Informationsextraktion
Information Retrieval

Hans Uszkoreit

Hans Uszkoreit Vorlesung
Informationswissenschaft
und Informationssysteme

WAS KANN MAN, UND WAS NICHT

- ☆ Man kann heute immer noch nicht menschliche Texte oder Reden automatisch verstehen.
- ☆ Aber man kann Texte klassifizieren, gesprochene Sprache in Text oder Text in gesprochene Sprache umwandeln, relevante Information aus Texten extrahieren, Übersetzern bei der Arbeit helfen, Rechtschreibung und Grammatik überprüfen und vieles mehr.
- ☆ Sprachtechnologie steckt heute in allen großen Suchmaschinen und in jedem kleinen Handy.

DIE HERAUSFORDERUNG SPRACHE

- ☆ Viele der Beziehungen stecken in unstrukturierten Daten
- ☆ Datenbankmethoden helfen hier nicht
- ☆ Sprachtechnologie kann Konzepte und Relationen in Texten erkennen

Mensch und Maschine

- ☆ Wir können den Menschen nicht ersetzen.
- ☆ Wir können seine Fähigkeiten aber ergänzen um Fähigkeiten, die er nicht besitzt.
  - z.B. Maschine filtert aus Tausenden von Texten sehr schnell Kandidaten für bestimmte Zielinformationen. Maschine strukturiert und visualisiert die Ergebnisse der Suche.
  - Mensch bewertet, filtert noch einmal und wertet aus.
- ☆ Vorteile der Maschine: quantitative Leistung, Geschwindigkeit, Gedächtnis für Namen, Begriffe und Faken, keine Ermüdung, Duplikation, Kenntnisse anderer Sprachen
- ☆ Vorteile des Menschen: Wissen, Inferenzfähigkeit, Kreativität

Informationsextraktion

- ☆ Wir können Texte nicht automatisch verstehen
- ☆ Das heißt, wir können nicht all die Information aus Texten ziehen, die Menschen aus diesen Texten gewinnnen können
- ☆ Wir können aber wenige Arten von Information mit zunehmender Sicherheit erkennen
- ☆ Für diese Technologie gibt es viele Anwendungen

APPLICATION BUSINESS INTELLIGENCE

- ☆ Die Reaktionszeit, d.h., die Zeit zwischen dem Entstehen neuer Anforderungen, Technologien, Marktsituationen, (Bedürfnisse, Pläne und Entwicklungen der Wettbewerber), sowie neuer Rahmenbedingungen und den notwendigen Maßnahmen ist heute mehr denn je wettbewerbsentscheidend

- ☆ Zur Business Intelligence gehören unter anderem die Beobachtung von Abnehmern, Wettbewerbern, Technologieentwicklungen, relevanten politischen und anderen gesellschaftlichen Prozessen, Zulieferer- und Rohstoffmärkten

SUCHE UND ANALYSE

- ☆ Die beiden Haupttätigkeiten sind Suche und Analyse.
- ☆ Dann kommt die Verdichtung, Aufbereitung, Zulieferung und Präsentation für die Entscheidungsträger
- ☆ Technologien können bei Suche und Analyse helfen, die menschliche Analyse aber nicht ersetzen

SUCHTECHNOLOGIE

- ☆ Heutige Technologie sucht nach Zeichenketten und Wörtern (strings and words)
- ☆ Die nächste Stufe sind Terme und Konzepte (terms and concepts)
- ☆ Die dritte Stufe sind Beziehungen und Zusammenhänge (relations)

Entitätenerkennung

- ☆ Namenerkennung ist die Erkennung von Eigennamen, z.B. Personennamen, Ortsnamen, Firmennamen, Produktnamen
- ☆ Dazu kommen dann andere Bezeichnungen, wie z.B. Kalenderdaten, Uhrzeiten, Geldbeträge, Maße

Entitätenerkennung

- ☆ Auch wenn wir Menschen einen Satz nicht verstehen, können wir doch Namen erkennen, die wir nie zuvor gesehen haben.
- ☆ Wir erkennen sie über Muster.
- ☆ Über Muster der Form und Muster des Kontexts.
- ☆ Manchmal reichen Muster der Form, manchmal Muster des Kontexts.
  - Er hatte Eier gefunden. vs. Er hatte Meier gefunden.
  - Er hatte den Hügel nicht gesehen. vs. Er hatte den Hügel nicht angerufen.

PROBLEM: IDENTITÄTSAUFLÖSUNG

- ☆ Identity Resolution
- ☆ Enes der schwersten Probleme im Informationsmanagement
- ☆ Auch für scheinbar feste Namen oder Bezeichnungen gibt es oft Vielzahl von Schreibweisen

Dr. J. Tsujii
Jun‘ichi Tsujii
Tsujii Jun’ichi
Jun Ichi Tsujii
Professor Tsujii
Junichi Tsujii
辻井潤一
Tsujii-sensei
Prof. Dr. J.I. Tsujii
Tsujii, J.I.
Jun-Ichi Tsujii

Form und Kontext

- ☆ Muster für die Form der Bezeichnung
  - Eralius van Dammendarp
  - Fjegor Jorassovitch Tschovtschenko
  - Dao Fa Jing
  - Dr. Magor Krast
- ☆ Muster für die Umgebung der Bezeichnung
  - Orrkor wurde als Lagerverwalter eingestellt.

NAMEN UND IDENTITÄTEN

- ☆ Erkennung, Auflösung und Analyse von Namen
- ☆ Am DFKI gibt es mehrere Systeme, die auf dem diesen Gebiet deutliche Fortschritte gemacht haben
  - ○ Automatische Erkennung der Herkunft von Namen
  - ○ Automatische Analyse komplexer Namen aus vielen Sprachen
  - ○ Auflösung verschiedene Schreibweisen arabischer Namen im Deutschen
  - ○ Auflösung verschiedener Schreibweisen westlicher Namen im Chinesischen

Named Entity Extraction

- ☆ *Personennamen*
- ☆ *Firmennamen*
- ☆ *chemische Verbindungen*
- ☆ *Uhrzeiten*
- ☆ *Geldbeträge*
- ☆ *Ortsnamen*

Tokenization

- ☆ - identifiziert die Textstruktur
  - ○ z.B. Paragraphen, Einrückungen, Titelzeile
- ☆ - identifiziert spezielle Zeichenketten (Tokens)
  - ○ z.B. Datums- und Zeitangaben, Abkürzungen,
  - ○ Wortgrenzen und Interpunktionszeichen

Morphological and Lexical Processing

- ☆ Bei der lexikalischen Verarbeitung erfolgt eine morphologische Analyse der potentiellen Wortformen

  --> Bestimmung der Wortart (Part-of-Speech, POS) und der Flexionsform (z.B. Plural und Singular)
- ☆ Analyse der Komposita und Hyphenkoordination (speziell für die deutsche Sprache)
- ☆ Anschließend werden morphosyntaktisch mehrdeutige Wörter mittels POS-Taggern disambiguiert
  - ○ z.B. Ich meine meine Tasche
- ☆ Eigennamenerkennung findet auch durch Behandlung von Referenzen zwischen Eigennamen (EN- Koreferenz) statt, um festzustellen, dass man im Text dieselbe Person bezeichnet.

Syntactic Analysis: Parsing

- ☆ in den meisten IE-Systemen wird keine vollständige syntaktische Analyse durchgeführt, sondern lediglch eine flache, fragmentarische Analyse (shallow paring, partial parsing) .

- ☆ - Die Parsingaufgabe wird stark modularisiert durch explizite Trennung von Phrasen- (NP, PP, VG) und Satzstruktur.
  - ○ eine domänenunabhängige Phrasenanalyse mit Regeln zur Erkennung von komplexen (Satz-) Einheiten.

Domain Analysis

- **Erkennung domänenrelevanter Muster**
  - Hier werden die Regeln definiert, die die Struktur der Templateinstanzen bestimmen
  - - Sie müssen Merkmale der Köpfe der extrahierten Phrasen überprüfen (z.B. syntaktische Eigenschaften, Eintrag im Domänenlexikon)

- **Template-Unifikation**
  - Ein einzelner Satz muss nicht alle notwendigen Informationen zur Instanziierung eines Templates enthalten
  - Daher ist es nötig, Informationen aus unterschiedlichen Templateinstanzen zu vereinigen

RELATIONEN IN DER SUCHE

- ☆ Manchmal suchen wir nach jeder Art von Information über bestimmte Begriffe: Firmen, Personen, Technologien, Länder, ...
- ☆ Meistens aber suchen wir nach Informationen zu Beziehungen zwischen Dingen:
  - Medizin gegen Schuppenflechte,
  - gute mehrsprachige Schulen im Freiburger Raum,
  - Wettbewerber von Yahoo in Japan
  - Acquisitions by IBM from 2005-2007

Relationen

☆ Relationen zwischen Entitäten:

z.B. Relationen zwischen Personen p und Firmen f

p ist Kunde von f

p ist Vorstand von f

z.B. Relationen zwischen Ereignis e und Datum d

e findet am d statt

INSTANZEN

☆ Entitäten-Klassen haben Instanzen

„Karl Sonntag“ ist eine Instanz von „Person“

☆ Relationen haben Instanzen

„Reinhard Selten erhielt 1994 den Nobelpreis für Wirtschaftswissenschaft“
ist eine Instanz der Ereignisklasse „Nobelpreisverleihung“

Nobelpreisverleihung (Empfänger, Jahr, Disziplin)

der Satz ist auch eine Instanz der Relation Auszeichnung

Auszeichnung(Empfänger, Auszeichnung, Jahr, Disziplin)

ERWÄHNUNGEN

- ☆ Die Stellen im Text, in denen Instanzen erwähnt werden, nennen wir Erwähnungen (mentionings).
- ☆ Für manche Aufgaben ist es wichtig, die Instanzen zu entdecken, in anderen die Erwähnungen.
- ☆ Eine Relationsinstanz kann über mehrere Sätze verteilt sein.
- ☆ Komplexe Relationen auch über mehrere Absätze oder gar Dokumente.
- ☆ Manchmal müssen die Erwähnungen alle auch gefunden werden, weil sie nur Anker für weiterführende Auswertungen sind.

Arten der Informationsextraktion

- ☆ Topic Detection - Themenextraktion
- ☆ Name Extraction - Namensextraktion
- ☆ Named Entity Extraction - Begriffsextraction
- ☆ Relation Extraction - Relationsextraktion
- ☆ Event Extraction - Ereignisextraktion
- ☆ Opinion Mining - Meinungsextraktion
- ☆ Sentiment Detection - Extraktion von Emotionen und subjektiven Bewertungen
- ☆ Answer Extraction - Antwortextraktion

Arten der Informationsextraktion

- ☆ Topic Detection - Themenextraktion
- ☆ Name Extraction - Namensextraktion
- ☆ Named Entity Extraction - Begriffsextraction
- ☆ Relation Extraction - Relationsextraktion
- ☆ Event Extraction - Ereignisextraktion
- ☆ Opinion Mining - Meinungsextraktion
- ☆ Sentiment Detection - Extraktion von Emotionen und subjektiven Bewertungen
- ☆ Answer Extraction - Antwortextraktion

Informationsextraktion aus Texten

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates
* Microsoft
Gates
* Microsoft

Bill Veghte
* Microsoft
VP
Richard Stallman
* founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

DAS PROBLEM DER GRAMMATIK

- ☆ Selbst wenn wir nicht jede Phrase oder jeden Satz analysieren, brauchen wir doch ein Regelsystem (oder Klassifikator), das die relevanten Passage erkennt und auflöst.

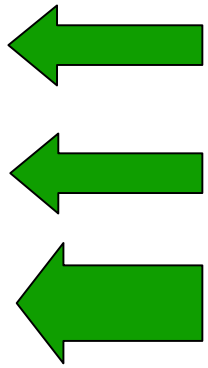
- ☆ Woher bekommen wir diese Regelwerke?

HAUPTMETHODEN

- ☆ instrospektive Handarbeit
- ☆ korpusgestützte Handarbeit
- ☆ überwachtes Lernen aus annotierten Texten
- ☆ Lernen aus Beispielen
- ☆ vollständig automatisches Lernen

HAUPTMETHODEN

- ☆ instrospektive Handarbeit (teuer, unvollständig)
- ☆ korpusgestützte Handarbeit (teuer, langsam)
- ☆ überwachtes Lernen aus annotierten Texten (teuer, langsam)
- ☆ Lernen aus Beispielen (vielversprechende Resultate)
- ☆ vollständig automatisches Lernen (meist nicht anwendbar)



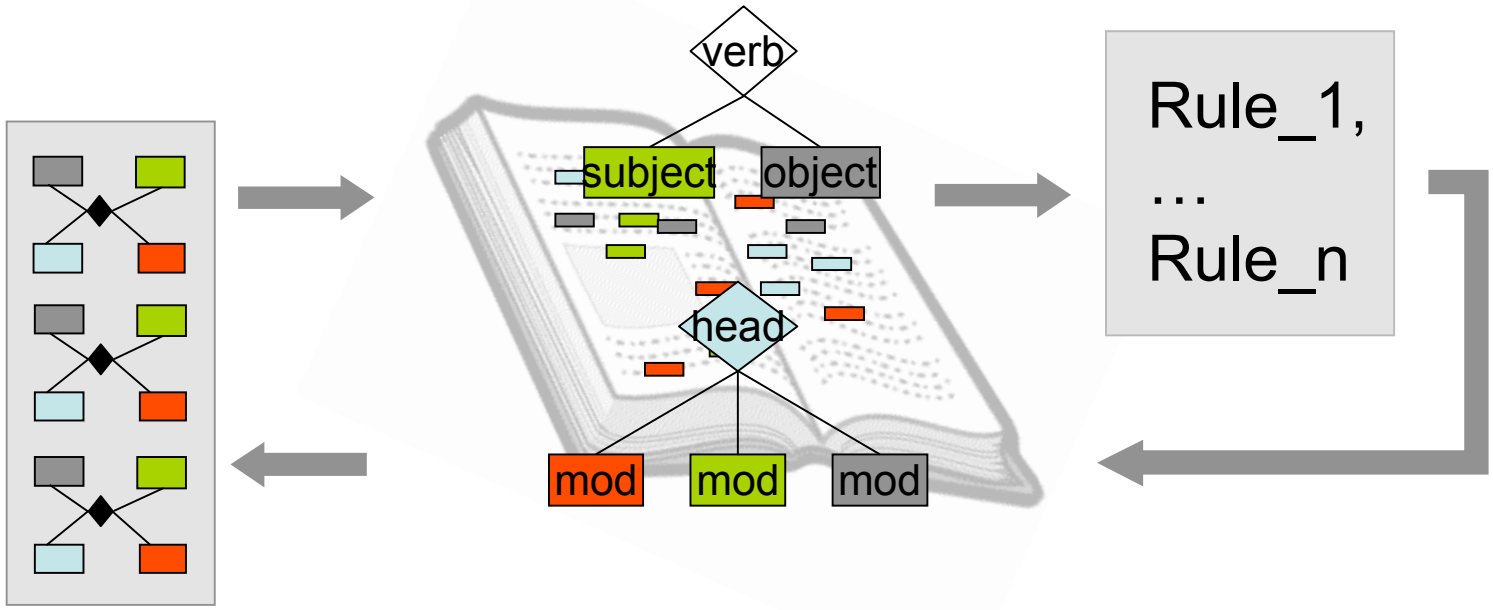
Ein selbstlernendes Verfahren

(entwickelt gemeinsam mit Feiyu Xu und Hong Li)

- ☆ In dem neuen Verfahren für das Erlernen von Relationsextraktionsmustern führen wir ältere Ideen von Brin und Yangarber fort.
- ☆ Wir haben das Verfahren aber in mehrere Richtungen erweitern und besitzen heute auch die hinreichend mächtige Analysetechniken.
- ☆ Wir geben dem System einige Beispiele für relevante Relationen.
- ☆ Dann sucht das System automatisch Erwähnungen, analysiert diese und lernt die Muster. Mit diesen Mustern werden wieder neue Instanzen gefunden. Mit diesen wieder mehr Muster usw.
- ☆ Nach 4-10 Durchläufen findet das System keine neuen Instanzen mehr.

LERNEN IN ZYKLEN AUS BEISPIELEN

DIPRE (Brin, 1998) und Snowball (Agichtein & Gravano, 2000)



Beispiel, Satz and Muster

Beispiel: Musikpreise

- ☆ Beispiel (Seed)

  <Madonna, Grammy, Best Long Form Music Video, 1992>

- ☆ Satz

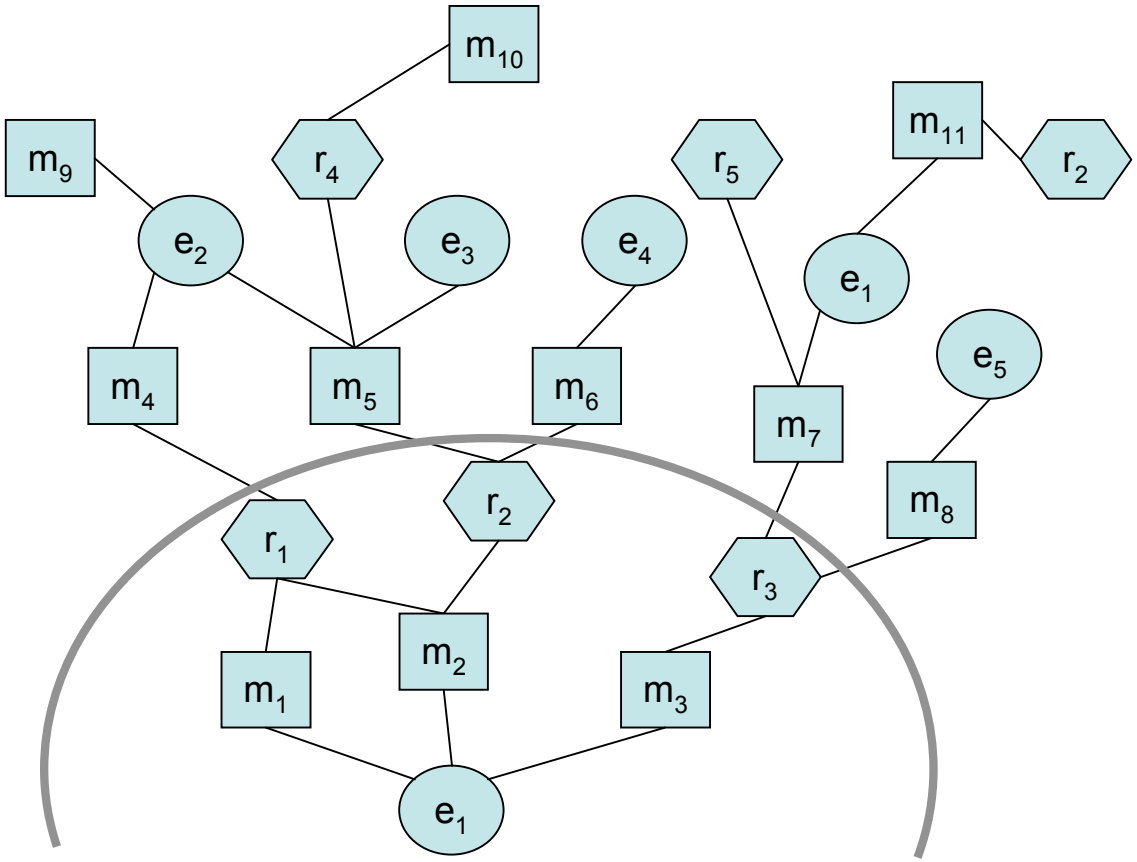
  Madonna won her first Grammy in 1992 in the Best Long Form Music Video category for the laserdisc release of her 1990 Blond Ambition Tour..

- ☆ Muster

  <subject: recipient> „win“ <object: prize> <mod: year> <mod: category>

LERNSTRUKTUR

e: event, Ereignis
m: mention, Erwähnung
r: rule, Regel



Beispiel: Nobelpreise

☆ Zielrelation

<recipient, prize, area, year>

☆ Beispielinstanz

< "Mohamed ElBaradei", "Nobel", "Peace", "2005">

Sentence: Mohamed ElBaradei won the 2005 Nobel Peace Prize on Friday for his efforts to limit the spread of atomic weapons.

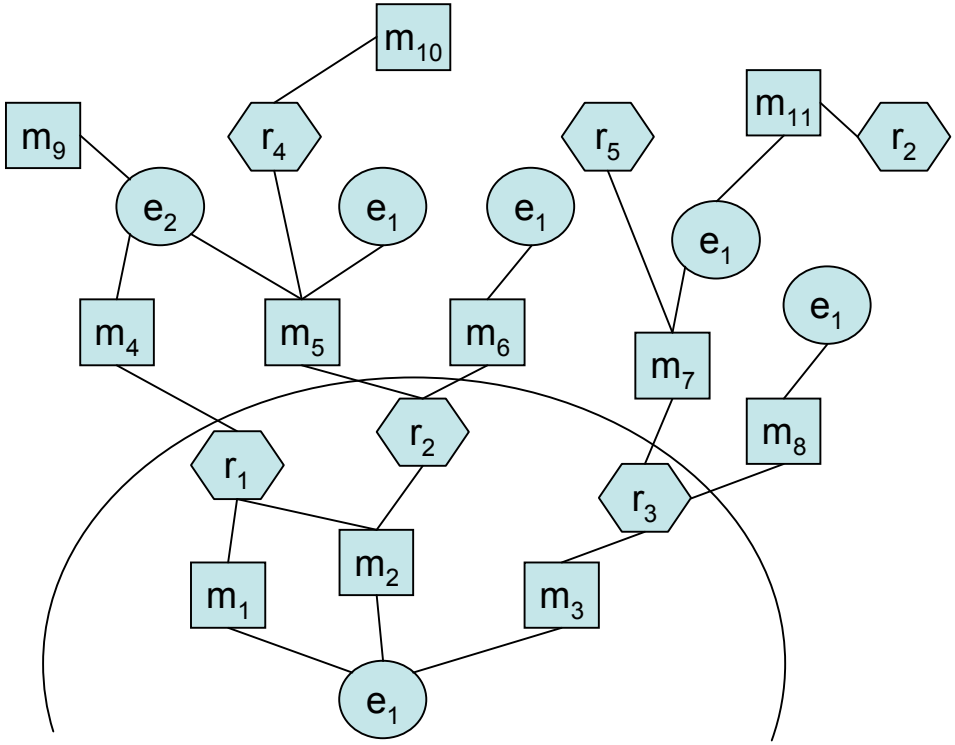
Anaphora Resolution

- ☆ Man hatte John Anderson für seine bahnbrechenden Arbeiten in der Virologie bereits 1962 für den Nobelpreis in Medizin vorgeschlagen.

- ☆ Für diese Leistungen erhielt er zehn Jahre später dann auch endlich den hohen Preis.

Lernen aus Beispielen

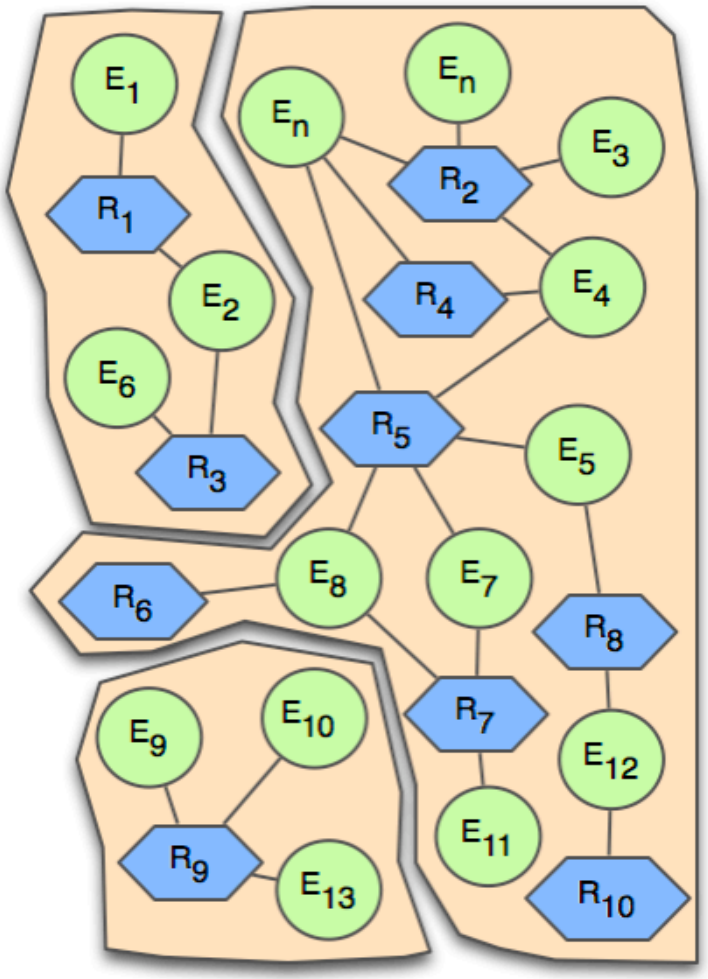
e: event, Ereignis
m: mention, Erwähnung
r: rule, Regel



GRENZEN DER METHODE

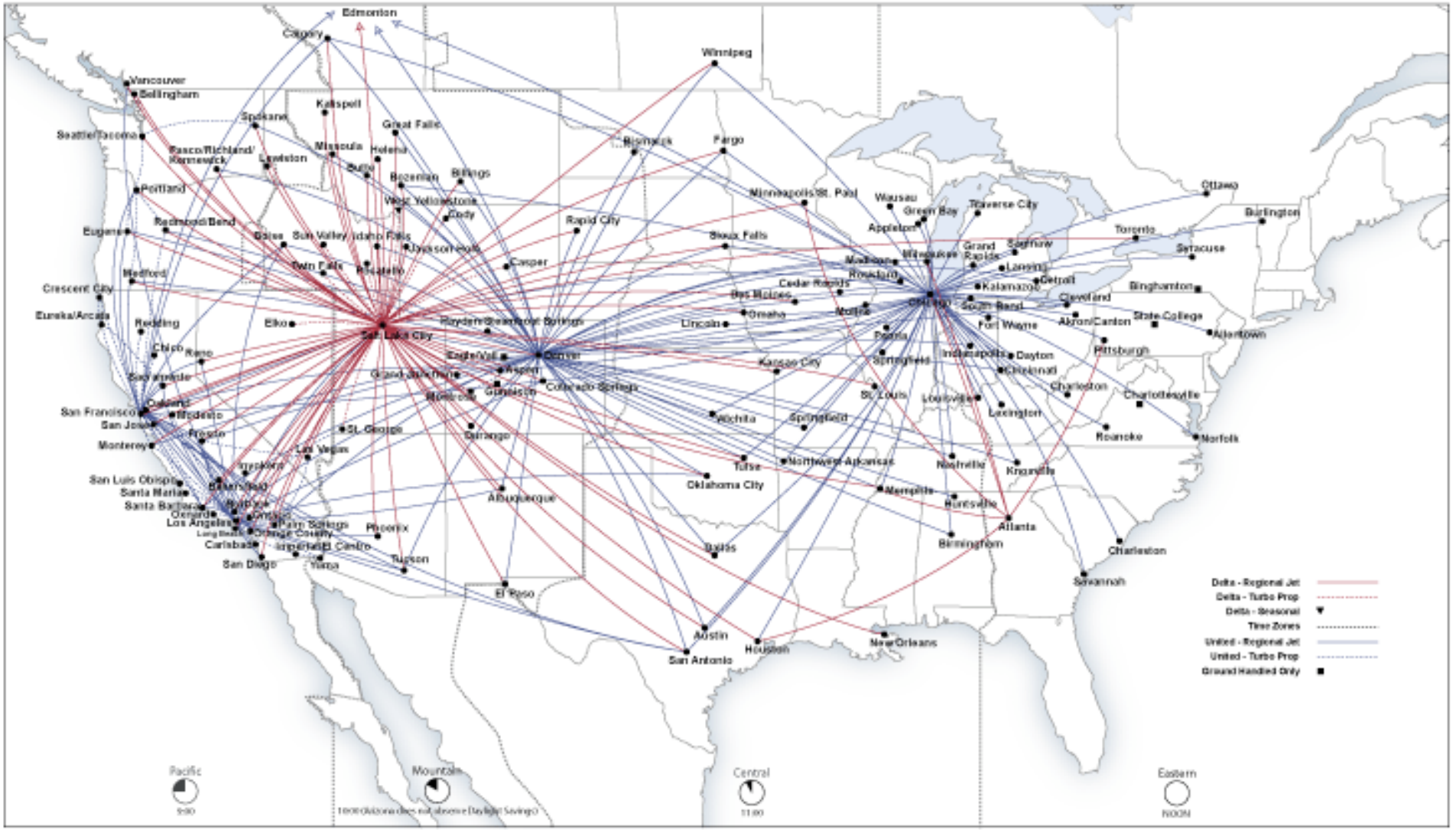
- ☆ E: event, Ereignis
- ☆ P: pattern, Muster

- ☆ Nicht von jedem Beispiel kommt man zu jedem anderen Ereignis

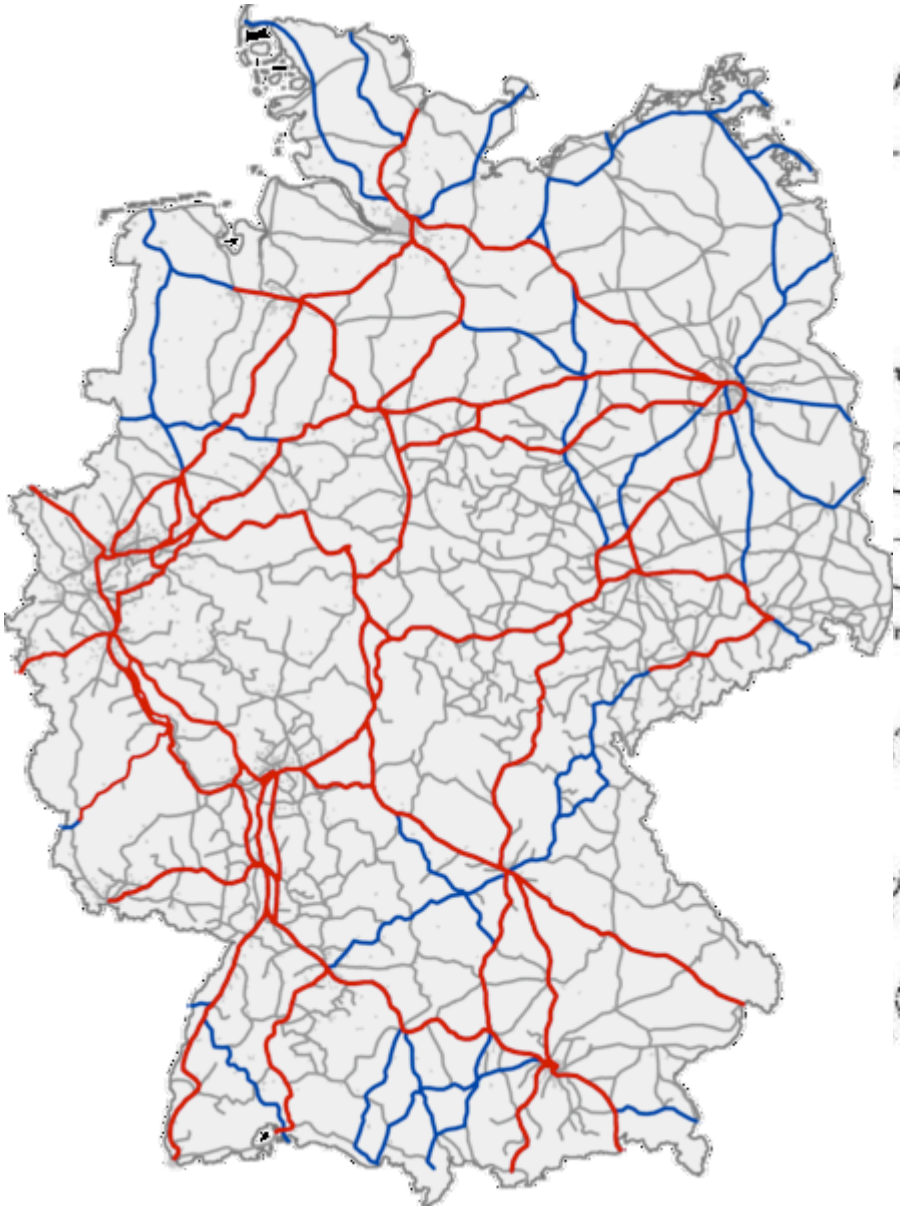


Kleine Welten

- ☆ Small World Property
- ☆ Wenn Graphen bestimmte Eigenschaften haben, sind die alle Knoten nur wenige Schritte voneinander entfernt
- ☆ Die Streckennetze von Fluggesellschaften weisen diese Eigenschaft auf



☆ Die Streckennetze von Eisenbahnen nicht.



Kleine Welten fürs Lernen

- ☆ Beim Lernen der Muster für die Relationsextraktion brauchen wir zwei Arten von zentralen Knotenpunkten oder „Hubs“ (Naben)
  - einige Muster, die sehr, sehr häufig verwendet werden, um die relevanten Klasse von Ereignissen zu beschreiben
  - einige Ereignisse, die sehr, sehr häufig erwähnt werden

ALTERNATIVE

- ☆ Für manche Anwendungen haben wir viel mehr Lernbeispiele als gesuchte Ereignisse.

- ☆ Beispiel: intelligence applications
- ☆ Beispiel: RASCALLI

KOSTENEINSPARUNG

- ☆ Wenn die Notwendigkeit der Beobachtung akzeptiert wird, dann kann man die Kosten für die manuelle, intellektuelle Beobachtung ohne zusätzliche Hilfsmittel ermitteln.
- ☆ Das ist schwer, weil es ohne technologische Hilfsmittel natürlich gar nicht geht.

RECALL vs. PRECISION

- ☆ Recall ist der Anteil der gefundenen relevanten Informationen im Verhältnis zu allen relevanten Infomationen

- ☆ Precision ist der Anteil der relevanten Informationen an allen gefundenen Informationen

ERKENNUNGSLEISTUNGEN

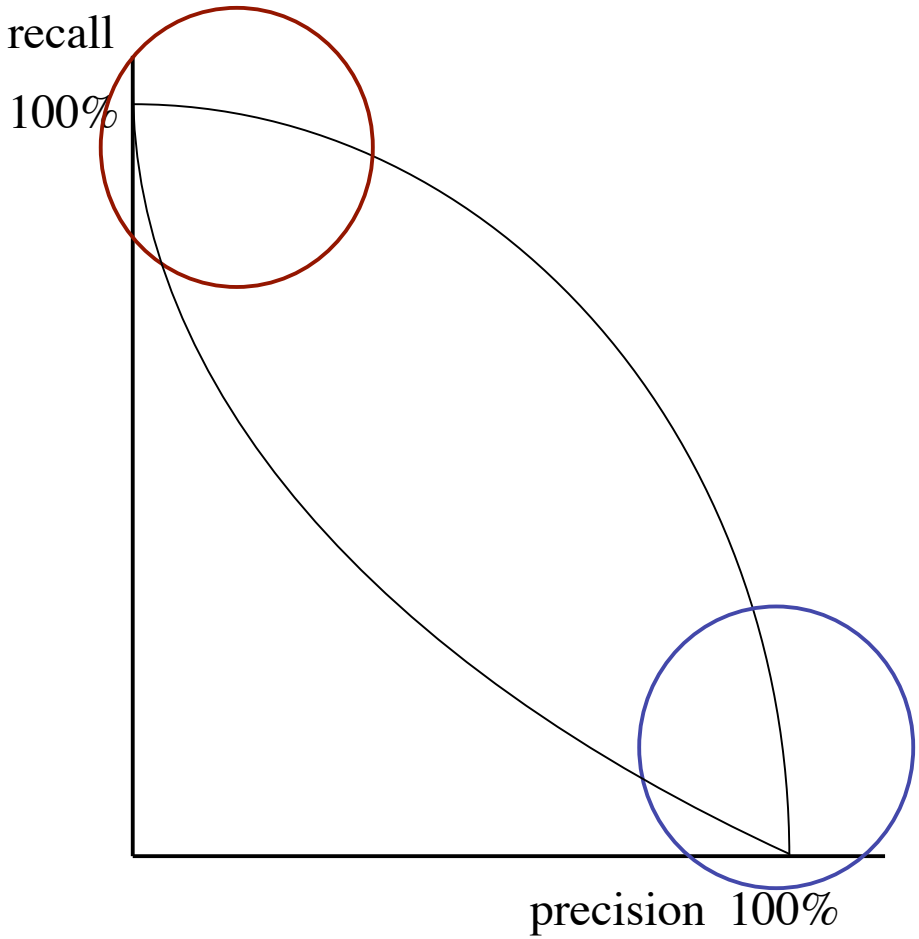
- ☆ Entitätenerkennung ja nach Klasse zwischen 60% und 92%

- ☆ Relationserkennung je nach Klasse zwischen 40% und 69%

VERSCHIEDENE AUFGABENBEREICHE

- Interessanter Bereich für den Sicherheitsbereich, für Business-Intelligence und alle Frühwarn-anwendungen

- Interessaner Bereich für vollautomatische Sammel- und Statistikaufgaben



EIN EINFACHES BEISPIEL

- ☆ Wenn jeden Tag 100.000 Dokumente durchsucht werden, und 100 gefunden werden, von denen 1 relevant ist, dann ist das bei 1% Precision immer noch ein sehr nützliches System, solange der Recall bei 100% liegt.

- ☆ Wenn dieses eine Dokument gefunden werden muss, dann ist die Kostenersparnis immens.

Kompetenzsuche

- ☆ Suche nach Kompetenzquellen
  - für Technologiemonitoring
  - für die Ideenbewertung
  - für die Ideenverfolgung

Kompetenzsuche und -analyse

- ☆ Differenzierte Suche
  - erfordert eine Taxonomie/Ontologie des Sachgebiets
  - eine Definition und Gewichtung von Indikatoren
  - eine Identifikation der relevanten Startquellen

Indikatoren

- ☆ Relevante Publikationen
- ☆ Zitationen
- ☆ Patente
- ☆ Mitgliedschaft in Gremien
  - ○ Konferenzen, Standardisierung, Beratende Gremien
- ☆ Keynote Lectures

Quellen

- ☆ Wissenschaftliche Zeitschriften
- ☆ Fachorgane
- ☆ Google Scholar
- ☆ Patentinformationsdienste
- ☆ Zitations-DBs
- ☆ Das Web

Kombination von Methoden

- ☆ Bibliometrie (Analyse von Publikationen, Zitationen, Links, etc.)
- ☆ Informationsextraktion (Suche nach den relevanten Indikator-Daten)
- ☆ Wissenstechnologie (Strukturierung des Bereichs durch Ontologien)
- ☆ Informationsvisualisierung (Visualisierung komplexer Zusammenhänge)
- ☆ Intelligente und semantische Suche
  - ○ Frage-Antwort-Systeme
  - ○ Semantische Suche (Web 3.0, Theseus-Projekt http://theseus-programm.de/)

Der ständige Radar

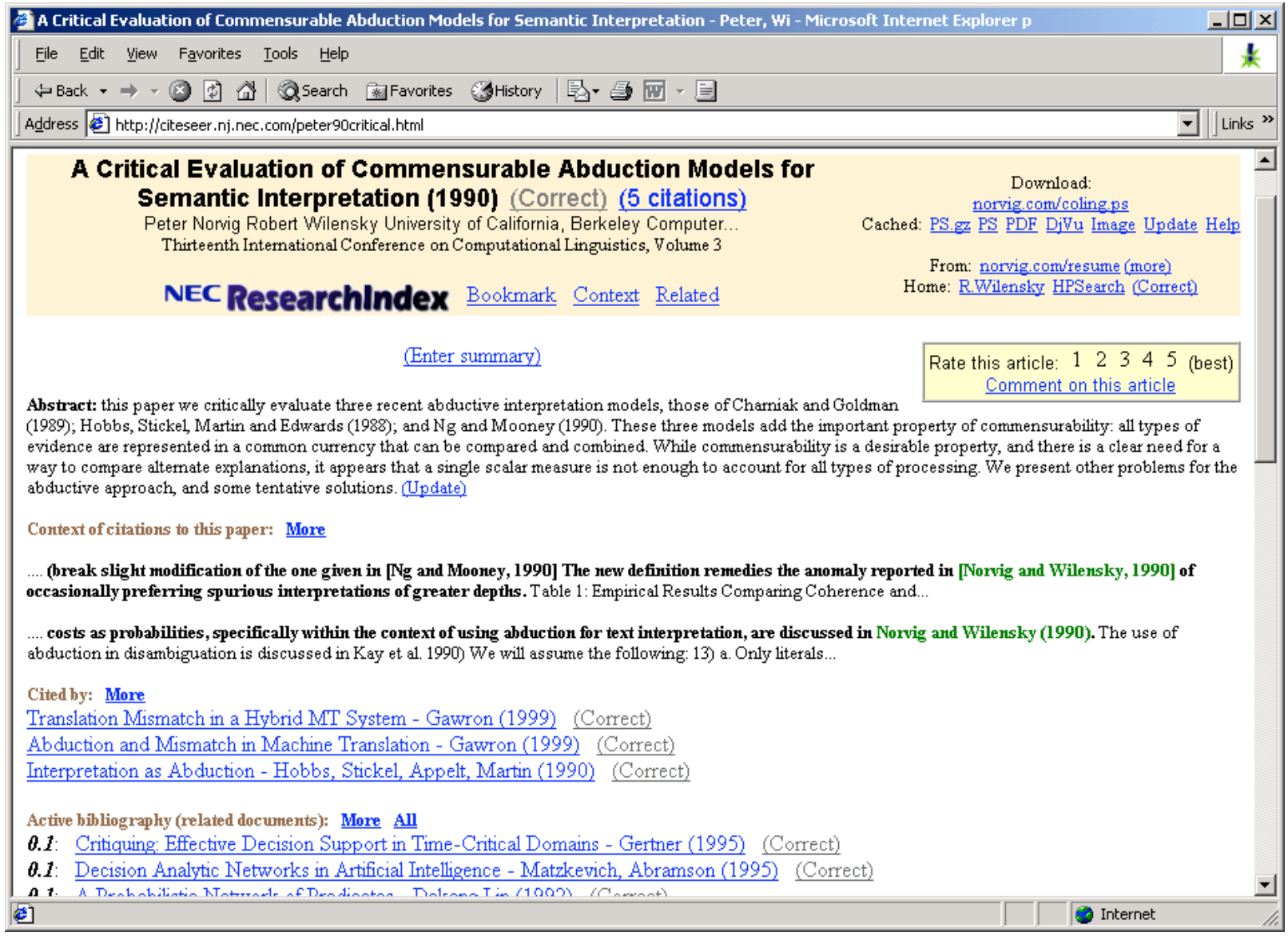
- ☆ Ein proaktiv handelndes Unternehmen identifiziert mögliche Chancen und Gefahren durch künftige Gegebenheiten oder Ereignisse
- ☆ Je früher der Eintritt dieser Gegebenheiten erkannt wird, desto besser sind die Möglichkeiten der effektiven Reaktion
- ☆ Das Unternehmen muss jetzt so lange auf den Radarschirmen nach Indikatoren für die Chancen und Risiken suchen, bis Entwarnung gegeben wird.
- ☆ Der Aufwand steigt mit der Zunahme der identifizierten Chancen und Gefahren und mit der Menge der zu scannenden Quellen
- ☆ Hier kann Technologie helfen!

Warum ist die Früherkennung wichtig?

1. Weil auch der Wettbewerb reagiert und weil die Schnelligkeit der Reaktion entscheidend sein kann

2. Weil die minimale Reaktionszeit eines Weltkonzerns auch bei proaktiver Vorbereitung meist noch recht lang ist (Je größer ein Schiff, desto früher müssen Hindernisse erkannt werden, um ausweichen zu können.)

Informationsextraktion aus Publikationen



Informationsextraktion aus semi-struktuierten Daten

Suche und Analyse von Innovationsindikatoren

- Relevante Schlagwörter, e.g.,

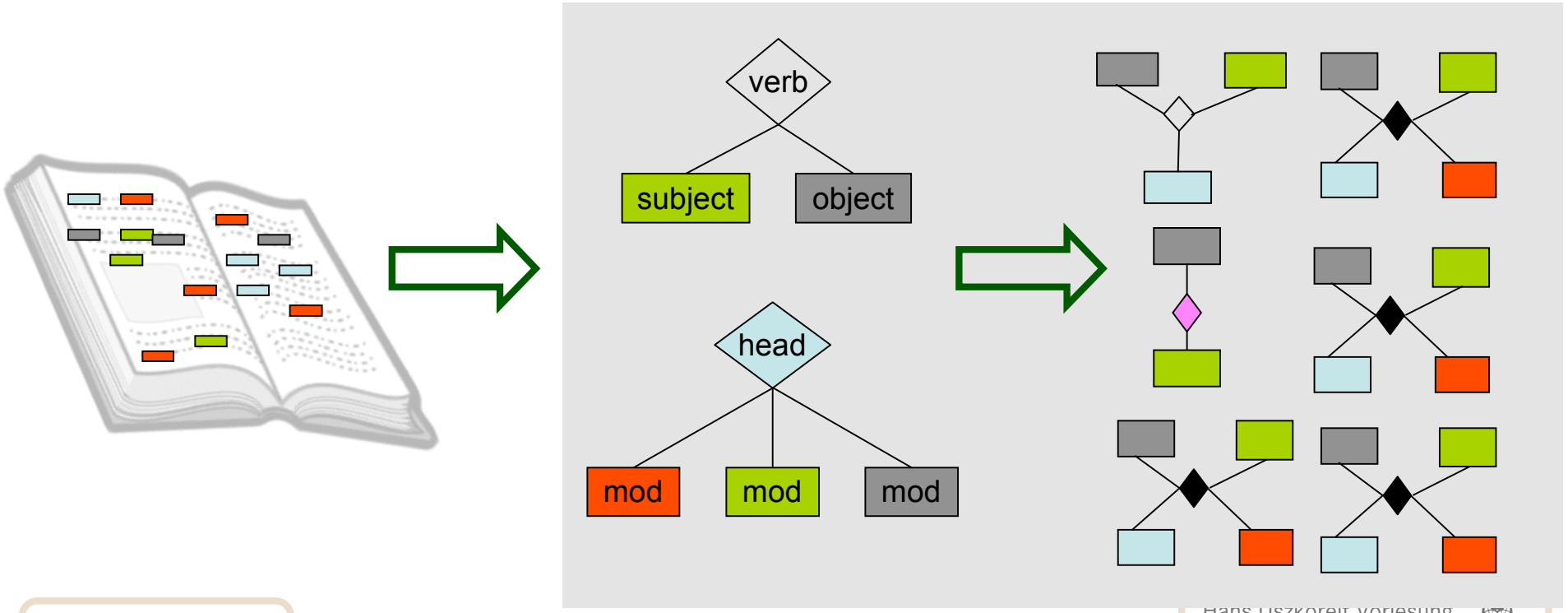
  „Technologie von Übermorgen“, „Zukunft“, „Entwicklung“

- Satzmuster, e.g.,
  - *„Lithium-Ionen-Akkus treiben Autos der Zukunft an.“*
  - *„Lithium-Ionen-Technik ist die Zukunft für Hybrid- und Elektrofahrzeuge.“*
  - *„Der Laser wird auch künftig immer für eine Überraschung gut sein“, sagt Peter Leibinger voraus.*

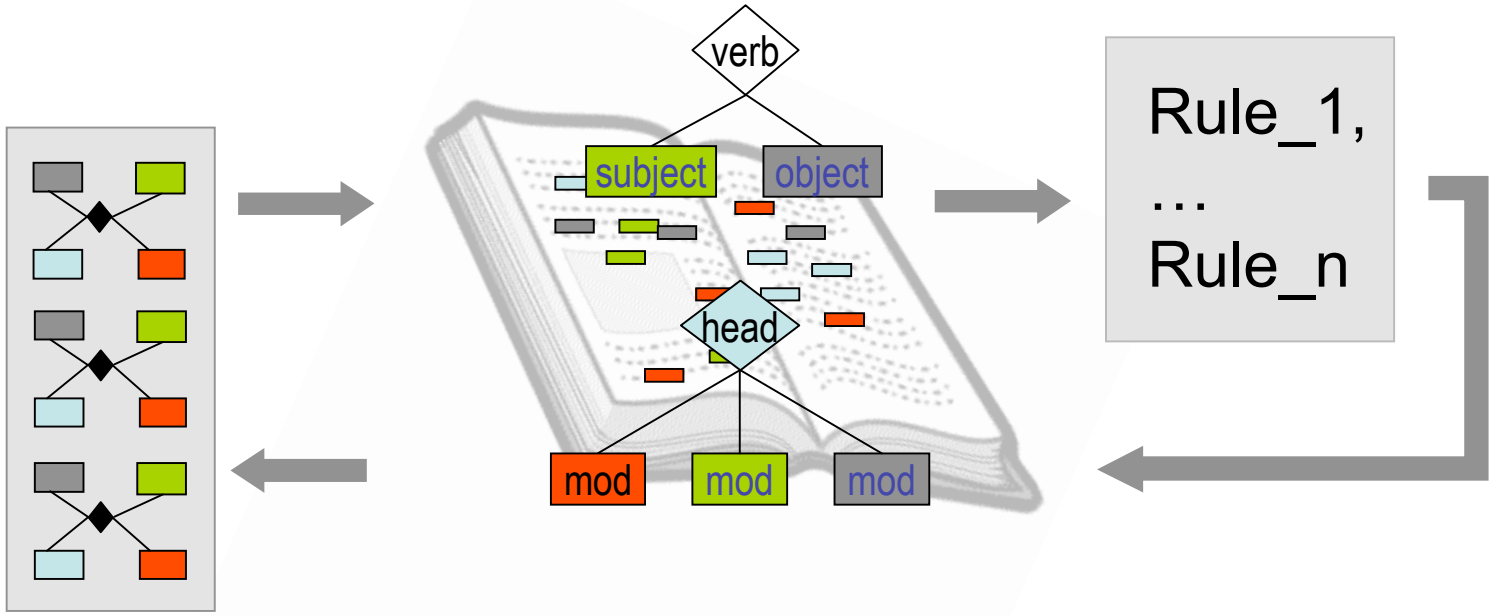
DFKI Informationsextraktionsplattform:

"A general framework for automatically learning mappings between linguistic analyses and target semantic relations, with minimal human intervention."

Ein selbst-lernendes System, um die Sprachmuster zu entdecken, die für gesuchte Information und Beziehungen relevant sind.

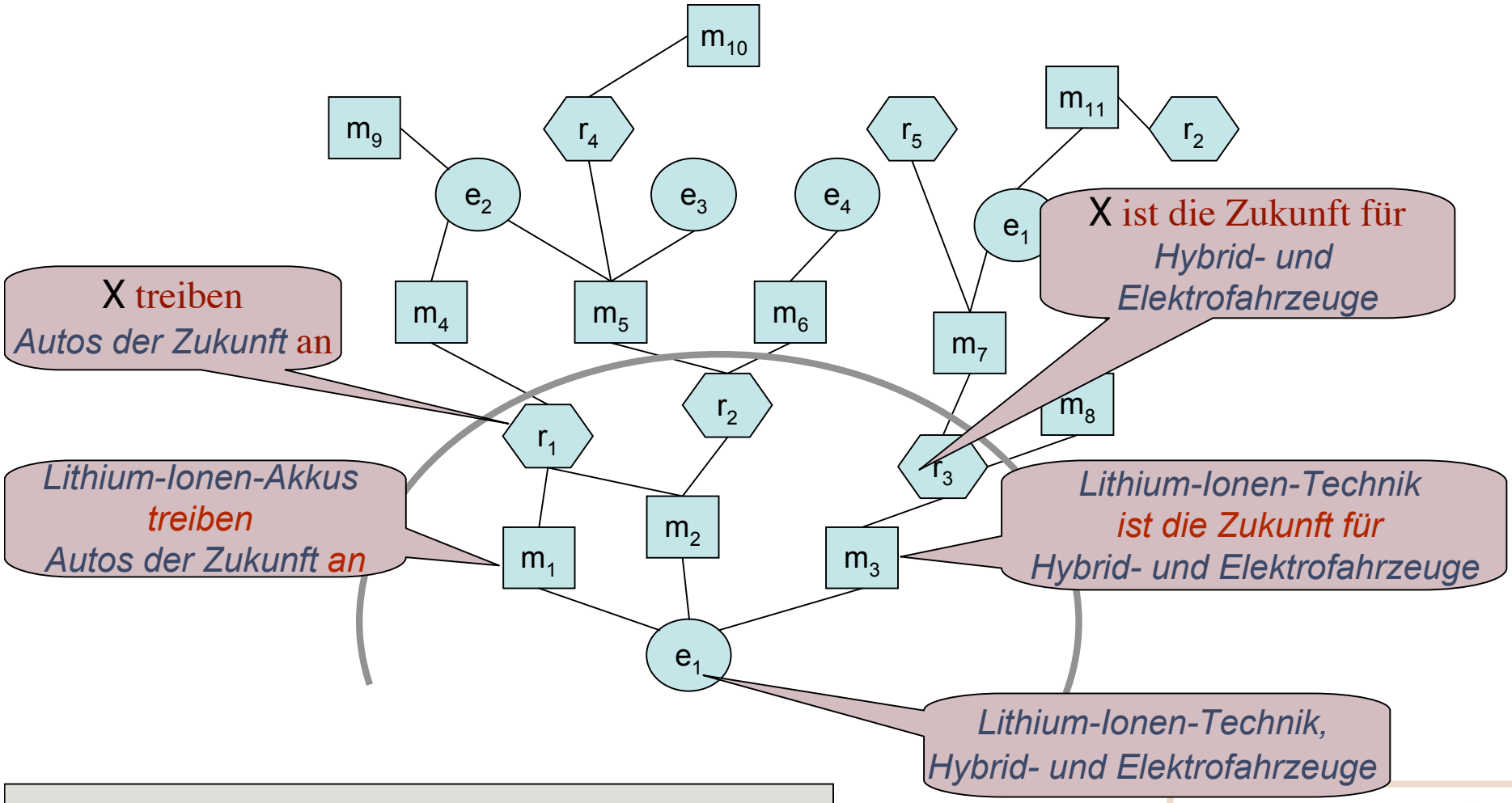


DARE
Bootstrapping Relation Extraction from Semantic Seed (http://dare.dfki.de)



Hier fängt man mit einer kleinen Menge von Beispielen an, um Sätze und dann die Sprachmuster zu finden.

Das Lernen ist ein Boostrappingprozess



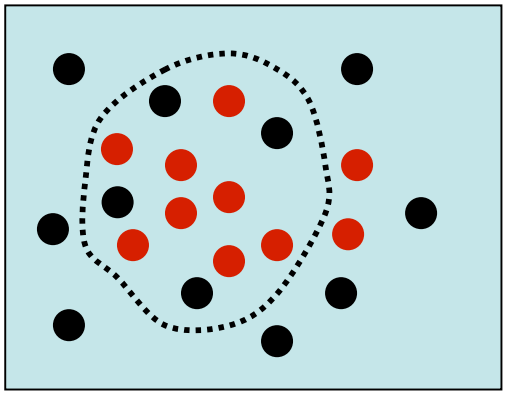
e: Beispiele; m: Sätze; r: Sprachmuster

RECALL vs. PRECISION

- ☆ Recall (Trefferquote) ist der Anteil der gefundenen relevanten Informationen im Verhältnis zu allen relevanten Infomationen

- ☆ Precision (Genauigkeit) ist der Anteil der relevanten Informationen an allen gefundenen Informationen

Beispiel:
8 von 10 gefunden = 80,0% Trefferquote
8 von 12 korrekt = 66,6% Genauigkeit



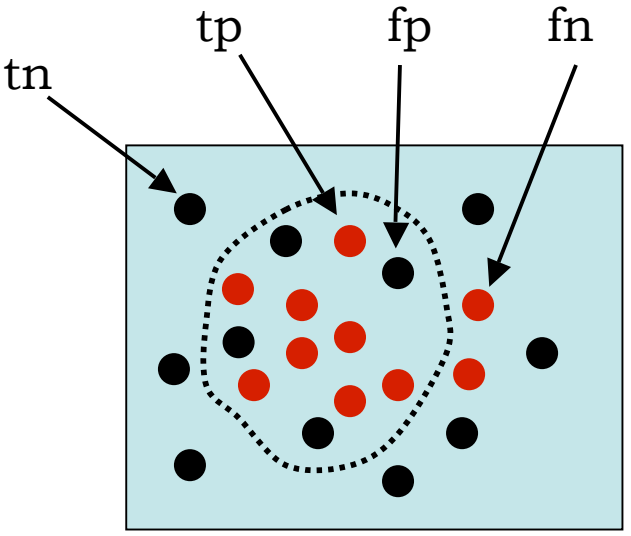
RECALL vs. PRECISION

- ☆ Recall (Trefferquote) ist der Anteil der gefundenen relevanten Informationen im Verhältnis zu allen relevanten Infomationen

- ☆ Precision (Genauigkeit) ist der Anteil der relevanten Informationen an allen gefundenen Informationen

  tp are true positives, fp are false positives
  tn are true negatives, fn are false negatives

$$\text{Recall} = \frac{tp}{tp+fn} \qquad \text{Precision} = \frac{tp}{tp+fp}$$



F -measure oder F-Maß

- ☆ ist das gewichtete harmonische Mittel von Trefferquote und Genauigkeit.
- ☆ Bei Gleichgewichtung der beiden Größen (F_1):

$$F = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall}).$$

- ☆ Generelles Maß für alle Gewichtungen α

$$F_\alpha = (1 + \alpha) \cdot (\text{precision} \cdot \text{recall})/(\alpha \cdot \text{precision} + \text{recall}).$$

Relevant Related Work and Inspirations

- ☆ (Brin, 1998) “Extracting Patterns and Relations from the World Wide Web”
  - DIPRE - Dual Iterative Pattern Relation Expansion
- ☆ Unsupervised and minimally supervised automatic acquisition methods for relation extraction patterns, e.g.,
  - (Agichtein & Gravano, 2000)
  - (Yangarber, et al., 2000), (Yangarber, 2003)
  - (Sudo et al., 2003)
  - (Greenwood & Stevenson, 2006)
  - (Xu, Uszkoreit & Li 2007)
  - (Xu, Uszkoreit & Li 2008)

Two Approaches to Seed Construction by Bootstrapping

- ☆ Pattern-oriented (e.g., ExDisco (Yangarber 2001))
  - too closely bound to the linguistic representation of the seed, e.g.,
    subject(company) v("appoint") object(person)
  - an event can be expressed by more than one pattern and by various linguistic constructions
- ☆ Relation and event instances as seeds, e.g., DIPRE (Brin 1998), Snowball (Agichtein and Gravano 2000), (Xu et al. 2006, 2007)

Our Approach: *DARE (1)*

seed-driven and bottom-up rule learning in a bootstrapping framework

- ☆ starting from sample relation instances as seeds
  - ○ complexity of the seed instance defines the complexity of the target relation
- ☆ pattern discovery is bottom-up and compositional, i.e., complex patterns are derived from simple patterns for relation projections
- ☆ bottom-up compression method to cluster and generalize rules
- ☆ only subtrees containing seed arguments are pattern candidates
- ☆ pattern rule ranking and filtering method considers two aspects of a pattern
  - ○ its domain relevance and
  - ○ the trustworthiness of its origin

Our Approach: *DARE (2)*

Compositional rule representation model

- support the bottom-up rule composition
- expressive enough for the representation of rules for various complexity
- precise assignment of semantic roles to the slot arguments
- reflects the precise linguistic relationship among the relation arguments and reduces the template merging task in the later phase
- the rules for the subset of arguments (projections) may be reused for other relation extraction tasks.

Two Domains

☆ Award Events

start with subdomain Nobel Prizes

reasons: good news coverage
complete list of all award events
good starting point for other award domains

☆ Management Succession Events

reason: comparison with previous work

Nobel Prize Awards

☆ Target relation

<*recipient*, *prize*, *area*, *year*>

☆ Example

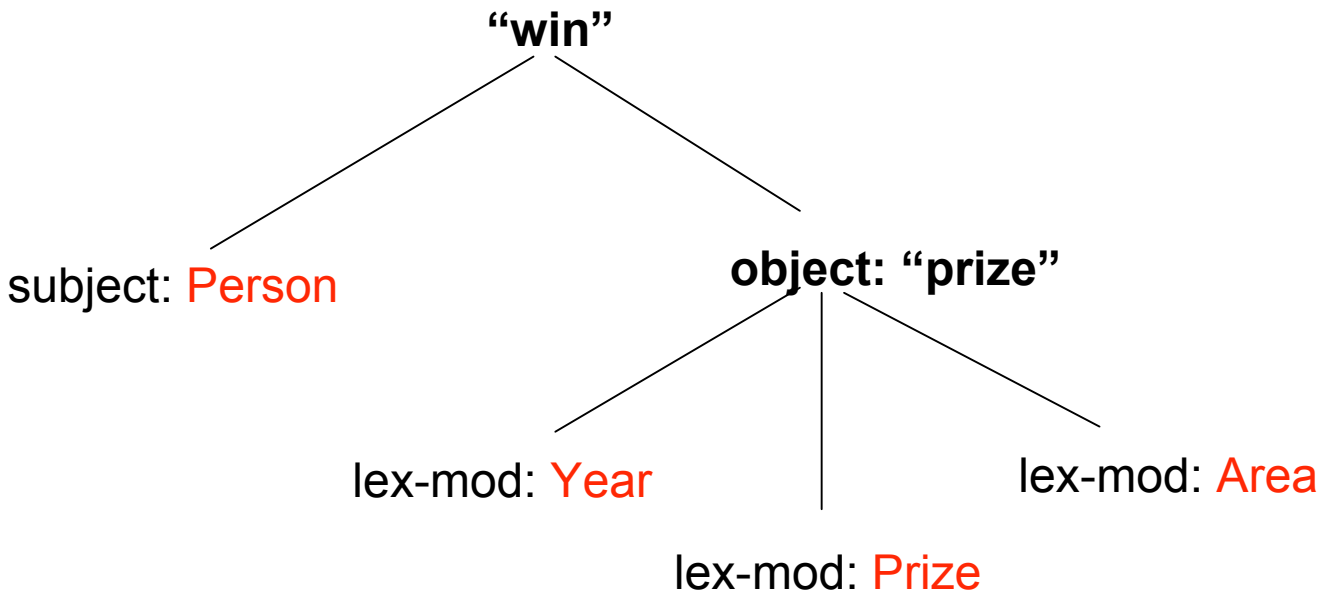
Seed: < "*Mohamed ElBaradei*", "*Nobel*", "*Peace*", "*2005*">

Sentence: Mohamed ElBaradei won the 2005 Nobel Peace Prize on Friday for his efforts to limit the spread of atomic weapons.

Rules are learned from the linguistic structure

We recognize named entities with the DFKI IE-system SProUT (Piskorski, Xu, et al.)

For sentence analysis we first employed MiniPar (Dekan Lin)



prize_area_year_1

Rule name:: prize_area_year_1

Rule body::

$$\left[\begin{array}{ll} \text{head} & \left[\begin{array}{ll} \text{pos} & \text{noun} \\ \text{lex-form} & \text{"prize"} \end{array}\right] \\ \text{daughters} & \left\langle \left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{3}\ \text{Year}\right]\right], \right. \\ & \left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{1}\ \text{Prize}\right]\right], \\ & \left.\left[\text{lex-mod} \quad \left[\text{head} \quad \boxed{2}\ \text{Area}\right]\right]\right\rangle \end{array}\right]$$

Output:: $\langle \boxed{1}Prize, \boxed{2}Area, \boxed{3}Year \rangle$

recipient_ prize_area_year_1

Rule name:: recipient_prize_area_year_1

Rule body::

$$\left[\begin{array}{ll} \text{head} & \left[\begin{array}{ll} \text{pos} & \text{verb} \\ \text{mode} & \text{active} \\ \text{lex-form} & \text{"win"} \end{array}\right] \\ \text{daughters} & \left\langle \left[\begin{array}{ll} \text{subject} & \left[\begin{array}{ll} \text{head} & \boxed{1}\ \text{Person} \\ \text{rule} & \text{recipient\_1::}\ \langle \boxed{1} Person \rangle \end{array}\right] \end{array}\right], \right. \\ & \left. \left[\begin{array}{ll} \text{object} & \left[\begin{array}{ll} \text{head} & \left[\text{lex-form} \quad \text{"prize"}\right] \\ \text{rule} & \text{prize\_area\_year\_1::}\ \langle \boxed{2} Prize, \boxed{3} Area, \boxed{4} Year \rangle \end{array}\right] \end{array}\right] \right\rangle \end{array}\right]$$

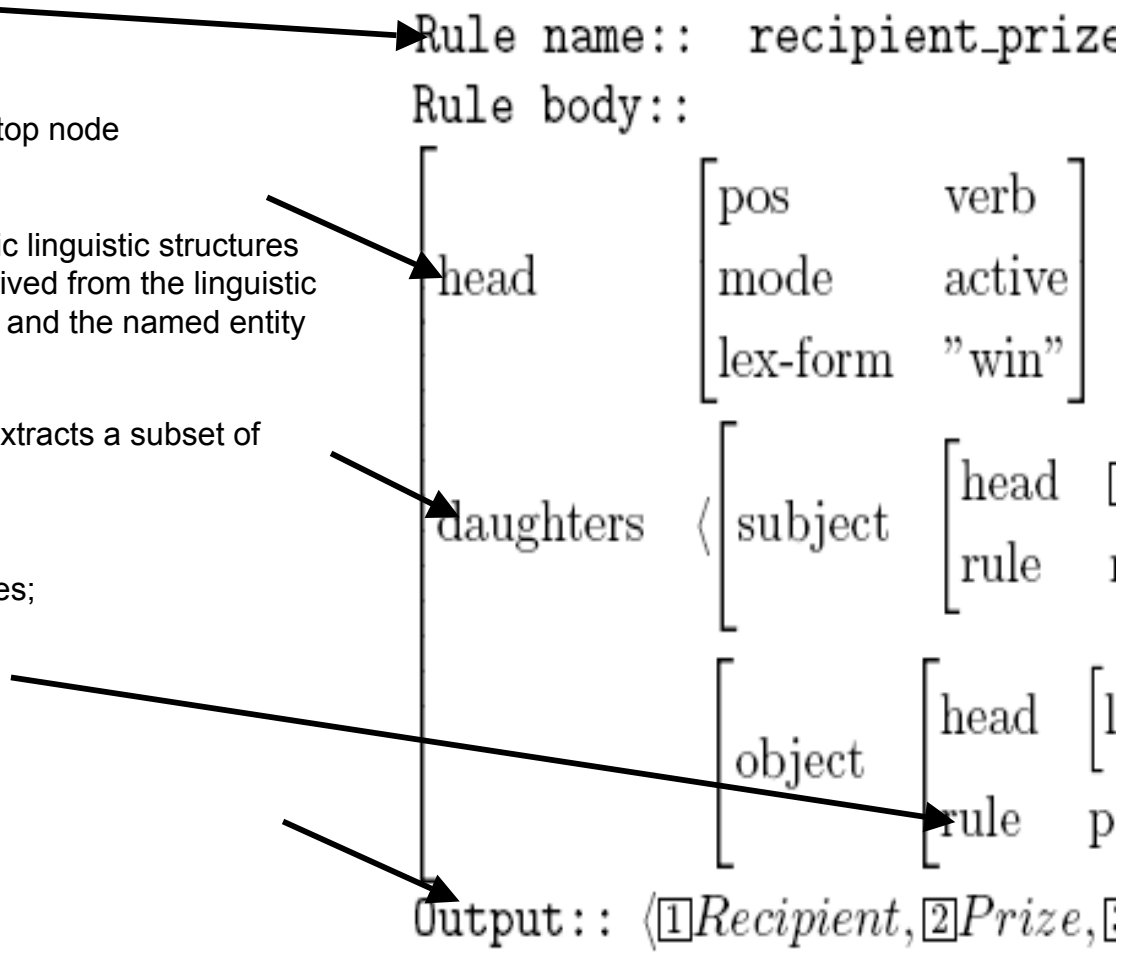
Output:: $\langle \boxed{1} Recipient, \boxed{2} Prize, \boxed{3} Area, \boxed{4} Year \rangle$

Rule Components

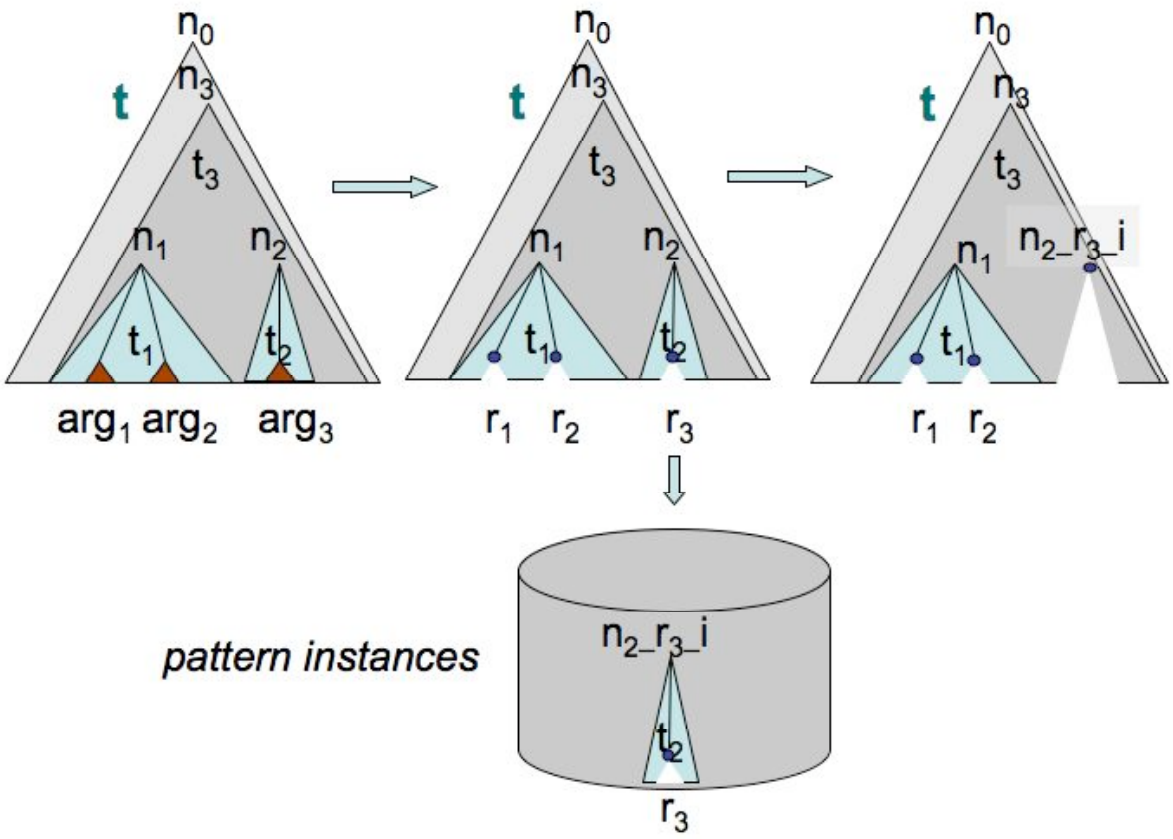
1. **rule name**: r_i;

2. **rule body:** in AVM format containing:
   - **head**: the linguistic annotation of the top node of the linguistic structure;
   - **daughters**: its value is a list of specific linguistic structures (e.g., subject, object, head, mod), derived from the linguistic analysis, e.g., dependency structures and the named entity information;
   - **rule**: its value is a DARE rule which extracts a subset of arguments of ***A***.

3. **output**: a set ***A*** contains the *n* arguments of the *n*-ary relation**,** labelled with their argument roles;

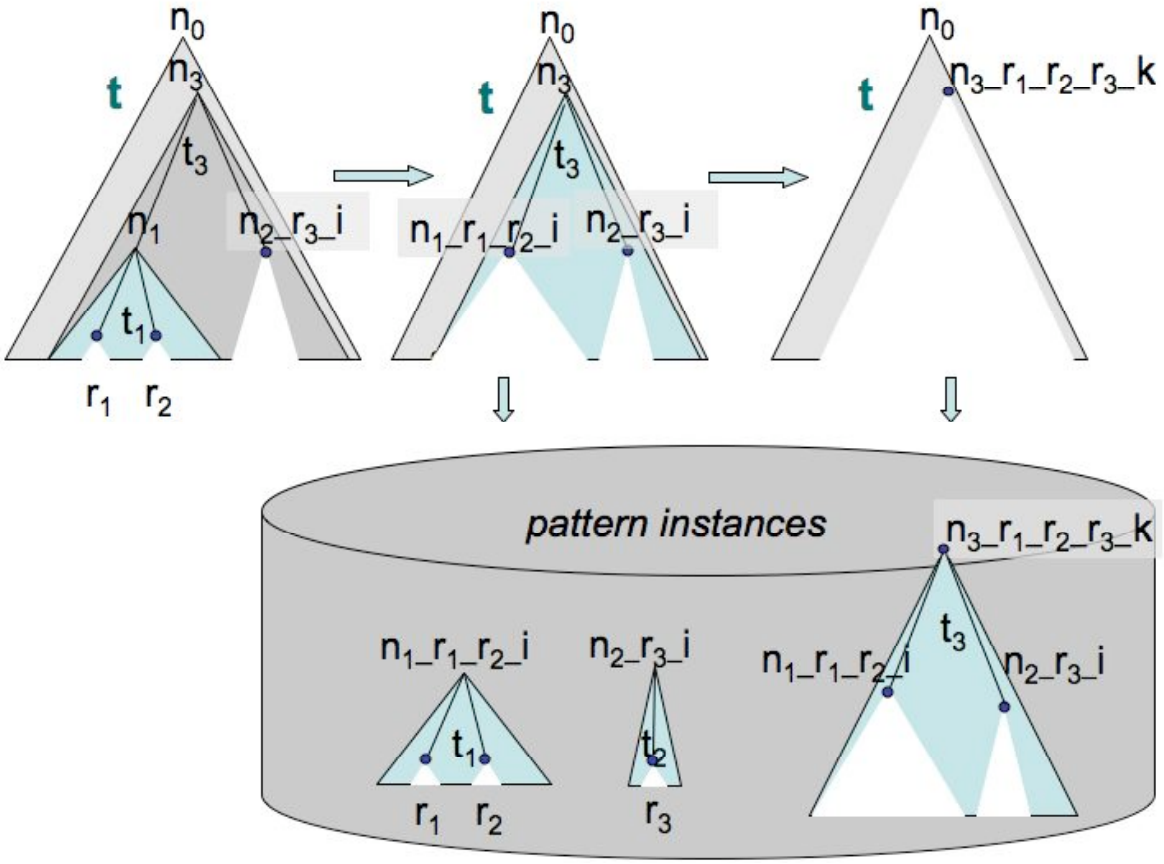


Pattern Extraction Step 1



1. replace all terminal nodes that are instantiated with the seed arguments by new nodes. Label these new nodes with the seed argument roles and their entity classes;
2. identify the set of the lowest nonterminal nodes N_1 in t that dominate only one argument (possibly among other nodes).
3. substitute N_1 by nodes labelled with the seed argument roles and their entity classes
4. prune the subtrees dominated by N_1 from t and add these subtrees into P. These subtrees are assigned the argument role information and a unique id.

Pattern Extraction Step 2



For i=2 to n

1. find a set of the lowest nodes N_i in t that dominate in addition to other children only i seed arguments;
2. substitute N_i by nodes labelled with the i seed argument role combination information (e.g., r_i_r_j) and with a unique id.
3. prune the set of subtrees T_i dominated by N_i in t;
4. add T_i together with the argument, role combination information and the unique id to P

Seed Complexity and Sentence Extent

☆ Which kind of sentences could represent an event?

complexity	matched sentence	event sentence	Relevant sentences in %
4-ary	36	34	94.0
3-ary	110	96	87.0
2-ary	495	18	3.6

Table 1. distribution of the seed complexity

Experiments

- Two domains
  - Nobel Prize Awards: *<recipient, prize, area, year>*
  - Management Succession: *<Person_In, Person_Out, Position, Organisation>*
- Test data sets

Data Set Name	Doc Number	Data Amount
Nobel Prize A (1999-2005)	2296	12.6 MB
Nobel Prize B (1981-1998)	1032	5.8 MB
MUC-6	199	1 MB

Evaluation of Nobel Prize Domain

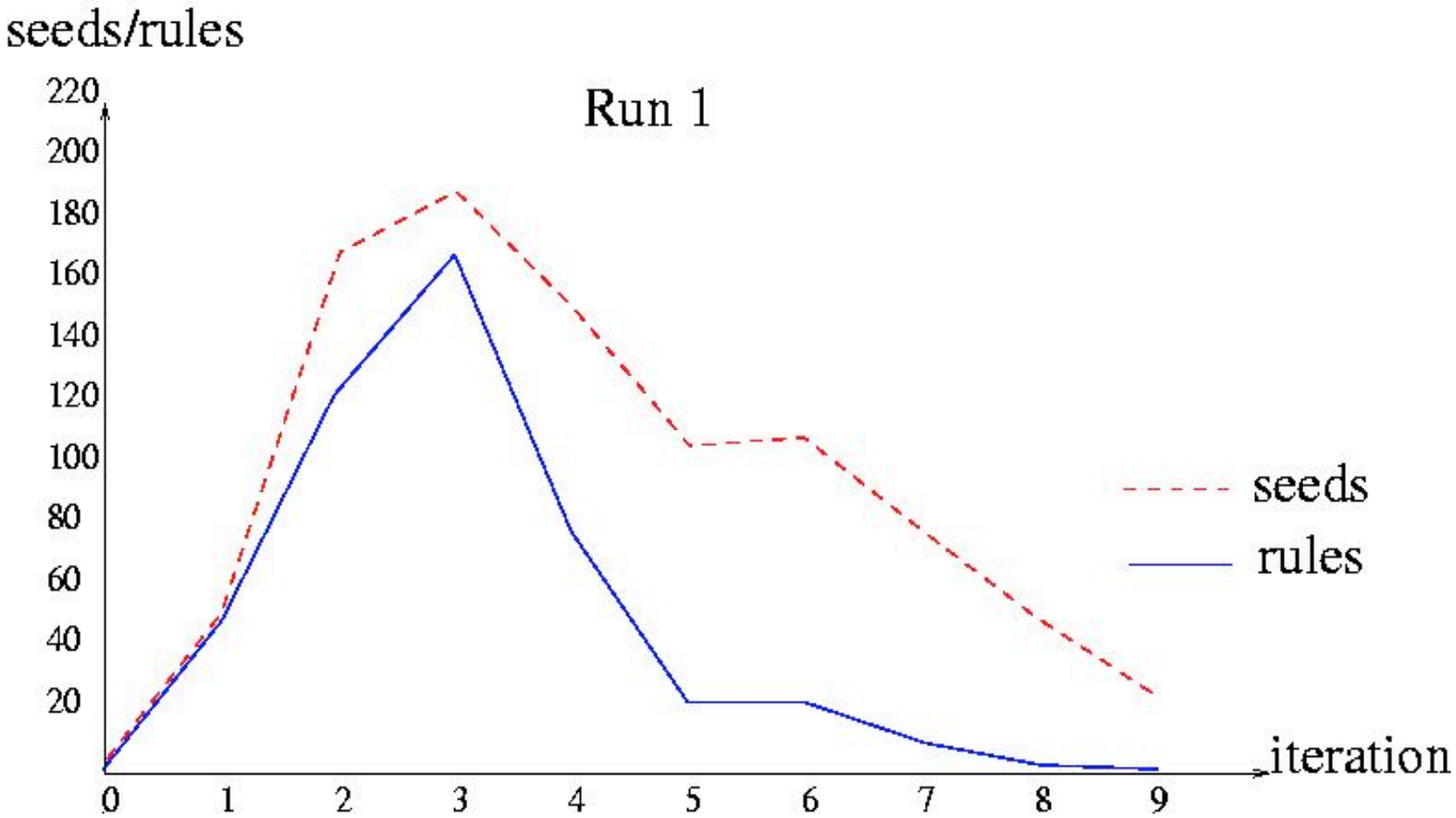
- ☆ Conditions and Problems
  - Complete list of Nobel Prize award events from online portal Nobel-e-Museum
  - No gold-standard evaluation corpus available

- ☆ Solution
  - our system is successful if we capture one instance of the relation tuple or its projections, namely, one mentioning of a Nobel Prize award event. (Agichtein and Gravano, 2000)
  - construction of so-called *Ideal* tables that reflect an approximation of the maximal detectable relation instances
    - The Ideal tables contain all Nobel Prize winners that co-occur with the word “Nobel” in the test corpus and integrate the additional information from the Nobel-e-Museum

Evaluation Against Ideal Tables

Data Set	Seed	Precision	Recall
Nobel Prize A	<[Zewail, Ahmed H], nobel, chemistry, 1999>	**71.6%**	**50.7%**
Nobel Prize B	<[Sen, Amartya], nobel, economics, 1998>	**87.3%**	**31.0%**
Nobel Prize B	<[Arias, Oscar], nobel, peace, 1987>	**83.8%**	**32.0%**

Iteration Behavior (Seed vs. Rule)



Management Succession Domain

Initial Seed #		Precision	Recall
1	A	12.6%	7.0%
	B	15.1%	21.8%
20		48.4%	34.2%
55		62.0%	48.0%

Comparison

Our result with 20 seeds (after 4 iterations)

- precision: 48.4%
- recall: 34.2%

compares well with the best result reported so far by (Greenwood and Stevenson, 2006) with the linked chain model starting with 7 hand-crafted patterns (after 190 iterations)

- precision: 43.4%
- recall: 26.5%

Reusability of Rules

- ☆ Prize award patterns
  - Detection of other Prizes such as *Pulitzer Prize*, *Turner Prize*
  - Precision: 86,2%

- ☆ Management succession
  - Domain independent binary pattern rules:
    Person-Organisation, *Person-Position*
  - Evaluation of top 100 relation instances
    - Precision: 98%

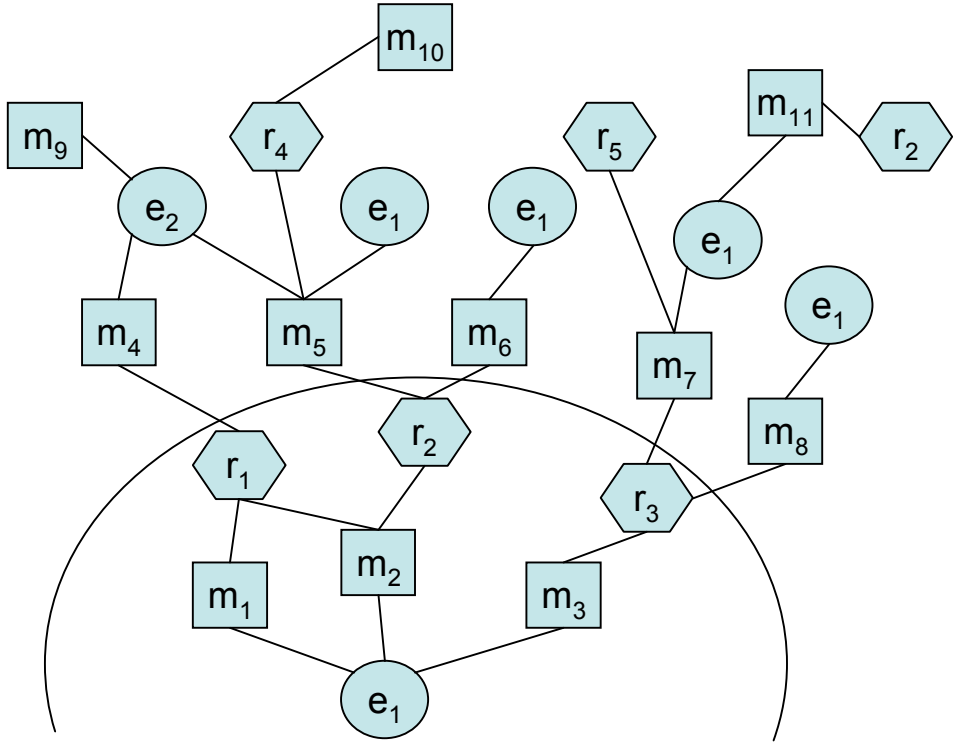
The Dream

- ☆ Wouldn‘t it be wonderful if we could always automatically learn most or all relevant patterns of some relation from one single semantic instance!
- ☆ Or at least find all event instances. (IDEAL Tables or Completeness)
- ☆ This sounds too good to be true!

Research Questions

- ☆ As scientists we want to know:
  - Why does it work for some tasks?
  - Why doesn‘t it work for all tasks?
  - How can we estimate the suitability of domains?
  - How can we deal with less suitable domains?

Start of Bootstrapping (simplified)



Abstraction

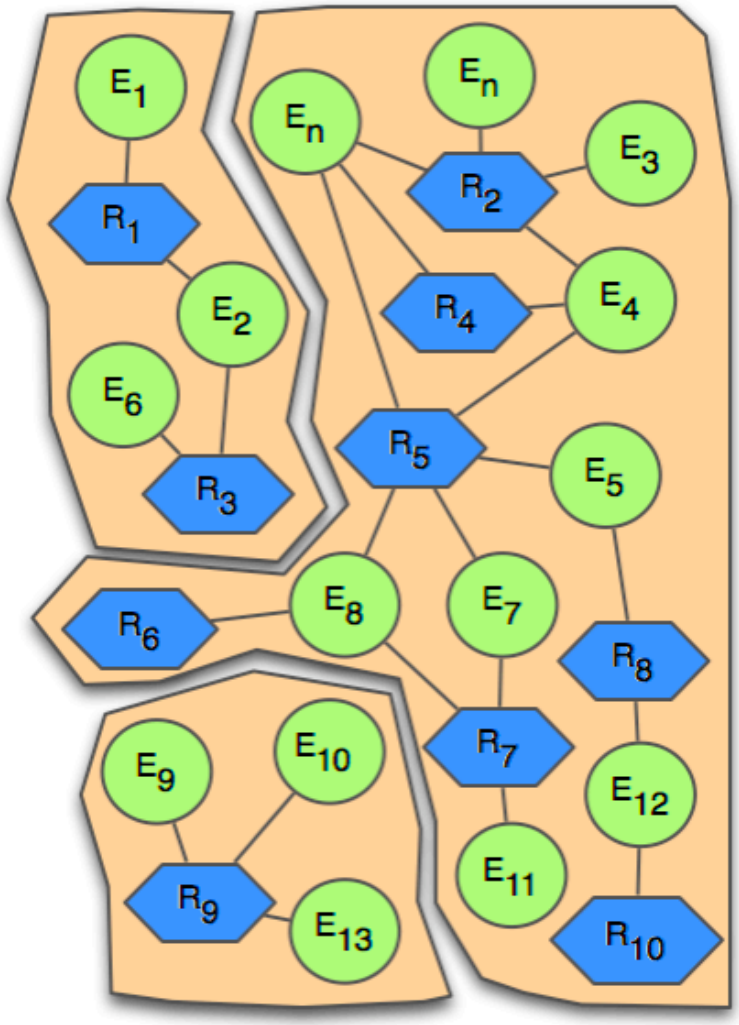
bipartite graph

two types of vertices

E_i = *event instance*
R_j = *derivable rule*

relevant properties:

- ☆ two degree distributions
- ☆ connectedness
- ☆ average and maximum path lengths between events



Questions

Can we reach all events in the graph?

By how many steps?

From any event instance?

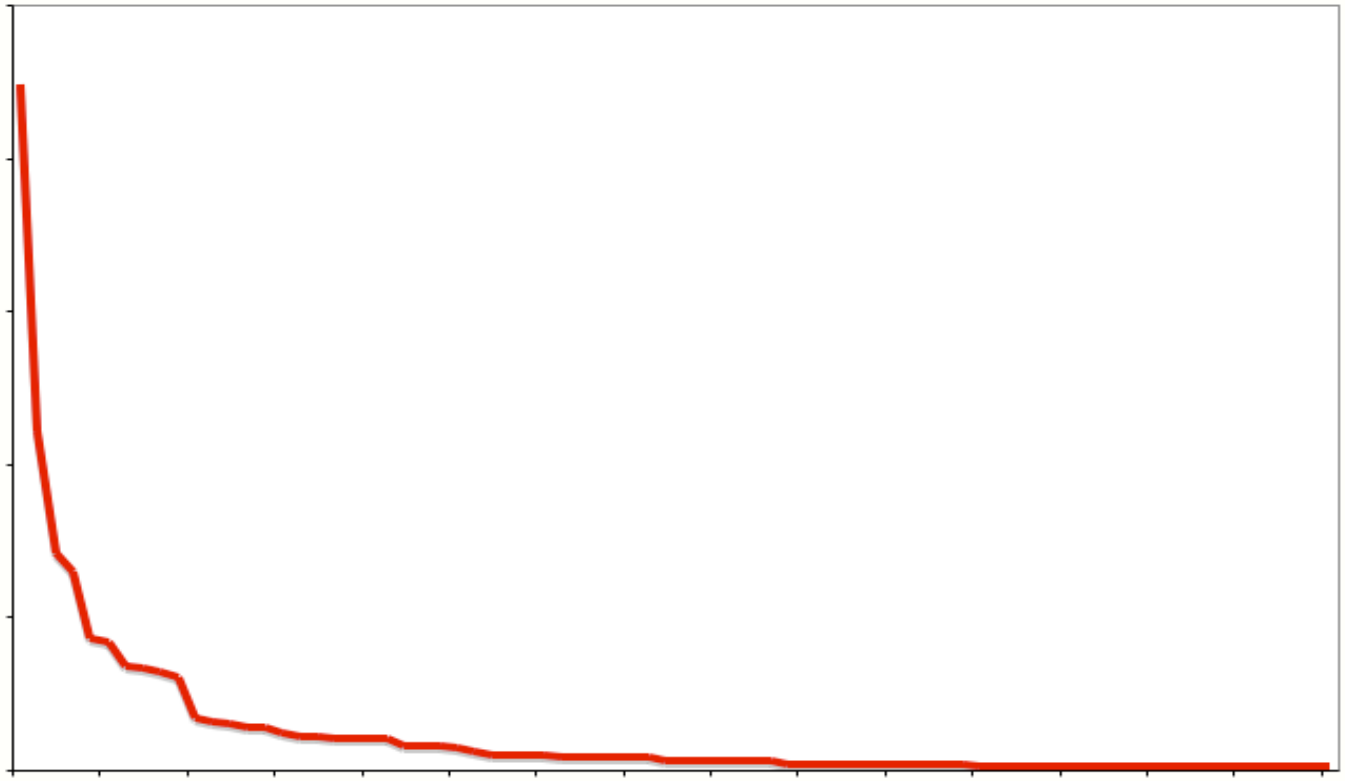
Two Distributions

1. Distributions of Pattern in Texts

2. Distribution of Mentionings to Relation Instances

Two Distributions

General distribution of patterns in texts probably follows Church‘s Conjecture: Zipf distribution (a heavy-tailed skewed distribution)



Distribution of Mentions to Events

- ☆ Distribution of mentionings to relation instances (events) differs from one task to the other.
- ☆ The distribution reflects the redundancy in textual coverage of events.
- ☆ Distribution depends on text selection, e.g. number of sources (newspapers, authors, time period)

  example 1: several periodicals report on Nobel Prize events

  example 2: one periodical reports on management succession events

Scale-Free Networks

☆ If both degree distributions are skewed, long-tailed, we may get so called scale-free networks.

$$p(k) = \sum_{v \in V \mid \deg(v)=k} 1$$

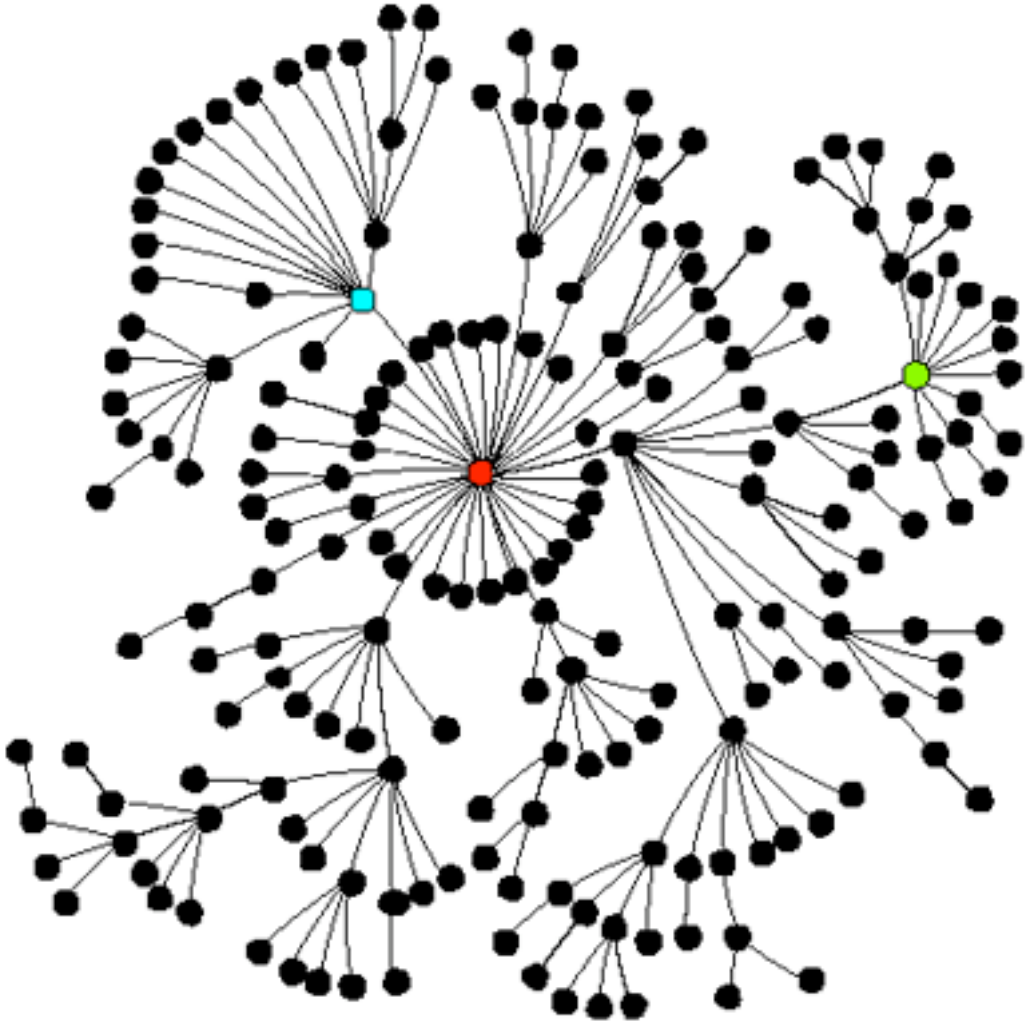
$$\mathbf{P(k) \sim k^{-\gamma}}$$

☆ with the nice properties that the usual distance between any two vertices is very short

Example of Scale-Free Nets

In scale-free networks, some nodes act as "highly connected hubs" (high degree), although most nodes are of low degree.

Scale-free networks' structure and dynamics are independent of the system's size N, the number of nodes the system has. In other words, a network that is scale-free will have the same properties no matter what the number of its nodes is.



Small-World Property

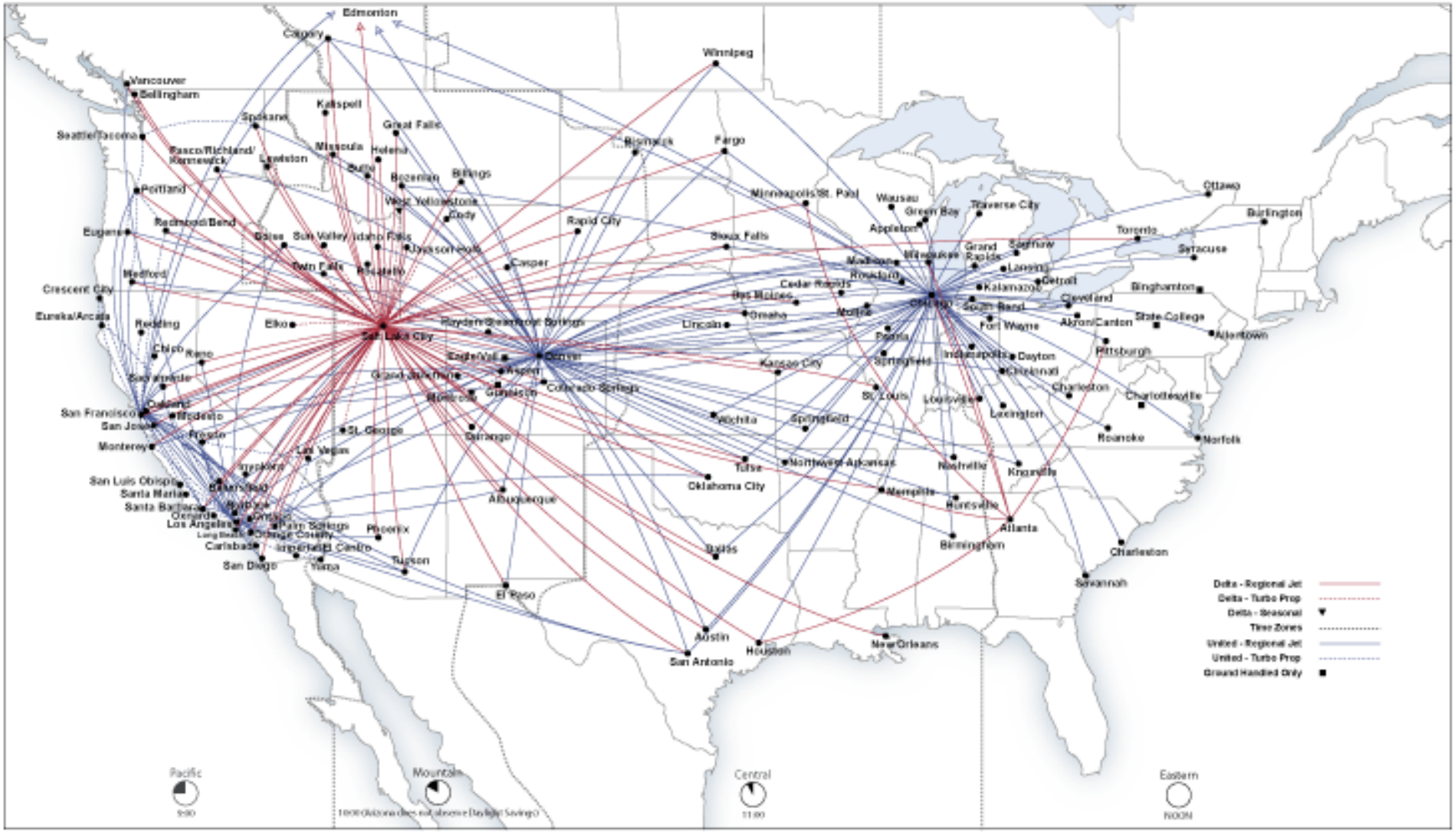
Networks exhibiting the small-world property

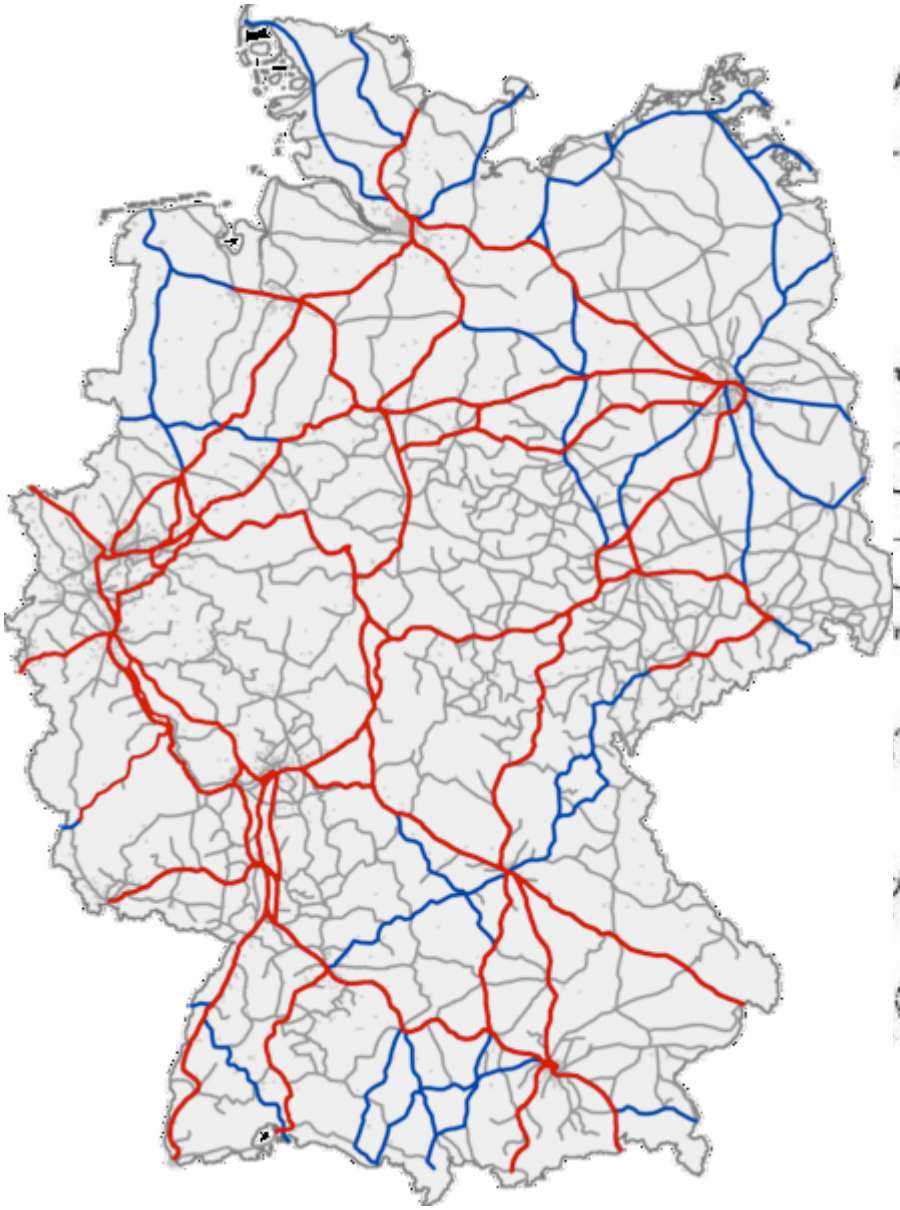
- social networks (max path-length 5-7)
- co-authorship networks (Erdös number)
- Internet
- WWW
- air traffic route maps (max. 3 hops)

Networks that do not exhibit the small-world property

- road networks
- railway networks
- kinship networks

Airline Route Networks

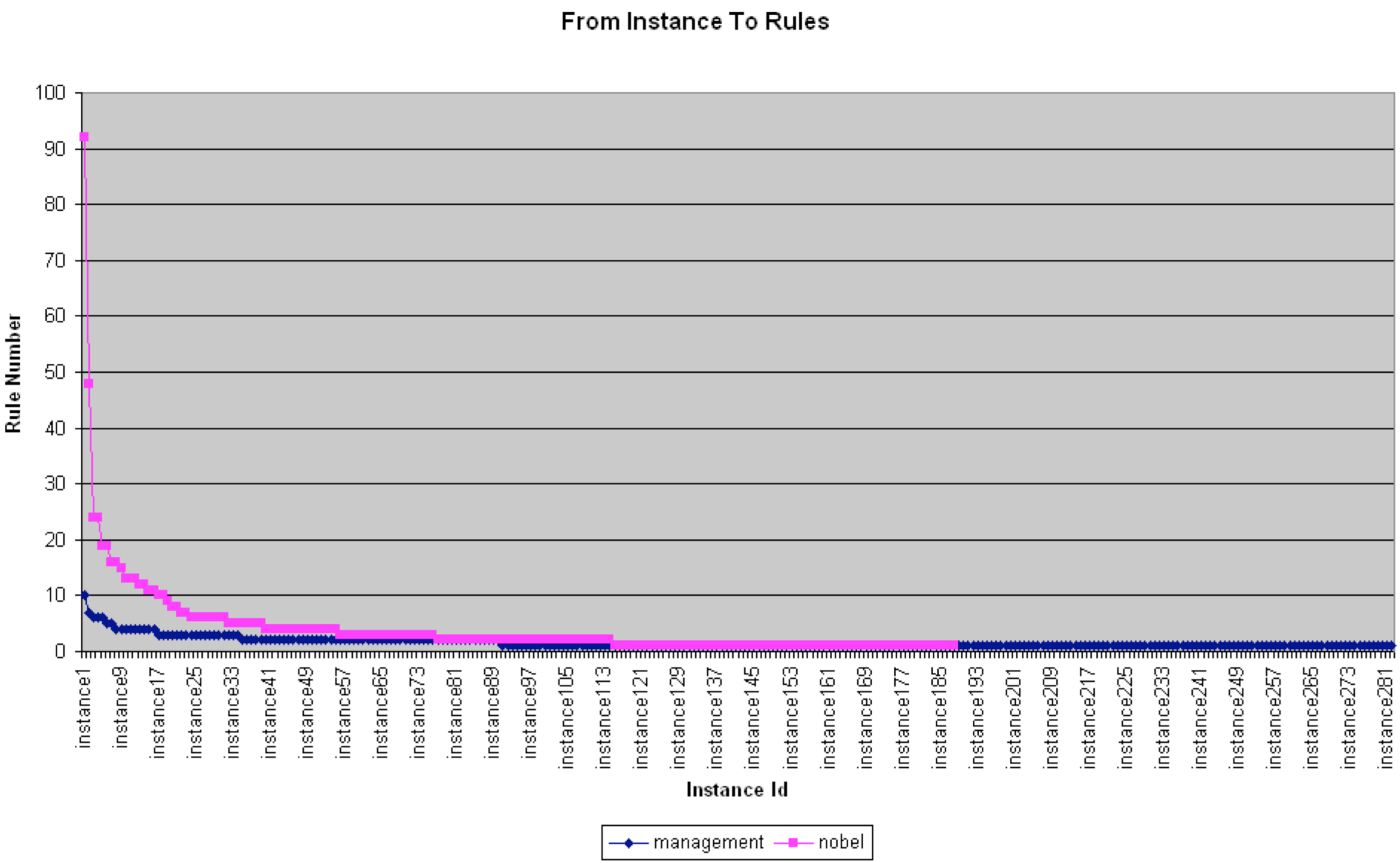




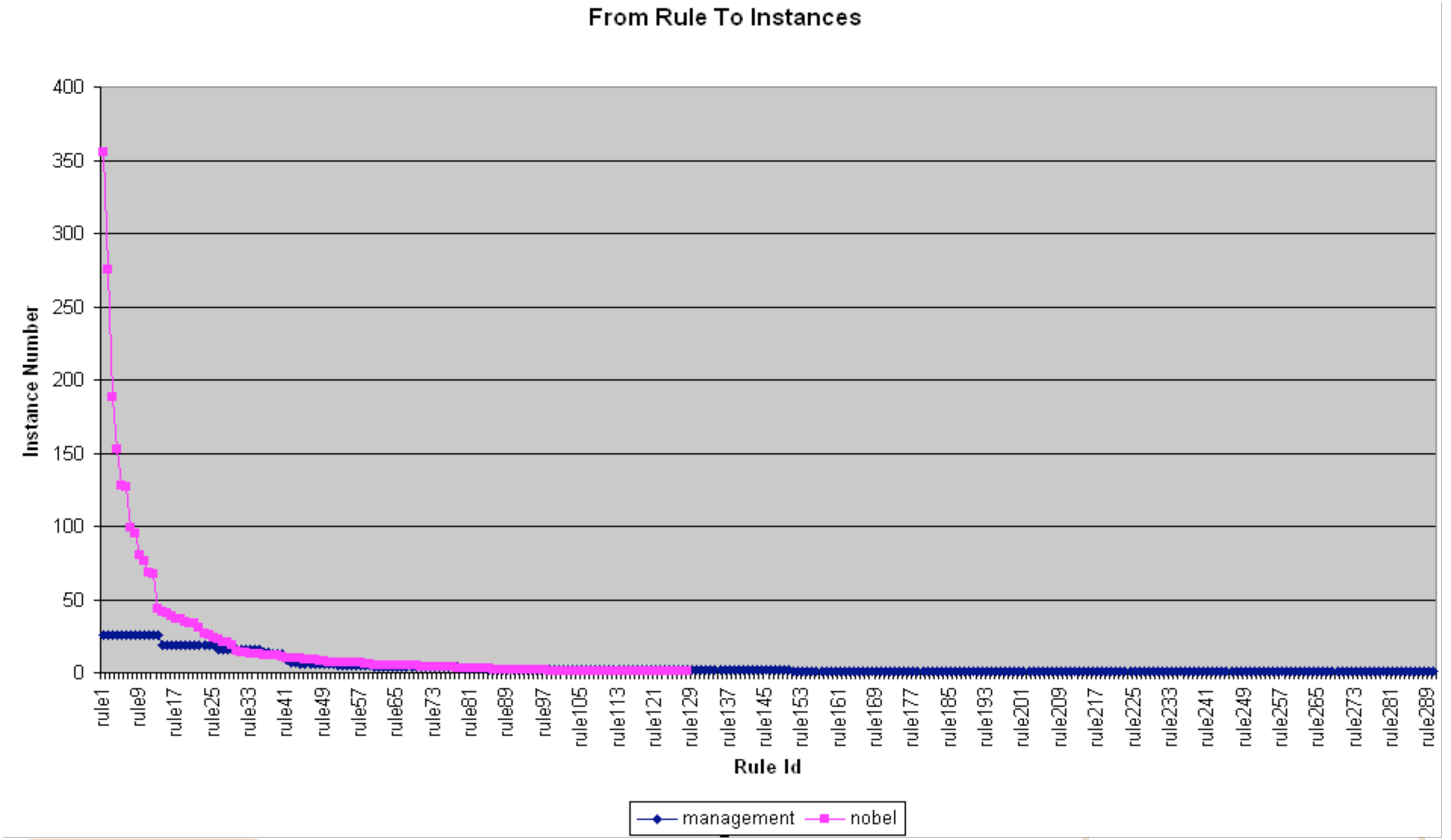
Small Worlds for Bootstrapping

- ☆ If both distributions follow a skewed distribution and if the distributions are independent from each other, then we get a scale-free network in the broader sense of the term.

- ☆ For each type of vertices we get strong hubs. This leads to very short paths (for most connections).



Rules to Instances



If we find a large world with continents and islands...

- ☆ What can we do if our domain or our data do not fit the nicely connected small world picture?
- ☆ Then we can give up and search for another domain...
- ☆ ...or we can try to change the data or relations in order to get benevolent learning graphs.

Approaches to Solve the Problem

☆ Enlarging the domain

Pulitzer Prize —> all Prizes

☆ selecting Carrier Domains (parallel learning domains)

Pulitzer Prize —> Nobel Prize
Ernst Winter Preis —> Nobel Prize
Fritz Winter Preis —> Nobel Prize

Other Discovered Award Events

Academy Award
actor % (Cannes Film Festival's Best Actor award)
American Library Association Caldecott Award
American Society
award
Blitzker
Emmy
feature % (feature photography award)
first % (the first Caldecott Medal)
Francesca Primus Prize
gold % (gold medal)
Livingston Award
National Book Award
Newbery Medal
Oscar
P.G.A
PEN/Faulkner Award
prize
reporting % (the investigative reporting award)
Tony
Tony Award
U.S. Open

But also:
nomination
$1 million
$29,000
about $226,000
praise
acclaim
discovery
doctorate
election

Further Approaches

☆ enlarging the text base for finding seeds and patterns

- New York Times MUC data –> general press corpora
- New York Times MUC data –> WWW

☆ enlarging the text base for finding new seeds

- New York Times MUC data –> WWW
- German Press Data –> English Press Data

Experiment with other Domains

- ☆ We tried to apply the method to another domain Pop Artist Gossip in an EU funded project RASCALLI

- ☆ First experiment: use the learned patterns for detecting other prize winning events such as Grammy and Music Awards

Observed Obstacles

- ☆ Obstacles to Recall
  - missing bridges between islands/continents
  - overly specific rules
  - spread over several sentences
    - missing coreferences
- ☆ Obstacles to Precision
  - intrusion of other relations
  - modality contexts

Overcoming these Obstacles

- ☆ Obstacles to Recall
  - missing bridges between islands/continents.... use of auxiliary data
  - overly specific rules..... better rule generalization
  - spread over several sentences
    - missing coreferences..... coreference resolution
- ☆ Obstacles to Precision
  - intrusion of other relations..... learning of negative rules
  - modality contexts..... learning of negative rules

Overcoming these Obstacles

- ☆ Obstacles to Recall
  - missing bridges between islands/continents.... use of auxiliary data
  - overly specific rules..... better rule generalization
  - spread over several sentences
    - missing coreferences..... coreference resolution 
- ☆ Obstacles to Precision
  - intrusion of other relations..... learning of negative rules 
  - modality contexts..... learning of negative rules 

Improving Recall: Coreference Resolution

- So far, we were missing patterns in which some role fillers are expressed as pronouns or as noun phrase anaphora
  - *He won the prize for his earth-shaking discoveries in genetic sequencing.*
  - *In the same year, the two biologists received the Nobel Prize in Medicine.*
- We included sentences in which potential role fillers occured some sentences before or after the pattern.
- We then used the domain ontology to determine whether the anaphorical phrase constituted a semantically suitable candidate for the relation and the coreference.

Coreference Resolution

domain	data size	#seed	without coref-resolution		coref-resolution Method 1		coref-resolution Method 2	
			prec.	recall	prec.	recall	prec.	recall
Nobel Prize	18.4 MB	1	86.5%	50.7%	82.7%	53.5%	83.9%	54.2%
MUC-6	1 MB	55	62.0%	48.0%	48.9%	51.6%	33.5%	52.9%

Ranking

For positive patterns and instances

- Variables
  - *i*: bootstrapping step
  - *alpha:* ranking parameter, default **0.8**
  - $rank_{ss}$ *:* rank value of the begin seed instances, default **10**
- Within the Bootstrapping
  - For each instance relation **R** from the step **I**

    $rank(R) = rank_{ss} * alpha^{i}$
- After the Bootstrapping
  - For each pattern **P**

    $rank(P) = sum(rank(R_{P\_from})) * number(R_{P\_to})$
  - For each instance relation **R**

    $rank(R) = rank(R) * log(sum(rank(P_{R\_from})))$

Ranking Method

- Definitions
  - R_{p_from} : The set of relation instances that lead to the pattern p
  - R_{p_to} : The set of relation instances extracted by using the pattern p
  - P_{r_from}: The set of pattern which extracts the relation r
  - *i*: bootstrapping step
  - *alpha:* ranking parameter (default 0.8)
  - $rank_{ss}$ *:* rank value of the begin seed instances, (default 10)

Ranking Method (1)

trustworthiness of origin

//initial variables
$rank_{ss}$ = 10
alpha = 0.8
i=0

noisy rules/seeds in the later iterations, i↑ rank↓

//within the Bootstrapping
while (|seeds|>0){
    patterns$_i$ = *getPatterns*(seeds)
    results$_i$ = *getRelations*(patterns$_i$);
    i++;
    ∀r∈results$_i$ rank(r)= $rank_{ss} * alpha^i$

    seeds = getSeeds(results$_i$)
}

Productivity:
• trustable instances to trustable rules
• trustable rules to trustable instances

connectivity

//After the Bootstrapping
patters = $\cup_i$ patterns$_i$, results = $\cup_i$ results$_i$

$\forall p \in patterns\ rank(p) = |R_{P\_to}| * (\sum_{r \in R_{p\_from}} rank(r))$

$\forall r \in results\ rank(r) = rank(r) * \log(\sum_{p \in P_{r\_from}} rank(p))$

Experiment: Who is married to whom

Iterate	Seed	Related Sentence	Rules	Extracted Relation
0	1	1	1	125
1	120	103	171	391
2	362	645	1001	1279
3	1204	1394	2102	1150
4	1071	702	970	309
5	262	126	198	75
6	75	36	53	22
7	19	7	9	0

Experiment: Who is married to whom

- ☆ Extraction Result
  - Total 3113
    - rank>=7.0: Total:233, Correct 190, Incorrect 43
      - Precision 81,55%
    - rank>=6.0: Total 270, Correct 226, Incorrect 44
      - Precision 83,70%

	r>=8.0	7.0<= r <8.0	6.0<= r <7.0	5.0<= r <6.0	4.0<= r <5.0	3.0<= r <4.0	2.0<= r <3.0	1.0<= r <2.0	r<=1
Total	120	113	37	376	345	881	565	510	166
Correct	117	73	36	-	-	-	-	-	-
Incorrect	3	40	1	-	-	-	-	-	-
Precision	97,50%	64,60*	97,30%	-	-	-	-	-	-

* The most errors are from the rule *meet (sub, obj)*

Extraction Result – Rules

- ☆ Total: 4505 Useful: 584
- ☆ Top Rules + ¯Ranking
  - *“wife” (mod: “his”<person>, mod: <person>)* 4248
  - *“star” (subj: <person> , dep: <person>)* 3744
  - *“husband” (mod: “his”<person>, mod: <person>)* 3680
  - *“meet” (subj: <person> , obj: <person>)* 2640
  - *“play” (subj: <person> , dep: <person>) 2293*
  - *“play” (subj: <person> , obj: <person>)* 1582
  - *“role” (mod: “his”<person>, mod: <person>)* 1474
  - *“father” (mod: “his”<person>, mod: <person>) 1292*
  - *“marry” (sub: <person>, obj: <person>)* 1270
  - *“appear” (subj: <person> , dep: <person>) 1126*
  - *“marry” (sub: <person>, dep(prep_to): <person>) 936*
  - *……*

Negative Examples

☆ *The talk has included speculation [that North Korean leader Kim Jong Il and South Korean President Kim Dae-jung might win the Nobel Peace Prize for their step toward reconciliation, the most promising sign of rapprochement since the Korean war ended with a fragile truce in 1953]*

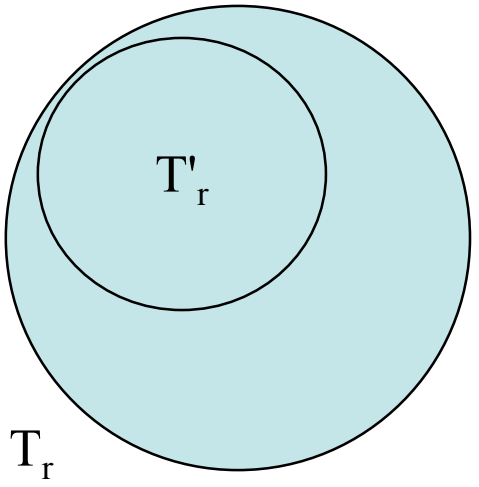
☆ *It's also possible [that O.J. Simpson will find the real killer, that Bill Clinton will enter a monastery and that Rudy Giuliani will win the Nobel Peace Prize.]*

Distinctiveness

☆ Let $r=<a_1,a_2,...,a_n>$ be an instance of an n-ary relation R.

☆ Then T_r the tuple-mention set of r is the set of all segments (i.e., sentences) in which mentions of all arguments are detected. T'_r is a subset of T_r containing exactly those segments that actually intensionally refer to the instance of r. The ratio T_r/T'_r we call distinctiveness. The more distinctive a seed, the better for precision.

$r=<a_1,a_2,...,a_n>$



KOSTENEINSPARUNG

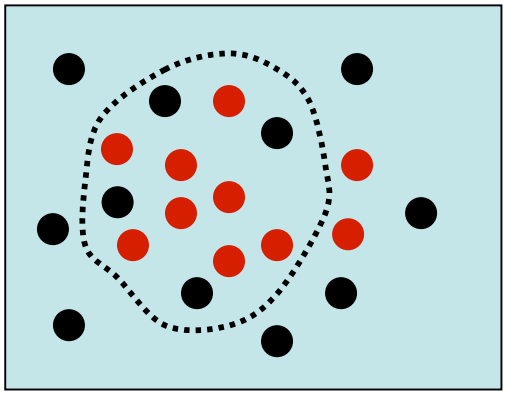
- ☆ Wenn die Notwendigkeit der Beobachtung akzeptiert wird, dann kann man die Kosten für die manuelle, intellektuelle Beobachtung ohne zusätzliche Hilfsmittel ermitteln.
- ☆ Das ist schwer, weil es ohne technologische Hilfsmittel natürlich gar nicht geht.

RECALL vs. PRECISION

- ☆ Recall (Trefferquote) ist der Anteil der gefundenen relevanten Informationen im Verhältnis zu allen relevanten Infomationen

- ☆ Precision (Genauigkeit) ist der Anteil der relevanten Informationen an allen gefundenen Informationen

Beispiel:
8 von 10 gefunden = 80,0% Trefferquote
8 von 12 korrekt = 66,6% Genauigkeit



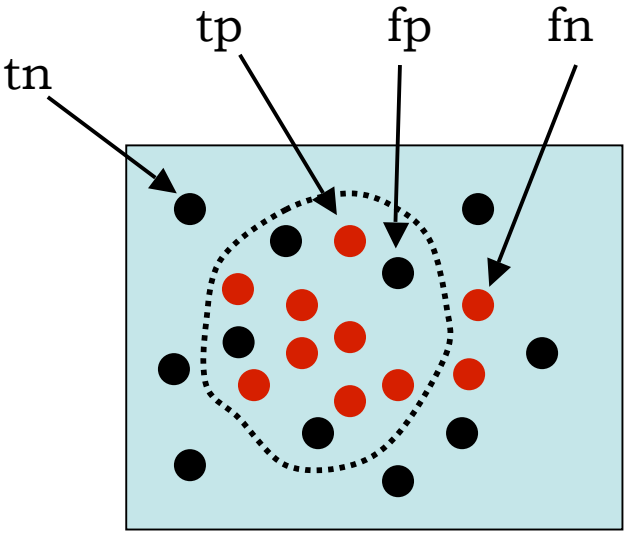
RECALL vs. PRECISION

- ☆ Recall (Trefferquote) ist der Anteil der gefundenen relevanten Informationen im Verhältnis zu allen relevanten Infomationen

- ☆ Precision (Genauigkeit) ist der Anteil der relevanten Informationen an allen gefundenen Informationen

  tp are true positives, fp are false positives
  tn are true negatives, fn are false negatives

$$\text{Recall} = \frac{tp}{tp+fn} \qquad \text{Precision} = \frac{tp}{tp+fp}$$



F -measure oder F-Maß

- ☆ ist das gewichtete harmonische Mittel von Trefferquote und Genauigkeit.
- ☆ Bei Gleichgewichtung der beiden Größen (F_1):

- ☆ Generelles Maß für alle Gewichtungen α

$$F = 2 \cdot (\mathrm{precision} \cdot \mathrm{recall})/(\mathrm{precision} + \mathrm{recall}).$$

$$F_\alpha = (1 + \alpha) \cdot (\mathrm{precision} \cdot \mathrm{recall})/(\alpha \cdot \mathrm{precision} + \mathrm{recall}).$$

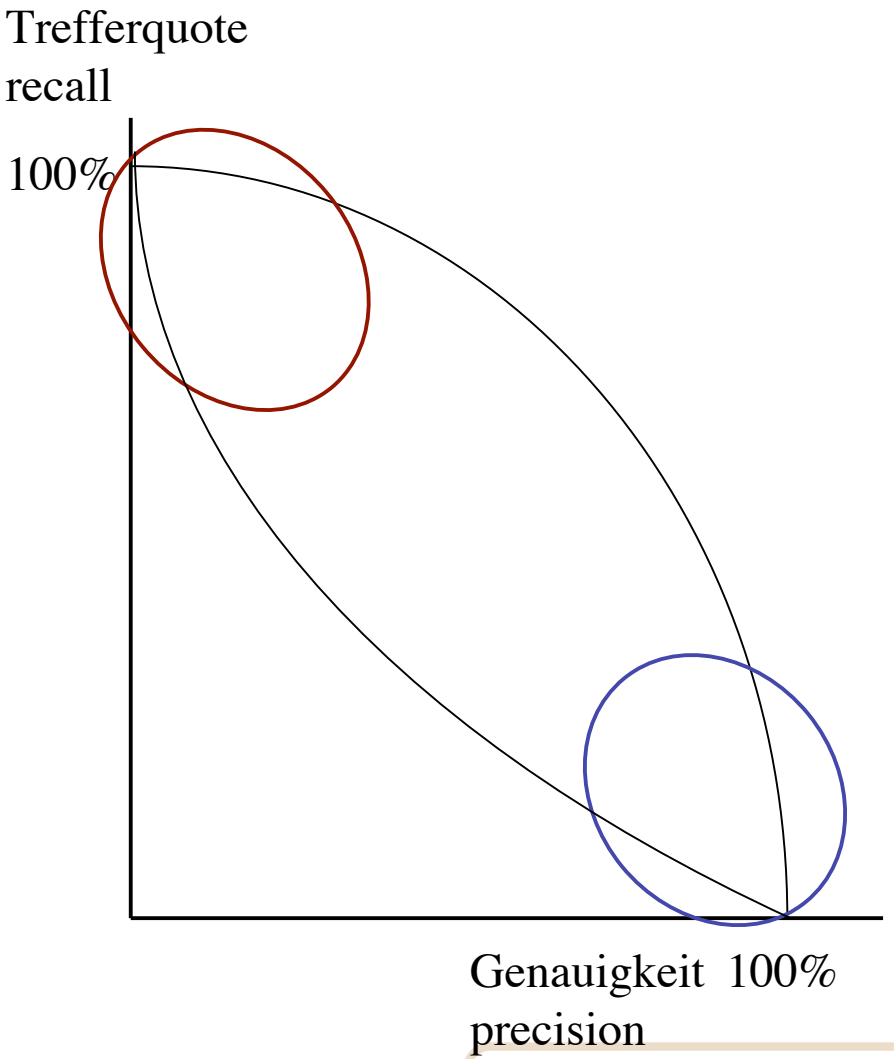
ERKENNUNGSLEISTUNGEN

- ☆ F-Maß (F-Measure) harmonisches Mittel aus Trefferquote und Genauigkeit
- ☆ Stand der Kunst:
  - Entitätenerkennung ja nach Klasse zwischen 60% und 92%
  - Relationserkennung je nach Klasse zwischen 40% und 69%

VERSCHIEDENE AUFGABENBEREICHE

- Interessanter Bereich für den Sicherheitsbereich, für Business-Intelligence und alle Frühwarn-anwendungen

- Interessaner Bereich für vollautomatische Sammel- und Statistikaufgaben, z.B. Trendanalysen



EIN EINFACHES BEISPIEL

- ☆ Wenn jeden Tag 100.000 Dokumente durchsucht werden, und 100 gefunden werden, von denen 1 relevant ist, dann ist das bei 1% Precision immer noch ein sehr nützliches System, solange der Recall bei 100% liegt.

- ☆ Wenn dieses eine Dokument gefunden werden muss, dann ist die Kostenersparnis immens.