

Aprenentatge Automàtic pel Processament del Llenguatge Natural

**Classificació
d'Entitats amb Nom**

Jordi Duran i Cals

Desembre del 2006



Aprenentatge Automàtic pel Processament del Llenguatge Natural

- Índex
 - Introducció
 - Aprenentatge Automàtic
 - Aplicació de l'Aprenentatge Automàtic en el Processament del Llenguatge Natural (Classificació d'Entitats amb Nom)

El per què d'aprendre

- Situacions complexes:
 - Capacitats humanes que no som capaços d'explicar (speech recognition)
 - Experiències humanes que no hem tingut (exploració d'altres planetes) és difícil i es necessita temps
- Tenim dades en grans quantitats i barates, per altra banda el coneixement és car i escàs
 - Crear sistemes manualment és difícil i es necessita temps

Aprendre a aprendre

La didàctica és la branca de la pedagogia que s'ocupa d'estudiar com ha de funcionar l'aprenentatge en els éssers humans de manera òptima, és a dir, quina és la millor manera d'ensenyar uns determinats continguts o habilitats.

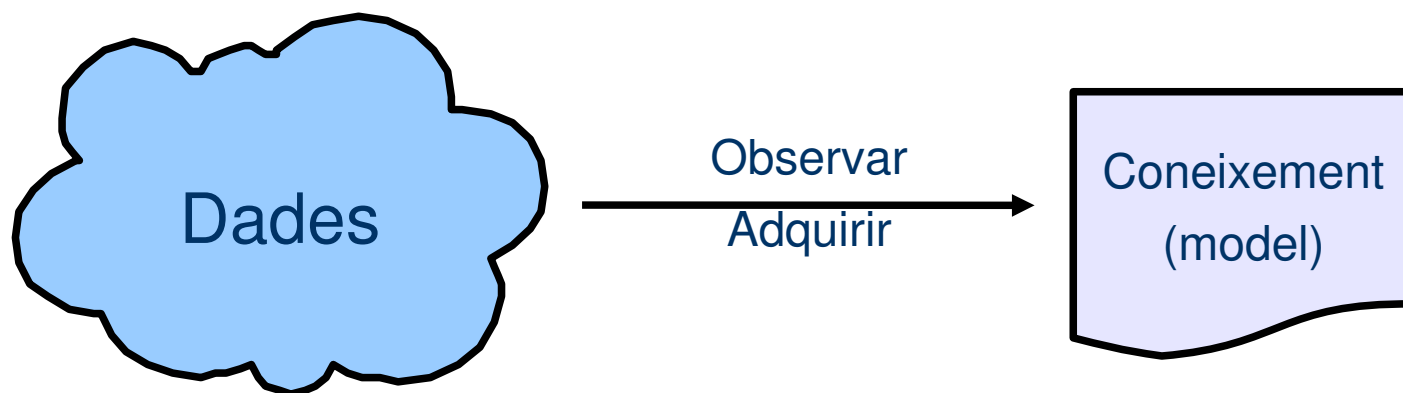
- Estratègies d'aprenentatge
 - Anàlisi i discussió de casos
 - Imitació de models →
 - Procediments d'interrogació

L'estratègia d'aprenentatge basada en la imitació, és sens dubte un dels procediments més naturals d'enfrontar-se a les coses... Els nens petits, i no tant petits..., fan servir els models més propers com a pauta d'acció - reflexió...

Com es veurà es vol simular el comportament humà

Aprentatge

- Adquirir coneixement des d'exemples concrets



- El coneixement adquirit (model) és una *bona aproximació* de les dades observades?

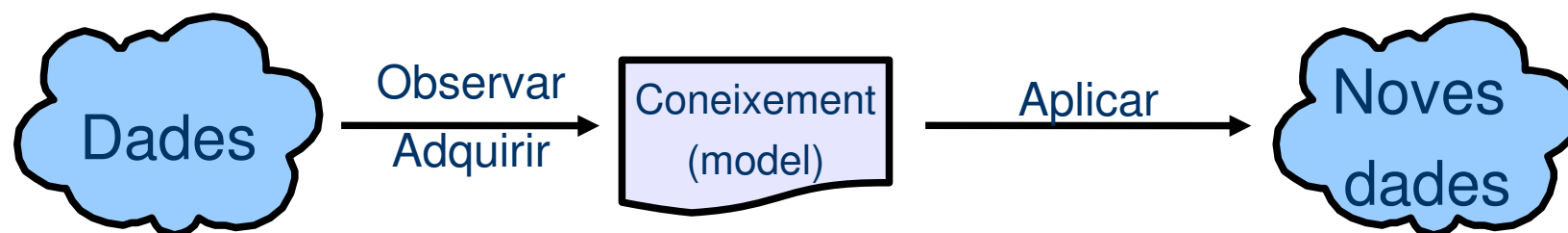
Es pot avaluar

Aprentatge Automàtic (ML)

- ML s'engloba dins de la Intel·ligència Artificial (IA)
- S'aplica en molts altres camps d'investigació
- Fer que els ordinadors adquireixin automàticament algun tipus de coneixement a partir de l'observació d'un determinat conjunt de dades
- Els ordinadors són el mitjà (suport)
- Els algorismes (programari) donen la funcionalitat de l'aprenentatge automàtic

Aprentatge Automàtic (ML)

- Obtenir una descripció d'un concepte en algun camp del processament del llenguatge natural que ens permeti mostrar observacions i ajudi a predir noves instàncies d'aquesta distribució

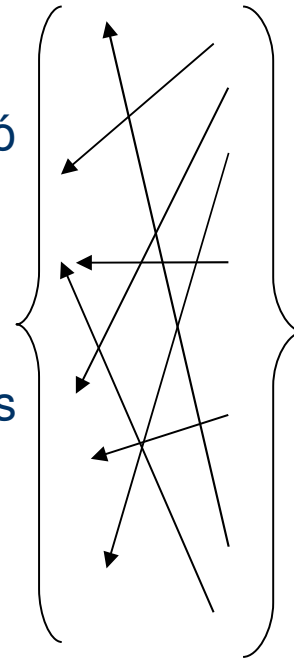


- L'estadística ens servirà per inferir a través de les mostres
- La computació ens permetrà crear algoritmes eficients per:
 - resoldre problemes d'optimització
 - Representar i avaluar els models

Aprentatge Automàtic (ML)

- Tipus de ML

- Aprentatge Supervisat:
Volem aprendre una relació entre unes i altres dades
- Aprentatge no Supervisat:
Tenim només unes úniques dades i volem trobar-hi regularitats entre elles

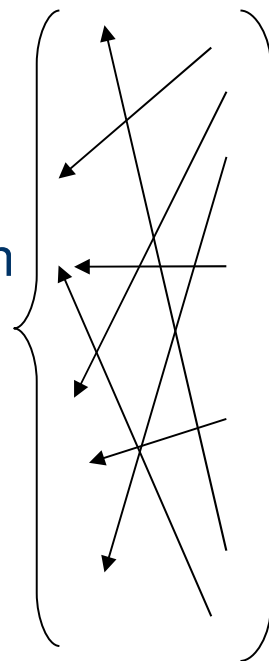


- Paradigmes de ML

- Arbres de decisió
- Llistes de decisió
- Clustering
- Inducció Lògica
- Algoritmes genètics
- Xarxes neuronals
- Maquines de Vectors de Suport
- etc..

Tasques de NLP

- Speech Recognition
- Spelling Correction
- Part-of-speech tagging
- Word-sense disambiguation
- Parsing (full/shallow)
- Information retrieval
- Information extraction
- Machine Translation
- NE Classification
- *I un llarg etc.*



Paradigmes de ML

- Arbres de decisió
- Llistes de decisió
- Clustering
- Inducció Lògica
- Algoritmes genètics
- Xarxes neuronals
- Maquines de Vectors de Suport
- etc..

Interacció entre ML i NLP

De ML a NLP

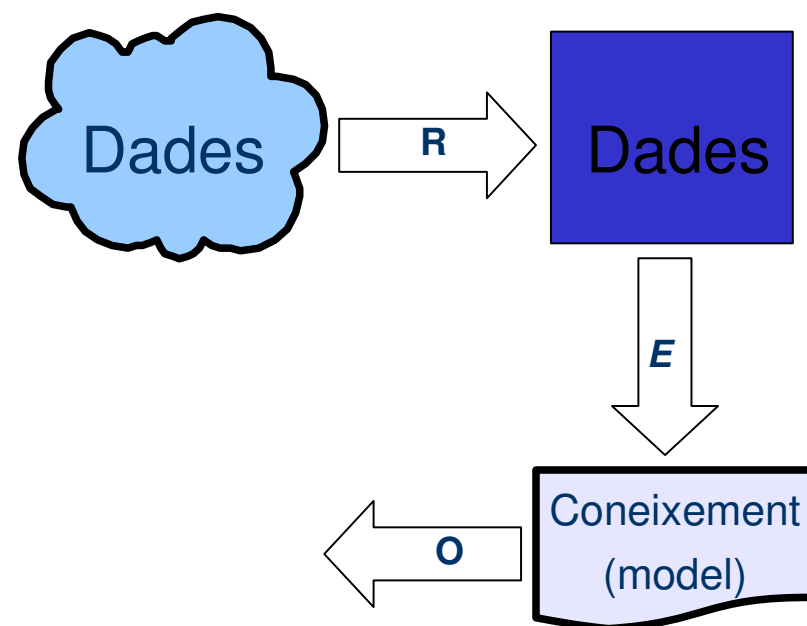
- Trobar la solució més apropiada per cada tipus de problema

De NLP a ML

- Problemes que plantegen reptes interessants ja que contenen característiques com ara: conjunts d'entrenament extremadament grans (o petits), alta dimensionalitat, atributs dependents, soroll en les dades, no només problemes de classificació, etc.

ML per NLP

- Formalització del problema
 - Representació
 - Cadenes de caràcters
 - Vectors de característiques
 - Tipus d'estructures
 - Etc. (camp molt obert)
 - Entrenament
 - Aplicar paradigma de ML
 - Objectiu
 - Classificar
 - Reconèixer
 - Detecció
 - Etc.

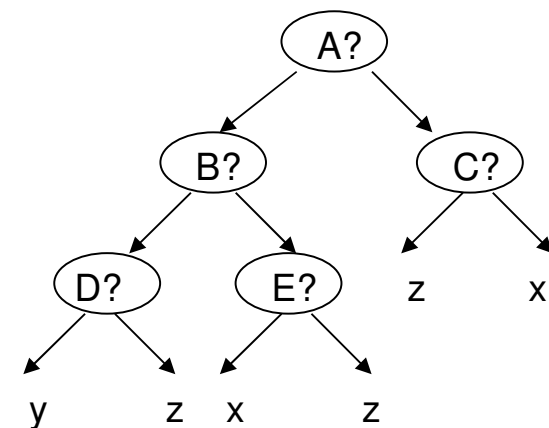
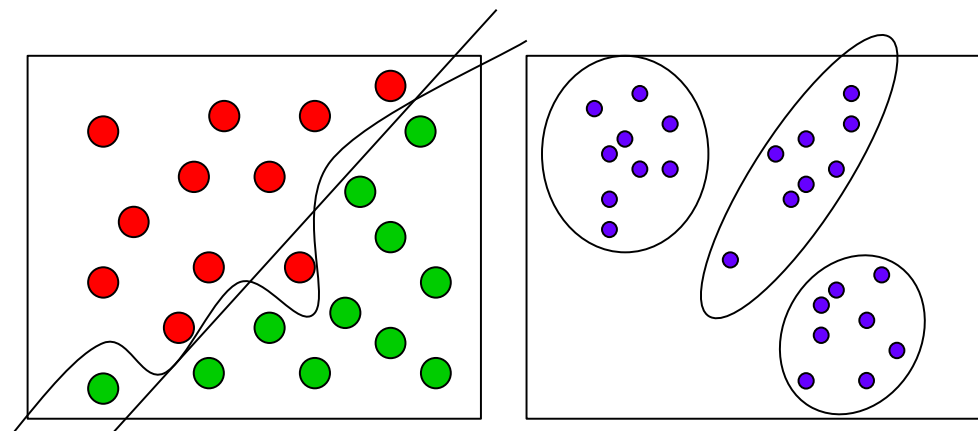


ML per NLP

• El Model

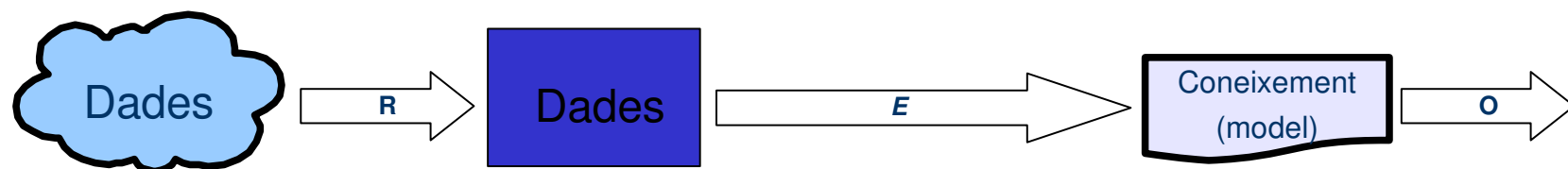
Representa la informació apresada en funció del paradigma utilitzat

- Regles
- Exemples d'una classe
- Etc.



Classificació d'Entitats amb Nom (NE)

- Aprentatge automàtic supervisat (Surdeanu et al, 2005; Màrquez et al, 2003)
- Aprentatge automàtic no supervisat (Collins, 1999)



- Etiquetatge manual
- Extracció de característiques

Forma Lema Forma[n-1..n] Forma[n-2..n] Forma[n-3..n] TextWithoutAlphabetic
 TextWithoutNumber isAllCap isAllCapOrDots isAllDigits isAllDigitsOrDotsComm
 isInitialCap PoS BIO

Classificació d'Entitats amb Nom (NE)

- Aprentatge automàtic supervisat (Surdeanu et al, 2005; Màrquez et al, 2003)
- Aprentatge automàtic no supervisat (Collins, 1999)

```

...
Creu eu reu Creu __nill__ Creu N N N N Y NCFS000 B-ORGANIZATION
Roja ja oja Roja __nill__ Roja N N N N Y AQ0FS0 I-ORGANIZATION
ha ha __nill__ __nill__ __nill__ ha N N N N N VAIP3S0 O
...

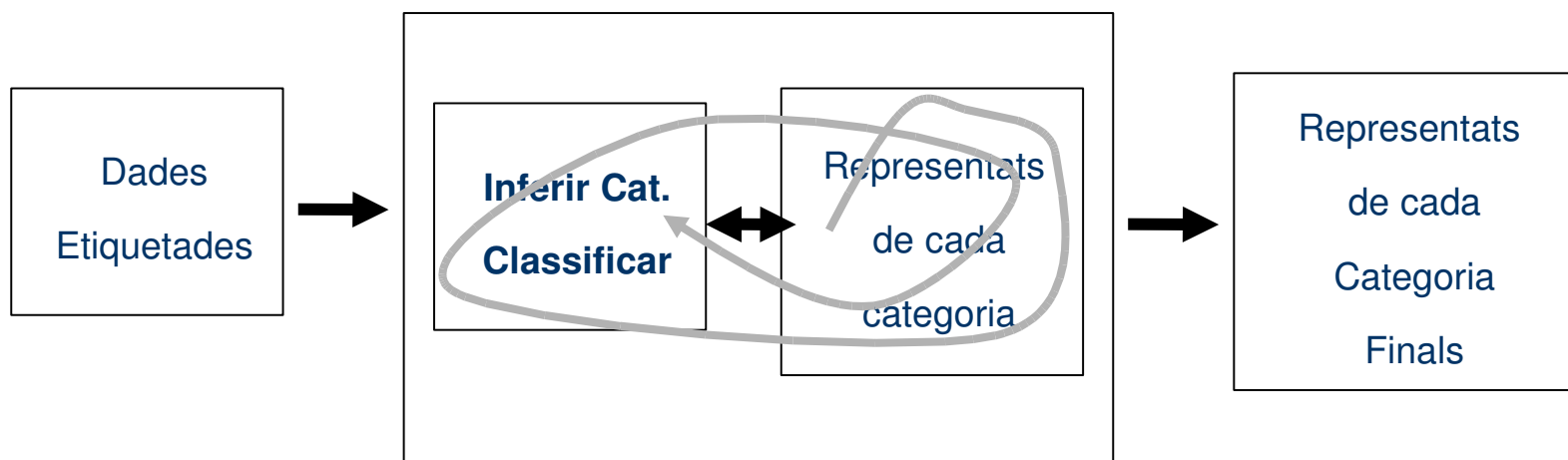
a a __nill__ __nill__ __nill__ __nill__ a N N N N N SPS00 O
106,52 106,52 52 ,52 6,52 106,52 , N N Y Y N Z B-MONEY
euros euro es nes enes __nill__ euros N N N N N NP00000 I-MONEY
per per er per __nill__ __nill__ per N N N N N SPS00 O
...

UNESCO unesco co sco esco __nill__ UNESCO Y Y N N Y NP00000
B-ORGANIZATION
...

```

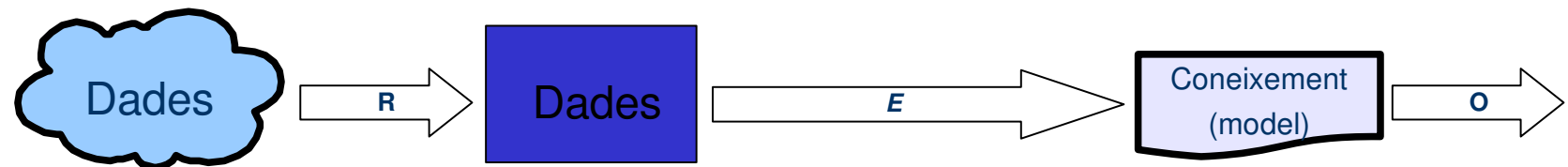
Classificació d'Entitats amb Nom (NE)

- Aprentatge automàtic supervisat (Surdeanu et al, 2005; Màrquez et al, 2003)
- Aprentatge automàtic no supervisat (Collins, 1999)



Classificació d'Entitats amb Nom (NE)

- Aprentatge automàtic supervisat (Surdeanu et al, 2005; Màrquez et al, 2003)
- Aprentatge automàtic no supervisat (Collins, 1999)



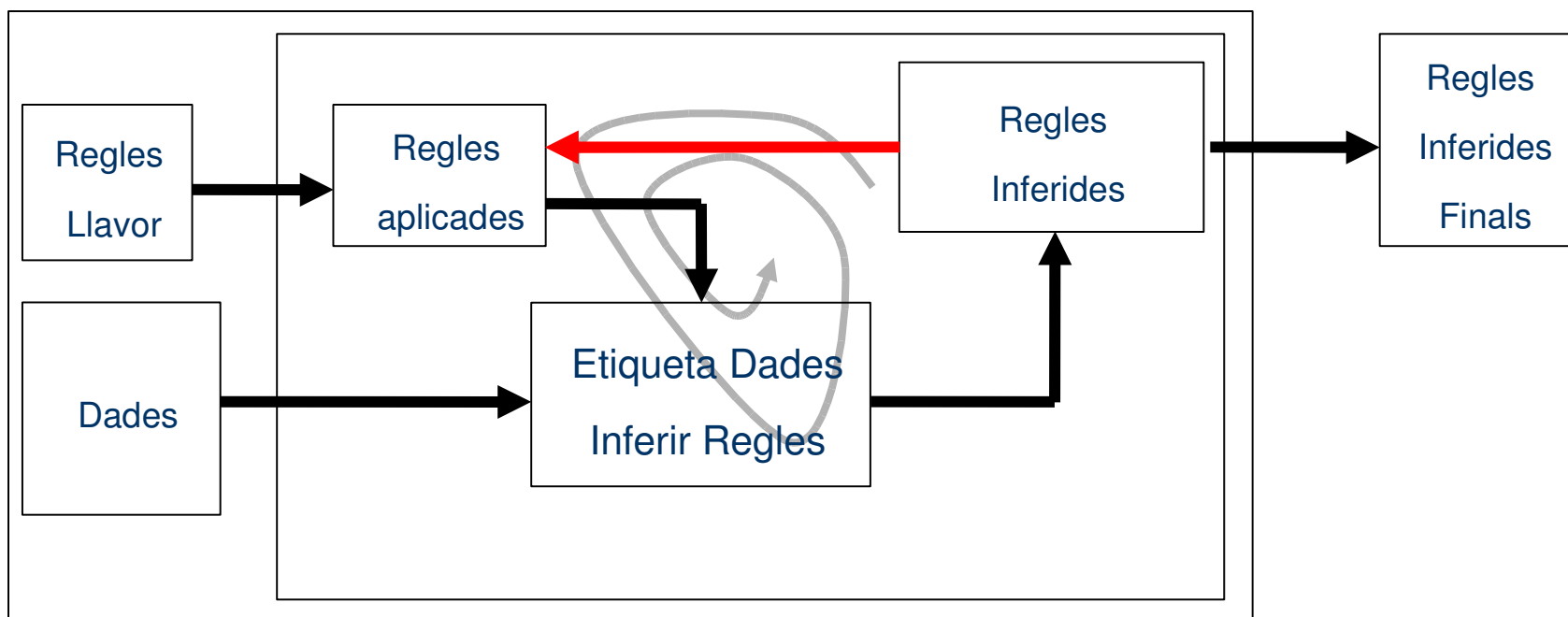
Llista de decisió

- Definició de regles llavor
- Extracció de característiques
 - Tipus de Sintagma
 - Conté
 - Trigger Word
 - Tipus de context (aposió, SP)

Text (New_York) → Lloc
 Text (Barcelona) → Lloc
 Conté (Sr.) → Persona
 Conté (Corporació) → Organització
 TotesMajúscules (si) → Organització
 ...

Classificació d'Entitats amb Nom (NE)

- Aprentatge automàtic supervisat (Surdeanu et al, 2005; Màrquez et al, 2003)
- Aprentatge automàtic no supervisat (Collins, 1999)



Bibliografia

- Machine Learning; Mitchell, 1997
- Machine Learning in Speech and Language Technologies; Roth, Fung, 2005
- Machine Learning Approaches for Natural Language Processing; Collins, 2003
- Projects in Machine Learning; Alpaydin, 2004
- Unsupervised Models for Named Entity Classification; Collins et al, 1999
- Low-cost Named Entity Classification for Catalan; Màrquez et al, 2005
- Mètodes Empírics pel processament del llenguatge natural; Doctorat en Intel·ligència Artificial (UPC), Ll. Màrquez