



Information Retrieval: between Natural Language, Texts and Meaning

Roberto Basili

(Università di Roma, Tor Vergata, basili@info.uniroma2.it)

Some slider borrowed from the tutorial «[Natural Language Understanding: Foundations and State-of-the-Art](#)», by [Percy Liang](#) (Stanford University).

Web Mining & Retrieval, a.a. 2021-22

Overview



- **Documents in Information Retrieval**
 - Information, Representation, (re)current challenges, success(and unsuccess)ful stories
- **Information and Content**
 - **Natural Language Processing: introduction to the linguistic background**
 - Natural Language and Content
 - NL Syntax
 - NL Semantics
 - **Document Representation and IR models**
- **Summary**

Semantics, Open Data and Natural Language

- Web contents, characterized by rich multimedia information, are mostly opaque from a semantic standpoint

Today is 2011年11月13日 星期日 顯示器最佳分辨率1024x768

今天天氣 加入最愛 設為首頁 大公網新版

2011 中国証券金紫荊獎 Golden Bauhinia Awards

看大公報 國際短信 讀者推薦 廣告業務

首頁 國內 國際 港澳 兩岸 評論 財經 體育 教育 科技 醫學 娛樂 文化 副刊 軍事 生活 旅遊 圖片 博客

關鍵詞: 欄目: 全部 最近三個月 三個月之前 檢索

▶ 手機新聞 ▶ 手機博客 ▶ 漢語學習 ▶ 新聞點點排行 ▶ 招聘啓事

滾動新聞:

胡總語特首:防範經濟金融風險
胡錦濤在夏威夷會見出席APEC峰會的曾蔭權。他祝賀香港區議會選舉成功,並充分肯定曾蔭權及港府工作,要求做好經濟金融風險防範

胡連會登場 共同宣示九二共識
胡錦濤第四度在APEC峰會期間會見連戰。他強調,認同「九二共識」是兩岸開展對話協商的必要前提,也是兩岸關係和平發展的重要基礎

西藏黨代會高調反「藏獨」 德國作家:外媒錯誤報道西藏
傳媒入日本福島核電站探訪 英國大裁軍 傷兵難倖免
滇礦難已30死 13人生還 礦工講述內幕 事故並不意外
范徐麗泰認民望跌最不熱 選舉再獲60提名表 累積逾千人
聖保羅中學本月底截止招 選舉再獲60提名表 累積逾千人
民調逆轉 藍高層:國親吵鬧地 秋門訴求多 向藍綠表不滿
世界新七奇觀 亞洲景佔四席 新奇觀選舉惹爭議
中國實體書店苦苦掙扎求 加入TPP 台密集會談探路
香港人家/蔡仕榮 人生導師 活出自我 香港人家/教導子女...
債務危機紓 港ADR幾全線造 歐元反彈 兌美元逼近1.38
入世十年/充分對接 華強北最 入世十年/挑戰「二次」...
抽股除「雜項」 工人險生 南亞漢命案 警拘日籍妻

即時新聞 更多>>

- 組國/河南全國太極拳錦標賽況
- 奧巴馬重申美不支持「台灣獨立」
- 巴基斯坦西北部兩起襲擊 16人死
- 圖文/胡錦濤會見美國總統奧巴馬 (圖)
- 兩岸30對愛侶在廈門集體證婚
- 中日韓衛生部長會議在青島舉行
- 面向中國遊客中英雜誌紐約創刊
- 「CEO聖經」成內地官員考試內容
- 斯滕恩:經紀人是勞資談判的障礙
- 香港冀成爲人幣國際化關鍵角色
- 日學者提出地核物質形態新假說
- 中國影視機構向國際大師「取經」

焦點關注 更多>>

- 區議會選舉
- 香港特首選舉
- 2011APEC 港黑金事件
- 2011施政報告
- 神八天宮對接
- 第七次陳江會
- 李克強訪港
- 9.1衝擊事件
- 中國航母試航
- 辛亥革命百年

http://www.takungpao.com.hk/news/11/11/13/2011_apec_xgbd-1423309.html

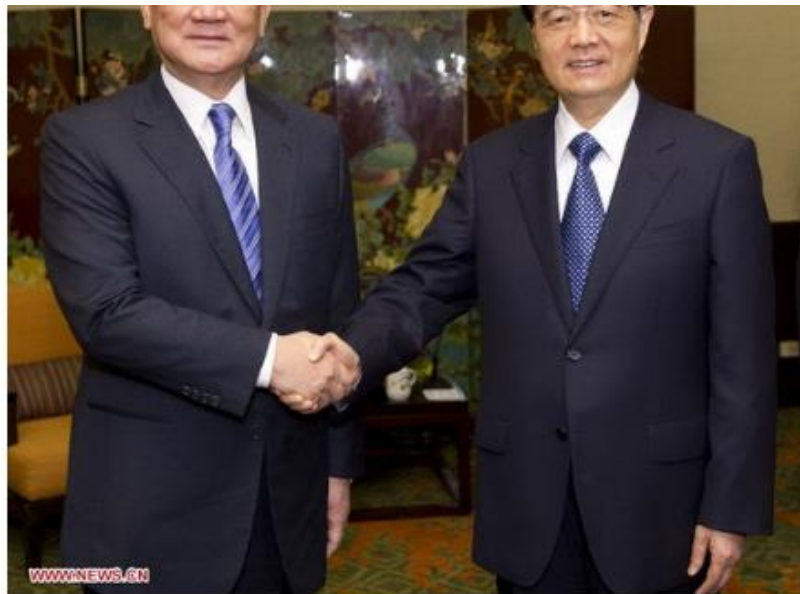
Information, Web and Natural Languages



Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011.

(Xinhua/Huang Jingwen)

HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central



Chinese President Hu Jintao (R) shakes hands with Honorary Chairman of the Chinese Kuomintang (KMT) Lien Chan, in Honolulu, Hawaii, the U.S., Nov. 11, 2011.

(Xinhua/Huang Jingwen)

HONOLULU, United States, Nov. 11 (Xinhua) -- Hu Jintao, general secretary of the Central



Who is Hu Jintao?

- 1 China in APEC: a mutually beneficial en...
- 2 Night life in Shanghai
- 3 China's 2011 foreign trade to grow 20 p...
- 4 Beijing house prices stumble 5.1 pct as...
- 5 Lama students start school in Tibet Col...
- 6 Police in central China crack phoney ca...
- 7 China-ASEAN cooperation sees notable pr...
- 8 Miao ethnic group celebrates Miao's New ...



Hu Jintao



Ricerca

Circa 725.000 risultati (0,09 secondi)

Tutto

Immagini

Mappe

Video

Notizie

Shopping

Più conte

Tutti i ri

Per argomento

Qualsiasi dimensione

Grandi

Medie

Icone

Maggiori di...

Dimensioni esatte...

Qualsiasi colore

A colori

Bianco e nero



Qualsiasi tipo

Volti

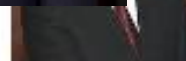
Foto

Clip art

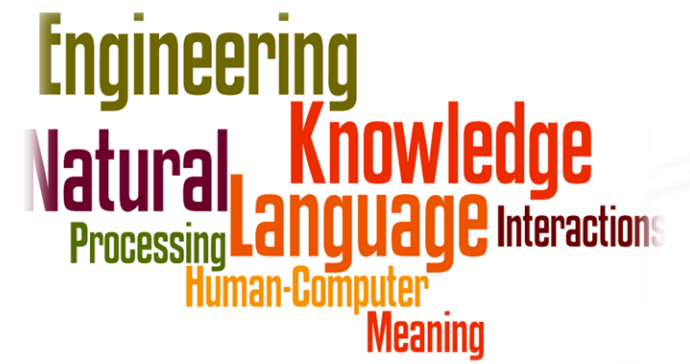
Disegni

Visual standard

Mostra dimensioni



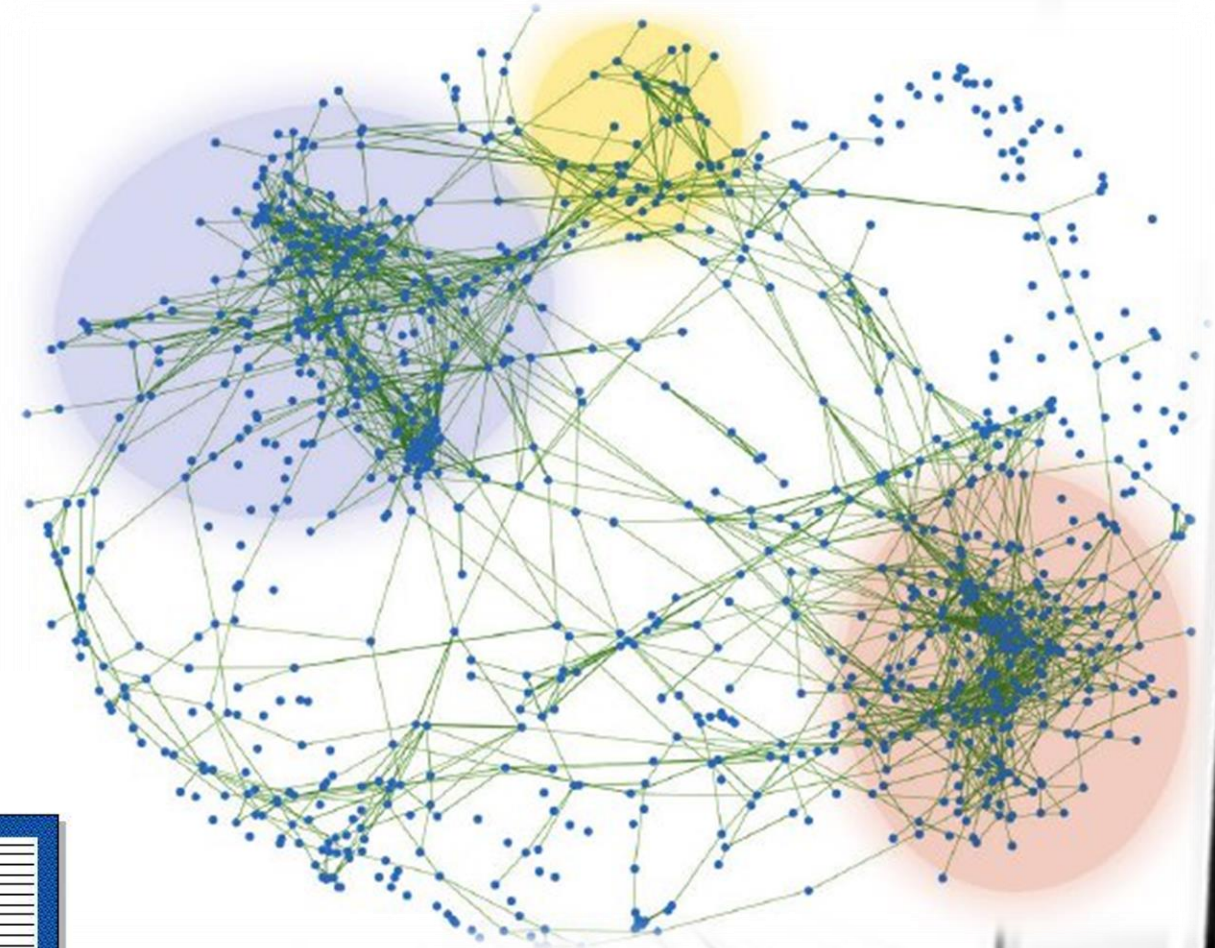
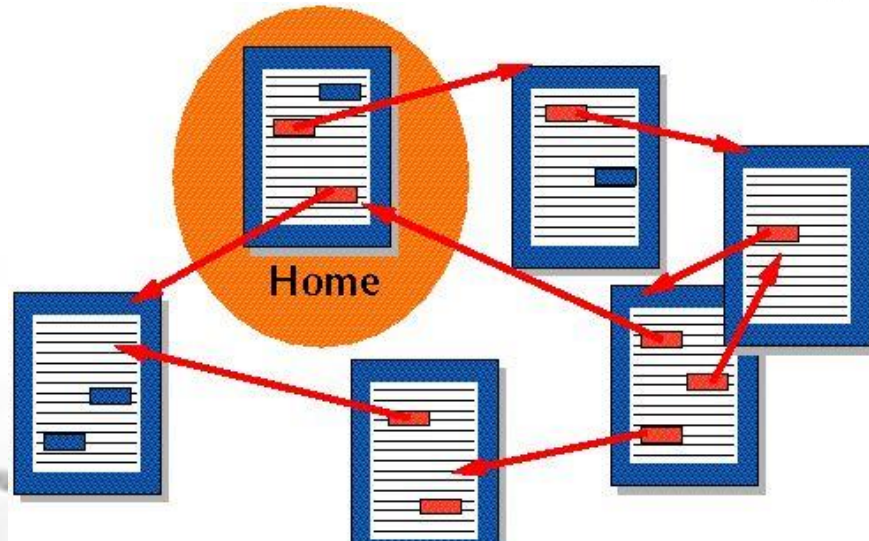
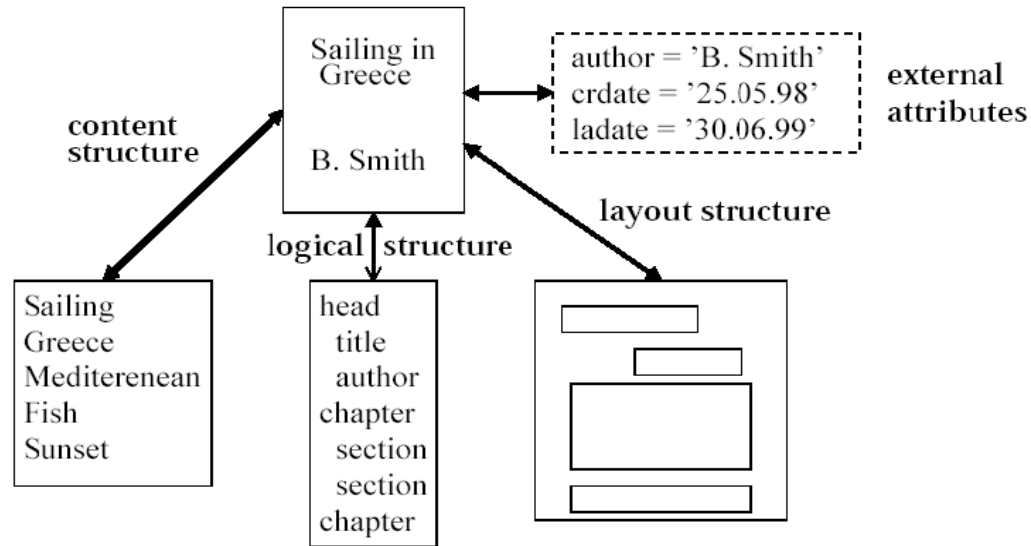
Content Semantics and Natural Language

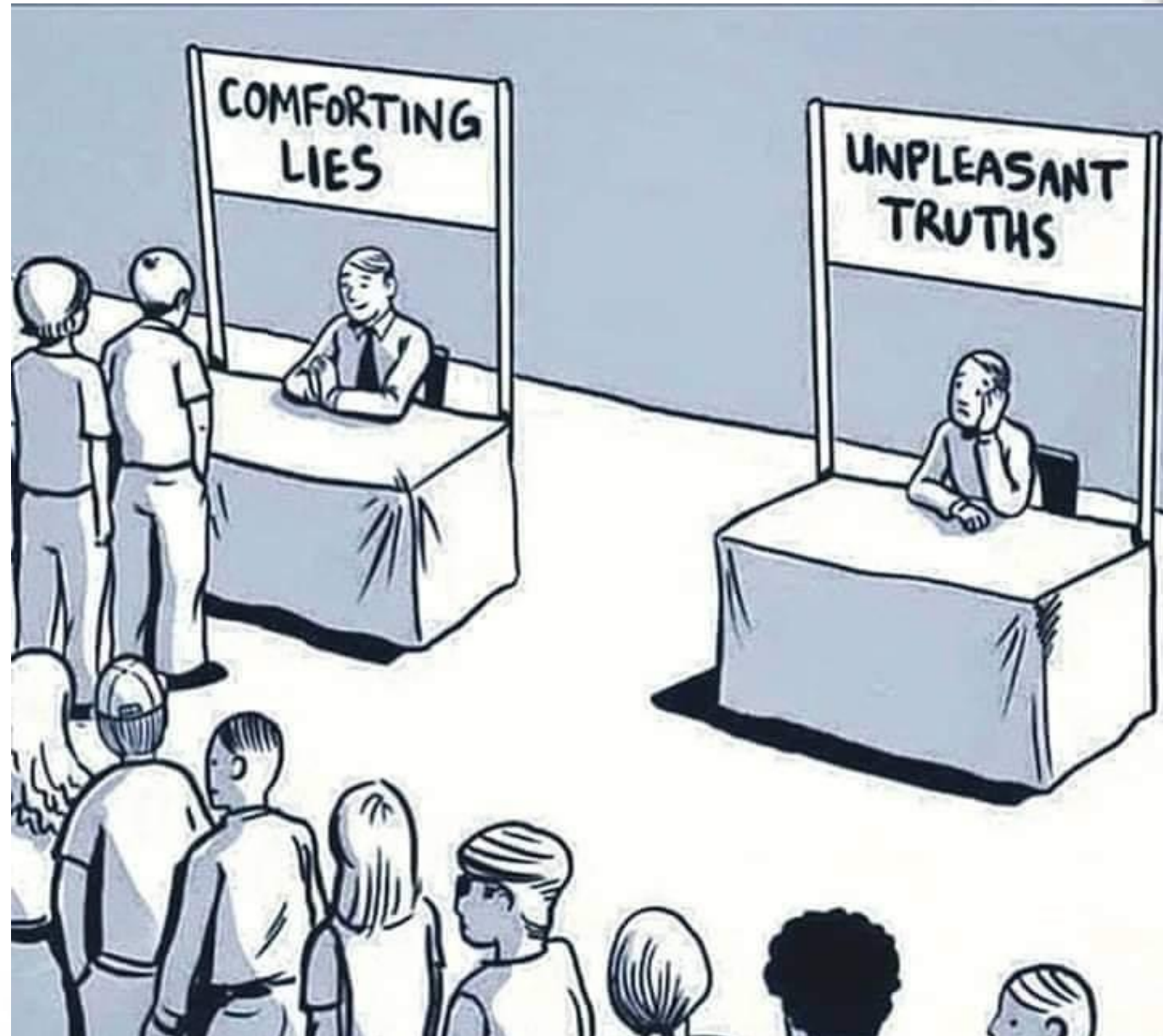


- Human languages are the main carrier of the information involved in processes such as retrieval, publication and exchange of knowledge as it is associated to the open Web contents
- Words and NL syntactic structures express concepts, activities, events, abstractions and conceptual relations we usually share through data
- “Language is parasitic to knowledge representation languages but the viceversa is not true” (Wilks, 2001)
- From Learning to Read to Knowledge Distillation as a(n integrated pool of) Semantic interpretation Task(s)

Texts, Information & Document Structures

What is a document?

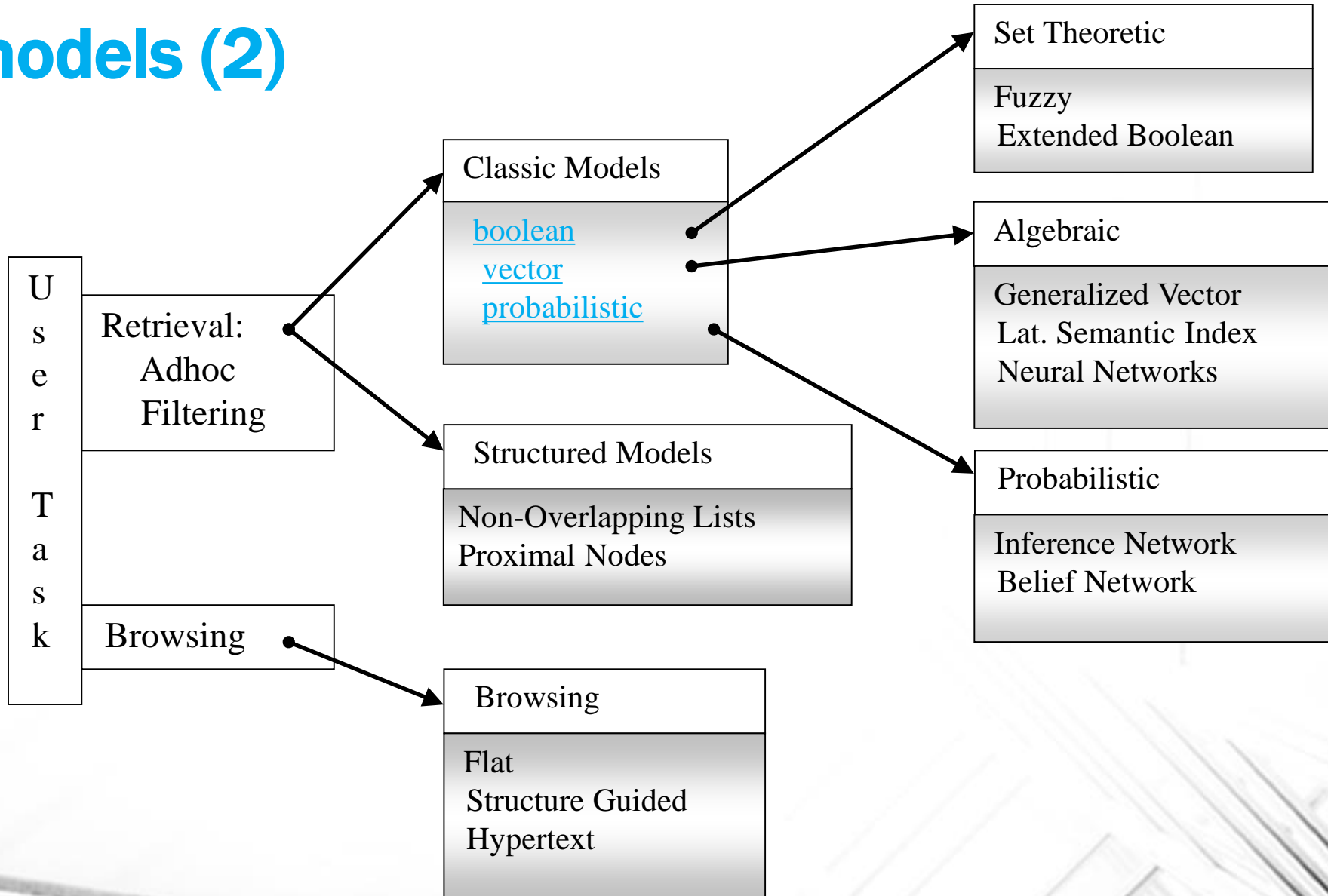




Information Retrieval Models

- An IR model must specify (at least) :
 - A representation of the document
 - A representation of individual queries
 - The retrieval function
- The model determines a specific notion of relevance.
- Relevance can be discrete (e.g. binary) or continuous (i.e. rank or relevance order).
- It is a perfect example of learnable function through induction from examples (see Google)

IR models (2)



Model Families for IR

- Boolean Models (set theoretic)
 - Standard boolean
 - Extended Boolean
- Vector Models (algebraic)
 - Generalized Vector Space
 - Latent Semantic Indexing
 - Neural models
- Probabilistic Models

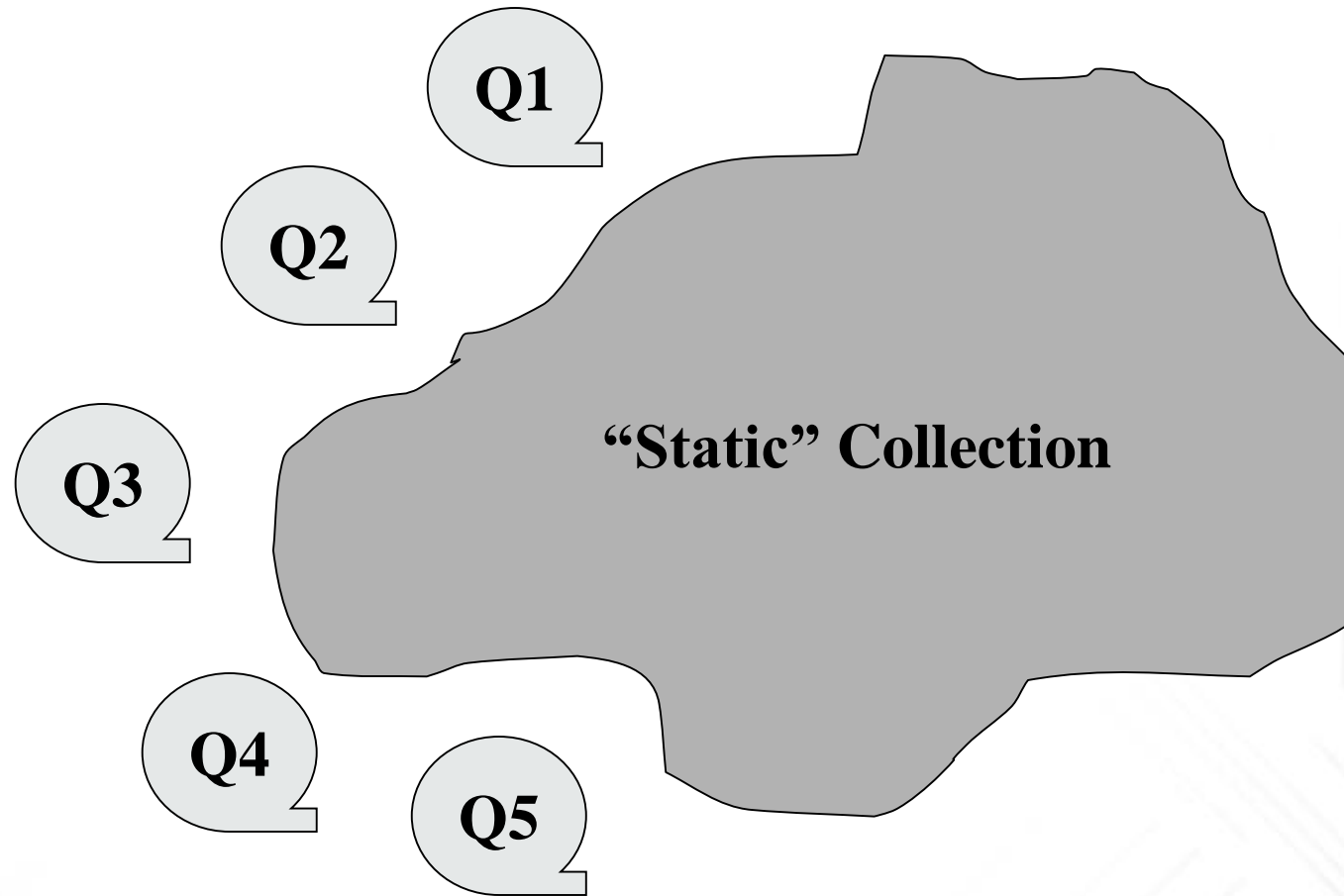
Other classification Dimensions

- Document Logical Model
 - Type of Indexes
 - Structures vs. Content
 - Metadata vs. Content
 - Full text as a model of the content
 - Full text viz. Document (Hypertextual) Structure
 - Declarative vs. operational semantics
- The role of user
 - Subjective vs. Objective forms of relevance
 - Operational environment
 - Search vs. Browsing

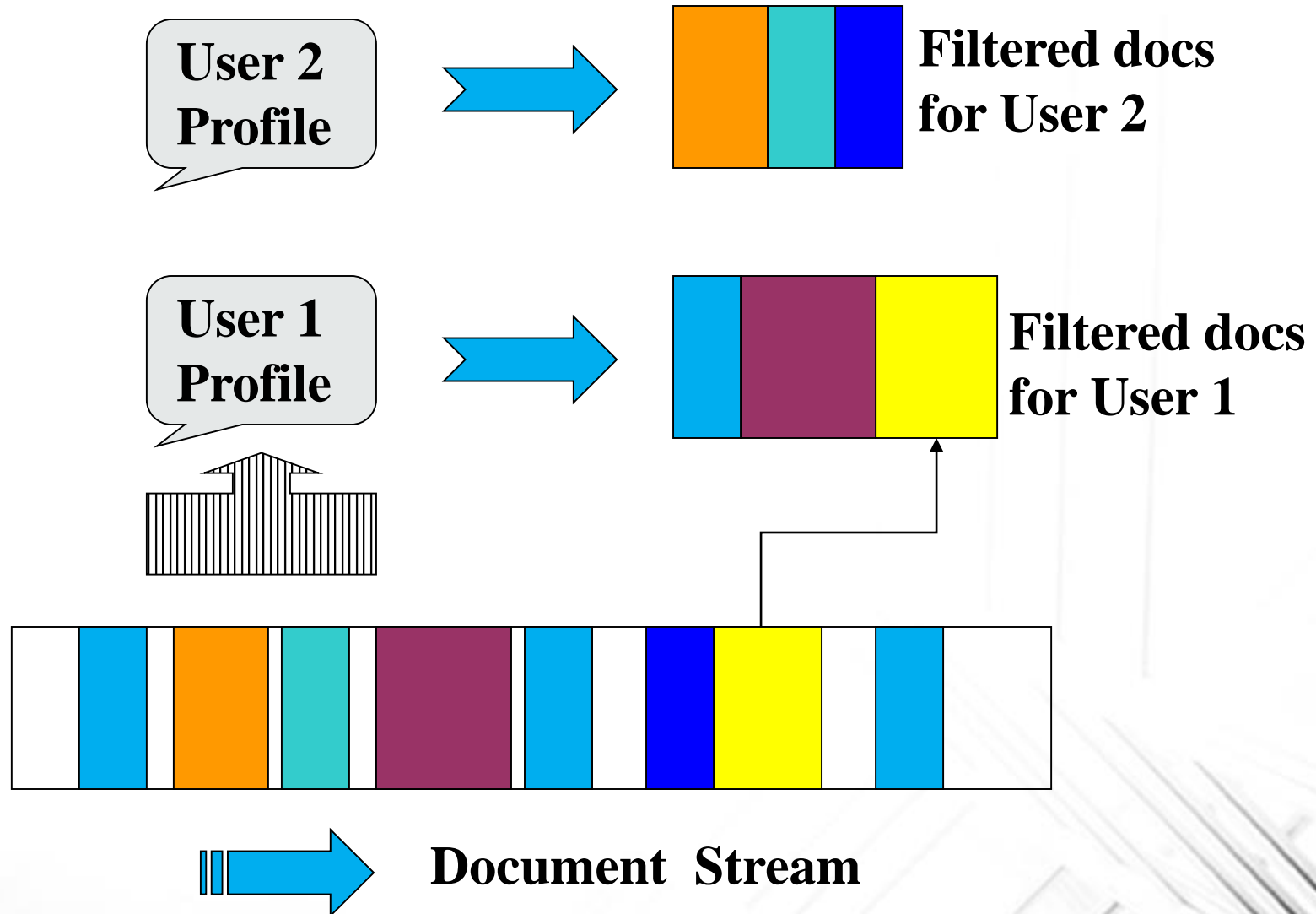
Retrieval Tasks

- **Ad hoc retrieval**
 - DEF. Relatively stable document collection vs. highly variable queries.
- **Information Filtering:**
 - DEF. Fixed Queries and continuous streams of documents
 - Type of Filtering
 - User Filter: static model of the subjective preferences
 - Category based filtering: static model of categories as domain preferences
 - Target Function: binary decision, in general
- **Information Routing:**
 - DEF. When filters define dynamic e non binary models of preference.

Ad Hoc Retrieval



Filtering



Learning and IR

- The task in IR and the need of modeling either documents and queries are strongly related to Machine Learning
- First, **no analytical function is available** for every domain, document collection, user and query is available
- Second, **unstructured data** (as much frequently occurring in Web applications) are hard to be modeled without resorting to a reference notion of **content**
- CHALLENGE: How to deal in an **efficient** manner with the tasks of **representing, querying, matching, filtering** and **sorting** the complex contents characterizing the arbitrarily **distributed and unstructured Web data**?
 - In the case of textual document: **how can we learn to formalize the vague notion of content for a document?**

Semantics, Natural Language & Learning: tasks

- In order to make contents explicit in an IR process they must be recognized in the contexts of their use
- All these process (also called **Learning to Read** or **Knowledge Distillation**) proceed as a (integrated pool of) **Semantic interpretation Task(s)**
 - **Information Extraction** (from text to machine readable concepts)
 - Entity Recognition and Classification
 - Relation Extraction
 - Semantic Role Labeling (Shallow Semantic Parsing)
 - **Estimation of Text Similarity** (from text to quantitative semantic measures)
 - Structured Text Similarity/Textual Entailment Recognition
 - Sense disambiguation
 - **Semantic Search, Question Classification** and **Answer Ranking**
 - **Knowledge Acquisition**, e.g. ontology learning
 - **Social Network Analysis: Opinion Mining, Recommending**

Two major objectives

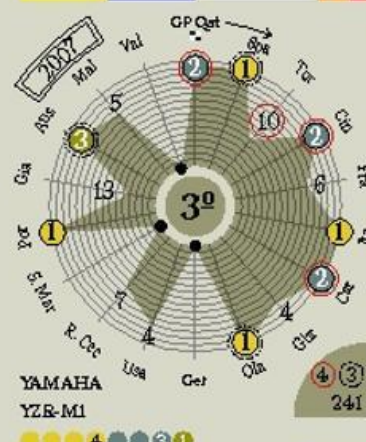
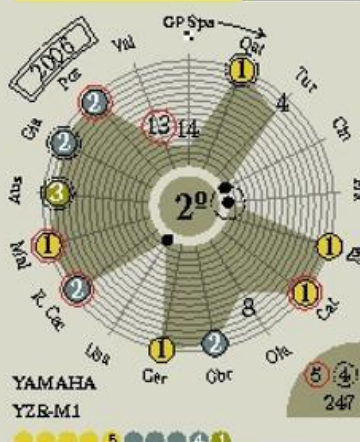
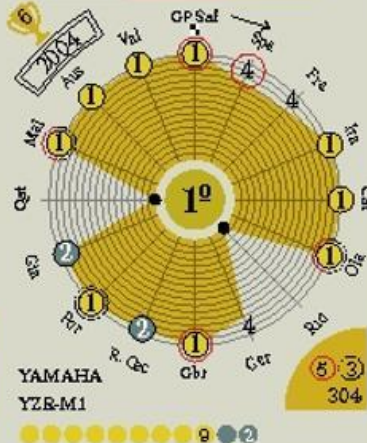
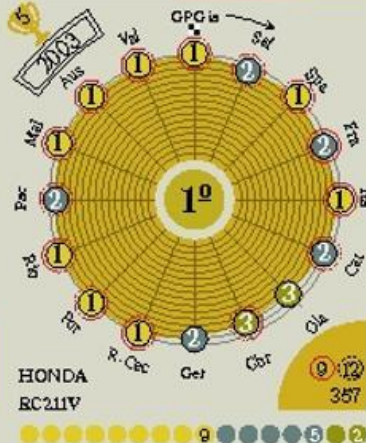
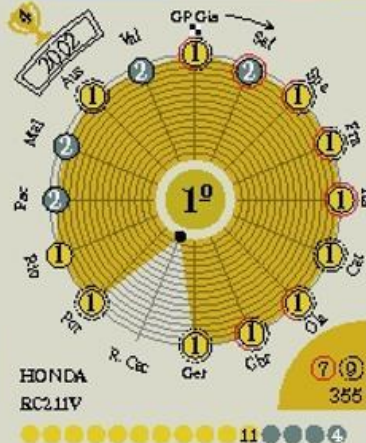
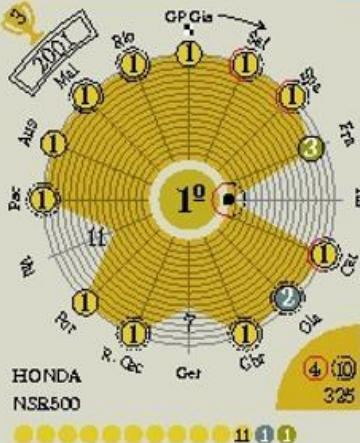
- Discuss the **nature of content** in unstructured data within a semantic perspective over natural language
 - What constitute a useful notion of content within unstructured data collections (that are largely made of linguistic information, e.g. Web pages or infographics)
 - What is **natural language semantics** and how can we model it formally?
 - What is the **meaning of a linguistic expression**?
- What is **the notion of document** that we can use within IR processes
 - Nature and role of document information
 - Relationship between a declarative view on content wrt an operational view of content
 - How this has to do with IR and ML?

Colari


LEGENDA

- ① ② ③ = posizione sul podio
- 4, 5, 6... = altre posizioni
- NP = non partito
- GA = gara annullata
- = ritirato
- = pole position
- (con linea) = giro veloce
- (con linea e punto) = pole + giro veloce
- 🏆 = numero titoli mondiale
- X^o = posizione finale (in giallo)
- X^o (con X) = totale pole giuri veloci
- X^o (con X) = totale punti stagione

Casa motociclistica → APRILIA
 Modello moto → RS 125 R
 Numero podi → 1 1
 Principali colori della tuta → [Colori: Giallo, Nero, Blu]



Overview

- Documents in Information Retrieval
 - Information, Representation, (re)current challenges, success(and unsuccess)ful stories
-  • Information and Content
 - Natural Language Processing: introduction to the linguistic background
 - Natural Language and Content
 - NL Syntax
 - NL Semantics
 - Document Representation and IR models
- Summary

Content in unstructured data

- Natural Language
 - Structure
 - Semantics
 - Types of semantics
 - Relationship with Machine Learning
- Examples:
 - NLU: natural language as a logic language
 - Providing more structure: Frame semantics
 - Logic, Frames and Scripts
 - The relationships between syntax and semantics
 - Semantic role labeling

Natural Language & Ambiguity



NOTICE	
	Dogs must be carried on escalator



Ambiguità

- *"Dogs must be carried on this escalator"*

can be interpreted in a number of ways:

- *All dogs should have a chance to go on this wonderful escalator ride*
- *This escalator is for dog-holders only*
- *You can't carry your pet on the other escalators*
- *When riding with a pet, carry it*

The NLP chain

Levels of linguistic analyses

Pragmatics: what does it do?

Semantics: what does it mean?

Syntax: what is grammatical?

natural language utterance

Analogy with artificial languages

- **Syntax**: formal correctness (e.g. no compiler errors)
- **Semantics**: complete interpretation (no implementation bugs)
- **Pragmatics**: the algorithm is right!

- Different **syntax**, same **semantics**: $2+3$ vs. $3+2$
- Same **syntax** but different **semantics**: ... $*a \Rightarrow$ the addressing vs. the multiplication
eco's book \Rightarrow the owner vs. the writer
- Good **semantics**, bad **pragmatics**: in *route planning*
 - correct brute force search algorithm vs. heuristics A* search

Ambiguity and Linguistic Levels

• Semantics



• Syntax



• Morphology



• Phonology



can/can

eat cake with fork

earth observation satellite
Eco's book



del (pane)
/del (libro)

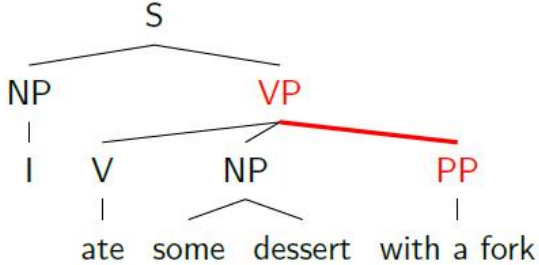
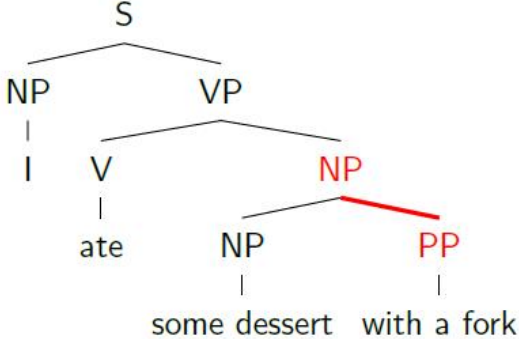
compro la borsa
in pelle

il timore dei manager



Grammars & Ambiguity

I ate some dessert with a fork.



Summary

- IR models necessary in Web mining depend on the ways unstructured data can be made available for filtering, classification, retrieval and ranking tasks
- A semantic model for the content of unstructured data is strongly dependent on the linguistic nature of these latter
 - Facts, Entities, Relations, Thematic areas, Subjective information are always rooted in a form of rather free linguistic description
- Studies in Linguistics have provided the basic notion for dealing with the meaning of Natural Language expressions
 - Levels
 - Basic paradigms: lexical description, grammars, logic as a meaning representation language

References

- AI & Robotics. «Robot Futures», Ilah Reza Nourbakhsh, MIT Press, 2013
- NLP & ML:
 - «Statistical Methods for Speech Recognition», F. Jelinek, MIT Press, 1998
 - «Speech and Language Processing», D. Jurafsky and J. H. Martin, Prentice-Hall, 2009.
 - «Foundations of Statistical Natural Language Processing, Manning & Schütze, MIT Press 2001.
- Sitografia:
 - SAG, Univ. Roma Tor Vergata: <http://sag.art.uniroma2.it/>
 - Reveal s.r.l.: <http://www.revealsrl.it/>

