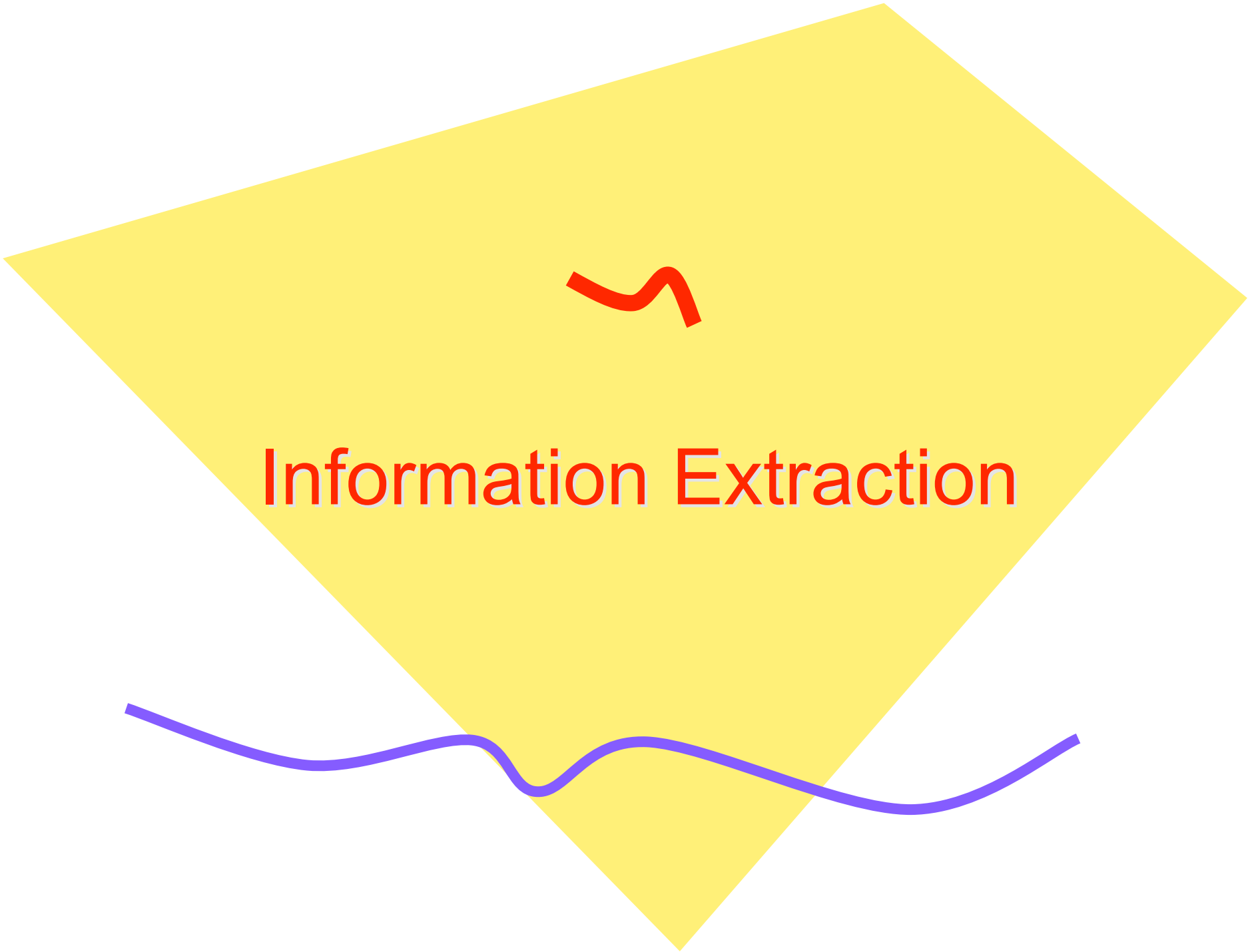

Information Extraction

“Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, providing additional aids to access and interpret the unstructured data by information systems.”

- We want to assign semantic meaning to content:
  - **Text**
  - Speech
  - Images
  - Video
  - Audio, ...

  e.g., in the form of **semantic labels**

Aims

- To understand information extraction **tasks**
- To understand the information extraction **methods**
- To illustrate with examples and have insight in the current **performance**
- To **link** information extraction **to information retrieval**
- To understand the current **problems** and research avenues in information extraction

Overview

1: Introduction
2: Symbolic and machine learning techniques
3: Current problems and possible solutions
4: IE in entity retrieval, exploratory search and question answering

- All parts are illustrated with examples, e.g., w.r.t. opinion mining, named entity recognition, semantic role labeling, extraction of temporal information, extraction of spatial information and scenario recognition

Why do we need IE?

- Huge amounts of unstructured data in a variety of media, languages and other formats
- Need for the machine to help:
  - Retrieval, search, question answering
  - (Web) mining
  - Summarizing and synthesis
  - Exploratory search and visualization

Anytime, anyplace

- How good is the machine already? Some examples

Named entity recognition

December 4, 2005: 12:19 AM EST

SYDNEY (Reuters) - Qantas Airways Ltd., which will seek board approval this week to spend up to A$20 billion ($15 billion) on new planes, said on Sunday the contest between rival manufacturers was the closest it has seen.

Planemakers Boeing and Airbus are set to end a record year for new plane orders with a decision from Australia's Qantas, which has said it might need as many as 100 new planes.

Qantas chief executive Geoff Dixon said management had not decided on a final recommendation for Wednesday's board meeting, but fleet renewal was essential for the carrier, which also wants to expand its low-cost Jetstar airline internationally …

…

Relation extraction

Motorola, Inc. (NYSE: MOT) announced that the company and General Instrument Corporation (NYSE: GIC) completed their previously announced merger following GIC shareholder approval at a special meeting held Wednesday.

Information extraction: more examples

- Named entity recognition and their relationships
  - e.g., Person **X** *works* for company **Y**, company **A** *merges* with company **B**
- Extraction of details of an event:
  - e.g., type of event, time, location, number of victims, symptoms of a disease
- Extraction of information on Web page:
  - e.g., e-mail, date of availability of a product, …
- Extraction of scientific data from publications:
  - e.g., localization of a gene, treatment of a disease, function of a gene
- Opinion mining:
  - e.g., recognition of positive or negative feelings expressed towards a consumer product
- Recognition of scenarios
  - e.g., recognition of a bank robbery

Beyond text ...

Birthday party

Happyness

Cake

Candles

Information extraction from text

- Relies on pattern recognition algorithms
- Relies on progress in general natural language processing
- Relies on increasing available computational power
- Relies on interest in biomedical domain, intelligence services, business intelligence, ...

Information extraction from text

- Information extraction relies on **pattern recognition** techniques:
  - **Features:**
    - **lexical**: e.g., words
    - syntactical (language dependent): e.g., POS-tags, parse tree information
    - semantic: e.g., from lexico-semantic resources, obtained in previous extraction tasks
    - discourse: e.g., discourse distance
    - other: e.g., HTML tags
  - of the information unit to be classified and its context

Information extraction

- Classification scheme = semantic labels and their relationships (external knowledge)
  - **Domain-independent**: e.g., semantic roles
  - Often **domain-dependent**: e.g., biomedical name classes
  - Cf. ontology

The symbolic approaches

- Symbolic approaches rely on **symbolic handcrafted knowledge**
  - drafted by a knowledge engineer, possibly helped by expert
  - based on moderate-sized corpus that is manually inspected
- Intuitive approach for extracting information from natural language texts

Early origin

- ◦ end 1960s and 1970s: [Schank 1972, 1975]:
  - defines all natural language words in terms of elementary primitives or predicates in an attempt of capturing the semantic content of a sentence
  - **conceptual dependency** representation specifies **semantic roles:** the **action** of the sentence (e.g., as reflected by the verbs of the text) and the **arguments** (e.g., agent, object) and **circumstances**
  - representations are ordered in a **script** or scenario which outlines sequences of events or actions

Script: human (X) taking the bus to go from LOC1 to LOC3

1. X **PTRANS** X **from** LOC1 **to** bus stop
2. bus driver **PTRANS** bus **from** LOC2 **to** bus stop
3. X **PTRANS** X **from** bus stop **to** bus
4. X **ATRANS** money **from** X **to** bus driver
5. bus driver **ATRANS** ticket **to** X
6. Various subscripts handling actions possible during the ride.
7. bus driver **PTRANS** bus **from** bus stop **to** LOC3
8. X **PTRANS** X **from** bus **to** LOC3

[Schank 1975]

X gives money to the bus driver. ATRANS is used to express a transfer of an abstract relationship, in this case the *possession* of money.

(3), (7), (8): mandatory

Frame-based approaches

- [Minsky 1975]: frame-based knowledge representations
  - frames are often triggered by the occurrence of a certain word or phrase
  - very **partial analysis** of the input text:
    - algorithm tries to match natural language sentences with particular frames by simply filling out the slots in accordance with the constraints placed on them
    - often top-down (expectation-driven): guided by the expected patterns to be found in the text
    - robust: ignoring of irrelevant information
  - **template frames** that outline the information can be used as **output**

Frame-based approaches

- Patterns to be identified can be encoded as regular expressions and recognized by finite state automaton
- Frames are often organized in a **script:**
  - because of their strict organization, scripts have good predictive ability useful in information extraction

- Examples of some famous information extraction applications:
  - FRUMP: Yale University
  - FASTUS: Stanford Research Institute

FASTUS

- Finite state automaton implementation: **set of cascaded, non-deterministic finite-state transducers**
  - application of symbolic rules in the form of hand-crafted regular expressions
  - cascade: output of finite state transducer is input for next finite state transducer

[Hobbs et al. 1996] [Hobbs JBioInformatics 2002]

Cascade of finite state transducers

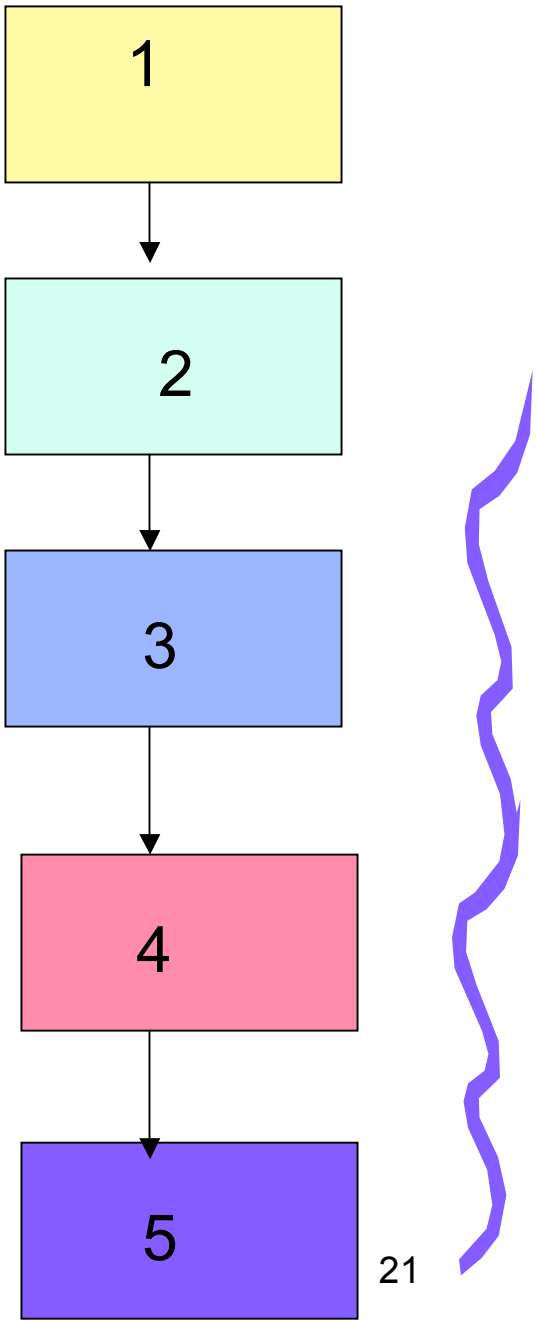
1. Recognition of compound words and named entities

2. Partial parse: recognition of verb, noun, prepositional phrases, actives, passives, gerunds

3. Recognition of complex noun groups

4. Resolution to active form, recognition of information to be extracted

5. Structure merging



- Example sentence:

  Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

Step 2

Company nam e	**Bridgestone Sports Co.**
Verb group	**s a i d**
Noun group	**Frid a y**
Noun group	**i t**
Verb group	**had set up**
Noun group	**a joint venture**
Preposition	**i n**
Location	**Taiwan**
Preposition	**with**
Noun group	**a local concern**
A n d	**a n d**
Noun group	**a Japanese trading house**
Verb group	**to produc e**
Noun group	**golf clubs**
Verb group	**to be shipped**
Preposition	**t o**
Location	**Japan**

Step 4

Extraction rules:

<Company/ies> {Set-up} {Joint-Venture} {with} <Company/ies>
{Produce} <Product>

Relation:	TIE-U P
Entities:	**Bridgestone Sports C o . a local concer n a Japanese trading house**
Joint Venture Company:	
Activity:	
Amount:	

Activity:	PRODUCTION
Company:	
Product:	**golf clubs**
Start Date :	

Symbolic techniques: results

- Successful systems, built and tested in many subject domains
- e.g., MUC-7 (1998): subject domain of air plane crashes:
  - performance of individual systems: largely similar
  - certain information much easier to extract than others
- **Problem:**
  - infinite variety of subject domains: very difficult to exhaustively implement the symbolic knowledge
  - very difficult to construct a script for every conceivable situation

Table 2: Maximum Results Reported in MUC-3 through MUC-7 by Task

Evaluation\Tasks	Named Entity	Coreference	Template Element	Template Relation	Scenario Template	Multilingual
MUC-3					R < 50% P < 70%	
MUC-4					F < 56%	
MUC-5					EJV F < 53% EME F < 50%	JJV F < 64% JME F < 57%
MUC-6	F < 97%	R < 63% P < 72%	F < 80%		F < 57%	
MUC-7	F < 94%	F < 62%	F < 87%	F < 76%	F < 51%	
Multilingual						
MET-1	C F < 85% J F < 93% S F < 94%					
MET-2	C F < 91% J F < 87%					

Legend: R = Recall P = Precision F = F-Measure with Recall and Precision Weighted Equally
E = English C = Chinese J = Japanese S = Spanish
JV = Joint Venture ME = Microelectronics

What do we learn from the symbolic approaches?

- They are very useful in case:
  - the knowledge can be easily manually crafted
  - the knowledge is stable and can be used in many applications
  - the knowledge patterns are unambiguous
- Illustrated with opinion mining

Opinion mining

- **User generated content** (e.g., on World Wide Web) often contains opinionated content:
  - Review sites, fora, discussion sites, blogs: word-of-mouth dissemination of information
  - Can be mined for opinions, sentiments, radicalism, defamation, etc.
- **Mine the opinions**:
  - Practical very useful
  - Challenging research

Opinion mining

- Often reliance on lists of handcrafted expressions that signal positive or negative feelings:

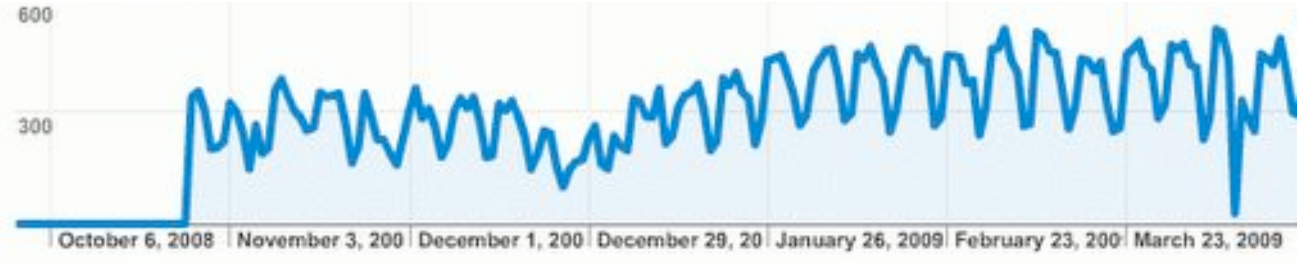
Ashamed	Bo	Beatific	Blessed	bused	Cynical
Beaten down	Co	Beatify	Blessing	fraid	Guarded
Cut down	Im	Beatitude	Bliss	ttacked	Skeptical
Criticized	Im	Beauteous	Bloom	efensive	Suspicious
Dehumanized	Inh	Beautiful	Blossom	rightened	Untrusted
Disrespected	Inv	Beautify	Bonafide	secure	Untrusting
Embarrassed	Fo	Benefaction	Bonanza	timidated	
Humiliated	Ma	Beneficial	Bonus	ver-protected	
Inferior	Ob	Befriend	Boost	cared	
Insulted	Ov	Benefit	Bountiful	errified	
Invalidated	Ov	Benevolent	Bounty	hreatened	
Labeled	Po	Beauty	Bright	nder-protected	
Lectured to	Pre	Beloved	Brighten	nsafe	
Mocked	Re	Best	Brill	iolated	
Offended	Su	Bestow	Brilliant		
		Better	Bubbly		
		Betterment	Budding		
		Big	Buddy		
		Bijou	Build		
		Bless			

Opinion mining

Land Rover Range Rover Sport Massively improved, right up there with the nicer sporty-SUVs.

Just one thing: it's not a Range Rover.
The best Evo ever, Makkinen included.
But at 50,000 you've REALLY got to want one.

Mining: dashboard



Opinion mining

- At the document level
- At the **sentence level**:
  - Recognizing the opinion expressed towards a certain product and its attributes
  - Not only extraction of the relation between an opinion and an object of interest, but also dealing with comparisons:
    - In one sentence: " ... *certainly more rewarding and comfortable than an Audi Q7* ..."
    - Also possible across sentences
  - Needs some form of linguistic processing

[Pang & Lee IR 2008]

Nokia 5800 XpressMusic

Nokia's first touchscreen phone is impressively good. It may not have quite the polished feel of the iPhone, but with so many good features, it's in many ways better than the iPhone. Headline features include the very high resolution touchscreen with 16 million colours, the 3.2 megapixel camera with Carl Zeiss optics, 3G video calling, an excellent web browser with support for Flash, the music player & FM radio, GPS positioning with Nokia Maps, WiFi, Bluetooth and an 8GB memory card. Available in black, red or blue.

★★★★★ Outstanding

[http://www.mobile-phones-uk.org.uk/]

FEATURED REVIEW

by Shelia -- December 3rd 2008 -- Burlington, NC

I have had my Samsung eternity for a week today. I have had other cell phones nothing is comparable to the Samsung eternity. I love how it feels in your hands and everything about it. I have wide fingers, and it is so easy to use, it is hard for me to put down. I think the name of this phone has a meaning, I plan to have this cell phone as long as I can. My rating is higher then a 10... My husband is planning to buy a Samsung eternity cell phone. This month.

Machine learning techniques

- IE in terrorism domain: experiment of Riloff (1996): automatic construction of dictionary of extraction patterns from an annotated training corpus achieved 98% of the performance of handcrafted patterns

=> **machine learning of extraction patterns**

Machine learning techniques

- Usually **supervised learning algorithms**:
  - e.g., learning of rules and trees, support vector machines, maximum entropy classification, hidden Markov models, conditional random fields
- **Semi-supervised algorithms**:
  - e.g., latent word models
- **Unsupervised learning algorithms**:
  - e.g., pattern mining

Supervised learning

- **Maximum entropy classifier:**

  Given *n* training examples $S = \{(\mathbf{x}, y)_1, \ldots, (\mathbf{x}, y)_n\}$.

  where $\mathbf{x}$ = feature vector and y = class

- We choose the model p^* that preserves as much uncertainty as possible, or which maximizes the entropy $H(p)$ between all the models $p \in P$ that satisfy the constraints enforced by the training examples:

$$H(p) = -\sum_{(\mathbf{x},y)} p(\mathbf{x},y) \log p(\mathbf{x},y) \qquad p^* = \arg\max_{p \in P} H(p)$$

[Berger et al. CL 1996]

It has been shown that $p^*(S)$ is unique and must be in the following form:

$$p^*(y|\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{k}\lambda_j f_j(\mathbf{x},y)\right), \quad 0 < \lambda_j < \infty$$

where

$f_j(\mathbf{x}, y)$ = one of the k binary-valued **feature functions**

λ_j = parameter adjusted to model the observed statistics

Z = normalizing constant

We need numerical methods in order to derive λ_j given the set of constraints

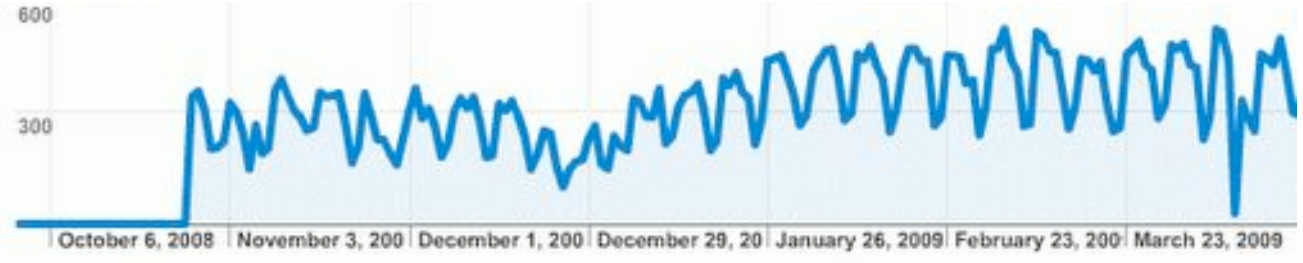
Opinion recognition

\- *The movie really seems to be spilling the beans on a lot of stuff we didnt think we hand if this is their warm up, what is going to get us frothing in December*

-+ *de grote merken mogen er dan patserig uitzien en massa's pk hebben maarals de bomen wat dicht bij elkaar staan en de paadjes steil enbochtig,dan verkies ik mijn Jimny .*

\+ *L'é tro bel cet voitur Voici tt ce ki me pasione ds ma petite vi!!!é tt mé pote é pl1 dotre truk!!!Avou de Dcouvrir*

Mining: dashboard



[Boiy & Moens IR 2009]

Table 2: Our best results in terms of accuracy, precision, recall and F-measure (F_1) using the English (a), Dutch (b) and French (c) corpora. For English, Dutch and French we implemented respectively an MNB, an SVM and an ME classifier – 10 fold cross-validation.

(a) English

Architecture	**Accuracy**	**Precision** pos/neg/neu	**Recall** pos/neg/neu	**F-measure** pos/neg/neu
Cascade with layers 1, 2 and 3	83.30	69.09/85.48/85.93	55.73/82.40/91.84	61.70/83.91/88.79
Cascade with layers 1 and 2	83.10	70.49/87.72/84.61	54.13/79.07/93.00	61.24/83.17/88.61
SC uni-lang	83.03	69.59/86.77/85.08	56.13/79.60/92.12	62.14/83.03/88.46
SC uni-lang-dist	80.23	60.59/78.78/86.57	59.87/82.67/85.60	60.23/80.68/86.08
SC uni	82.73	68.01/85.63/85.53	58.40/78.67/91.24	62.84/82.00/88.29

(b) Dutch

Architecture	**Accuracy**	**Precision** pos/neg/neu	**Recall** pos/neg/neu	**F-measure** pos/neg/neu
Cascade with layers 1,2 and 3	69.03	63.51/53.30/72.20	42.93/31.20/88.20	51.23/39.36/79.40
Cascade with layers 1 and 2	69.80	66.60/58.31/71.66	41.73/29.47/90.32	51.31/39.15/79.92
SC uni-lang	69.05	60.39/52.59/73.63	49.60/33.87/85.44	54.47/41.20/79.10
SC uni-lang-dist	68.85	61.08/54.52/72.20	43.73/30.53/87.88	50.97/39.15/79.27
SC uni	68.18	58.73/49.58/73.24	48.00/31.73/85.16	52.82/38.70/78.75

(c) French

Architecture	**Accuracy**	**Precision** pos/neg/neu	**Recall** pos/neg/neu	**F-measure** pos/neg/neu
Cascade with layers 1, 2 and 3	67.68	50.74/55.88/71.90	27.47/38.67/88.44	35.64/45.71/79.32
Cascade with layers 1 and 2	67.47	52.69/53.96/71.56	26.13/38.13/88.68	34.94/44.69/79.21
SC uni-lang	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni-lang-dist	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni	65.83	45.67/50.82/72.23	28.80/41.33/84.28	35.32/45.59/77.79

[Boiy & Moens 2009]

Supervised learning

- **Context dependent classification** = the class to which a feature vector is assigned depends on:
  1) the feature vector itself
  2) the values of other feature vectors
  3) the existing relation among the various classes

- Examples:
  – conditional random field

- When processing text, many variables are interdependent (often dependent on previous content in the discourse):
    - e.g., there is a higher likelihood to find the name of a drug in the neighborhood of the name of a disease than in the neighborhood of the name of a car brand

Conditional Random Field

- **Linear chain conditional random field:**
  - Let X = ($\mathbf{x}_1$, ... , $\mathbf{x}_T$) be a random variable over data sequences to be labeled and Y a random variable over the corresponding label sequences
  - All components y_i of Y are assumed to range over a finite label alphabet $\sum$
  - We define $G = (V, E)$ to be an **undirected graph** such that there is a node $v \in V$ corresponding to each of the random variables representing an element y_v of Y
  - If each y_v obeys the Markov property with respect to G, then the model (Y,X) is a conditional random field

Conditional Random Field

- In an information extraction task, X might range over the words or constituents of a sentence, while Y ranges over the semantic classes to be recognized in these sentences

- In theory the structure of graph G may be arbitrary: e.g., template based or **general CRF**, where you can define the dependencies in the Markov network or graph

[Lafferty et al. ICML 2001]

Conditional random field

- **Feature functions** depend on the current state or on the previous and current states

$$s_i(y_j, X, j) = \begin{cases} 1 \text{ if the observation at position } j \text{ is the word "run"} \\ \quad \text{and } y_j = \text{movement} \\ 0 \text{ otherwise} \end{cases}$$

$$t_i(y_{j-1}, y_j, X, j) = \begin{cases} 1 \text{ if } y_{j-1} = \text{"person" and } y_j = \text{"movement"} \\ 0 \text{ otherwise} \end{cases}$$

- We use a more global notation f_i for a feature function where $f_i(y_{j-1}, y_j, X, j)$ is either a state function $s_i(y_j, X, j) = s_i(j_{i-1}, y_j, X, i)$ or a transition function $t_i(y_{j-1}, y_j, X, j)$.

Conditional random field

- To classify a new instance $P(Y|X)$ is computed as follows:

$$P(Y|X) = \frac{1}{Z}\exp\left(\sum_{t=1}^{T} F(y_{t-1}, y_t, X)\right)$$

where $F(y_{t-1}, y_t, X)$ is a potential function that captures the degree to which the assignment y_t to the output variable fits the transition y_{t-1} and is often computed as $\sum_{j=1}^{k} \lambda_j f_j(y_{t-1}, y_t, X, t)$

$f_j(y_{t-1}, y_t, X, t)$ = one of the k binary-valued **feature functions**

λ_j = parameter adjusted to model the observed statistics,

Z = normalizing constant

The most probable label set Y^* for input sequence X is:

$$Y^* = \underset{Y}{argmax}\, P(Y|X)$$

Conditional Random Field

- **CRF training:**
  - Like for the maximum entropy model, we need numerical methods in order to derive λ_j given the set of constraints
  - E.g., linear-chain CRF: variation of the Baum-Welch algorithm
  - In general CRFs we use approximate inference (e.g., Markov Chain Monte Carlo sampler)

Conditional Random Field

- **CRF and information extraction**:
  - Advantages:
    - Successful: context-dependent classification and training following the maximum entropy principle
    - One of the current most successful information extraction techniques
  - Disadvantage:
    - Training is computationally expensive, especially when the graphical structure is complex

Conditional Random Field

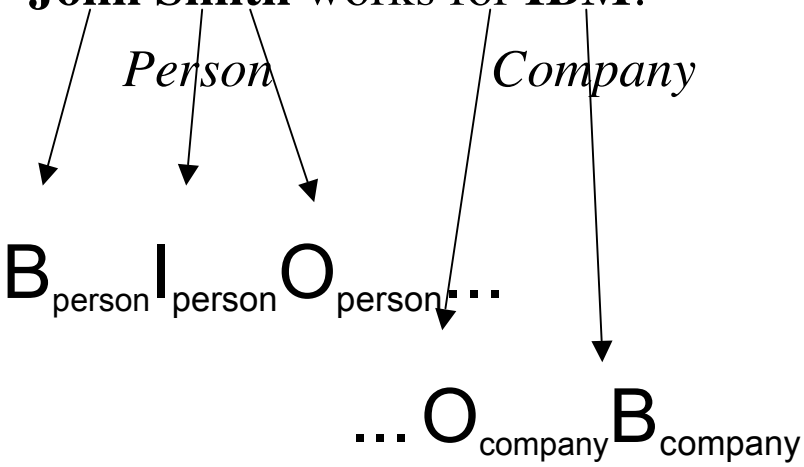
- **CRF and information extraction**:
  - Advantages:
    - Successful: context-dependent classification and training following the maximum entropy principle
    - One of the current most successful information extraction techniques
  - Disadvantage:
    - Training is computationally expensive, especially when the graphical structure is complex

Named entity recognition

- Named entity recognition recognizes and classifies named expressions in text (such as person, company, location or protein names):

Example:

John Smith works for **IBM**.



Named entity recognition

- Two problems: **Segmentation + Classification**
- **Constituent based processing**:
  - Constituents are first identified (constituency parser or phrase chunker)
  - Constituents are classified
- **Use of BIO format**:
  - B= Begin, I = Inside, O = Outside labels per class
  - Words or tokens are classified

Here illustrated for NER, but similar approach for other extraction tasks (e.g., relation extraction)

Table 4.1. Typical features in a named entity recognition task of the candidate entity name *i* that occur in the context window of *l* words.

FEATURE	VALUE TYPE	VALUE
Short type	Boolean	True if *i* matches the short type *j*; False otherwise.
POS	Nominal	Part-of-speech tag of the syntactic head of *i*.
Context word	Boolean or real value between 0 and 1; Or nominal.	True if the context word *j* occurs in the context of *i*; False otherwise; If a real value is used, it indicates the weight of the context word *j*. Alternatively, the context word feature can be represented as one feature with nominal values.
POS left	Nominal	POS tag of a word that occurs to the left of *i*.
POS right	Nominal	POS tag of a word that occurs to the right of *i*.
Morphological prefixes/suffixes	Nominal	Prefix or suffix of *i*.

F1 scores on the CoNLL Dataset						
Approach	LOC	ORG	MISC	PER	ALL	Relative Error reduction
Bunescu and Mooney (2004) (Relational Markov Networks)						
Only Local Templates	-	-	-	-	**80.09**	
Global and Local Templates	-	-	-	-	**82.30**	11.1%
Finkel et al. (2005)(Gibbs Sampling)						
Local+Viterbi	88.16	80.83	78.51	90.36	**85.51**	
Non Local+Gibbs	88.51	81.72	80.43	92.29	**86.86**	9.3%
Our Approach with the 2-stage CRF						
Baseline CRF	88.09	80.88	78.26	89.76	**85.29**	
+ Document token-majority features	89.17	80.15	78.73	91.60	86.50	
+ Document entity-majority features	89.50	81.98	79.38	91.74	86.75	
+ Document superentity-majority features	89.52	82.27	79.76	92.71	**87.15**	12.6%
+ Corpus token-majority features	89.48	82.36	79.59	92.65	87.13	
+ Corpus entity-majority features	89.72	82.40	79.71	92.65	87.23	
+ Corpus superentity-majority features (All features)	89.80	82.39	79.76	92.57	**87.24**	13.3%

Named entity recognition: 2-stage approach: 1) CRF with local features; 2) local information and output of first CRF as features. Comparison against competitive approaches. Baseline results are shown on the first line of each approach.

[Krishnan & Manning 2006]

Supervised learning

- **Support vector machine**: see **Text categorization**

[Christianini & Shawe-Tailor 2000]

Kernel function

- The decision function $f(\mathbf{x})$ we can just replace the dot products with kernels $K(\mathbf{x}_i,\mathbf{x}_j)$:

$$h(\mathbf{x}) = sign(f(\mathbf{x}))$$

$$f(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i y_i \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) \rangle + b$$

$$f(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i y_i K(\mathbf{x}_i,\mathbf{x}) + b$$

- Many valuable kernels for information extraction:
  - String kernel
  - Syntactic kernel
  - Semantic tree kernel
  - Relation kernel (see further)

[Moschitti & Quarteroni IP & M 2011]

Relation extraction

- World Wide Web = huge source of unlabeled examples !
- **Relation extraction**:
  - Search sentences on the Web:
    - containing 2 entities for which target relation holds => **positive bag**, but might contain some negative examples
    - containing 2 entities for which target relation does not hold = > **negative bag**, assumed to contain only negative examples

  ⇒input for **Multiple Instance Learning** algorithm

[Bunescu & Mooney ACL 2007]

+/S_1: Search giant **Google** has bought video-sharing website **YouTube** in a controversial $1.6 billion deal

-/S_2: The companies will merge **Google**"s search expertise with **YouTube**"s video expertise, pushing what executives believe is a hot emerging market of video offered over the internet

+/S_3: **Google** has acquired social media company, **YouTube** for $1.65 billion in a stock-for-stock transaction as announces bu Google Inc. on October 9, 2006

+/S_4: Drug giant **Pfizer Inc.** has reached an agreement to buy the private biotechnology form **Rinat Neuroscience Corp.**, the companies announced Thursday.

-/S_5: Ha has also received consulting fees from Alpharma, Eli Lilly and Company, **Pfizer**, Wyeth Pharmaceuticals, **Rinat Neuroscience**, Elan Pharmaceuticals, and Forrect Laboratories

WLS-SVM: errors in positive and negative bags are weighted differently to comply with MIL setting

Minimize over *w,b* and *e:*

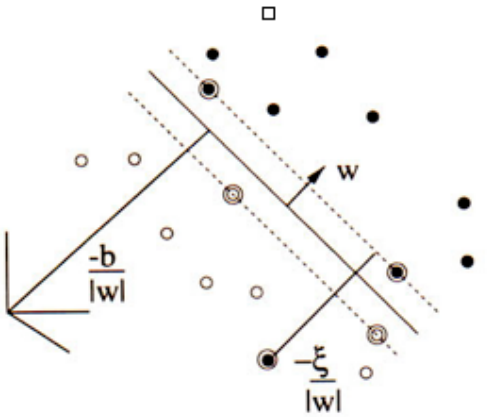
$$J(w,b,e) = \tfrac{1}{2}\|w\|^2 + \tfrac{C}{L}\left(v_p \Xi_p + v_n \Xi_n\right)$$

$$\Xi_p = \sum_{B_i \in B^+} \sum_{x \in B_i} e_x^2$$

$$\Xi_n = \sum_{B_i \in B^-} \sum_{x \in B_i} e_x^2$$

$$v_p = \frac{1}{\gamma c_p}$$

$$v_n = \frac{1}{\gamma c_n}$$



Classical SVM

Subject to constraints:

$$w\phi(x) + b = +1 - e_x, \qquad \forall x \in B_i \in B^+,$$

$$w\phi(x) + b = -1 + e_x, \qquad \forall x \in B_i \in B^-,$$

[Suykens et al. 2002]

[De Belder et al.SIM 2009]

Relational kernel

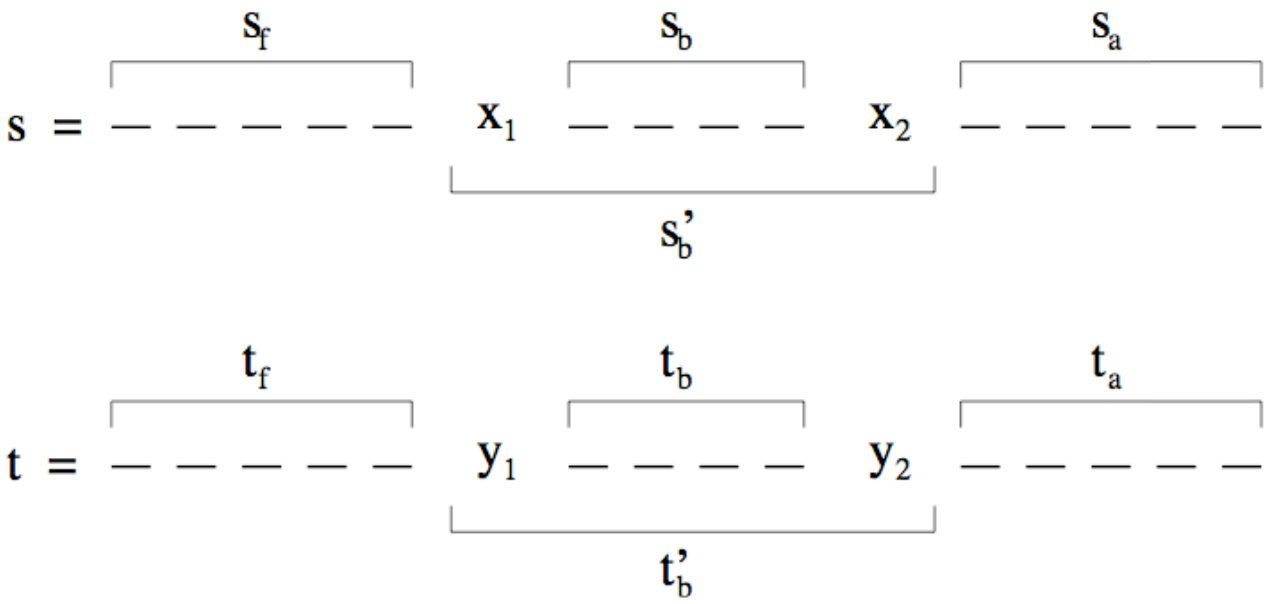


Figure 2: Sentence segments.

Relational kernel

Subsequence kernel for relation extraction (SSK-T1):

The relation kernel $rK(s,t)$ of sequences s and t is the sum of 3 subkernels:

$$rK(s,t) = K^{fb}(s,t,\lambda) + K^{b}(s,t,\lambda) + K^{ba}(s,t,\lambda)$$

where $K^{fb}(s,t)$ = number of common fore-between pairs in segment spanned by "fore" (s_f, t_f) + "between" (s_b, t_b) entities (x_1,y_1) and (x_2,y_2) included, i.e., (s_b', t_b')

$K^{b}(s,t)$ = number of common between pairs in segment spanned by (s_b', t_b')

$K^{ba}(s,t)$ = number of common between-after pairs in segment spanned by (s_b', t_b') + "after" (s_a, t_a)

Relational kernel

$$K(s,t,\lambda) = \sum_{k} \sum_{u \in \Sigma^k_{\cup}} \sum_{\mathbf{i}: u \prec s[\mathbf{i}]} \sum_{\mathbf{j}: u \prec \mathbf{t}[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})}$$

- where u = sparse subsequence of length k common to s and t (i.e., $u \prec s[\mathbf{i}]$, $u \prec t[\mathbf{j}]$)
  $\lambda^{l(\mathbf{i})+l(\mathbf{j})} \leq 1$: weight of u
  k ranges usually between 1 and 4
  $\mathbf{i}$ = sequence of $|\mathbf{i}|$ indices in s (idem $\mathbf{j}$)
  $l(\mathbf{i})$ = length of the index sequence $\mathbf{i}$ in s

Relational kernel adapted with weighting scheme that takes into account frequency of a Word (SSK-T2)

[De Belder et al.SIM 2009]

	SSK	BOW	SSK-T1	BOW-T2	SSK-T2
WLS:	80,41	33,46	85,80	55,08	**87,08**
QP:	77,62	25,50	83,93	44,77	86,09

Own dataset

	SSK	BOW	SSK-T1	BOW-T2	SSK-T2
WLS:	52,00	22,90	71,00	39,56	**86,28**
QP:	51,18	23,63	67,16	30,84	62,65

Dataset Raymond Mooney

Results “Acquisition” relation

	SSK	BOW	SSK-T1	BOW-T2	SSK-T2
WLS:	84,90	17,54	95,08	52,30	94,95
QP:	88,51	52,61	94,00	77,96	**96,07**

Own dataset

	SSK	BOW	SSK-T1	BOW-T2	SSK-T2
WLS:	68,07	8,76	87,54	18,92	89,38
QP:	84,77	18,69	88,40	20,00	**90,02**

Dataset Raymond Mooney

Results “Born in” relation

Evaluation of the supervised learning methods

- **Results approach the results of using hand-crafted patterns**
- But, for some tasks the results fall short of human capability:
  - **both for the hand-crafted and learned patterns**
  - explanation:
    - high variation of natural language expressions that form the context of the information or that constitute the information, patterns are not seen in the training data
    - ambiguous patterns and lack of discriminative features
    - lack of world knowledge not made explicit in the text

Too much variation

Google owns YouTube

Google has acquired YouTube

YouTube is bought by Google

...

Too ambiguous

- The meaning of a word depends on the company it keeps

Take a right at the green plant which produces solar energy.

Tom Mitchell Center for Automated Learning and Discovery.

Evaluation of the supervised learning methods

- Annotating: tedious task ! => reduce as much as possible
  - Learning the extraction patterns of generic semantic classes that can be used for many texts
  - Reusable knowledge resources: possibly acquired by machine learning
  - Ways of reducing annotations:
    - **Semi-supervised methods**
    - **Unsupervised methods**

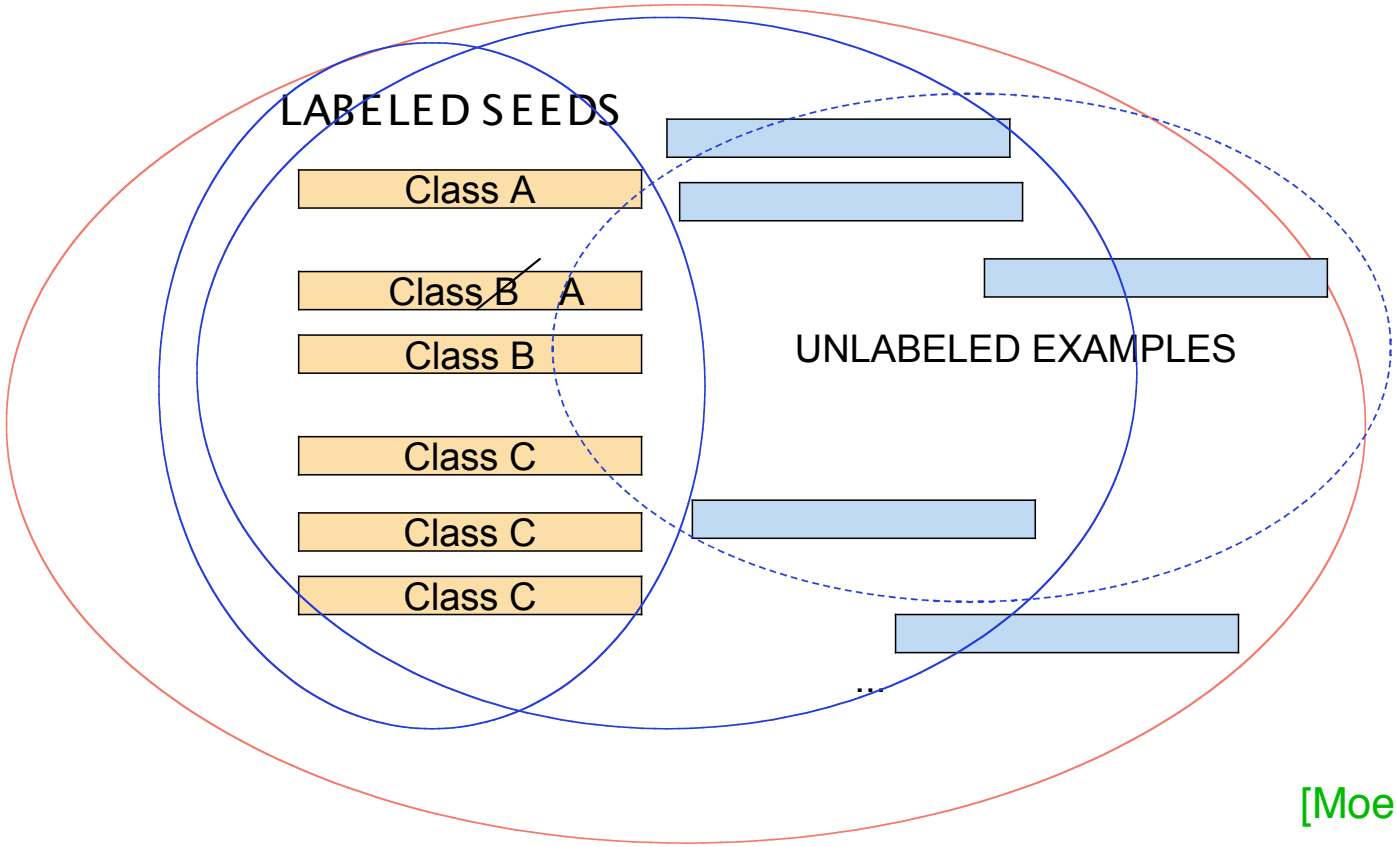
- **Can we learn patterns for the semantic labeling with a minimum of supervision** taking into account the large variation of patterns and ambiguity?

- In the following we focus on text, but many approaches can be ported to other media

A classifier is trained from labeled and unlabelled examples

- = **Semi-supervised learning**: most known forms are:
  - Self-learning: iterative retraining after labeling of data points for which the current model is most confident
  - Transductive inference: no general decision rule is inferred, only the labels of the unannotated examples are predicted according to a most likely model

Self-training



[Moens 2006]

Fig. 6.3. Self-training: A classifier is incrementally trained (blue line), first based on the labeled seeds, and then based on the labeled seeds and a set of unlabeled examples that are labeled with the current classifier. The dotted blue line represents the set of all unlabeled examples that were considered for labeling in this step.

Benefit of semi-supervised learning

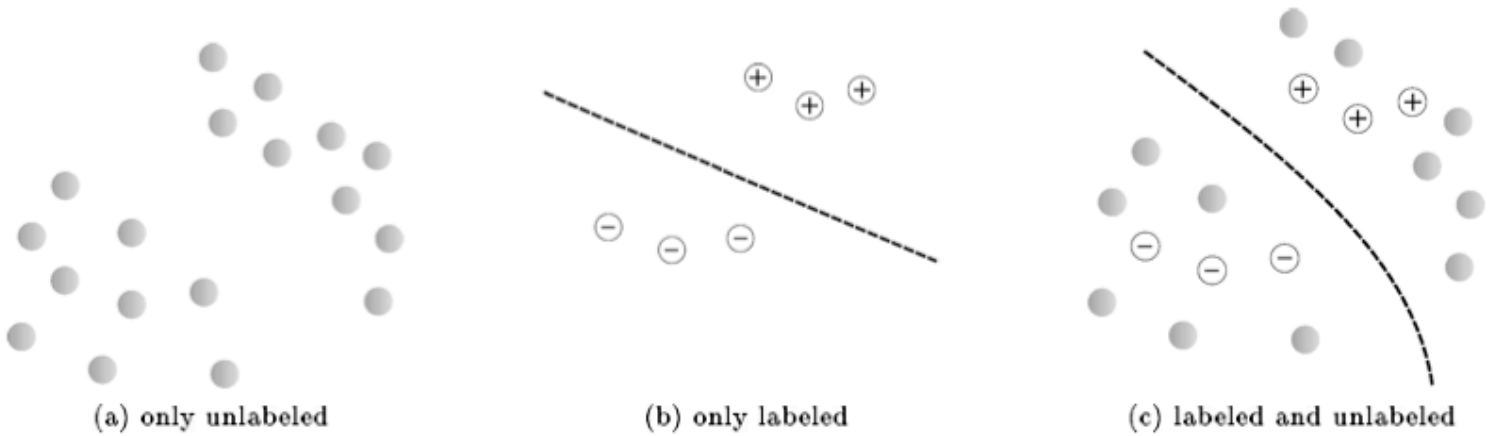


Figure 2: Schematic figure illustrating how unlabeled data might improve a supervised classifier. Grey dots are unlabeled data, white dots labeled data and the dotted line the classification boundary.

But does this hold for all types of text classification?

Google injects search savvy into display ad system

By MICHAEL LIEDTKE, AP Technology Writer - Fri Sep 18, 2009 4:28AM EDT

Add articles about technology to your My Yahoo! MY Y!

SAN FRANCISCO - Google Inc. is counting on the crown jewel of its online advertising empire to burnish a diamond in the rough.

Hoping to take an even bigger bite out of ad budgets, Google has melded the technology powering its lucrative search marketing network with a system that it bought 18 months ago to sell online billboards and other more visual commercials, including video.

The long-awaited combination poses another threat to Yahoo Inc., whose profits have been sliding the past three years. Yahoo is the Internet's largest seller of display advertising, a mantle that Google has set its sights on. Microsoft Corp. and Time Warner Inc.'s AOL also operate large exchanges that help manage display ads.

The upgrade announced Friday has been something Google has been working toward since it bought DoubleClick Inc. for $3.2 billion a year-and-a-half ago. Google prized DoubleClick largely for its tools for selling and serving display ads.

Technology

Canadian Press

Report: Gender tests on runner done in SAfrica

AP – 2 hrs 33 mins ago

JOHANNESBURG – A South African newspaper has published what it says are e-mails showing local track officials authorized gender tests done in the country on runner Caster Semenya. Full Story »

Sports

Sanofi buys Merck's half of animal health business

Associated Press

Buzz up! 0 votes | Send | Share | Print

By LINDA A. JOHNSON, AP Business Writer – 23 mins ago

TRENTON, N.J. – French drugmaker Sanofi-Aventis SA has completed its $4 billion purchase of Merck & Co.'s half interest in their veterinary medicine business, Merial Ltd., the companies said Friday.

The move was required by regulators before Merck can close its $41 billion purchase of New Jersey neighbor Schering-Plough Corp., which also sells animal health products.

Merial, a joint venture founded in 1997, sells two widely used pet medicines, flea-and-tick blocker Frontline and chewable heartworm preventer Heartgard. It also sells Ivomec, which kills parasites in hogs and cattle, and other medicines and vaccines for livestock.

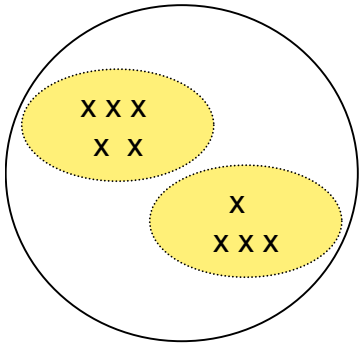
RELATED QUOTES

MRK	32.29	+0.28
PFE	16.64	+0.32
SAN	52.64	+0.68
SGP	28.31	+0.18
WYE	48.16	+0.31

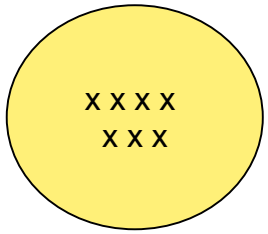
Business

⇨ News filtering

- **Text categorization**: good results:
  - Generative model: e.g., self learning with Naive Bayes and Expectation Maximization, e.g., ca. average 95% accuracy on standard Reuters text categories, but use of multiple mixture components per class [Nigam, McCallum & Mitchell SL 2006]
  - Discriminative model: transductive learning with SVM e.g., better results when few training data, approach results SVM with more training data (accuracy > 80 %) [Joachims SL 2006]



Semi-supervised smoothness assumption: if two points in a high density region are close, so should be the corresponding classes



Cluster assumption: the points of each class tend to form a cluster

=>These assumptions do not necessarily hold for fine-grained text classification tasks !

Manifold assumption: curse of dimensionality: many features → many training examples

[Chapelle, Schölkopf & Zien SL 2006]

When does semi-supervised learning work?

- For semi-supervised learning to work: it is an important prerequisite that the distribution of examples, which the unlabeled examples help elucidate, is relevant for the classification problem [Chapelle, Schölkopf & Zien SL 2006]

- Model learned from the labeled examples should be rather correct [Cozman & Cohen in Chapelle, Schölkopf & Zien SL 2006]

- => evidenced by own research for semantic role labeling [Deschacht & Moens Technical report 2009]

Semantic role labeling

Recognizing the basic event structure of a sentence
("**who**" "**does what**" "**to whom/what**" "**when**" "**where**", …)

fall.01

Arg1: Logical subject, patient, thing falling

Arg2: Extent, amount fallen

Arg3: Start point

Arg4: End point, end state of Arg1

Ex1: [$_{Arg1}$ Sales] *fell* [$_{Arg4}$ to $251.2 million] [$_{Arg3}$ from $278.7 million].

Ex2: [$_{Ag1}$The average junk bond] *fell* [$_{Arg2}$ by 3.7%].

By how much has fallen the average junk bond?

Question answering search: see below

Latent Words Language Model

- Generative model of natural language
- Latent variable (**hidden word**) models *words* that have a *similar meaning* in a *specific* left and right *context*

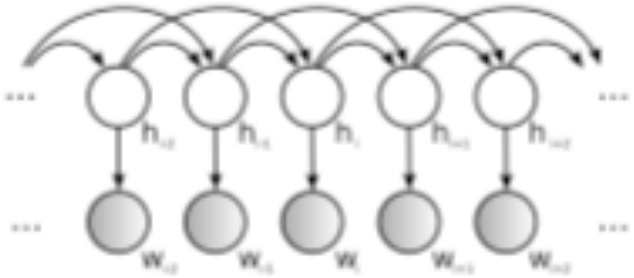


Figure 4: BN of the latent words language model. The words w_i (gray nodes) are observed and the hidden words h (white nodes) are hidden variables.

- Predicts how likely a word will be present in a given context

[Deschacht & Moens EMNLP 2009]

Latent Words Language Model

- Model is trained on large representative corpus:
  - Model is trained with variation of Baum-Welch algorithm or with Gibbs sampling

Examples

Compuserve	corp	said	Tuesday	it	anticipates	a	loss
Microsoft	inc	told	Friday	they	expects	the	profit
Crysler	corp.	reported	Thursday	he	expected	some	gain
Oracle	ltd	added	Monday	she	assumes	an	deficit
Software	co	say	Wednesday	this	doubts	another	earnings

A	Japanese	electronics	executive	was	kidnapped	in	Mexico
the	U.S.	tobacco	director	is	abducted	on	Usa
its	German	sales	manager	we	killed	at	UK
an	British	consulting	economist	are	found	of	Australia
one	Russian	electric	spokesman	be	abduction	into	Canada

Latent words for SRL

- Latent words are used as probabilistic features in MEMM for classification (results as F_1 - CoNLL dataset):

	5%	20%	50%	100%
Supervised	40.49%	67.23%	74.93%	78.65%
LWFeatures	**60.29%**	**72.88%**	**76.42%**	**80.98%**
ClusterFeatures	59.51%	66.70%	70.15%	72.62%

Table 7: Results (in F1-measure) on the CoNLL 2008 test set, comparing the standard supervised classifier (*Supervised*) with the classifiers employing latent words (*LWFeatures*) or semantic clusters (*ClusterFeatures*) as extra features. Best results are in bold.

→ **Word expansion** improves recall

→ **Word sense disambiguation** improves precision

→ **Easy to use** in many other NLP applications

[Deschacht & Moens EMNLP 2009]

Few examples for training, but additional constraints

- Knowledge of language or cognitive knowledge on how people understand text might help:
  - For selecting seed labeled examples
  - For adding additional constraints
- Illustrated with the recognition of temporal information in text

Recognition of temporal relations

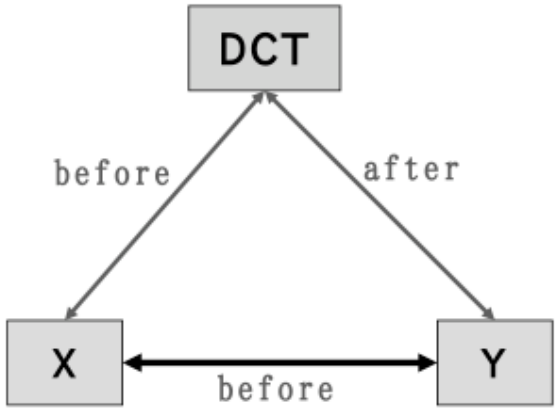
- Between an event and time expressions that occur within the same sentence
- Between document creation time and an event
- Between the main events of adjacent sentences

⇒ Markov Logic model that jointly identifies these relations:

  ⇒ Explicit incorporation of (soft) temporal constraints

  ⇒ Cf. Maximum entropy model, CRF: but feature functions can also be in the form of first-order logical expressions manually drafted

[Yoshikawa et a. ACL 2009]



Constraints help in **resolving ambiguity**: reported 2% rise in accuracy compared to state of the art (average accuracy over above tasks = 68.9% on the TimeML dataset)

=> **But still problem of the variety of patterns in language**

Unsupervised learning

- **Use of sequence mining** on World Wide Web documents: Recognition of typical scenarios

1. look at menu 2. decide what you want 3. order at counter 4. pay at counter 5. receive food at counter 6. take food to table 7. eat food
1. walk to the counter 2. place an order 3. pay the bill 4. wait for the ordered food 5. get the food 6. move to a table 7. eat food 8. exit the place
1. walk into restaurant 2. find the end of the line 3. stand in line 4. look at menu board 5. decide on food and drink 6. tell cashier your order 7. listen to cashier repeat order 8. listen for total price 9. swipe credit card in scanner 10. put up credit card 11. take receipt 12. look at order number 13. take your cup 14. stand off to the side 15. wait for number to be called 16. get your drink

Figure 1: Three event sequence descriptions

[Chamber & Jurafsky ACL 2009]

[Regneri et al. ACL 2010]

Unsupervised learning

- Techniques of clustering of sequences, where the similarity metric relies on the results of a sequence alignment algorithm

[Regneri et al. ACL 2010]

row	s_1	s_2	s_3	s_4
1	⊘	walk into restaurant	⊘	enter restaurant
2	⊘	⊘	walk to the counter	go to counter
3	⊘	find the end of the line	⊘	⊘
4	⊘	stand in line	⊘	⊘
5	look at menu	look at menu board	⊘	⊘
6	decide what you want	decide on food and drink	⊘	make selection
7	order at counter	tell cashier your order	place an order	place order
8	⊘	listen to cashier repeat order	⊘	⊘
9	pay at counter	⊘	pay the bill	pay for food
10	⊘	listen for total price	⊘	⊘
11	⊘	swipe credit card in scanner	⊘	⊘
12	⊘	put up credit card	⊘	⊘
13	⊘	take receipt	⊘	⊘
14	⊘	look at order number	⊘	⊘
15	⊘	take your cup	⊘	⊘
16	⊘	stand off to the side	⊘	⊘
17	⊘	wait for number to be called	wait for the ordered food	⊘
18	receive food at counter	get your drink	get the food	pick up order
19	⊘	⊘	⊘	pick up condiments
20	take food to table	⊘	move to a table	go to table
21	eat food	⊘	eat food	consume food
22	⊘	⊘	⊘	clear tray
22	⊘	⊘	exit the place	⊘

Figure 2: A MSA of four event sequence descriptions

	Scenario	Precision			Recall			F-Score			
		sys	$base_{cl}$	$base_{lex}$	*sys*	$base_{cl}$	$base_{lex}$	*sys*	$base_{cl}$	$base_{lex}$	upper
MTurk	pay with credit card	0.52	0.43	0.50	0.84	0.89	0.11	**0.64**	0.58	• 0.17	0.60
	eat in restaurant	0.70	0.42	0.75	0.88	1.00	0.25	**0.78**	• 0.59	• 0.38	• 0.92
	iron clothes I	0.52	0.32	1.00	0.94	1.00	0.12	**0.67**	• 0.48	• 0.21	• 0.82
	cook scrambled eggs	0.58	0.34	0.50	0.86	0.95	0.10	**0.69**	• 0.50	• 0.16	• 0.91
	take a bus	0.65	0.42	0.40	0.87	1.00	0.09	**0.74**	• 0.59	• 0.14	• 0.88
OMICS	answer the phone	0.93	0.45	0.70	0.85	1.00	0.21	**0.89**	• 0.71	• 0.33	0.79
	buy from vending machine	0.59	0.43	0.59	0.83	1.00	0.54	**0.69**	0.60	0.57	0.80
	iron clothes II	0.57	0.30	0.33	0.94	1.00	0.22	**0.71**	• 0.46	• 0.27	0.77
	make coffee	0.50	0.27	0.56	0.94	1.00	0.31	**0.65**	• 0.42	○ 0.40	• 0.82
	make omelette	0.75	0.54	0.67	0.92	0.96	0.23	**0.83**	• 0.69	• 0.34	0.85
	Average	0.63	0.40	0.60	0.89	0.98	0.22	**0.73**	0.56	0.30	0.82

Figure 4: Results for paraphrasing task; significance of difference to *sys*: • : $p \leq 0.01$, ○ : $p \leq 0.1$

	Scenario	Precision			Recall			F-Score			
		sys	$base_{cl}$	$base_{lex}$	*sys*	$base_{cl}$	$base_{lex}$	*sys*	$base_{cl}$	$base_{lex}$	upper
MTurk	pay with credit card	0.86	0.49	0.65	0.84	0.74	0.45	**0.85**	• 0.59	• 0.53	0.92
	eat in restaurant	0.78	0.48	0.68	0.84	0.98	0.75	**0.81**	• 0.64	0.71	• 0.95
	iron clothes I	0.78	0.54	0.75	0.72	0.95	0.53	**0.75**	0.69	• 0.62	• 0.92
	cook scrambled eggs	0.67	0.54	0.55	0.64	0.98	0.69	0.66	**0.70**	0.61	• 0.88
	take a bus	0.80	0.49	0.68	0.80	1.00	0.37	**0.80**	• 0.66	• 0.48	• 0.96
OMICS	answer the phone	0.83	0.48	0.79	0.86	1.00	0.96	0.84	• 0.64	**0.87**	0.90
	buy from vending machine	0.84	0.51	0.69	0.85	0.90	0.75	**0.84**	• 0.66	○ 0.71	0.83
	iron clothes II	0.78	0.48	0.75	0.80	0.96	0.66	**0.79**	• 0.64	0.70	0.84
	make coffee	0.70	0.55	0.50	0.78	1.00	0.55	**0.74**	0.71	○ 0.53	○ 0.83
	make omelette	0.70	0.55	0.79	0.83	0.93	0.82	0.76	○ 0.69	***0.81***	• 0.92
	Average	0.77	0.51	0.68	0.80	0.95	0.65	**0.78**	0.66	0.66	0.90

Figure 5: Results for happens-before task; significance of difference to *sys*: • : $p \leq 0.01$, ○ : $p \leq 0.1$

[Regneri et al. ACL 2010]

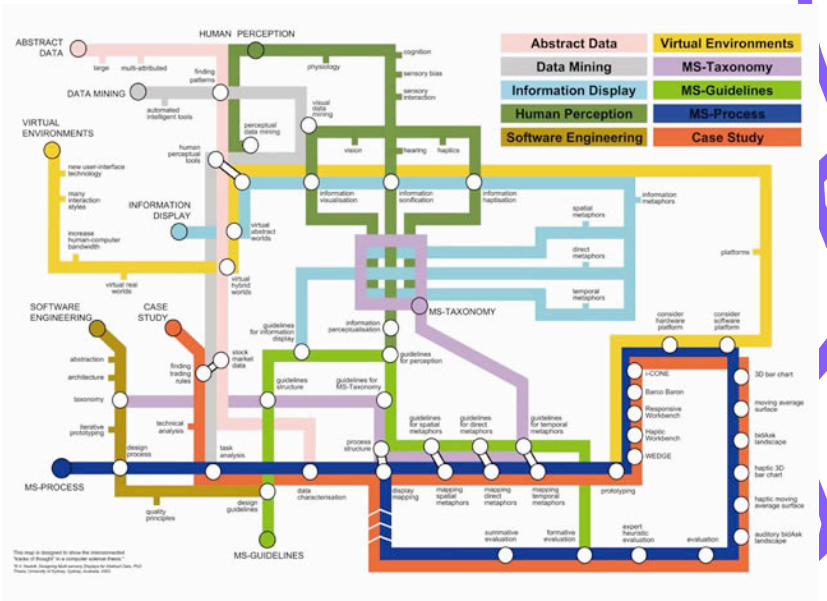
Exploratory search

- Exploratory search:
  - does not necessary rely on keywords to initiate a search, but allows a user to navigate a document collection in a guided way
  - Users who might not be familiar with the field, will more easily find their way in large collections and users who do not have a very specific information question in mind, might freely explore the information
  - Information extraction is an important step in the realization of exploratory search: for the construction of visualizations, menus, etc.

[Marchionini CACM 2006]

Extraction of factoid information

- Current automated techniques are successful in identifying factoid information:
  - E.g., named entities, relations between entities
  - Possibly link these factoids across documents for clustering, ontology population, social network analysis, visualization, ...
  - Enhances exploratory search

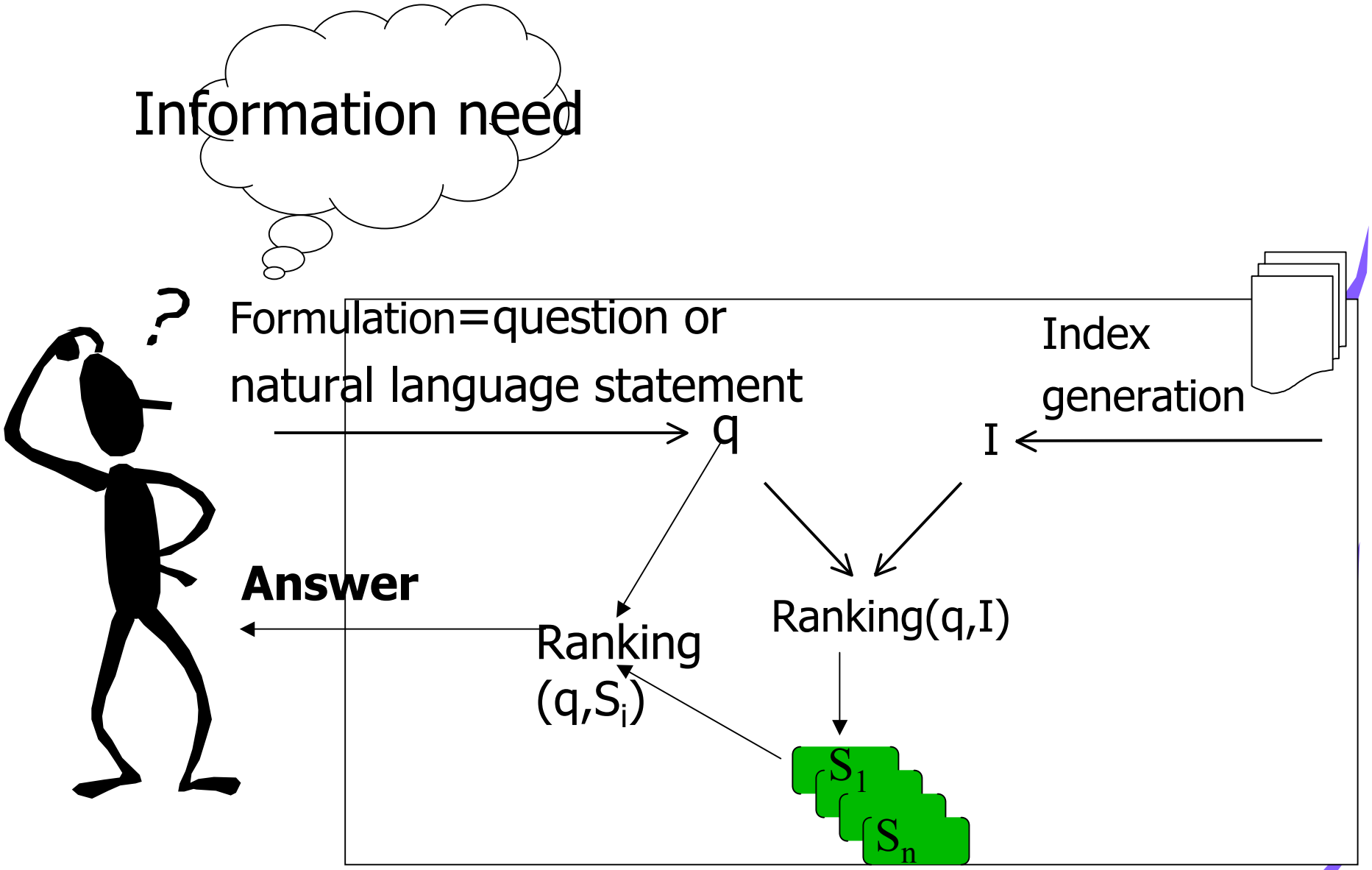


Source: Keith V. Nesbitt

Question answering

- **Automatic question answering**:
  - Single questions are automatically answered by using a collection of documents as the source of data for the production of the answer
  - Interest in the Text REtrieval Conferences (TREC) and Knowledge and Reasoning for Answering Questions (KRAQ) conferences

Question answering system



- Example:

question: “Who is the architect of the Hancock building in Boston?”

answer: “I.M. Pei”

extracted from:

“The John Hancock Tower was completed in 1976 to create additional office space for the John Hancock Life Insurance Co. It was designed by the renowned architect I.M. Pei.”

“Designed by world renowned architect I.M.Pei, the John Hancock Tower is the highest in New England.”

- Example:

Natural language query: “Show me a video fragment where a red car takes a right turn on Saint-John’s square.”

answer:

keyframes of video fragment

extracted from:

video indexed with entities their attributes and relations including spatial and temporal relations

Question answering

- **General procedure:**

1. Analysis of the question
   - selection of key terms for retrieval
   - identification of the question type: e.g. “Who” -> person
   - linguistic analysis of the question: e.g., POS tagging, parsing, recognition of verbs and arguments, semantic role detection and named entity recognition
2. Retrieval of subset of the document collection that is thought to hold the answers and of candidate answer sentences
3. Linguistic analysis of the candidate answer sentences: cf. question

Question answering

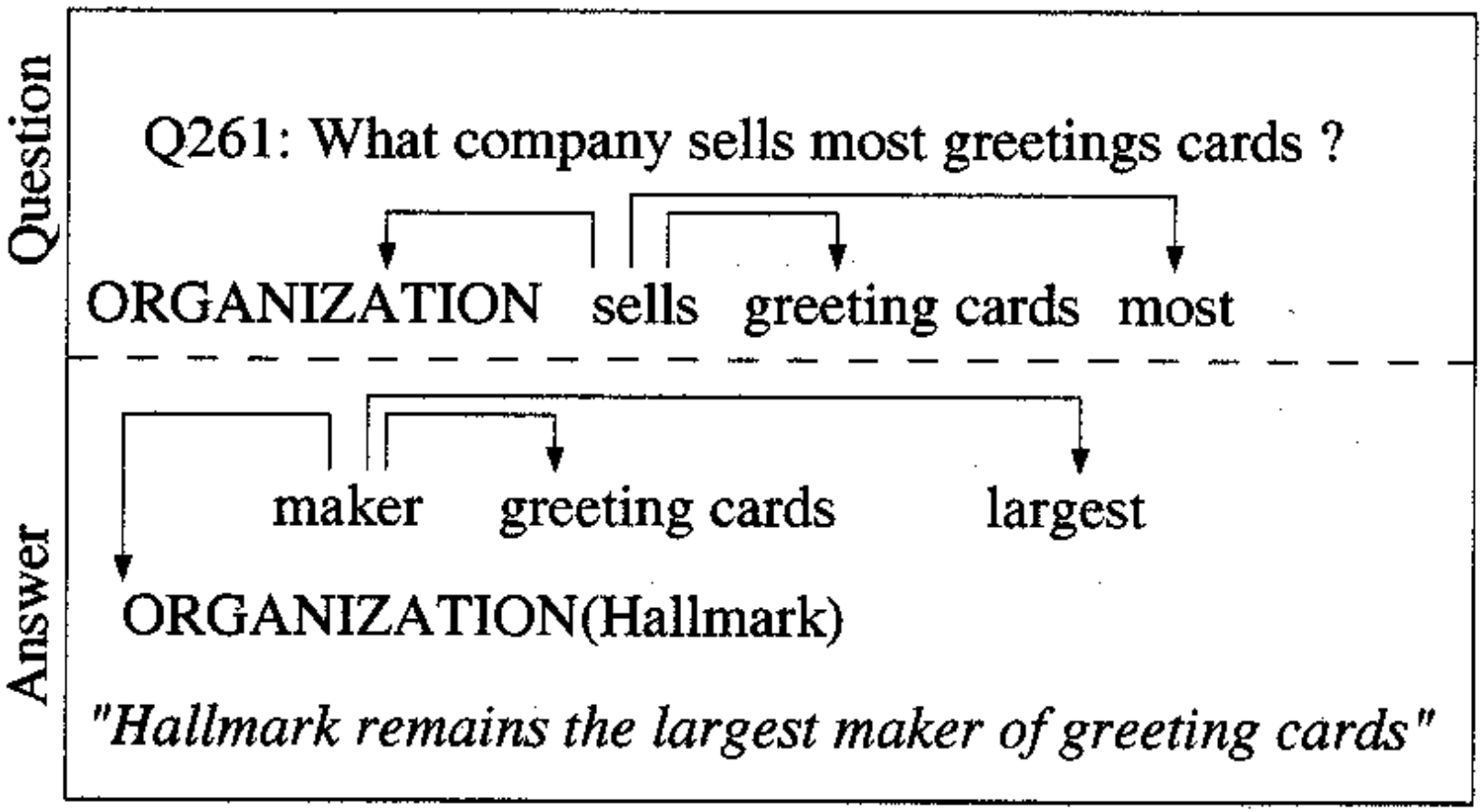
4. Selection and ranking of answers:
   - candidate sentences are scored usually based on the number of matching concepts and the resolution of an empty slot (**expected answer type**), namely the variable of the question
   - answers can be additionally ranked by frequency
5. Possibly answer formulation

Question answering

- **To improve recall** (e.g., no answers found):
  - lexico-semantic alterations of the question based on thesauri and ontologies (e.g., WordNet)
  - morpho-syntactic alterations of the query (e.g., stemming, syntactic paraphrasing based on rules)
  - translation into paraphrases, paraphrase dictionary is learned from comparable corpora
  - incorporation of domain or world knowledge to infer the matching between question and answer sentences

Question answering

- **To improve precision** (e.g., too many answers found):
  - extra constraints: extracted information (e.g., named entity classes, semantic relationships, temporal or spatial roles) from the question and answer sentence must match
  - use of logical representation of question and answer sentences and logic prover selects correct answer

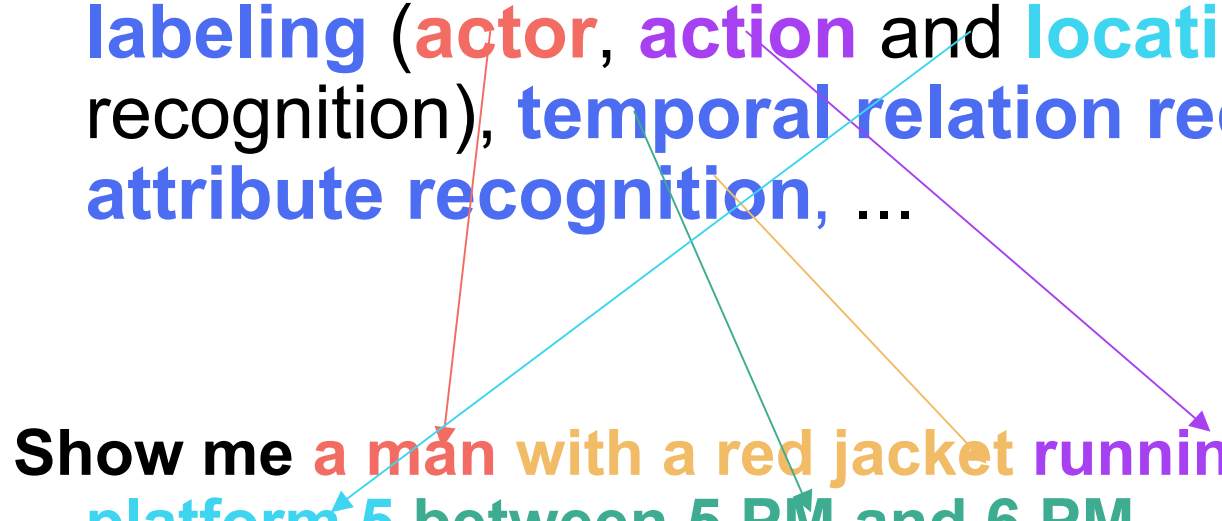


[Pasca & Harabagiu SIGIR 2001]

- Increased role of:
  - **Information extraction**:
    - Understanding the question
    - Content recognition in documents: multimedia content recognition (e.g. in computer vision): information also in text will increasingly be semantically labeled
    - Cf. Semantic Web
  - **Automated reasoning**:
    - To infer a mapping between question and answer statement
    - For temporal and spatial information processing

Understanding the question

- Understanding the question or natural language statement => **semantic role labeling** (**actor**, **action** and **location** recognition), **temporal relation recognition**, **attribute recognition**, ...



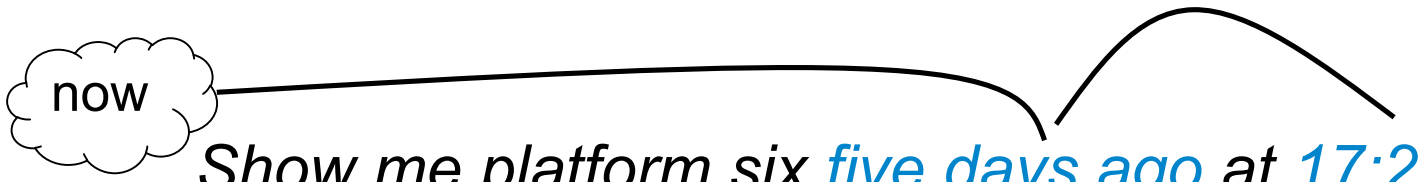
Temporal information processing

- Semantic processing: supervised machine learning techniques (conditional random fields, maximum entropy Markov models, possibly augmented with language models e.g., latent word language model)
- Temporal information is often important:
  - **Temporal expression recognition**
  - **Temporal expression normalisation** (ISO standard)
  - **Temporal relation recognition**

[Boguraev & Ando IJCAI 2005] [Mani et al. COLING-ACL 2006] [Kolomiyets & Moens KI 2008]

Temporal information processing

- Steps
  - Recognition – identify which phrases are temporal and which are not
  - Normalization – estimate an standardized temporal value to a string-based phrase
  - Transformation for querying – standard query languages still do not support temporal values derived from natural language

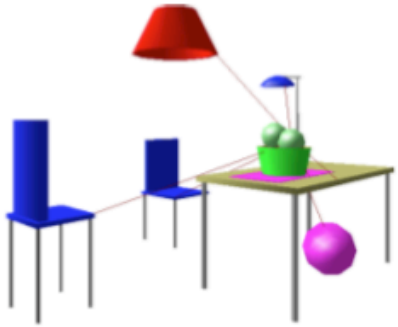


Temporal information processing

- Challenges:
  - Number of temporal units (days, months, parts of day etc.) and magnitudes
  - Vagueness of the language (*17:20* = 5:20 pm.~ *late afternoon* ~ *around 5:30 pm*)
  - Context-based normalization (*17:20* of the day five days ago from *now*, and not of *today*)
  - Augmenting results for querying (an hour ago in spoken language is not strictly 10:45 am.)

Spatial relation recognition

[Kordjamshidi et al. COSLI 2010]



A TABLE-TOP SCENE :"THE APPLES ARE ON THE TABLE"

A STREET-SCALE SCENE :"THE CAR IS IN FRONT OF THE CHURCH"

- Semantic processing: supervised machine learning techniques (conditional random fields, maximum entropy Markov models, possibly augmented with language models e.g., latent word language model)
- Spatial reasoning with the output of the recognizers or already incorporated with the learners?

Example 1

A	woman	and	a	child	are	walking	over	the	square.
none	**trajector**	**none**	**none**	**trajector**	**none**	**none**	**spatial_indicator**	**none**	**landmark**

Example 2

Go	under	the	bridge	.
none	**spatial_indicator**	**none**	**landmark**	

- Emerging information extraction task
- Relevant for question answering of visual data

Question answering

- Difficult task: requires a substantial degree of natural language understanding of the question and of the document texts
- Classes of questions:
  1. **Factual questions**:

     -"When did Mozart die?"

     -Answer verbatim in text or as morphological variation

  2. Questions that need **simple reasoning techniques**:

     - "How did Socrates die? ": "die" has to be linked with "drinking poisoned wine"

     -Needed: ontological knowledge

Question answering

3. Questions that need **answer fusion from different documents:**

   - e.g., “In what countries occurred an earthquake last year?”

   - Needed: reference resolution across multiple texts

4. **Interactive QA systems:**

   - Interaction of the user: integration of multiple questions, referent resolution

   - Interaction of the system: e.g., “What is the rotation time around the earth of a satellite? “ -> “ Which kind of satellite: GEO, MEO or LEO” ?:

      - needed: ontological knowledge

      - cf. expert system

Question answering

5. Questions that need **analogical reasoning:**
   - Speculative questions: “Is the US moving towards a recession?”
   - Most probably the answer to such questions is not found in the texts, but an analogical situation and its outcome is found in the text
   - Needs extensive knowledge sources, case-based reasoning techniques, temporal, spatial and evidential reasoning
   - Very difficult to accomplish due to the lack of knowledge

- Internationale competities:
  - **Message Understanding Conference** (MUC): 1980-1990
  - **Automatic Content Extraction** (ACE) (1999 - 2006): http://www.ldc.upenn.edu/Projects/ACE/
  - **Text Analysis Conference** (TAC) (2008 -): http://www.nist.gov/tac/

What have we learned about information extraction?

- The **older symbolic techniques** learn us:
  - IE = feasible
  - but, knowledge acquisition bottleneck
- The **supervised machine learning** techniques:
  - achieve comparable results
  - offer more flexibility
  - but, still annotation bottleneck
- The **semi-supervised machine learning** techniques
  - artificially creating extra training examples based on external knowledge
- The **unsupervised machine learning** techniques
  - current pattern mining and alignment techniques offer interesting research tracks
- IE becomes important component of IR: e.g., **exploratory search** and **question answering**

References

- Berger, Adam, Stephen A. Della Pietra & Vincent J. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22 (1), 39-71.
- Boguraev, Branimir & Rie Kuboto Ando (2005). TimeML-compliant text analysis for temporal reasoning. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (pp. 997–1003).
- Boiy Erik & Marie-Francine Moens (2009). A machine learning approach to sentiment analysis. *Information Retrieval*, 12, 526-558.
- Bunescu, Razvan C. & Raymond J. Mooney (2006). Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems,* 18.
- Bunescu, Razvan C. & Raymond J. Mooney (2007). Learning to extract relations from the Web using minimal supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL.
- Chambers, Nathanael & Dan Jurafsky. (2009) Unsupervised learning of narrative schemas and their participants. In *Proceedingsof the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (pp. 602–610*).* ACL.
- Chapelle, Olivier, Bernhard Schölkopf & Alexander Zien (Eds.) (2006). *Semi-supervised Learning*. MIT Press.
- Christianini, Nello & John Shawe-Taylor (2000). *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods.* Cambridge, UK: Cambridge University Press.

- De Belder, Jan, Wim De Smet, Rachel Mochales Palau & Marie-Francine Moens (2009). Does Google own Youtube? Entity relationship extraction with minimal supervision. In *Proceedings of SIM 2009, Joint Conference: SRL ILP MLG, K.U.Leuven*.
- Deschacht, Koen & Marie-Francine Moens (2009). Semi-supervised semantic role labeling using the Latent Words Language Model. In *Proceedings of EMNLP 2009* (pp. 21–29). ACL.
- Hobbs, Jerry R. (2002). Information extraction from biomedical text. *Journal of Biomedical Informatics*, 35, 260-264.
- Hobbs, Jerry R., Douglas Appelt, David Israel, Megumi Kameyama, Mark Stickel & Mabry Tyson (1996). FASTUS: A cascaded finite-state transducer for extracting information from natural language text. In Emmanuel Roche and Yves Schabes (Eds.), *Finite State Devices for Natural Language Processing* (pp. 383-406). Cambridge, MA: The MIT Press.
- Joachims, Thorsten (2006). Transductive support vector machines. In Olivier Chapelle, Bernhard Schölkopf & Alexander Zien (Eds.) (2006). *Semi-supervised Learning* (pp. 115-135). Cambridge, MA: MIT Press.
- Kolomiyets, Oleksandr & Marie-Francine Moens (2009). Comparing two approaches for the recognition of temporal expressions. In *Künztliche Intelligenz* (LNAI 5803) (pp 225–232). Berlin: Springer.
- Kordjamshidi, Parisa,Martijn Van Otterlo, & Marie-Francine Moen. (2010). From Language towards Formal Spatial Calculi. In *Proceedings of Computational Models of Spatial Language Interpretation at Spatial Cognition 2010.*

- Lafferty, John, Andrew McCallum & Fernando C.N. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282-289). San Francisco, CA: Morgan Kaufmann.
- Mani, Inderjeet, Marc Verhagen, Ben Wellner, Chon Min Lee & James Pustejovsky (2006). Machine learning of temporal relations. In *Proceedings of COLING-ACL 2006* (pp. 753-760). East Stroudsburg, PA: ACL.
- Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49 (4), 41-46 (2006).
- Marquez, Luis Xavier Carreras, Ken Litkowski & Suzanne Stevenson (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34 (2), 145-159.
- Minsky, Marvin (1975). A framework for representing knowledge. In Patrick H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211-277). New York: McGraw-Hill.
- Moens, Marie-Francine (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context* (The *Information Retrieval Series* 21). New York: Springer.
- Moschitti, Alessandro & Silvia Quarteron (2011). Linguistic kernels for answer re-ranking in question answering systems. *Information Processing & Management.*
- Nigam, K., McCallum, A. & Mitchell, T. (2006). Semi-supervised classification using EM. In Olivier Chapelle, Bernhard Schölkopf & Alexander Zien (Eds.) (2006). *Semi-supervised Learning* (pp. 33-55). Cambridge, MA: MIT Press.

- Pang, Bo & Lillian Lee (2008). Opinion mining and sentence analysis. *Foundations and Trends in Information Retrieval* 2 (1-2), 1–135.
- Regneri, Michaela, Koller, Alexander and Manfred Pinkal (2010). Learning script knowledge with Web experiments. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 979–988). ACL.
- Riloff, E. (1996). An empirical study for automated dictionary construction for information extraction in three domains. *Artificial Intelligence* 85, 101-134.
- Schank, R.C. (1975). *Conceptual Information Processing.* Amsterdam: North Holland.
- Schank, Roger C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology,* 3 (4), 532-631.
- Schank, Roger C. (1975). *Conceptual Information Processing.* Amsterdam: North Holland.
- Suykens, J., De Brabanter, J., Lukas, L & Vandewalle, J. (2002). Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing,* 48, 85-105.
- Yoshikawa, Katsumasa, Sebastian Riedel, Masayuki Asahara & Yuji Matsumoto (2009). Jointly identifying temporal relations with Markov logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 405–413). ACL.