Text Mining, Information and Fact Extraction
Part 1: Introduction and Symbolic Techniques

Marie-Francine Moens
Department of Computer Science
Katholieke Universiteit Leuven, Belgium
sien.moens@cs.kuleuven.be

Text

Text

- Text = constellation of linguistic elements meant for human interpretation in a communicative act
- Written (here focus) and spoken text

Sharapova beats Ivanovic to win Australian Open

A year after being on the wrong end of one of Russian didn't drop a set in seven matches at Melbour g wins the most-lopsided losses in a Grand Slam final, Sharapo r third major title with a 7-5, 6-3 victory over fourth-seeded Ana Ivanovic on Saturday.The 20-year-old over three of the top four ranked players, erasing 12 months worth of painful memories in the wake of her 6-1, 6-2 loss to Serena Williams last year.After Ivanovic sprayed a forehand wide on match point, Sharapova dropped to her knees and appeared t k tears as she waved and blew kisses to the crowd.Then she dropped her racket in her chair before heading to shake hands and exchange high-fives with her father and supporters.

Great victory!

Tennis 2008

4: AMID — NIH cDNA clone, Links

Official Symbol: AMID **and Name:** apoptosis-inducing factor (AIF)-like mitochondrion-associated inducer of death [*Homo sapiens*]

Other Aliases: PRG3, RP11-367H5.2

Other Designations: 5430437E11Rik; apoptosis-inducing

Date: Fri, 25 Jan 2008 11:21:15 +0100 (CET)
From: Maarten Logghe <maarten@yoru.be>
To: Sien Moens <sien.moens@cs.kuleuven.be>
Subject: Re: Op zoek naar een medewerker

Hoi Sien,

Ik vrees dat je nog wat verder zult moeten zoeken... ik kan het niet doen
en ik heb wat rondgevraagd bij vrienden en ex-collega's enzo maar helaas..

veel succes,
maarten

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

To mine

- = to dig under to gain access
- = to extract from a source
- = ...

[Webster]

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."

[Moens 2006]

Fact extraction

- = particular type of information extraction, focusing on factual information
- fact = actual occurrence

Aim of the course

Text mining, information and fact extraction

- To learn the latest techniques of fact extraction from text
- To broaden the scope to more advanced information extraction and text mining
- To illustrate with current applications
- To integrate these techniques in **retrieval models**

Overview of the course

- Part 1: Introduction and symbolic techniques
- Part 2: Machine learning techniques
- Part 3: Machine learning techniques (continued)
- Part 4: Applications
- Part 5: Integration in retrieval models

Overview of part 1

- Examples from text, multimedia
- Examples of typical information extraction tasks
- The role of natural language processing
- The role of machine learning
- Evaluation
- Going back in the history: the symbolic approaches and how to blend them in future approaches

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 and searching
  - Mining
  - Summarizing and synthesis

- How good is the machine already? Some examples

Named entity recognition

Boeing, Airbus vie for Qantus contract

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 …

Noun phrase coreference resolution

Clinton was the third-youngest president, behind Theodore Roosevelt (the youngest president) and John F. Kennedy (the youngest to be *elected* president). He was the first baby boomer president. ...

The husband of Hillary Clinton ...

Relation recognition

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

Event detection and linking

- **REDMOND, Wash., May 3, 2008** Microsoft Corp. (NASDAQ: MSFT) today announced that it has withdrawn its proposal to acquire Yahoo! Inc. (NASDAQ: YHOO).✗We continue to believe that our proposed acquisition made sense for Microsoft, Yahoo! and the market as a whole. Our goal in pursuing a combination with Yahoo! was to provide greater choice and innovation in the marketplace and create real value for our respective stockholders and employees,✝ said Steve Ballmer, chief executive officer of Microsoft. ...

- Breaking: Microsoft Withdraws Yahoo Bid; Walks Away From Deal (Updated)

  Michael Arrington

  Microsoft will announce shortly that they have withdrawn their offer to acquire Yahoo. Talks between the two companies and **their advisors** broke down earlier today, according to a source close to Microsoft, after a failure to come to agreement on price and other terms. ...

...

June 3: During a legal action against the Yahoo board for its alleged failure to act in shareholders' interest, documents reveal that Yahoo's management drew up plans to reject a Microsoft takeover three months before the $45bn offer was made. ...

Sharapova beats Ivanovic to win Australian Oper

A year after being on the wrong end of one of Russian didn
seven matches at I
including wins the most
a
Grand Slam final, Sharapova wrapped up her third major tit
vi
eeded Ana Ivanovic on Saturday.The 2
th
anked players, erasing 12 months worth of painful
m
e of her 6-1, 6-2 loss to Serena Williams last year.After
Iv
orehand wide on match point, Sharapov
r
knees and appeared to be f
as she waved and blew kisses to
the crowd.Then she dropped
chair before heading to shake
hands and exchange high-fives with her father and supporters.

Australian Open won by Maria Sharapova

Great victory!

Location: Melboune Park

Tennis 2008

Sports

Date:
From:
To:
Subject: Re: Op zoek naar een medewerker

Declining job offer

Hoi Sien,

Ik vrees dat
t moeten zoeken... ik kan het niet doen
en ik heb wa
enden en ex-collega's enzo maar helaas..

Non-spam

veel succes,
maarten

4: AMID NIH cDNA clone, Links
Official Symbol: AMID **and Name:** apoptosis-inducing factor (AIF)-like mitochondrion-associated inducer of death [*Homo sapiens*]
Other Aliases: PRG3, RP11-367H5.2
Other Designations: 5430437E11Rik; apoptosis-inducing

L'é tro bel cet voitur Voici tt ce ki me pasione
ds ma petite vi!!!é
1 dotre
truk!!!Avou de Dc

Positive sentiment

Information extraction: more examples

- Extraction of details of an event:
  - e.g., type of event, time, location, number of victims, symptoms of a disease, etc.
- Extraction of information on Web page:
  - e.g., e-mail, date of availability of a product, …
- Extraction of opinions on certain topic
- Extraction of scientific data from publications:
  - e.g., localization of a gene, treatment of a disease, function of a gene,...
- See **PART 5**

- AND NOT ONLY TEXT !!!

Birthday party

Happyness

Cake

Candles

Information extraction from text

- Discipline in between NLP and IR in terms of difficulty and emphasis:
  - NLP: theoretical basis: description of structural properties of language:
    - syntactic, semantic and conceptual structure
    - features, classification scheme
  - IR: theoretical basis: mainly statistics and probability theory

- Tested in ARPA's Tipster Text Program and in the past **Message Understanding Conferences (MUC)**, **Automatic Content Extraction (ACE)** and current **Text Analysis Conference** (**TAC**) (National Institute of Standards and Technology, NIST)

Role of natural language processing

- Object of study = text written in natural language
- Number of properties (features, attributes) that can be exploited
  - Lexical features
  - Morpho-syntactic features
  - Semantic features
  - Discourse features

SWARM INTELLIGENCE

Following a trail of insects as they work together to accomplish a task offers unique possibilities for problem solving.

By Peter Tarasewich & Patrick R. McMullen

Stemming

Even with today's ever-increasing computing power, there are still many types of problems that are very difficult to solve. Particularly combinatorial optimization problems continue to pose challenges. An example of this type of problem can be found in product design. Take as an example the design of an automobile based on the attributes of engine horsepower, passenger seating, body style and wheel size. If we have three different levels for each of these attributes, there are 3^4, or 81, possible configurations to consider. For a slightly larger problem with 5 attributes of 4 levels, there are suddenly 1,024 combinations. Typically, an enormous amount of possible combinations exist, even for relatively small problems. Finding the optimal solution to these problems is usually impractical. Fortunately, search heuristics have been developed to find good solutions to these problems in a reasonable amount of time.

Removal of stopwords

heuristic: weight = 2

Over the past decade or so, several heuristic techniques have been developed that build upon observations of processes in the physical and biological sciences. Examples of these techniques include Genetic Algorithms (GA) and simulated annealing…

Lexical features

- Tokenization = converting input stream of characters into a stream of words or tokens
  - **space-delimited languages** (most European languages):
    - word = string of characters separated by white space
  - **unsegmented languages** (e.g., Chinese, Thai, Japanese):
    - e.g., use of a word list (MRD = machine-readable dictionary)

Lexical features

- Difficulties:
  1. use of special characters:
     - e.g., period in abbreviation can be confused with words that end with a full stop at the end of sentence
     - apostrophes
     - hyphens
     - => need for language specific rules
  2. normalization of numbers

Lexical features

- Use of a finite state automaton or finite state machine:
  - often integrates:
    - transformations (e.g., case of letters, abbreviations and acronyms)
    - removal of stopwords (caution!)

Morphological transformations

- **Lemmatization** = finding the lemma or lexeme of an inflected word form (the lemma is the canonical dictionary entry form of the word)
- Lookup of terms and their lemma in a **machine-readable dictionary (MRD)**:
  - correct (e.g., ponies -> pony)
  - often large lists that need to be searched efficiently
  - not always available or not all words covered

Morphological transformations

- **Stemming** = reducing the morphological variants of the words to their stem or root
- **Affix removal algorithms**:
  - language dependent rules to remove suffixes and/or prefixes from terms leaving a stem
  - possible language dependent transformation of the resulting stem
  - examples:
    - Lovins stemmer (1968)
    - **Porter algorithm** (1980)

POS tagging and sentence parsing

- **Part-of-speech** (POS) or syntactical word class:
  - contributes to the meaning of the word in a phrase
  - distinct component in syntactic structure

  - **content words:** nouns, adjectives, verbs, adverbs
  - **function words**: have functional properties in syntactic structure: act as determiners, quantifiers, prepositions and connectives (e.g., articles, pronouns, particles, …)

POS tagging and sentence parsing

```
(S1 (S (NP (NNP Lady) (NNP Hera))
   (VP (VBD was)
    (NP (NP (DT a)
            (ADJP (JJ jealous)
             (, ,)
             (JJ ambitious)
             (CC and)
             (JJ powerful))
            (NN woman))
     (SBAR (WHNP (WP who))
         (S (VP (VBD was)
            (ADVP (RB continually))
            (VP (VB irated)
             (PP (IN over)
               (NP (NP (NP (NNP Zeus) (POS ')) (NN pursuit))
                (PP (IN of)
                    (NP (JJ mortal) (CC and) (JJ immortal) (NN woman)))))))))
   (. .)))
```

POS tagging and sentence parsing

- Parse features:
  - **Path**: describes the path through the parse tree from the current word to the position of the predicate (verb), e.g., NNP↑NP↑PP↑NP↑S↓VP↓VBG:
    - ↑= indicates going up one constituent
    - ↓= indicates going down one constituent
  - **Split Path**: the path feature causes an explosion of unique features: reduced by splitting the path in different parts and use every part as a distinct feature: e.g.: split of the above path in different features: NNP; ↑NP; ↑NP;↑PP; ↑S; ↓VP; ↓VBG

Other natural language features

- **Semantic features**: e.g., from lexico-semantic resources, obtained in previous extraction tasks
- **Discourse features**: e.g., discourse distance
- And features typical for digital documents: e.g., HTML tags
- These features describe the information unit to be classified and its context

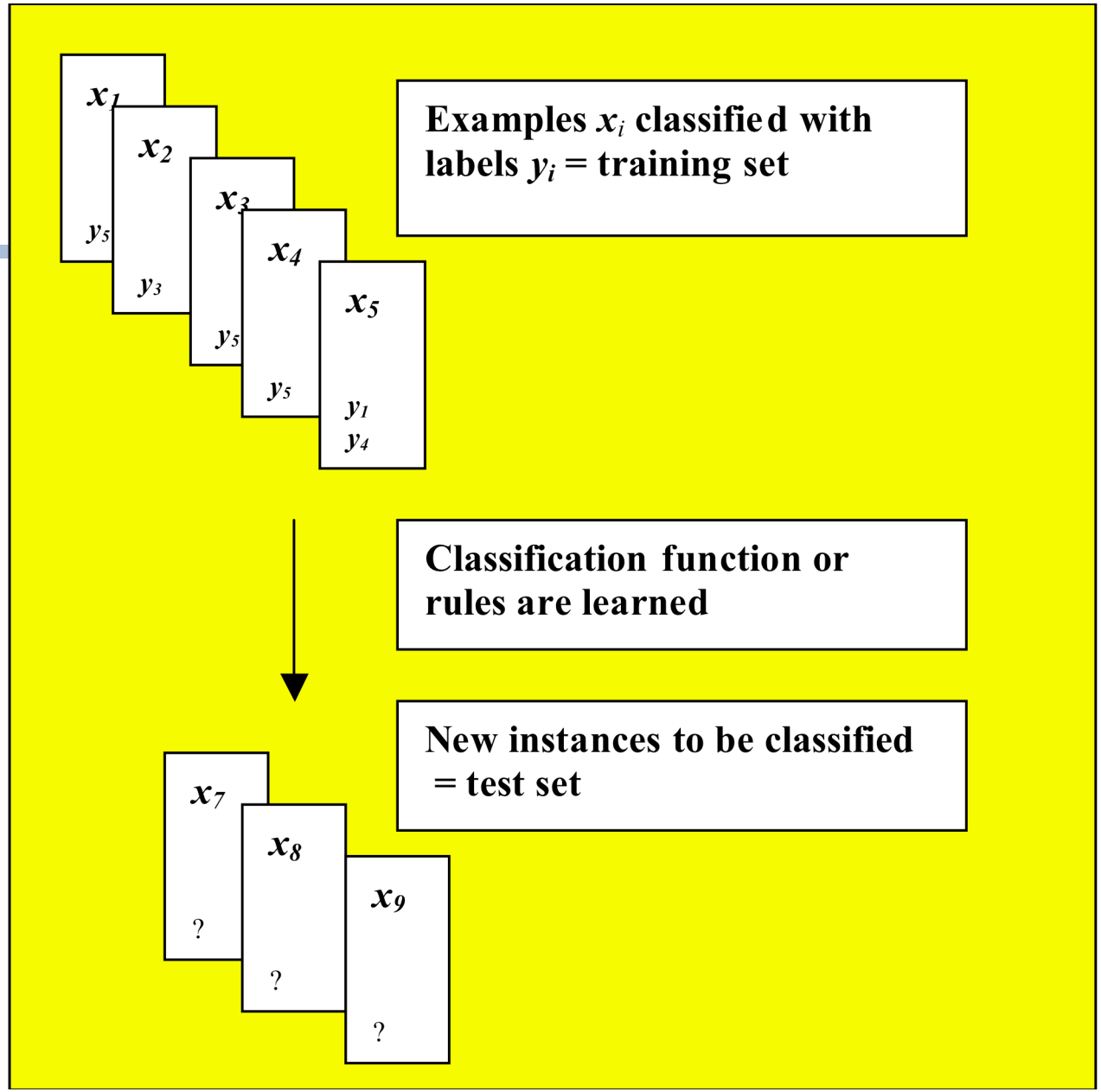
The role of machine learning

- Major advances in pattern classification today:
  - Classification of objects (e.g., images, text, ...) into a number of categories:

    => **recognition of content**
  - Many of the techniques involve machine learning and statistical classification

Supervised learning

- **Techniques of supervised learning***:*
  - **training set**: example objects classified by an expert or teacher
  - detection of general, but high-accuracy classification patterns (function or rules) in the training set based on object features and their values
  - patterns are predictable to correctly classify new, previously unseen objects in a **test set** considering their features and feature values

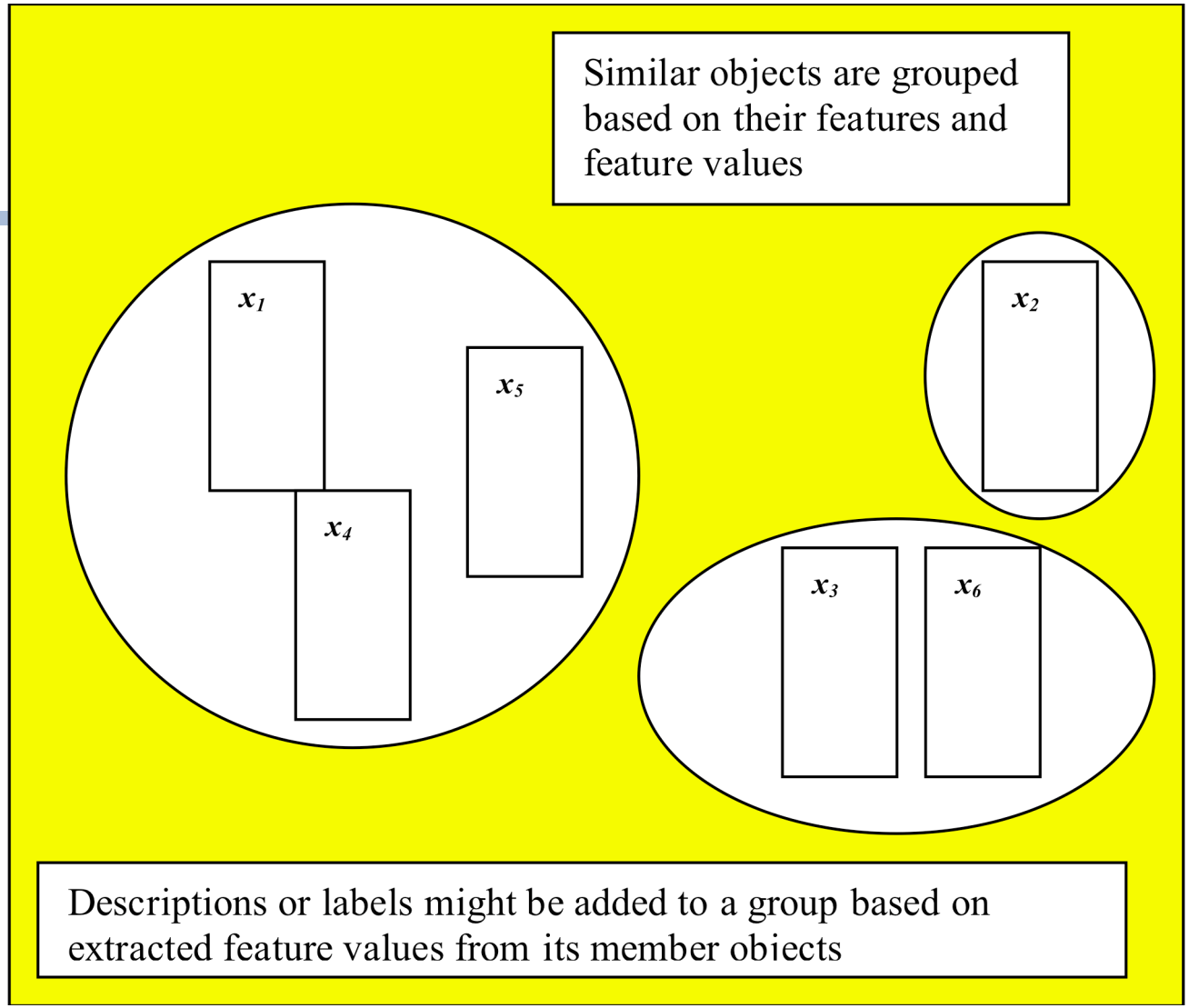


Supervised learning

- Text classification can be seen as a:
  - two-class learning problem:
    - an object is classified as belonging or not belonging to a particular class
    - convenient when the classes are not mutually exclusive
  - single multi-class learning problem
- Result = often **probability** of belonging to a class, rather than simply a classification

Unsupervised learning

- **Techniques of unsupervised learning** :
  - natural groupings of similar objects are sought based object features and their values
  - often use of simple hard and fuzzy clustering techniques



Weakly supervised learning

- Techniques of **weakly supervised learning**
  - supervised learning starting from a limited set of classified **seed** objects
  - exploit knowledge from set of unlabeled examples
  - often iterative learning until results on validation set cannot anymore be improved
  - E.g., self-training, co-training

In this course

- Information extraction originally developed for very limited domains, but interest in
  - Portability of the techniques to other domains or to open domain
  - Focus on **supervised learning**, because of high cost of hand-crafted patterns
- Large interest in **semi-supervised learning**, because of high cost of labeling examples
  - but, for natural language data difficult

Information extraction

- Classification scheme = semantic labels and their relationships (external knowledge)
  - **Domain-independent**: e.g., coreferent relations
  - **Domain-dependent**: e.g., biomedical name classes

  - Form:
    - list
    - hierarchy
    - binary scheme
    - **ontology**, labels can have relationships (e.g., hierarchically organized)

Evaluation: confusion matrix

- Column: gives number of instances classified by system in the specific class
- Row: gives number of instances classified by expert in the specific class
- Easy to see if system confuses two classes
- Built for binary and multi-class classification problems

Evaluation: confusion matrix

- Confusion matrix of binary classification decisions (e.g., for intrinsic evaluation of e.g., text categorization, information extraction, classification in relevant - non-relevant documents):

	System says yes	System says no
Expert says yes	*tp*	*fn*
Expert says no	*fp*	*tn*

where

tp = true positives

fn = *false negatives*

fp = false positives

tn = *true negatives*

Evaluation: confusion matrix

recall = $tp / (tp + fn)$

precision = $tp / (tp + fp)$

error rate = $(fp + fn) / (tp + fp + fn + tn)$

accuracy = $(tp + tn) / (tp + fp + fn + tn)$

Evaluation: F-measure

- **F-measure**: combines recall and precision

$$F = \frac{(\beta^2 + 1)\ \text{precision x recall}}{\beta^2\ \text{precision} + \text{recall}}$$

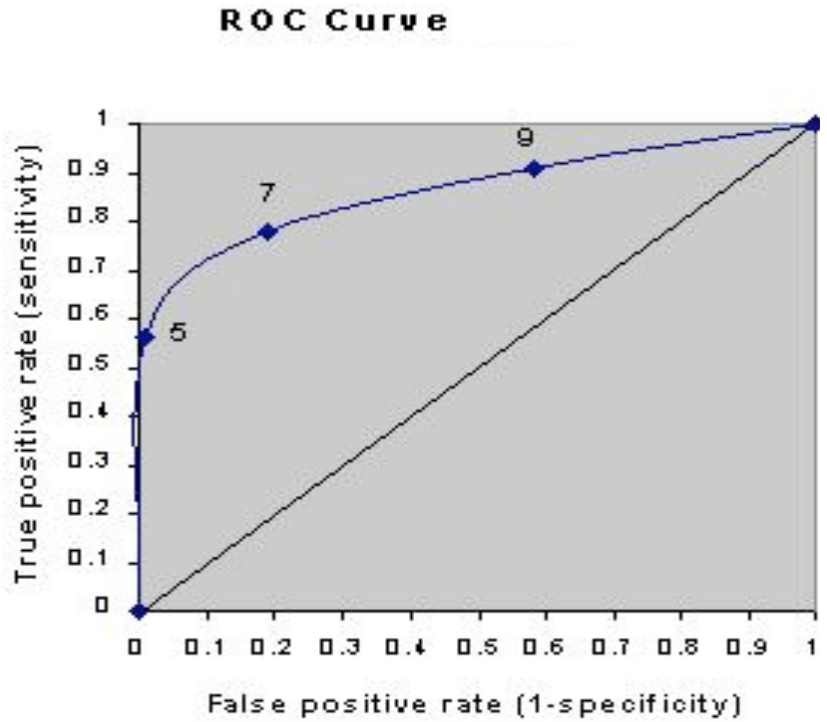
where

β = a factor that indicates the relative importance of recall and precision

- ideally close to 1
- when β = 1: also called harmonic mean= F_1

ROC curve

- **Receiver Operating Characteristic curve**: area under curve should be maximized



1-specificity (= $fp/(fp+tn)$)
sensitivity (= $tp/(tp+fn)$)

The symbolic approaches

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

FMLN terrorists in retaliation for recent arrests attempted to kill 5 policemen with car bombs.

- Task: to extract : the perpetrators (FMLN terrorists), the victims (5 policemen) and the weapons (car bombs).
- The following extraction patterns would do the job:

<perpetrator> attempted to kill
<subj> verb infinitive

attempted to kill <victim>
verb infinitive <obj>

to kill with <weapon>
infinitive prep <np>

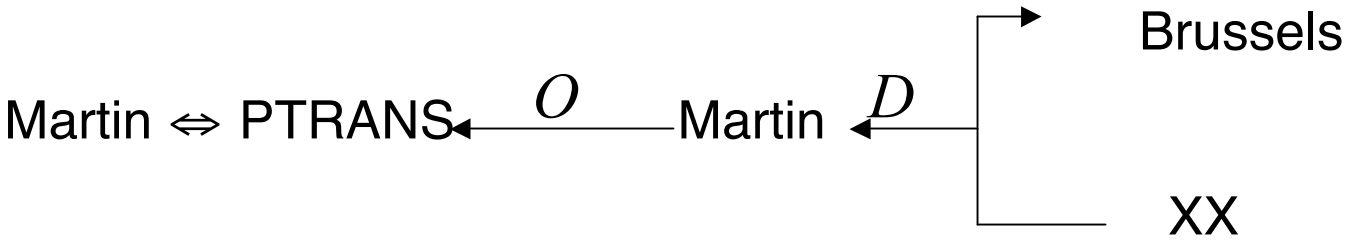
Early origin

- ° end 1960s and 1970s: Schank:
  - 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,** which outlines sequences of events or actions

Sentence:

Martin goes to Brussels.

will be graphically represented in **CD theory** as follows:



meaning that Martin performs the act of changing the location (PTRANS) of Martin from an unknown location (indicated by XX) to Brussels

where *O* and *D* indicate an objective and a directive relationship respectively

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

- Implementation:
  - Linguistic preprocessing of the text:
    - POS tagging, parsing, named entity recognition, ...
  - Mapping of the frames to the texts :
    - feature slots: labels, fixed for a particular frame
    - feature values: fill the slots with extracted information, certain constraints can be placed
  - Frames can be connected in a semantic net
    - advantages:
      - default values, inherited values

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

Frame-based approaches

- Examples of some famous information extraction applications:
  - FRUMP (DeJong, 1982): Yale University
  - FASTUS (Hobbs et al., 1996): 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. IJCAI 1993] [Hobbs JBioInformatics 2002]

Cascade of finite state transducers

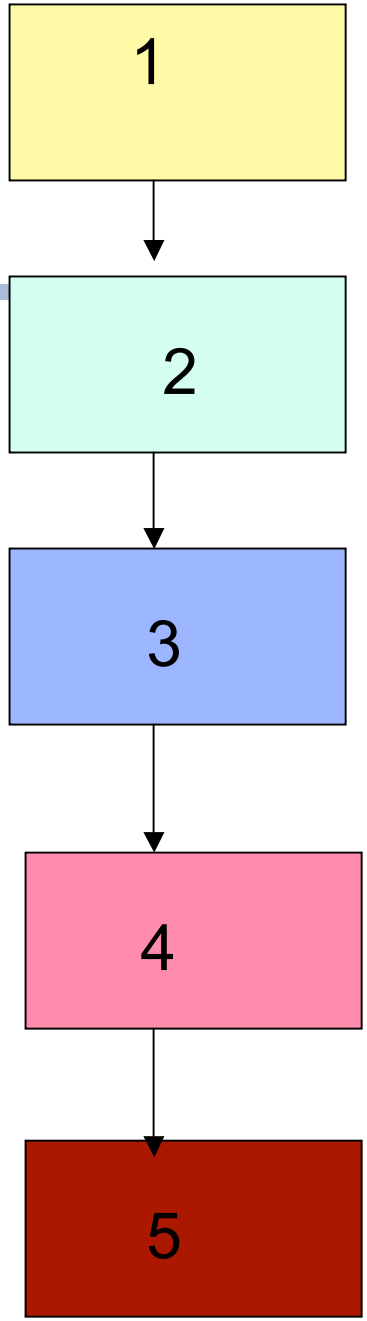
1. Recognition of compound words and named entities

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	**J a p a n**

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 to learn from the symbolic techniques?

- 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

  - Examples:
    - syntactic reformulation rules
    - rules for stemming

Today

- Cf. example above: Similar pipelined structure, but (supervised) machine learning models:
  - Named entity recognition (NER)
  - Syntactic analysis of the sentence (e.g., part-of-speech tagging, sentence parsing)
  - Recognition of relations between entities

References

DeJong, G. (1982). An overview of the FRUMP system. In W.G. Lehnert & M.H. Ringle (Eds.), *Strategies for Natural Language Processing* (pp. 149-176). Hillsdale: Lawrence Erlbaum.

Hobbs, J. H. et al. (1996). FASTUS: a cascaded finite-state transducer for extracting information from natural-language text. In *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge MA.

Minsky, Marvin (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211-277). New York: McGraw-Hill.

Moens, M.-F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context* (*The Information Retrieval Series* 21). New York: Springer.

MUC-7 (1999). *Message Understanding Conference Proceedings MUC-7.*

Schank, R.C. (1975). *Conceptual Information Processing.* Amsterdam: North Holland.