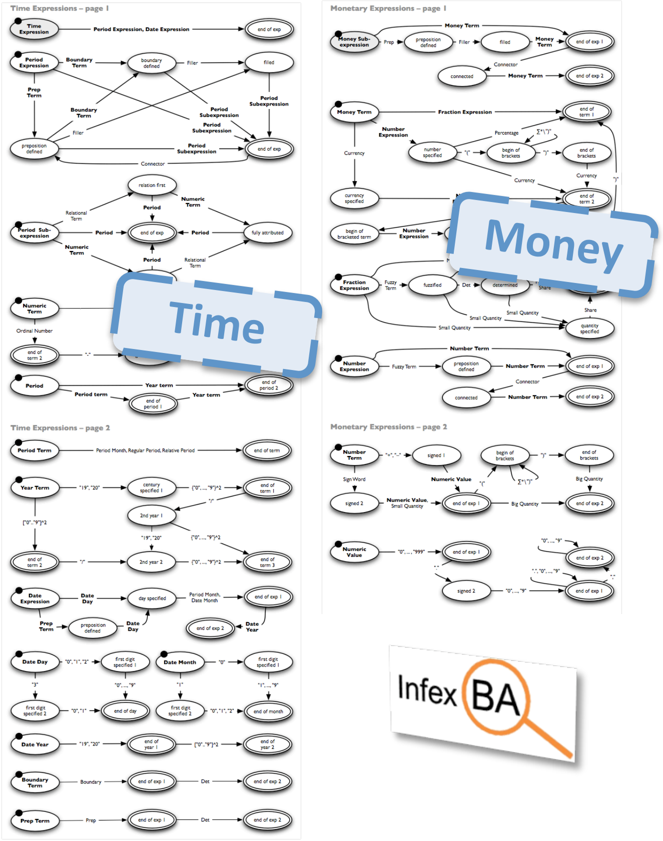
More on this labeling in the talk of Enes Yigitbas

More on NER in the talk of Michael Meier

Numeric Entity Recognition*

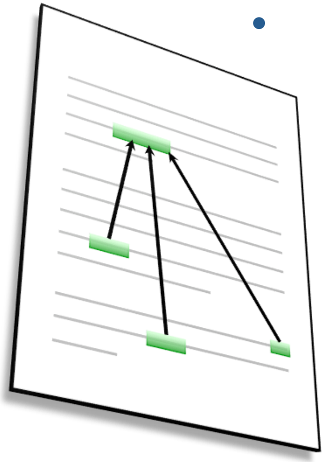
- **Numeric entities** can be time and money expressions, quantities, percentages, and the like
  - Sometimes also counted as named entities
  - Some standards distinguish **30 classes** of number expressions

- Numeric Entity Recognition can be done with Sequence Labeling or with rule-based approaches
  - In InfexBA, we found more than 96% of all time and money information with **regular expressions**


Time Expressions – page 1
Time Expressions – page 2
Monetary Expressions – page 1
Monetary Expressions – page 2
Time
Money

* also often only called **Temporal Expression Recognition (TER)**

Coreference Resolution



- Resolving coreferences is the task of **finding and unifying** different identifiers for the same referent
  - **Anaphoric:** relations between pronouns (etc.) and names
  - **Proper-noun:** different spellings and compoundings of the same object

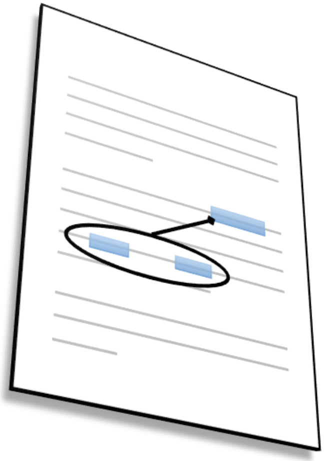
- **Example** from above: *LOEWE AG:VORLÄUFIGE NEUN-MONATS-ZAHLEN 2007*

  Kronach, 6. November 2007 – Auf vorläufiger Basis konnte das Ergebnis des Loewe Konzerns in den ersten neun Monaten deutlich verbessert werden. (...)

  Beim Umsatz strebt der Chef des Konzerns (...)

- Both **hand-crafted and statistical** approaches used in literature:
  - Identifying same head words, considering noun phrase types, distance features...

Relation Extraction and Event Detection



- **Relation Extraction (RE)** is on the recognition of relationship between entities
- **Event Detection** seeks for the events entities take part in
- Both tackle the identification of certain relations and classifying roles of entities:
  - Identification can be based on supervised learning (bag of words, etc.)
  - **Syntatic parsing** may then be used to determine roles

- Since parsing is quite inefficient, we need to minimize the set of candidate relations:
  - **Example:** Discard the following sentence, because it‘s not on revenue!
    Auf vorläufiger Basis konnte das Ergebnis des Loewe Konzerns in den ersten neun Monaten deutlich verbessert werden.

Template Filling

- **Compose** all information that refers to an instance of the scenario of interest
  - Use **entities** and **relations** from the previous stages
  - Maybe find further **attributes**, **properties** etc. (with the help of word lists, classifiers, etc.)

- **Example:** In InfexBA, we do the following:
  1. Extract named entities, time and money information
  2. Identify sentences that represent statements on revenue
  3. Determine type of statement and the relation between time and money
  4. Find subject, scope, statement time, etc.

Normalization

- **Resolve** in-text references
  - Example: *„Kronach, 6. November 2007 – Das Ergebnis des Loewe Konzerns konnte in den ersten neun Monaten deutlich verbessert werden. (...)“*

Master thesis possible!

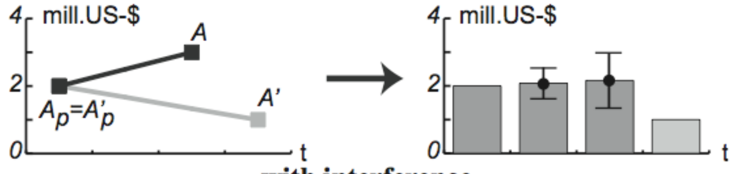
- **Transform** numeric expressions into a normalized form
  - Example: *„um 3 Millionen auf 21 Millionen Euro“* → (21.0, 16.67)
  - Example: „in den ersten neun Monaten 2007“ → (01.01.2007, 30.09.2007)
- **Match and unify** synonymous entity identifiers
  - Example: *„Loewe“* and *„Loewe AG“* → („Loewe AG“)

Aggregation

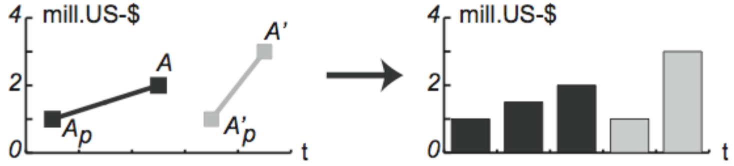
- With aggregation, we mean **merging, deduplication, and inference** on the normalized information

Example: *Market Forecast Information*

- Chronological merge values, compute average and deviations

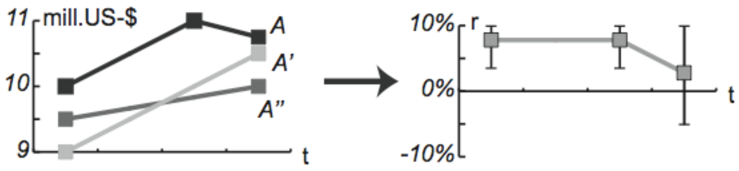


with interference



without interference

- Infer information on given topic from information on subtopics
  - e.g., from companies to markets



Bachelor thesis possible.
Tell your friends!

Visualization

- Once normalization and aggregation is done, visualization is straightforward

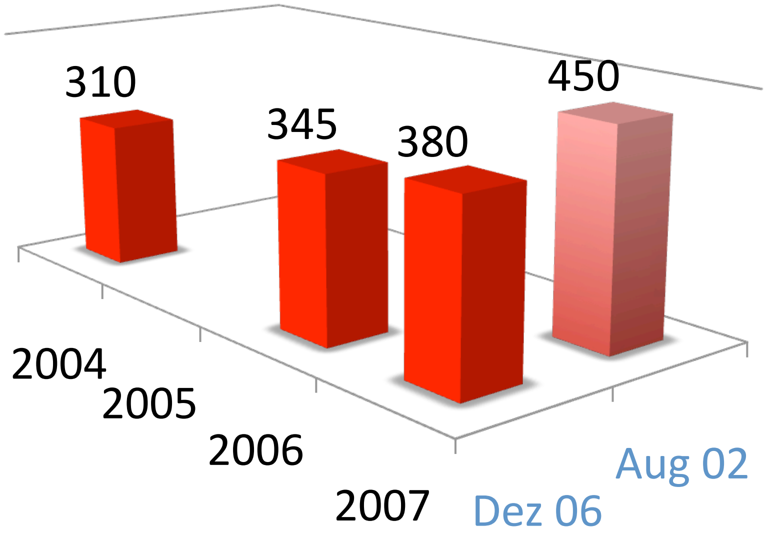
Loewe AG

Dez 06:
(01.01.2006, 31.12.2006): 345,0 Mio. €
(01.01.2007, 31.12.2007): 380,0 Mio. €

Dez 04:
(01.01.2004, 31.12.2004): 310,0 Mio. €

Aug 02:
(01.01.2007, 31.12.2007): 450,0 Mio. €

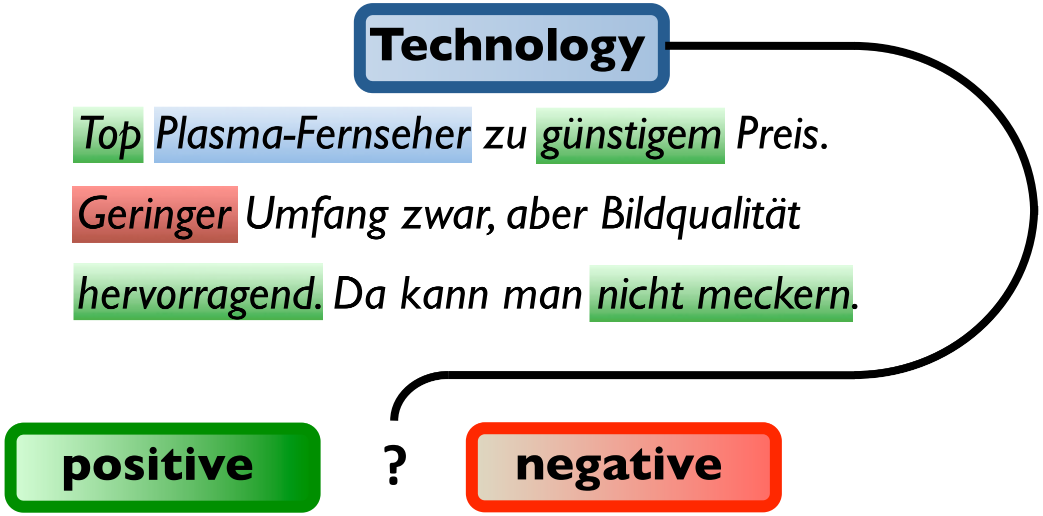
Bachelor thesis possible.
Tell your friends!



Loewe AG: past/predicted revenues

Text classification in IE

- Besides sequential classifiers, the classification of texts may be connected to Information Extraction
- Techniques like **genre analysis** or **opinion mining** are related to IE in that they often include the extraction of specific information:



Technology

Top Plasma-Fernseher zu günstigem Preis. Geringer Umfang zwar, aber Bildqualität hervorragend. Da kann man nicht meckern.

Outline, the 3rd

What is IE? [1]
Introduction, definitions and applications

[2] Techniques
A common IE pipeline and its algorithms

Evaluation [3]
Why, what and how to evaluate

[4] Outlook
Problems, current state and conclusion

Evaluation

Discover points of failure

Prove that your ideas are good

Why evaluate?

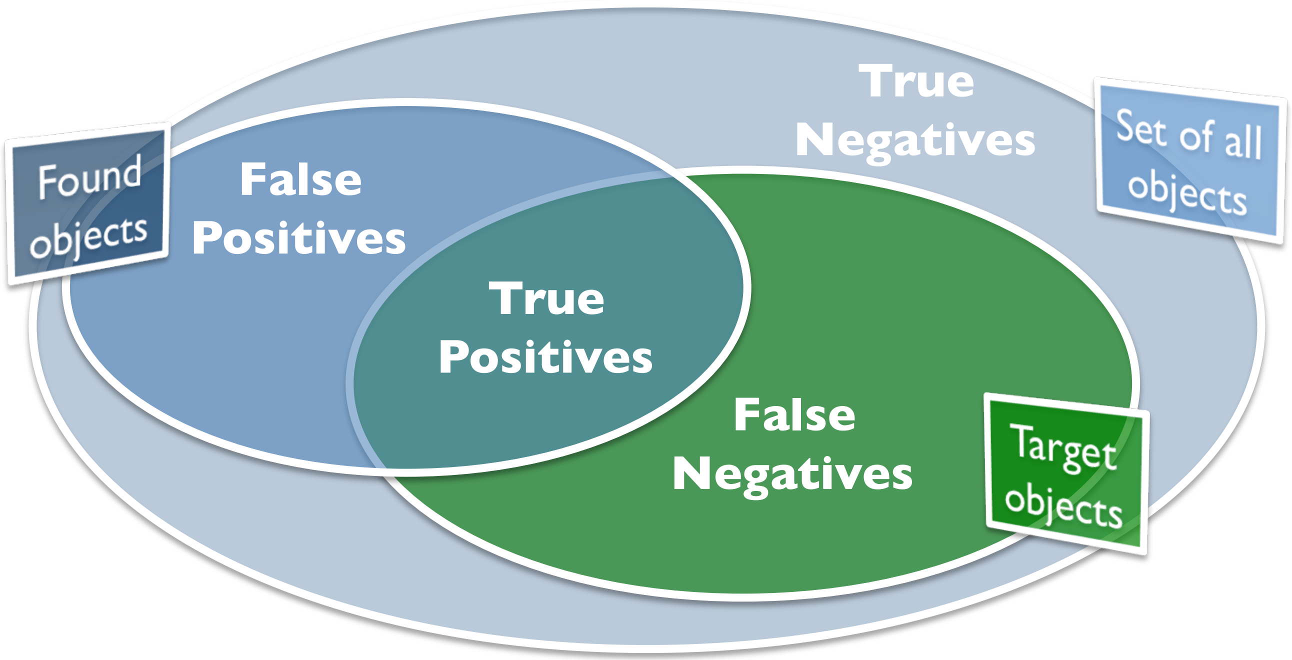
Decide between alternative methods

Tune / Optimize your system

How to evaluate? Some recommendations...

- **Clarify what you want to show:** efficiency vs. exactness vs. completeness vs. the best algorithm vs. proof of concept vs. ...
  - Can you make your results comparable?
- **Design your experiments in a scientific way:** Use test sets different from your training data, and make explicit what you‘ve done!
  - Could someone else reproduce your results from your documentation? How many variables are there? ...
- **Decide how to measure:** performance measures, correctness criteria, scope of evaluation (e.g. single vs. overall algorithms)...
  - Often standard measures such as precision and recall are used (next slide...)
  - What is a match (exact vs. overlap)? Is there only one correct text span?

Performance measures: Precision and Recall



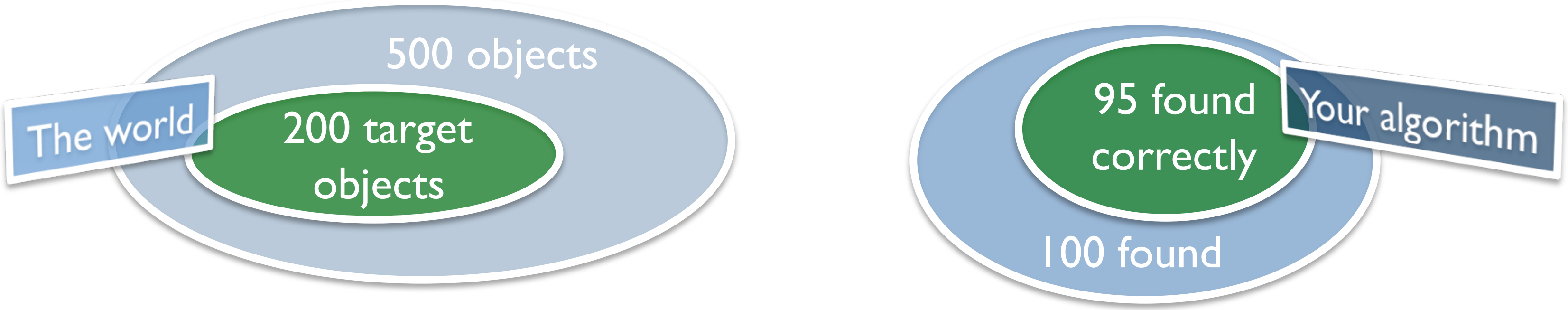
Precision (P) =
TP / (TP + FP)

Recall (R) =
= TP / (TP + FN)

F-Measure =
(2 · P · R) / (P+R)

- **Precision** is a measure for the exactness of your algorithm: How many of the found objects are correct?
- **Recall** is a measure for the completeness of your algorithm: How many of the target objects have been found?

Evaluation: example



Precision = 95 / (95+5) = 0.95

Recall = 95 / (95+105) = 0.475

F-Measure =
(2·0.95·0.475) / (0.95+0.475) = 0.62

Confusion Matrix

	Target	**Other**
Found	**95** (TP)	**5** (FP)
Not found	**105** (FN)	**295** (TN)

What to do with your results?

- Analyze your **false positives and false negatives**

  - Can you avoid some false extractions or classifications and can you find some of the missed true positives?
  - In consequence, introduce **better features/rules** and redo experiments

- In case results are new, innovative, best ever seen, scientifically important, or whatever: **publish them ;-)**

- Choose the best of your alternatives
  - **In theory,** the one that fits best to the problem at hand
  - **In practice,** the best-performing algorithm even if it‘s simple

State-of-the-art

- State-of-the-art **NER systems** achieve near-human performance:
  - F-Measure > 93% (humans ~ 96%)
  - **For English only!**
  - In German: F-Measure < 80%... WHY?
  - **Your task:** Improve German state-of-the-art ;-)
    It‘s all about feature engineering...

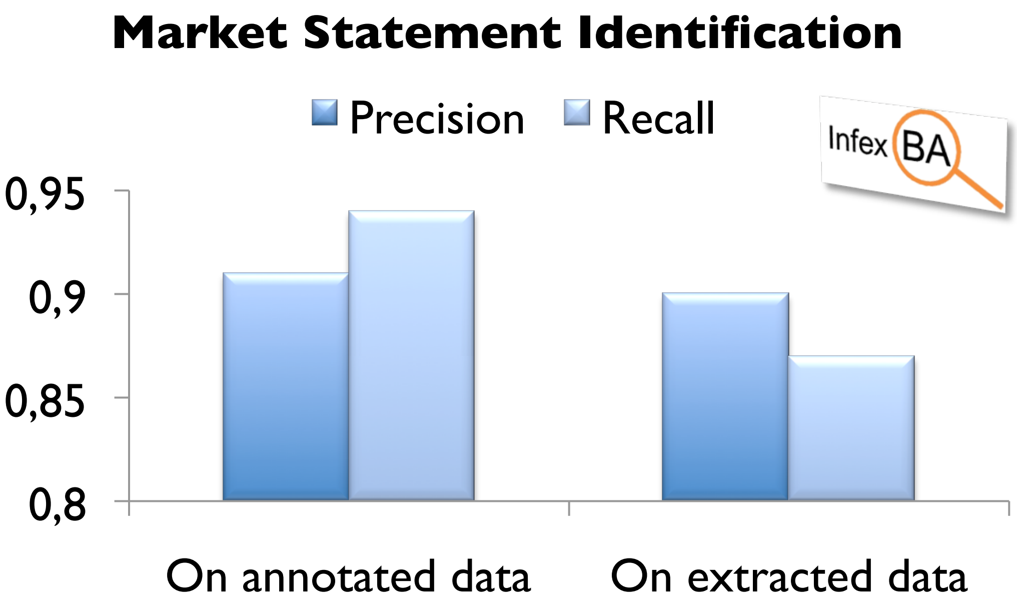
- **Coreference Resolution:** Results very domain-dependent, in hard cases only 50%-60% may be relied upon
- **Relation Extraction:** Good Systems have F-Measure > 75%
- **Template Filling:** Systems score around 60% while humans achieve 80%+

Error propagation

- **Imagine** you have a pipeline with 7 algorithms
  - Each has 95% F-Measure and is only based on the preceding one
  - What is your **overall F-Measure**?

- **In the worst case:**
  F-Measure = 0.95^7 = 0,698

- **In practice**, things are often not that bad. The problem is obvious, though:



Outline again

What is IE? [1]
Introduction, definitions and applications

[2] Techniques
A common IE pipeline and its algorithms

Evaluation [3]
Why, what and how to evaluate

[4] Outlook
Problems, current state and conclusion

The big problems

- **Limited accuracy** is actually often not that a big problem:
  - *„Even if results are not always accurate, they can be valuable if linked back to the original text“ (Diana Maynard, University of Sheffield)*
  - However, we know how to work on improvements of accuracy

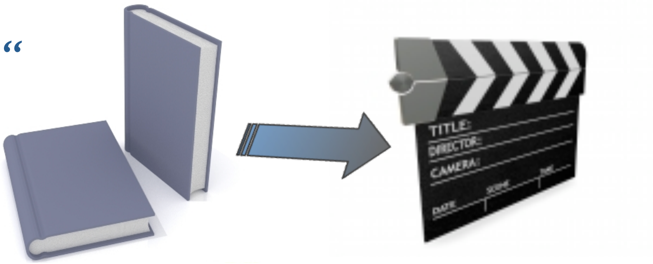
- The **need for training data** is a problem of time and resources and leads to the really big problems:

Domain dependency

Language dependency

Domain dependency

- Most IE techniques are more or less domain-dependent!
- **Example** from opinion mining: „*READ THE BOOK!*“
  - Positive statement on a **book?**
  - Positive statement on a **movie?**



- General **problems** refer to style of writing, specific scenarios of interest, etc.
- We need **annotated text corpora** for training and evaluation from the target domain
  - Collection of documents representative for the domain
  - Manually annotated by domain experts (for the project group: **YOU**)
- **Alternative:** research on domain adaptation, but evaluation problem remains

Language dependency

- Many rules or features tend to rely on language-specific properties even if no word lists are used
  - **German**: capitalization, words are spaced, etc., **English**: no capitalization
  - **Japanese**: words aren‘t even spaced
- Adaptation to new languages is the focus of much current work
  - Called **multilingual** or **language-independent** IE
  - **Needed:** language-independent features, bilingual dictionaries, annotated or at least unannotated data, support for non-latin scripts, different encodings, ...
  - Maybe language adaptation is just a special case of domain-adaptation?
- However, this is **not the focus** of the project group!

Frameworks, toolkits and the like

general architecture GATE for text engineering

Unstructured Information Management Architecture
An Apache Incubator Project

WEKA
The University of Waikato

openNLP

LingPipe

MALLET
machine learning for language toolkit

Natural Language Toolkit

Stanford Named Entity Recognizer (NER)

SVM*light*

TreeTagger

CALAIS
Powered by Thomson Reuters

What you should have learned...

- IE is about **extracting specific information** from natural language text
  - Uses linguistic and statistical phenomena
  - IE is „hot“ for the current industry
- A common IE pipeline consists of **several text analysis algorithms** such as
  - Entity Recognition and Relation Extraction
  - Normalization and aggregation
- **Evaluation is necessary** to show that your algorithms work

- IE is **often domain-specific** and thus needs the creation of and the work with annotated text corpora

Thanks for your attention!

Questions?