NLP Applications

- Two main areas:
  - Massive management of textual information sources:
    - for human use
    - for automatic collection of linguistic resources
  - Person/Machine interaction

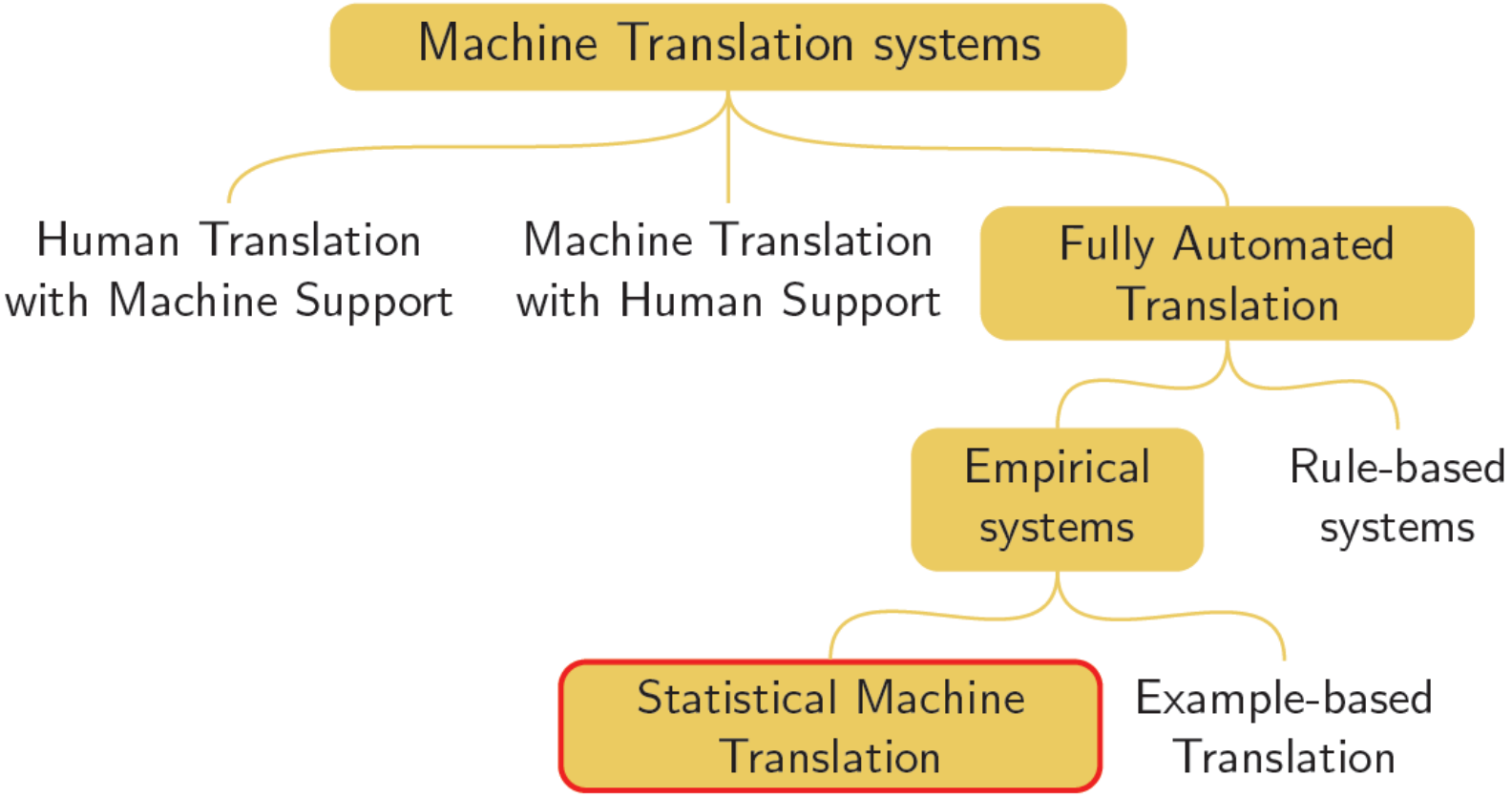
NLP Applications

- Massive management of textual information sources
  - Machine Translation (MT)
  - Information Retrieval (IR)
  - Question Answering (Q&A)
  - Information Extraction (IE)
  - Document Classification and Clustering

Machine Translation 1

- Process of translating a text from a source language to a target language preserving some properties
  - The main property to preserve (but not the only one) is the meaning
- MT textual vs oral
- Different degrees of human intervention

Machine Translation 2



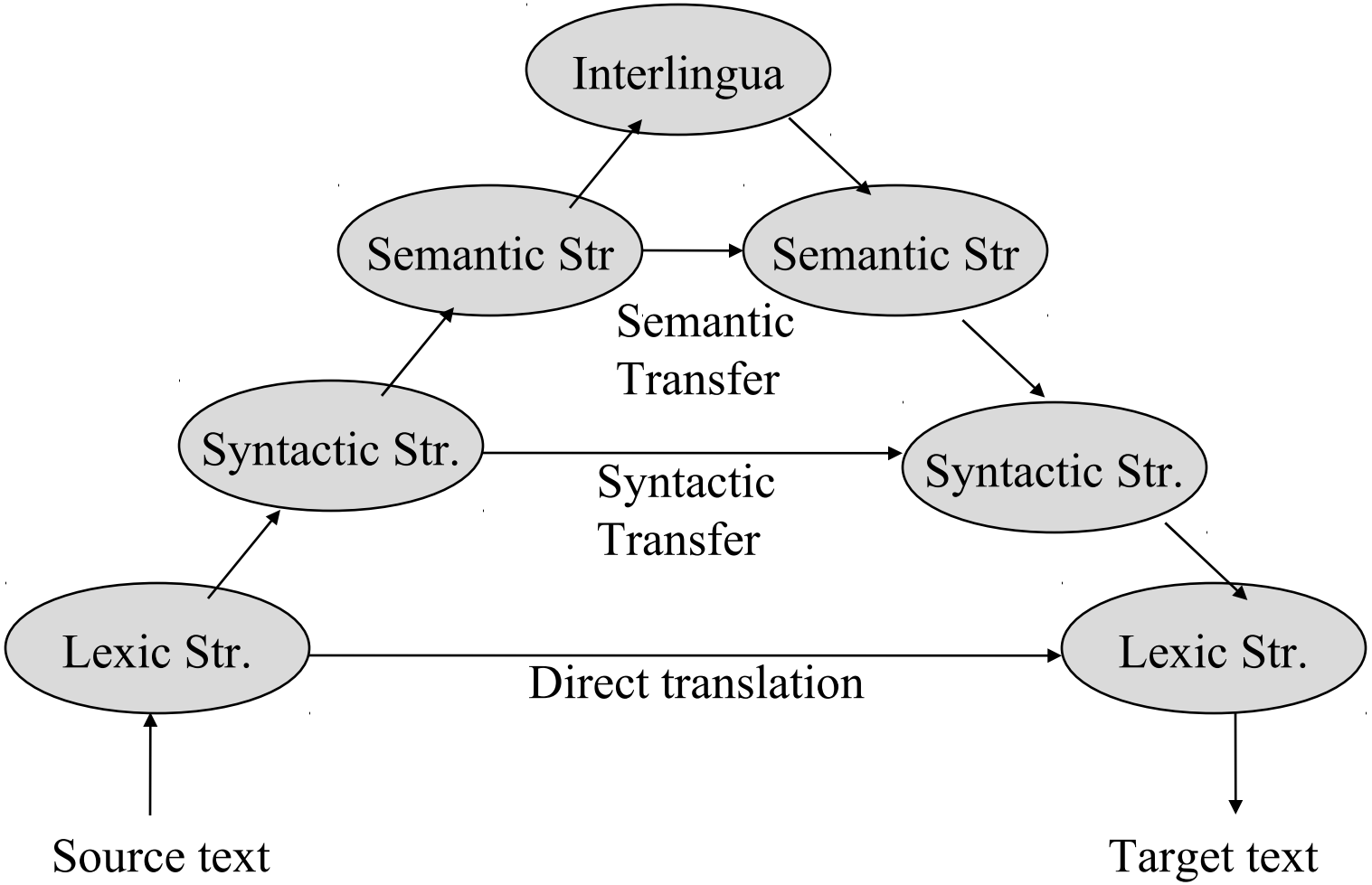
Machine Translation 3

- Some readings
  - General
    - Juan Alberto Alonso (2000) *La Traducció automàtica* chapter 4 of *Les tecnologies del llenguatge*, M.A.Martí (ed) UOC
  - SMT
    - Kevin Knight (1999)
      - http://www.isi.edu/natural-language/people/knight.html
    - Cristina España (2012) Introduction to Statistical Machine Translation
    - Software:
      - Giza++, Moses
    - Projects:
      - MOLTO, OpenMT

Machine Translation 4

- Basic approaches
  - Direct MT
  - Transfer-based
  - Interlingua-based
  - Translation Memories
- Statistic vs symbolic approaches

Machine Translation 5



Aligned parallel corpora numbers

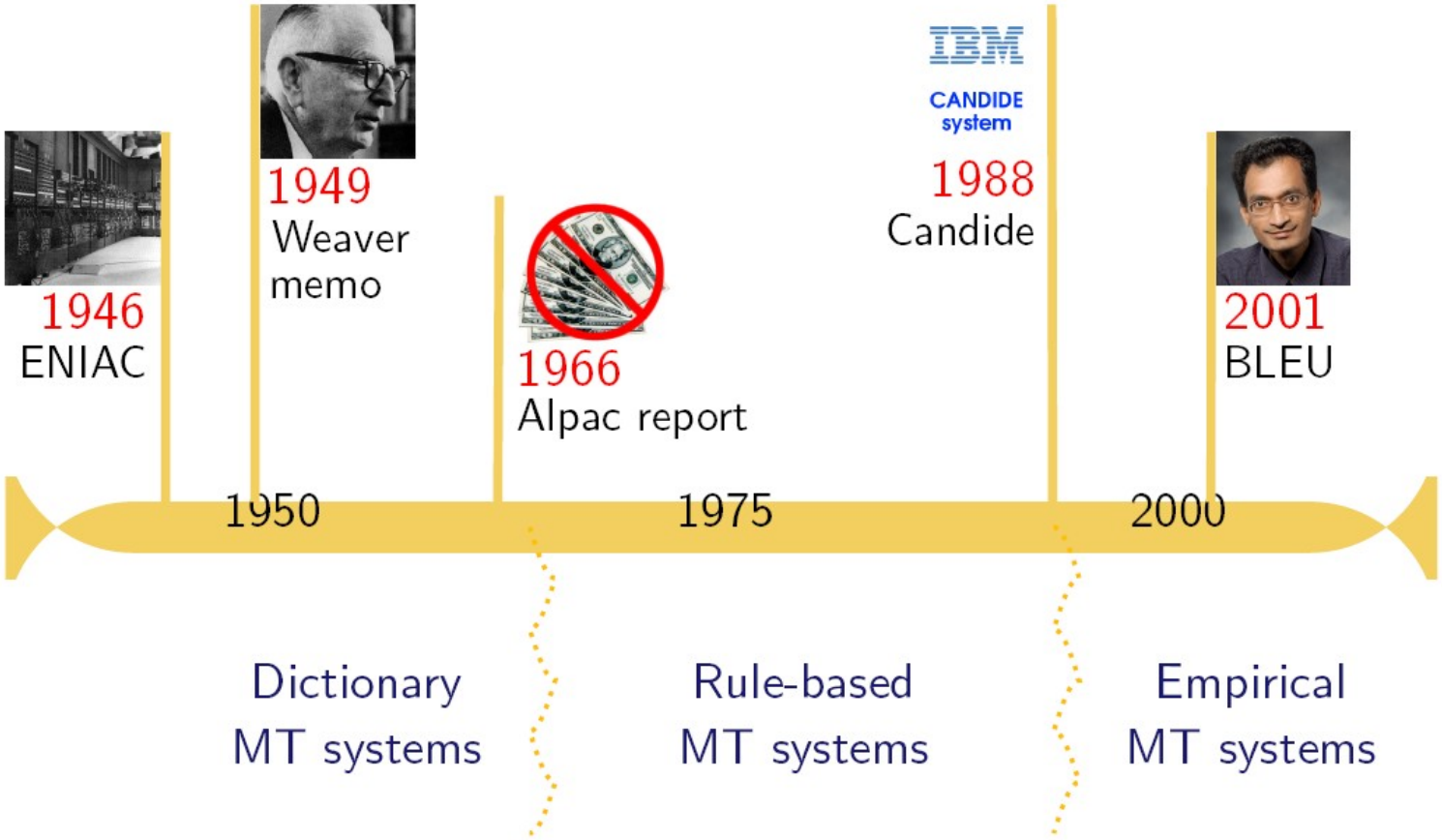
Corpora

Corpus	# segments (app.)	# words (app.)
JRC-Acquis	$1.0 \cdot 10^6$	$30 \cdot 10^6$
Europarl	$1.5 \cdot 10^6$	$45 \cdot 10^6$
United Nations	$3.8 \cdot 10^6$	$100 \cdot 10^6$

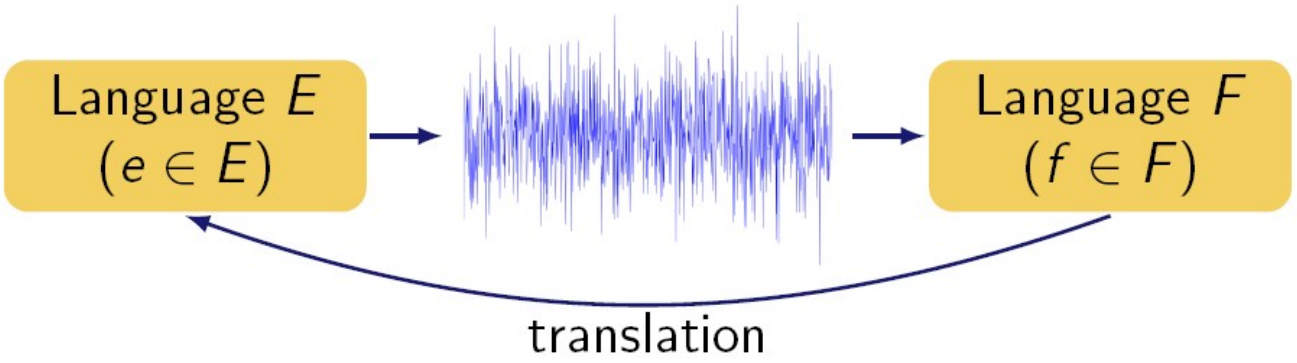
Books

Title	# words (approx.)
The Bible	$0.8 \cdot 10^6$
The Dark Tower series	$1.2 \cdot 10^6$
Encyclopaedia Britannica	$44 \cdot 10^6$

Machine Translation 7



Statistical Machine Translation 1



Mathematically:

$$P(e|f) = \frac{P(e)\,P(f|e)}{P(f)}$$

$$T(f) = \hat{e} = \text{argmax}_e\ P(e|f) = \text{argmax}_e\ P(e)\,P(f|e)$$

Statistical Machine Translation 2

$$T(f) = \hat{e} = \text{argmax}_e \; P(e)\, P(f|e)$$

Language Model

- Takes care of fluency in the target language
- Data: corpora in the target language

Translation Model

- Lexical correspondence between languages
- Data: aligned corpora in source and target languages

argmax

- Search done by the *decoder*

Statistical Machine Translation 3

Language model

$$T(f) = \hat{e} = \text{argmax}_e \; P(e)\, P(f|e)$$

Estimation of how probable a sentence is.

Naïve estimation on a corpus with N sentences:

Frequentist probability of a sentence e:

$$P(e) = \frac{N_e}{N_{sentences}}$$

Problem:

- Long chains are difficult to observe in corpora.
  $\Rightarrow$ Long sentences may have zero probability!

The n-gram approach

The language model assigns a probability $P(e)$ to a sequence of words $e \Rightarrow \{w_1, \ldots, w_m\}$.

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1})$$

- The probability of a sentence is the product of the conditional probabilities of each word w_i given the previous ones.
- Independence assumption: the probability of w_i is only conditioned by the n previous words.

Statistical Machine Translation 5

- Translation Model P(f|e)
  - source: $f = f_1 f_2 ... f_m$
  - target: $e = e_1 e_2 \ldots e_l$
  - alignment: $a = a_1 a_2 \ldots a_m$
  - in general
    - $a \in \{1,\ldots,m\} \times \{1,\ldots,l\}$
  - usually
    - $a: \{1,\ldots,m\} \rightarrow \{0,\ldots,l\}$
      - $a(j) \neq 0$ f_j is mapped into $e_{a(j)}$
      - $a(j) = 0$ f_j is not aligned
  - A(f,e) is the set of possible aligments (2^{lm})

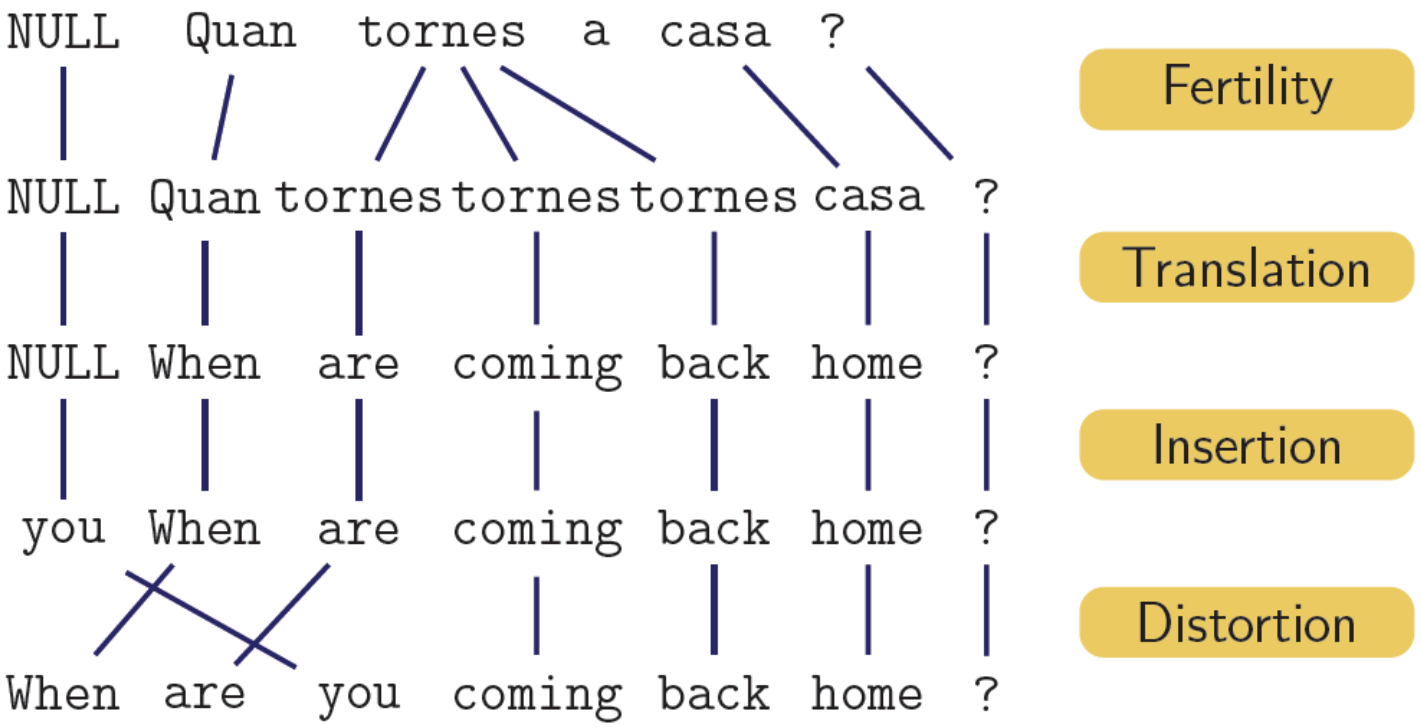
Statistical Machine Translation 5

- Translation Model P(f|e).
  - One should at least model for each word in the source language:
    - Its translation,
    - the number of necessary words in the target language,
    - the position of the translation within the sentence,
    - and, besides, the number of words that need to be generated from scratch.

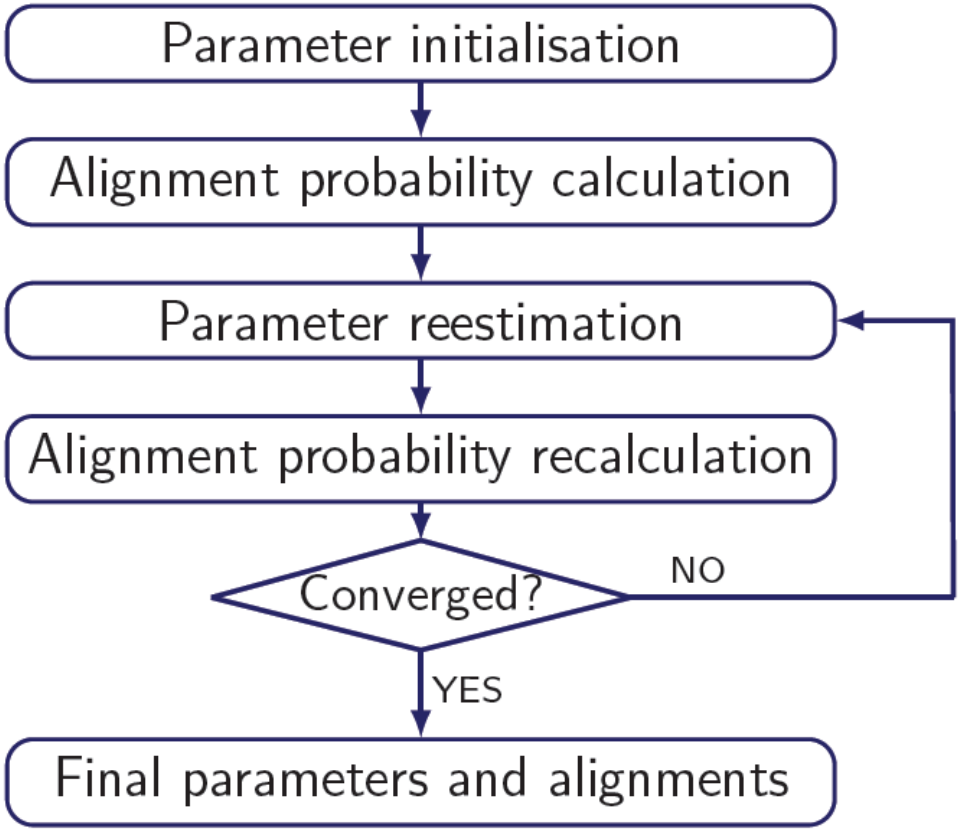
Statistical Machine Translation 6

- Word-based models: the IBM models
  - They characterise P(f |e) with 4 parameters: t, n, d, p1.
    - Lexical probability t
      - t(Quan|When): the prob. that Quan translates into When.
    - Fertility n
      - n(3|tornes): the prob. that tornes generates 3 words.
    - Distortion d
      - d(j ji ;m; n): the prob. that the word in the j position generates a word in the i position. m and n are the length of the source and target sentences.
    - Probability p1
      - p(you|NULL): the prob. that the spurious word you is generated (from NULL).

Statistical Machine Translation 7



Expectation-Maximisation algorithm



Statistical Machine Translation 9

Decoder

$$T(f) = \hat{e} = \text{argmax}_e \; P(e)\, P(f|e)$$

Responsible for the search in the space of possible translations.

Given a model (LM+TM+...), the decoder constructs the possible translations and looks for the most probable one.

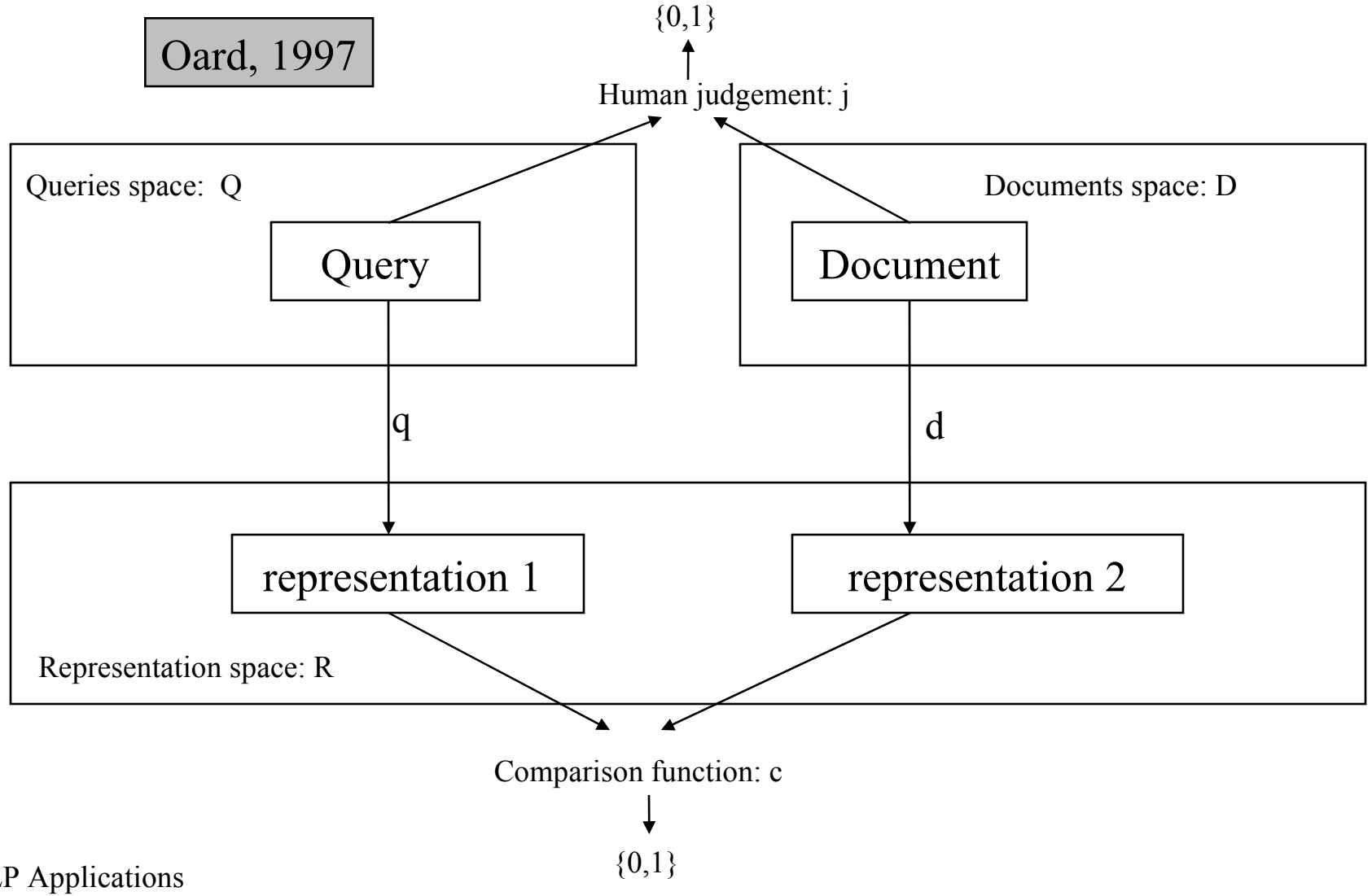
In our context, one can find:

- Greedy decoders. Initial hypothesis (word by word translation) refined iteratively using hill-climbing heuristics.
- Beam search decoders.

Information Retrieval 1

- Input
  - A collection of documents
    - The Web
    - A corporate document collection
    - ...
  - A user need represented as a query
- Output
  - The documents of the collection that satisfy the user needs.

Information Retrieval 2

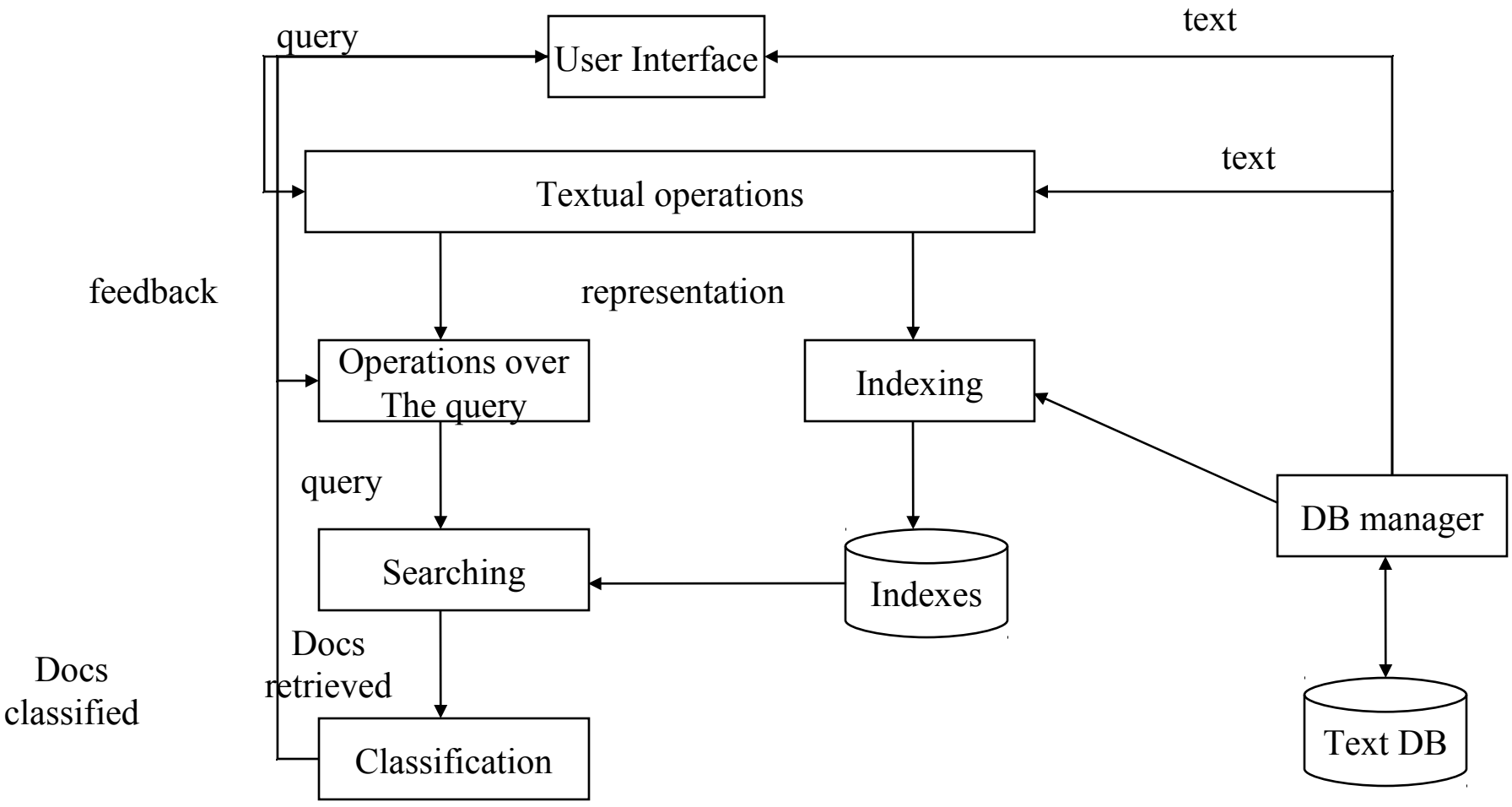


Information Retrieval 3

Ideal setting

$$c(q(query), d(doc)) = j(query, doc)$$
$$\forall query \in Q$$
$$\forall doc \in D$$

Information Retrieval 4



Information Retrieval 5

IR types

- Type of information
  - Text, speech, structured information
- Query language
  - Exact, ambiguous
- Matching
  - Exact, aproximate
- Kind of information needed
  - Loose, precise
- Relevance:
  - Usefulness of information according to user needs

Information Retrieval 6

Operations on texts & queries

- Preprocess
  - Lexical analysis, estandardization
    - non estandard forms, dates, numbers, acronyms, abbreviations, idioms, ...
  - lematization
    - Morphological analysis, stemming (Porter's stemmer)
  - filtering
  - Stopwords
- Classification
  - manual
  - Automatic
  - Classification vs clustering
- Compression

Information Retrieval 7

Indexing

- manual vs automatic
- indicators
  - objetive: structural
  - subjective: textual (content)
- indexing pre-coordinate vs post-coordinate
- Simple terms vs Complex terms (multiwords)

Most frequent : Bag of simple words

Information Retrieval 8

Representing documents

- Classical Models
  - Full text
  - Boolean
  - Vectorial
  - Probabilistic
- Variants of the Probabilistic Model
  - Bayesian
  - Statistic Graphical Models
- Other paradigms
  - Generalized vectorial model
  - Extended Boolean Model
  - Latent Semantic Indexing
  - Neural Nets

Information Retrieval 9

Simple Boolean Model

Boolean expressions over terms occurring in the document (key words).
Logical connectors: AND, OR, NOT
parenthesis

Extensions

distance constraints (at paragraph or sentence level)
Fixed or variable window

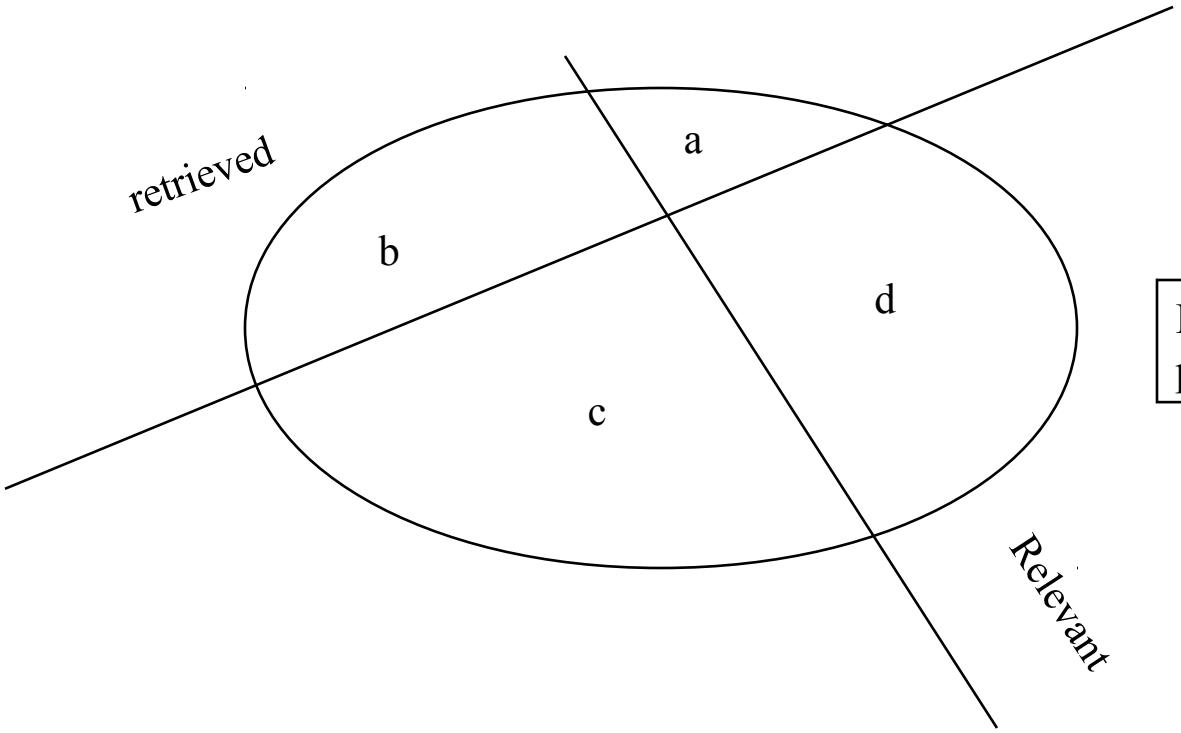
Extended Boolean Model

Term weighting: term frequency in the document, in the collection, normalization

Query expansion

-Use of external knowledge sources (e.g. WN) extension with synonyms and/or hyponyms
- Morphological generalization
- Relevance
- Feedback

IR quality measures



retrieved = a + b
relevants = a + d
recall = a / (a + d)
precision = a / (a + b)

F: weighted harmonic mean of precision and recall

$$F = \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}$$

When the result is not a Boolean but an ordered list of documents with an associated relevance score (ranked) measures can be vectors of precision at (usually) 3, 5, 7, 9, 11 points of recall (e.g. at 0, 0.25, 0.5, 0.75, 1)

Boolean Model

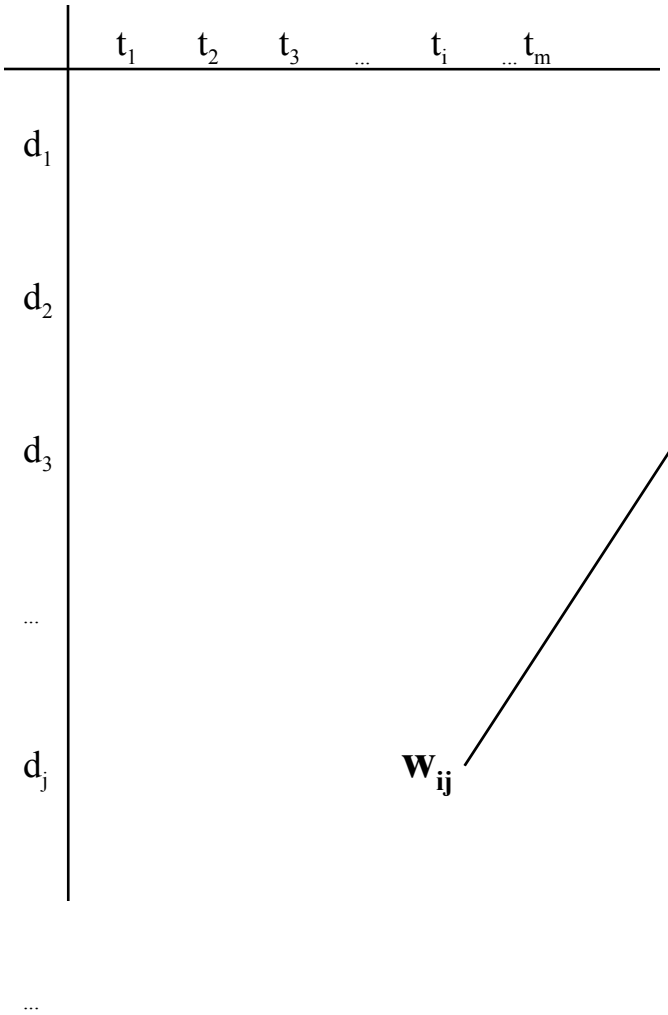
	t_1	t_2	t_3	...	t_i	... t_m
d_1	0	1	0			
d_2	1	0	1	0		
d_3						
...						
d_j						
...						

attributes: all the terms (words, lemmas, multiwords, ...) occurring in the collection (except stopwords). Sometimes only the most frequent.

rows: each document represented by a vector of Booleans (1 if the term occurs in the document, 0 otherwise). For *n* documents

columns: each term represented by a vector of Booleans . For m terms

Vectorial Model



w_{ij} weight (relevance) of term *j* in document *i*

Most used way of computing relevance: TF*IDF

tf_{ij} frecuency of term t_j in the document d_i

df_j # documents containing t_j

idf_j $\log (N / df_j)$

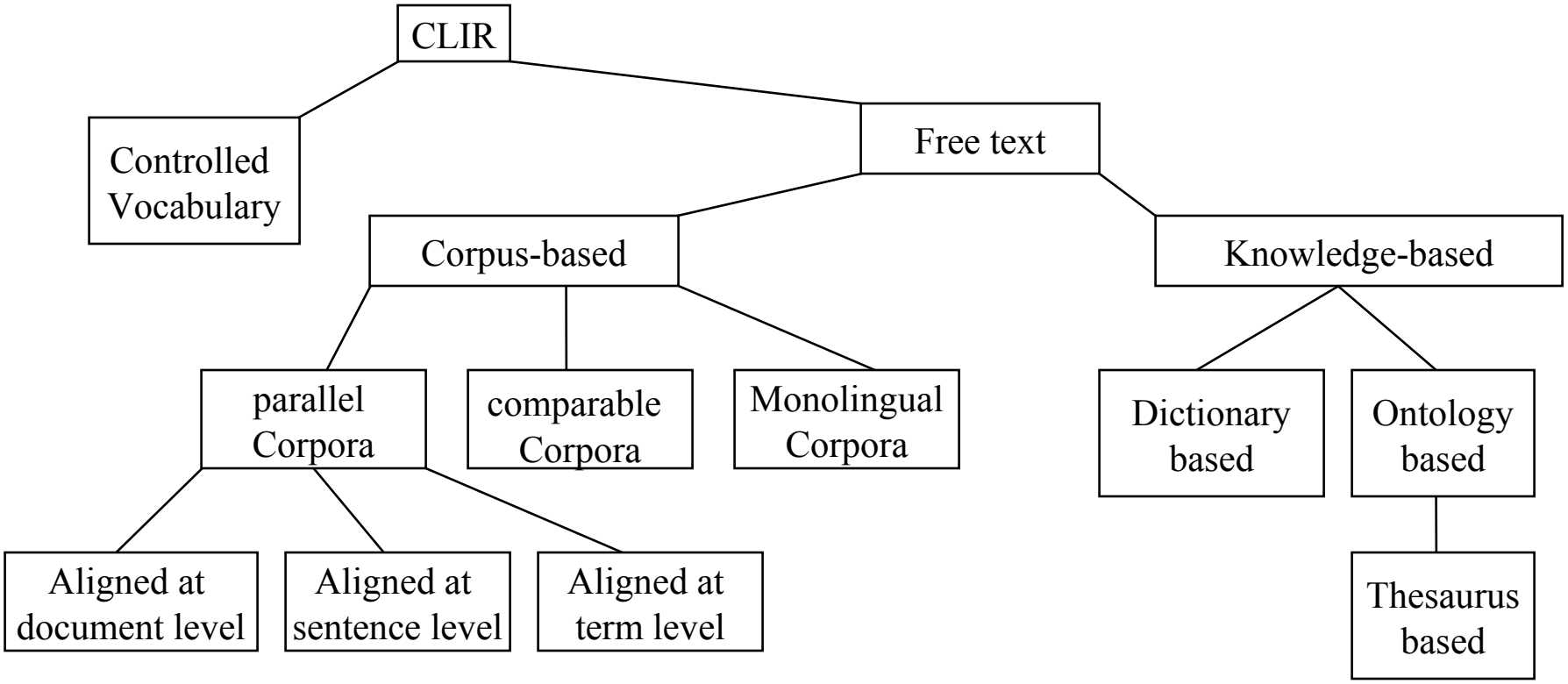
$w_{ij} = tf_{ij} * idf_j$

IR and NL

- NL Resources
- NL Processors
  - Indexing
    - words, stems, lemmas, senses, multiterms
    - phrases, …
    - problems:
      - Named entities
      - Unknown words
      - Non standard units
      - polysemy
    - => Only slight improvement over using forms
  - Retrieval
    - Query expansion

Cross Language Information Retrieval 15

CLIR, Oard, 1997



Question Answering 1

- Natural extension of IR
- A QA system receives a query expressed in NL and tries to provide not a document containing the answer but the proper answer (usually a fact).
- QA systems need to use NLP techniques for both processing the question and looking for the answer.

Question Answering 2

- Some QA systems that can be accessed through the Web:
  - Webclopedia
    - http://www.isi.edu/natural-language/projects/webclopedia/
  - AskJeeves
    - http://www.ask.com
  - LCC
    - http://www.languagecomputer.com/

Question Answering 3

- Starting in TREC challenges from del TREC-8 (1999)
- Later CLEF challenges
- Related Disciplines
  - Answer Finding
    - Given a collection of questions and answers the task consists on looking for the question(s) closest to the one formulated by the user in order to provide its answer.
    - FAQ Finder
  - NL Interfaces to databases
  - Information Integration, II
  - Information Extraction, IE
  - Answer Validation Exercise (AVE)

Question Answering 4

- Factual QA
  - Who? When? Where?
- List QA
  - Which are the last 10 presidents of USA?
- Domain independent vs domain restricted QA
- QA with complex queries:
  - Which are the USA republican presidents after world war II?
- Linked queries

Question Answering 5

Some readings

- Horacio Rodriguez (2001)

  http://www.lsi.upc.es/~horacio/varios/qaBuenosAires.zip

- Documentos de las conferencias TREC

  http://trec.nist.gov/pubs/trec8/t8_proceedings.html

  http://trec.nist.gov/pubs/trec9/t9_proceedings.html

  http://trec.nist.gov/pubs/trec10/t10_proceedings.html

http://www.isi.edu/natural-language/projects/webclopedia/
http://www.languagecomputer.com/
http://www.dlsi.ua.es/~vicedo/

Question Answering 7

Most QA systems consist on 4 processes

Question Processing

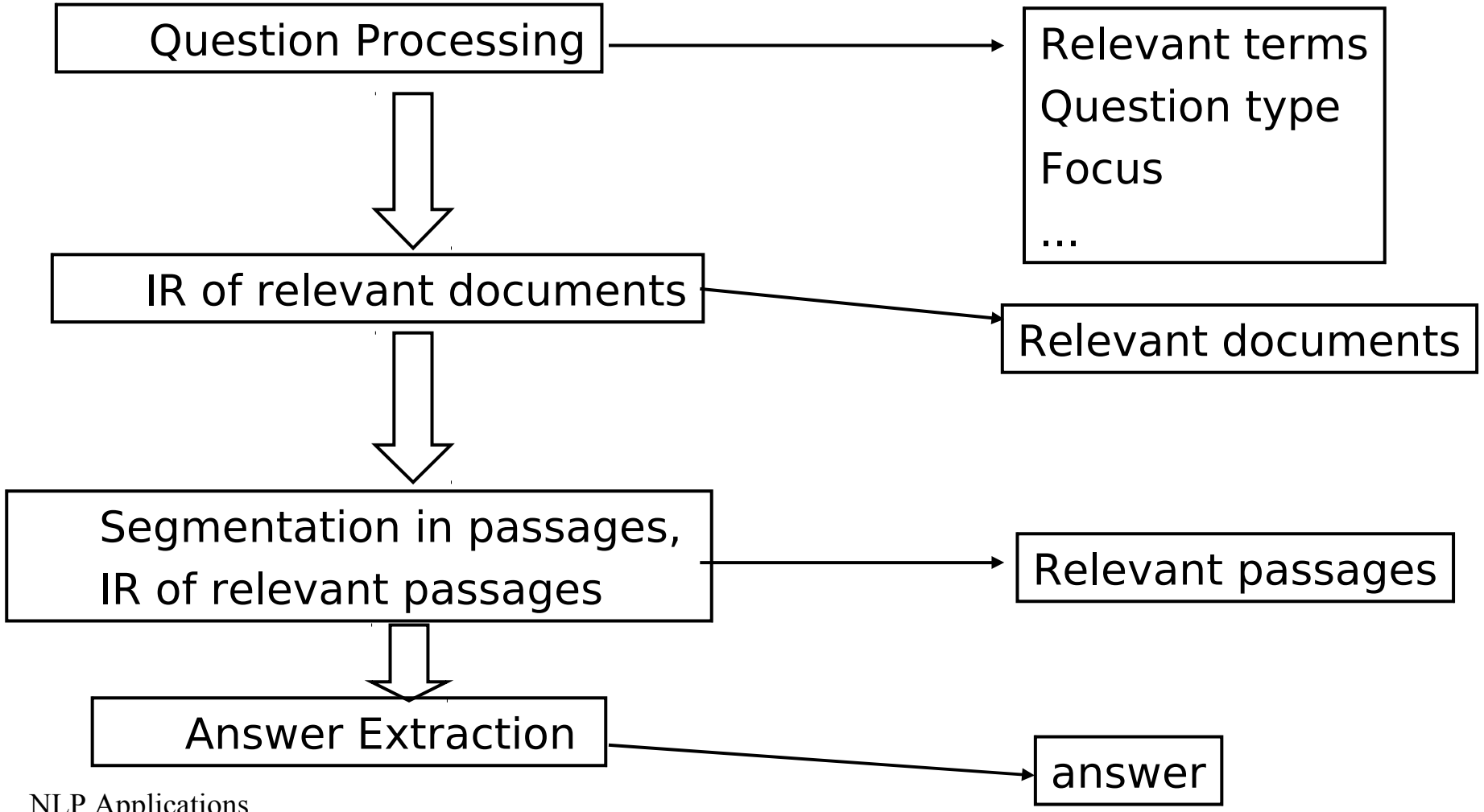
IR of relevant documents

Segmentation in passages,
IR of relevant passages

Answer Extraction

Question Answering 9

Frequently performed sequentially



Automatic Summarization 1

- A summary is a reductive transformation of a source text into a summary text by extraction or generation
  - Sparck-Jones, 2001

Automatic Summarization 2

- Look for the relevant parts of a document and produce a summary of them
- Summarization vs IE
  - IE
    - What has to be extracted is defined a priori
      - "I am interested on this, look for it"
  - Summarization
    - An a priori definition of what is relevant is not always defined

Automatic Summarization 3

Some readings

- Tutorial
  - E.Hovy, D. Marcu (1998)
- Horacio Rodriguez (2001) Summarization

  http://www.lsi.upc.es/%7Ehoracio/varios/alicante2007.zip

Automatic Summarization 4

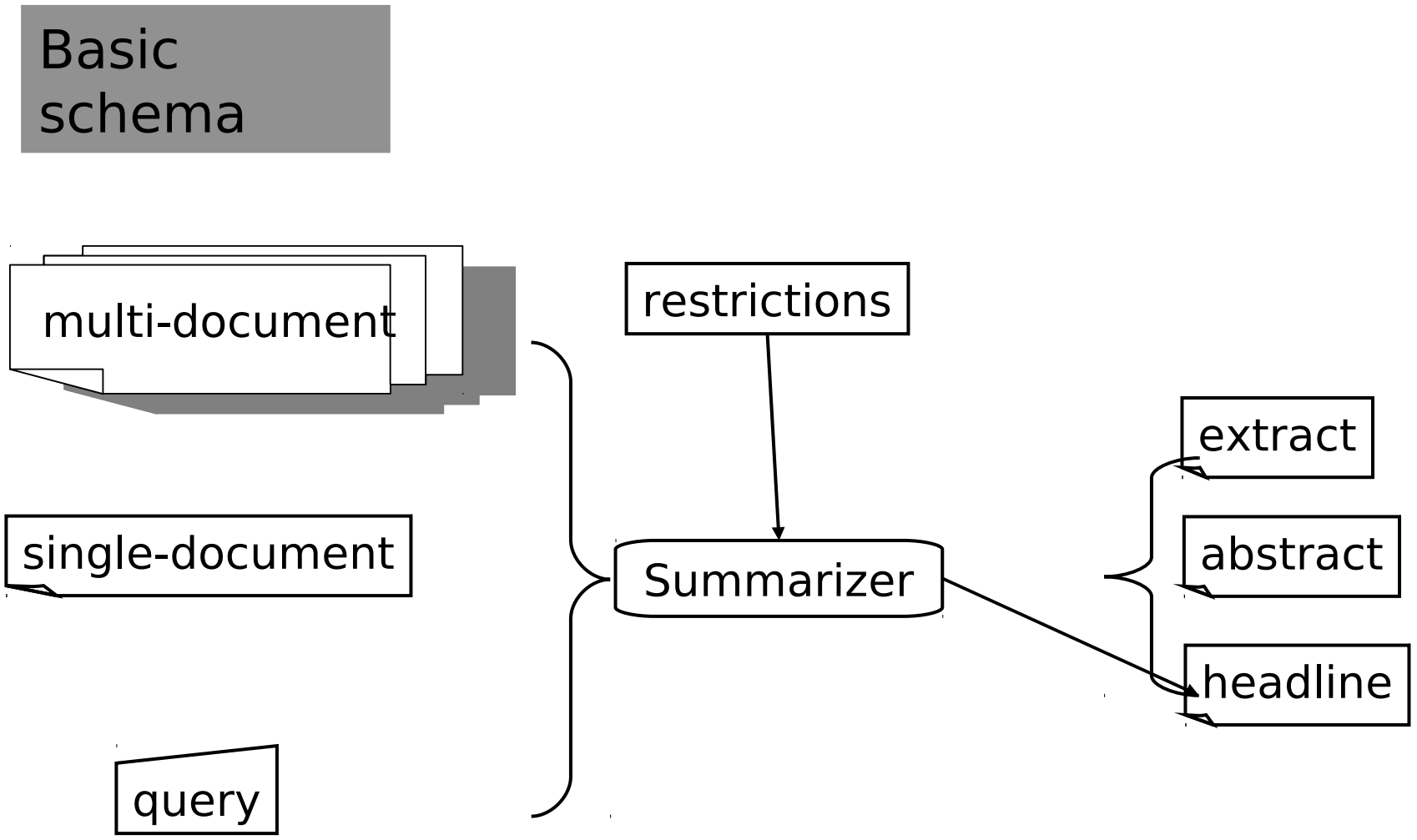
Types of summarization

- Type
  - Indicative vs informative
  - Extract vs Abstract
  - Generic vs query based
  - Background vs just-the-news
  - Single-document vs multi-document
  - general vs domain restricted
  - textual vs multimedia
- Input
  - domain, genre, form, size

Automatic Summarization 5

- Related disciplines
  - IE, IR, Q&A, Topic identification (TI), Document Classification (DC), Event (topic) detection and tracking (TDT)
- Evaluation
- Applications
  - Biographies
  - Medical reports
  - E-mails
  - Web pages
    - Word spotters
  - News
  - Headlines extraction
    - Automatic subtitle generation
  - IR enhancements
  - Meeting interventions

Automatic Summarization 5



Automatic Summarization 6

Techniques

- Lexical chains
  - [Barzilay, 1997], [Fuentes, 2008]
- Coreference chains
  - [Baldwin, Morton, 1998]
  - [Bagga, Baldwin, 1998]
- Alignment techniques
  - [Banko et al, 1999]
- Compression, reduction or simplification of sentences (cut & paste)
  - [Jing, 2000]
  - [Jing, McKeown, 1999]

Automatic Summarization 7

- Statistical models
  - modelos estadísticos de la lengua
    - [Berger, 2001], [Berger, Mittal, 2000]
  - modelos bayesianos
    - [Kupiec et al, 1995], [Schlesinger et al, 2001]
  - cadenas ocultas de Markov
  - Regresión logística
    - [Conroy et al, 2001]
- Machine Learning
  - Decision trees
  - ILP
    - [Knight, Marcu, 2000], [Tzoukerman et al, 2001]
- Similarity (and distance) measures
  - MMR
    - [Carbonell, Goldstein, 1998]

Automatic Summarization 8

- IE
  - [Kan, McKeown, 1999]
- Topic Detection
  - [Hovy, Lin, 1999]
  - [Hovy, 2000]
- Topic Signatures
  - [Lin, Hovy, 2001]
- Document's rethoric structure
  - [Marcu, 1997]
- Combination
  - [Goldstein et al, 1999], [Kraaij et al, 2001],
  - [Muresan et al, 2000], [White et al, 2001].

Multidocument Summarization (MDS) 1

Objectives

- Summary of a collection content
- Briefing
  - concise summary of the factual matter of a set of news articles on the same or related events (SUMMONS, Radev,1999)
- Actualization of already known information

MDS 2

SDS vs MDS

More challenging

- Compression
- Redundancy
- Temporal terms
- Correference

MDS 3

Requirements

- Clustering of documents and passages
- Recall
- Anti-redundancy
- Summary cohesion
- quality
  - readable
  - relevant
  - context
- Inconsistency of sources
- Actualization

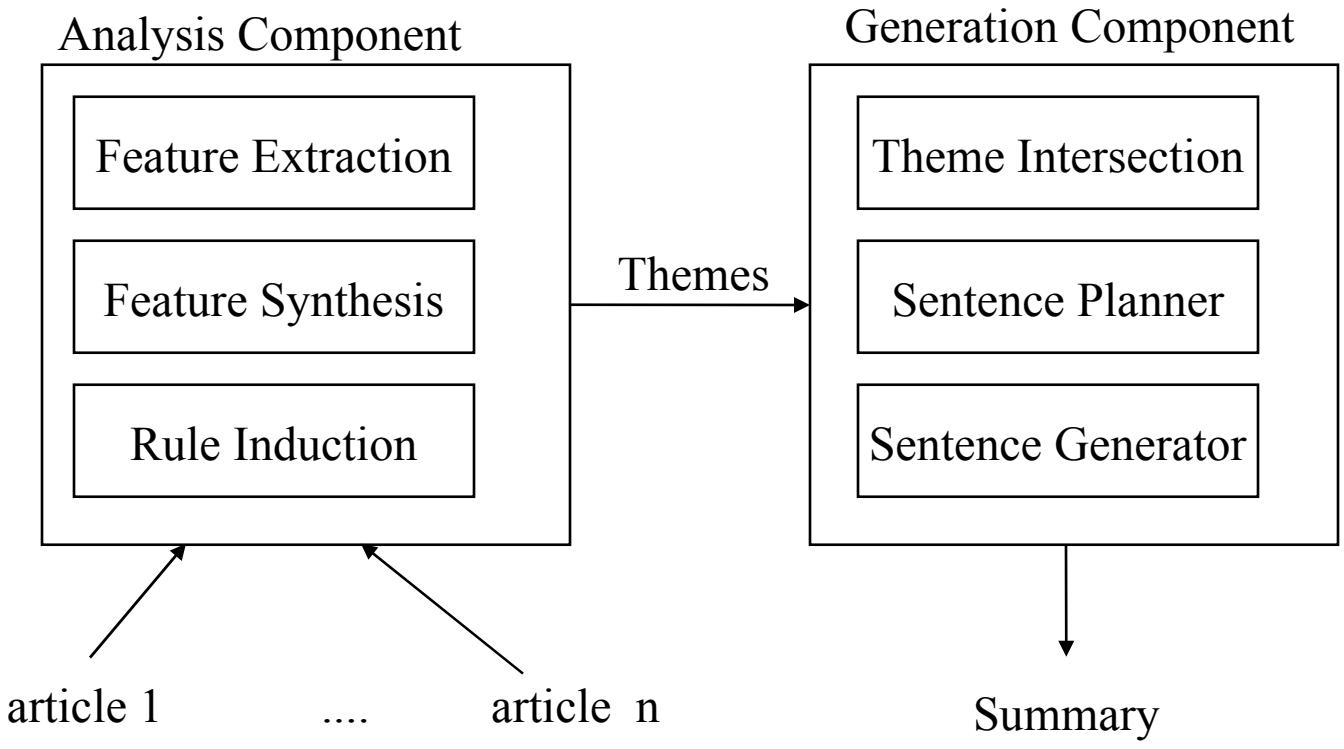
MDS 4

Approaches

- From the common sections of all the documents of the collection
- Common sections + unique sections
- Centroids
- Centroids + outliers
- Last document + outliers
- Common sections + unique sections + time weighting factor

MDS 6

Mc.Keown et al, 1999
MULTIGEN

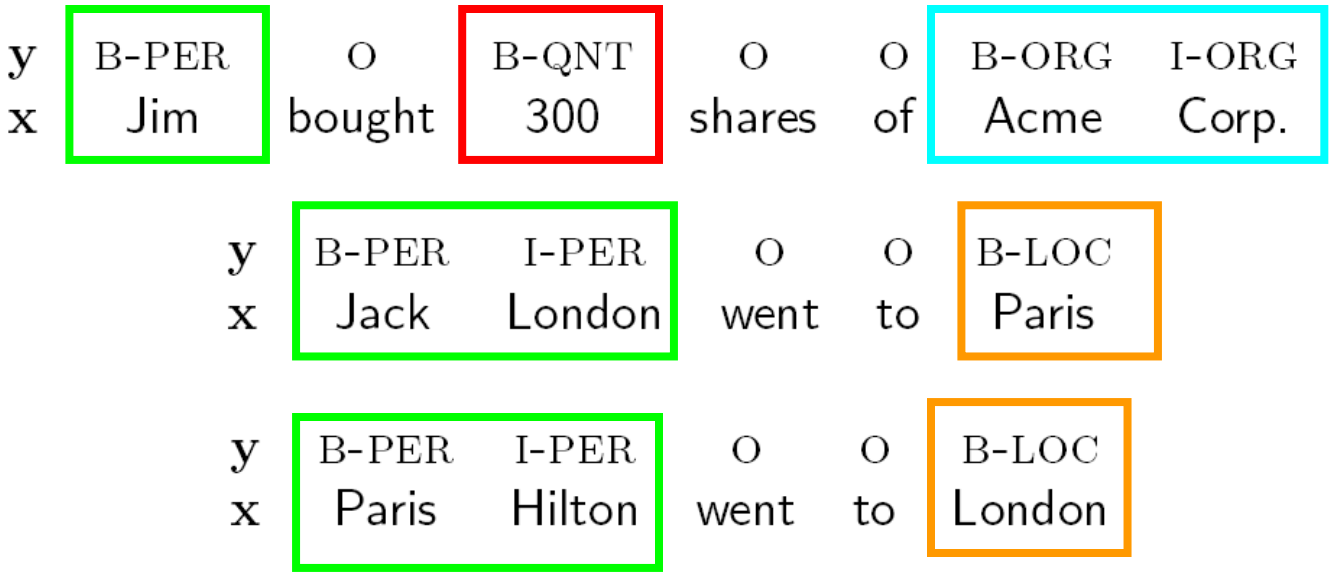


Information Extraction 1

- Extracting useful information from free text
- MUC, ACE, TAC challenges
- Named Entity Recognition (NER)
- Named Entity Classification (NEC)
- Both tasks together (NERC)
- Slot Filling
- Relation Extraction

Information Extraction 2

NERC



Information Extraction 3

Slot Filling

- Set of relevant slots
- ML
  - Supervised Learning
  - Unsupervised Learning
    - Distant learning
  - Semisupervised Learning
    - Active Learning
- Rule-based systems

Information Extraction 4

Relation Extraction

- Labeled vs unlabeled relations
- Binary vs n-ary relations
- Properties:
  - Simetric, transitive, reflexive
- Constraints over source and target
  - NE, PER, ORG, LOC,

Information Extraction 5

Relation Extraction

- ML
  - Supervised Learning
  - Unsupervised Learning
  - Semisupervised Learning

Document Classification 1

- Classification vs. Clustering
- Assign each document to one or more class(es) belonging to a predefined tagset
- Examples:
  - Spam filtering
  - Language identification
  - Level of relevance, urgency, ...
  - Thematic domain

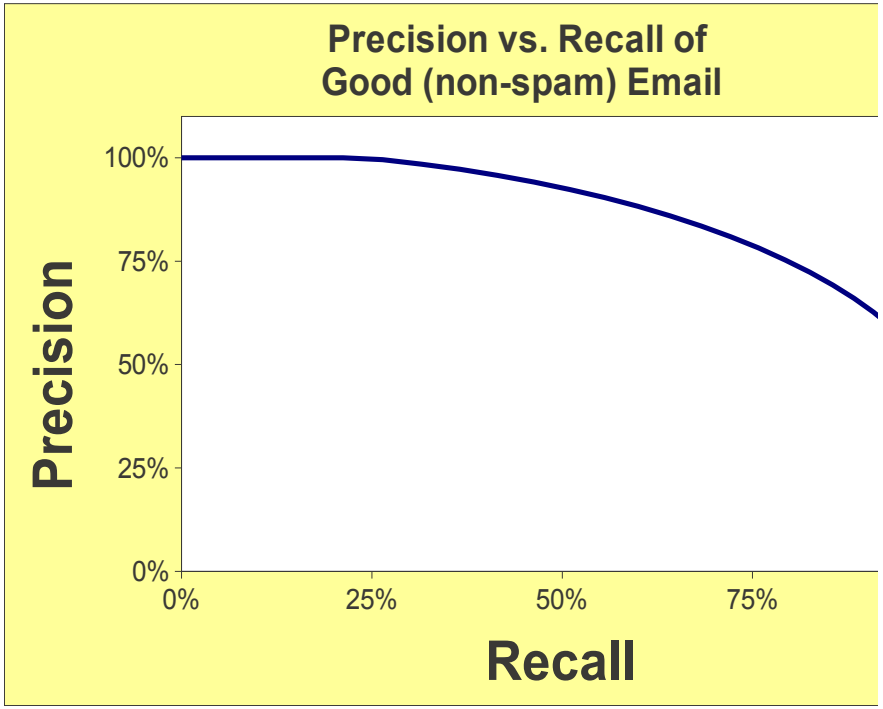
Document Classification 2

- Extensions:
  - Multiclass
    - A document can be assigned to more than one class
  - Rank
    - A document is a assigned to different classes acording a probabilistic distribution.
- Features
  - Textual content
  - Metadata

Document Classification 3

- Approaches
  - Vectorial
    - Categorize each class with a reference document (Topic Signature, Lexical Profile, ...)
    - Represent the document to classify with VSM (Vector Space Model)
    - Using a similariry measure for comparing the vector associated to the document with the reference document of each of the classes.
    - Choose the best or rank them
      - e.g. k-means
  - ML
    - Naive Bayes, decision lists, decision trees, maximum entropy, SVM, boosting, ...

Document Classification 4



- Precision = $\frac{\text{good messages kept}}{\text{all messages kept}}$
- Recall = $\frac{\text{good messages kept}}{\text{all good messages}}$