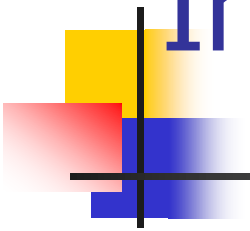
Sistemi di Elaborazione dell'informazione II

Corso di Laurea Specialistica in Ingegneria Telematica
II anno – 4 CFU
Università Kore – Enna – A.A. 2009-2010

Alessandro Longheu
http://www.diit.unict.it/users/alongheu
alessandro.longheu@diit.unict.it

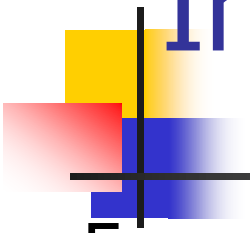
Information Extraction



Information Extraction (IE)

- La Information Extraction (IE) identifica frammenti di informazione specifici in testi parzialmente strutturati (es. XML) o non strutturati (es. testo puro) e trasforma l'informazione estratta in un database strutturato.
- Si applica ai domini più diversi:
  - Articoli di giornale
  - Pagine Web
  - Letteratura scientifica
  - Messaggi di Newsgroup
  - Annunci economici o di lavoro
  - Cartelle cliniche

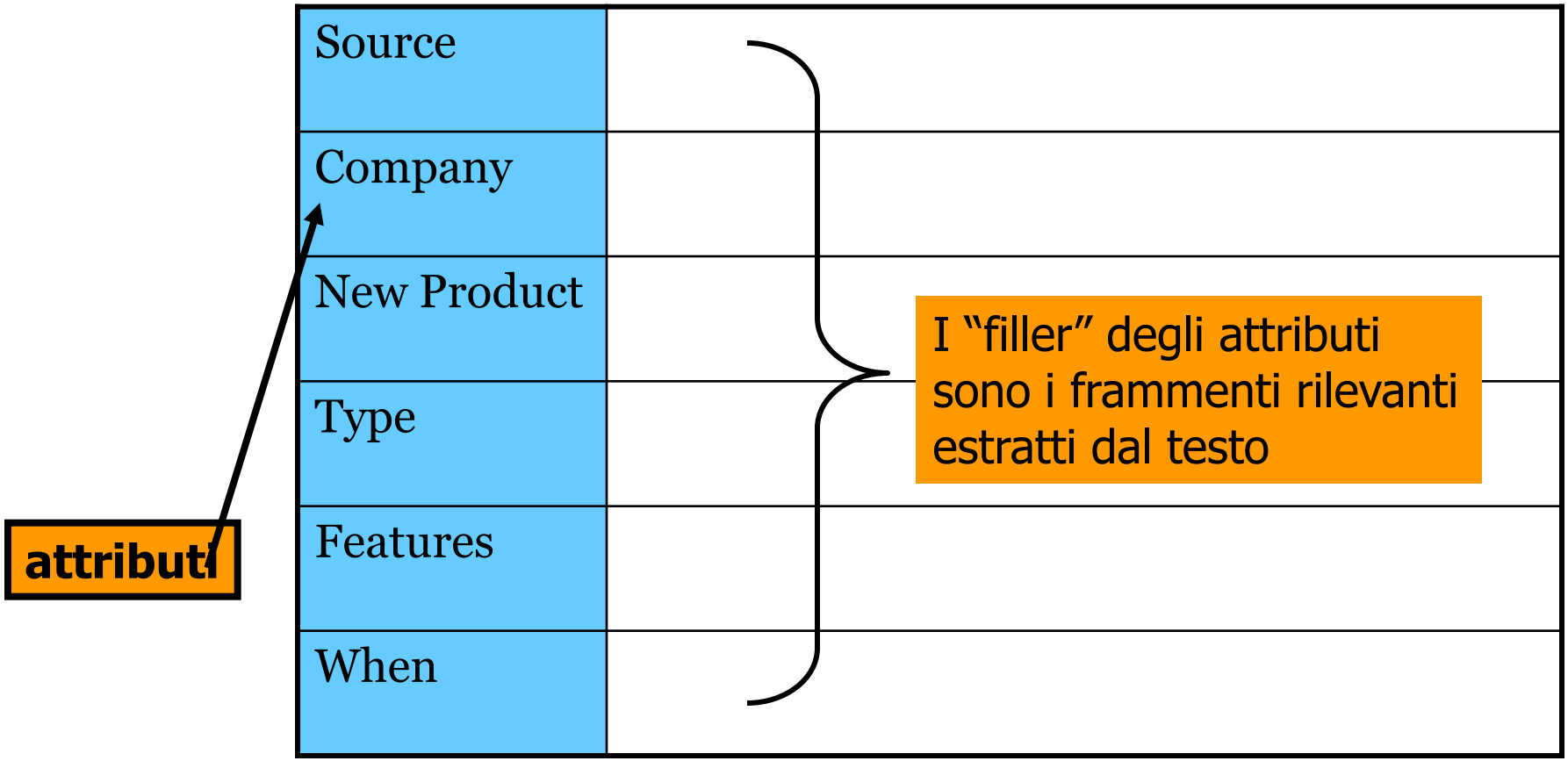
Information Extraction (IE)



- Formally, an **IE task is defined by its input and its extraction target.** The input can be unstructured documents like free text that are written in natural language or the semistructured documents that are pervasive on the Web, such as tables or itemized and enumerated lists
- The extraction target of an IE task can be a relation of k-tuple or it can be a complex object with hierarchically organized data. A generic term to identify such object is "**template**"
- Programs that perform the task of IE are referred to as **extractors or wrappers**

Information Extraction (IE)

- **Template**: uno "schema" dell'informazione di interesse



Information Extraction (IE)

Source	http://www.semtech.com/press/products/semtech_adds
Company	Semtech Corp (SMTC)
New Product	SC4525A, SC4524A, SC454B
Type	Three new step-down (buck) regulators
Features	High input voltage and programmable frequency
	February 26, 2008

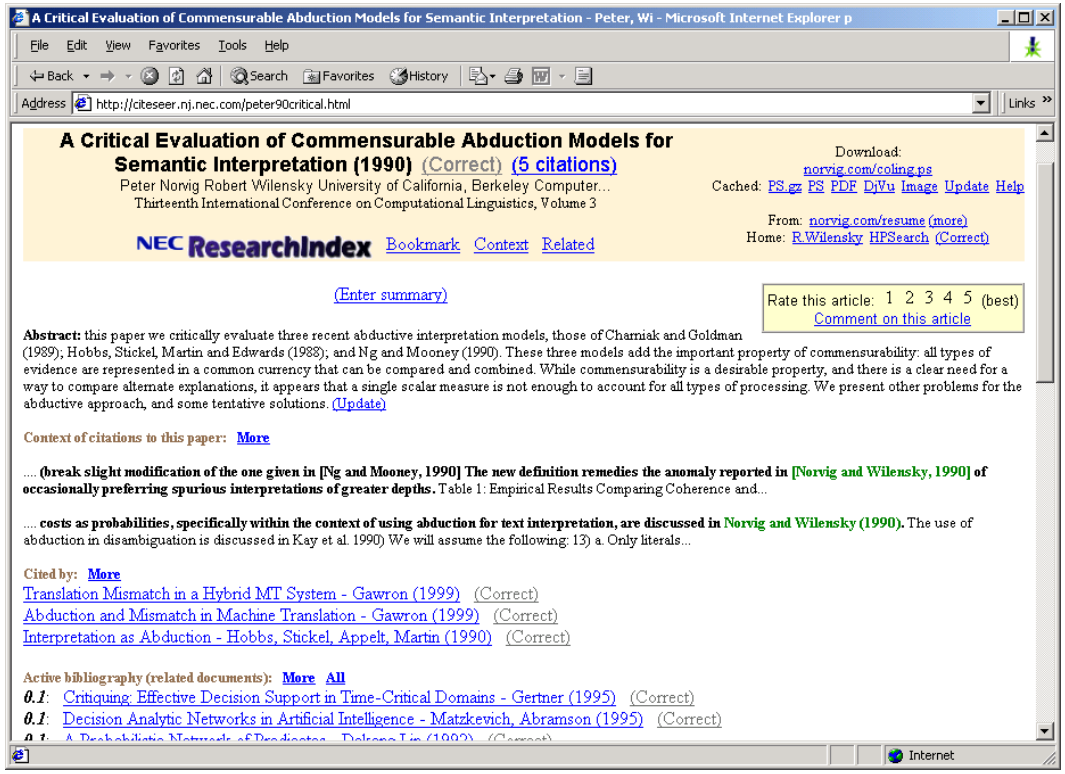
Camarillo, California - **February 26, 2008**

Semtech Corp. (Nasdaq: SMTC), a leading supplier of analog and mixed-signal semiconductors, today announced the SC4525A , SC4524A and SC4524B, three new step-down (buck) regulators with high input voltage and programmable frequency needed for networking and digital consumer applications.

Template filling

Information Extraction (IE)

Un altro esempio: IE da articoli di ricerca



Information Extraction (IE)

- DARPA (Defense Advanced Research Projects Agency) ha finanziato ricerche sul Message UnderStanding in IE dal 1990. **Message Understanding Conference (MUC)** è la conferenza-gara del settore. Dal 2000 sostituita (inglobata) dalle TREC conferences
- Generalmente le gare hanno come tema l'estrazione di notizie da giornali: Eventi terroristici, Joint venture, Cambi di management
- I task oggetto della MUC sono stati diversi: named entity recognition (marcare nel testo le stringhe che rappresentano persone, organizzazioni, luoghi, date, ecc.), coreference resolution (identificare le coreferenze alla stessa entità all'interno del testo), template element construction, template relation construction, and scenario template production.

Information Extraction (IE)

Applicazioni:

- Job postings:
  - Newsgroups: Rapier da austin.jobs
  - Pagine Web : Flipdog
- Annunci di lavoro:
  - BurningGlass
  - Mohomine
- Annunci di Seminari
- Notizie societarie sul web
- Corsi sul web (continuing education)
- Informazioni e annunci universitari sul web
- Annunci di affitto appartamenti
- Informazioni di biologia molecolare su MEDLINE

Information Extraction (IE)

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

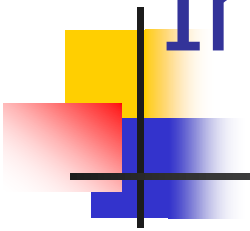
Esempio "Offerte di lavoro"

Information Extraction (IE)

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Struttura estratta (template)

Information Extraction (IE)

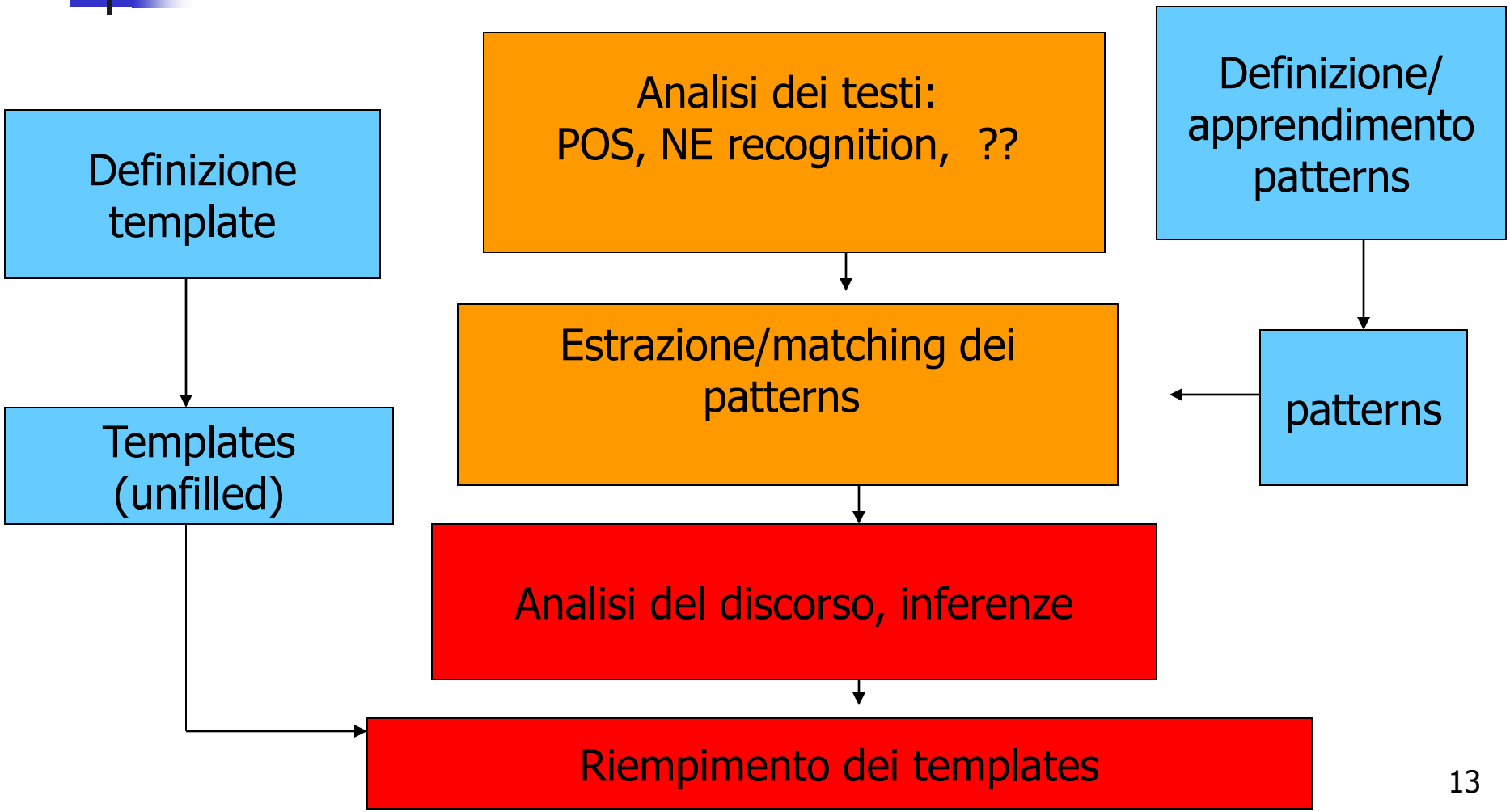


- Volendo **confrontare IE ed IR**, sono operazioni complementari
- Data una base documentale:
  - L'IR restituisce il sottoinsieme di documenti rilevanti per una certa interrogazione in input. L'informazione di interesse sarà quindi cercata dall'utente
  - L'IE estrae in modo strutturato le informazioni rilevanti per l'utente. IE produces structured data ready for postprocessing, which is crucial to many applications of Web mining and searching tools.

Information Extraction (IE)

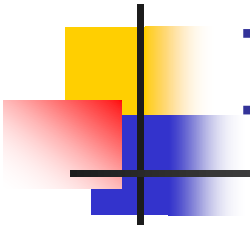
- In generale, tutti i sistemi di IE (non strutturati) hanno la seguente struttura:
  - Definizione dello schema dei templates (manuale o automatica)
  - Analisi del testo (se web, la cosa è più complicata per la presenza di figure, frames.., ma se il testo è semi-strutturato (xml) può anche essere più semplice)
  - Estrazione dei "fillers", con metodi di ML o pattern matching
  - Riempimento dei templates

Information Extraction (IE)



Information Extraction (IE)

- POS=Part of speech, ossia la classificazione dei termini in categorie grammaticali (nomi, avverbi, verbi ecc.)
- NE=Named Entities. NE involves **identification** of *proper names* in texts, and **classification** into a set of predefined categories of interest. Three universally accepted categories: **person**, **location** and **organisation**



Information Extraction (IE)

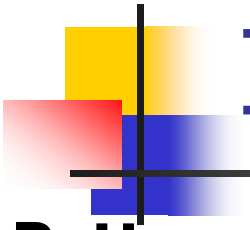
- La prima fase è la Template extraction
- Nel caso si voglia estrarre da documenti semi-strutturati (es. Amazon) l'estrazione di templates è relativamente semplice, inoltre gli slot fillers seguono un ordine predeterminato:
  - Title
  - Author
  - List price
  - …
- Molto più complesso nel caso di testi liberi.

Information Extraction (IE)

- I template sono "Record" di coppie **attributo** (slot) **valore** (filler). Valori sono parti del testo con cui riempire lo slot.
- Gli slot vanno riempiti con stringhe la cui natura (lessicale, sintattica, semantica) è in genere predeterminata in modo più o meno specifico
  - Terrorist act: threatened, attempted, accomplished.
  - Job type: clerical, service, custodial, etc.
  - Company type: codice SEC
- Alcuni slot possono accettare elementi di una classe, es:.
  - Programming language (JAVA, Prolog, C, ecc.)
  - Company names (IBM, ACE Inc, ...)
- In alcuni domini si devono estrarre più template da uno stesso documento, ad esempio una lista di appartamenti in vendita, in un unico avviso

Information Extraction (IE)

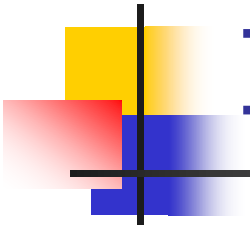
- Lo slot filling prevede due metodi:
  - Pattern matching: costruire espressioni regolari più o meno generalizzate che catturino la regolarità di certe stringhe.
  - Machine learning: imparare ad assegnare stringhe di testo ai vari slot (cioè classificarle come ad es: purchaser, seller, location, ...).



Information Extraction (IE)

Pattern matching

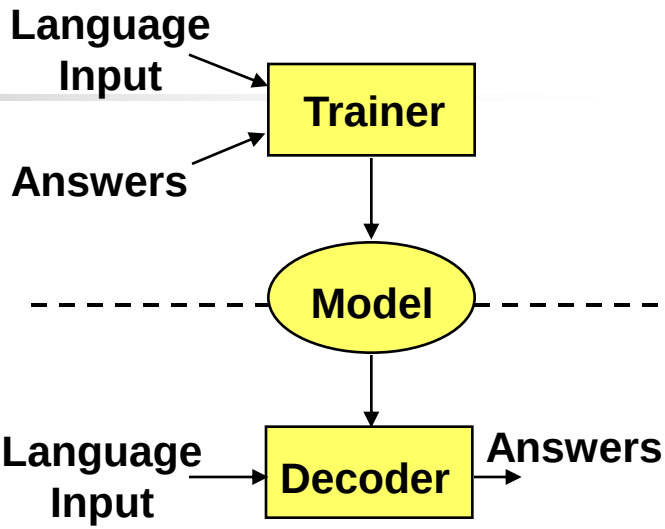
- Nel caso si estraggano i pattern da pagine web automaticamente generate, bastano espressioni regolari.
- In caso contrario, occorre utilizzare alcune tecniche di NLP.
  - Part-of-speech (POS) tagging
  - Syntactic parsing
  - Categorie semantiche (es da WordNet)
    - KILL: kill, murder, assassinate, strangle, suffocate
- I pattern possono usare categorie lessicali, sintattiche, semantiche.
  - Crime victim:
    - Prefiller: [POS: V, Hypernym: KILL]
    - Filler: [Phrase: NP]



Information Extraction (IE)

- L'aspetto più critico è la scrittura di pattern specifici per ogni dominio e template
- Scrivere delle regex accurate richiede tempo ed è una attività domain-dependent (non ri-usabile).
- L'alternativa è usare tecniche di machine learning:
  - Si parte da un set di apprendimento in cui esperti umani evidenziano i pattern di interesse (es. si sottolineano i filler degli slots).
  - Impara un modello generalizzato degli slot-fillers (cioè un pattern) usando algoritmi di ML.

Information Extraction (IE)



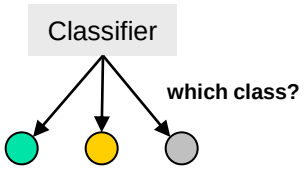
- Automatic pattern learning
- Vantaggi:
  - Portabile a vari dominii
  - I pattern hanno una copertura più ampia
  - Non serve rivolgersi a knowledge engineers
- Svantaggi:
  - Bisogna annotare un campione ampio di documenti.
  - Non funziona sicuramente meglio di un sistema in cui i pattern siano scritti a mano
- Esempi: ELIE (Ucd) , Riloff et al., AutoSlog (UMass); Soderland WHISK (UMass); Mooney et al. Rapier (Utexas)

Information Extraction (IE)

Machine Learning K. Nigam, *Machine Learning for Information Extraction*

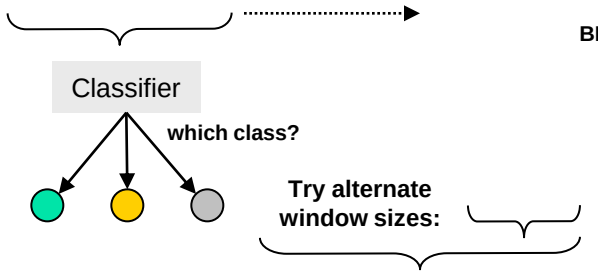
Classify Candidates

Abraham Lincoln was born in Kentucky.



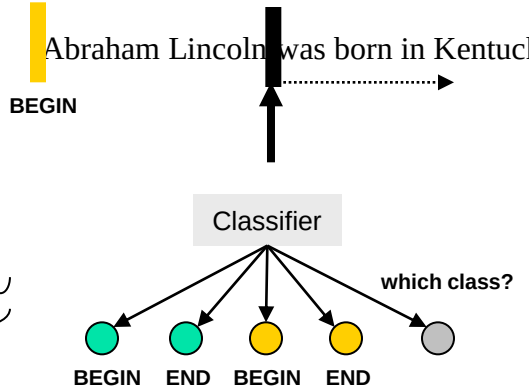
Sliding Window

Abraham Lincoln was born in Kentucky.



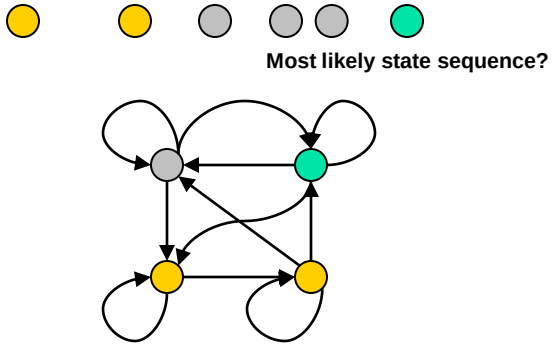
Boundary Models

Abraham Lincoln was born in Kentucky.



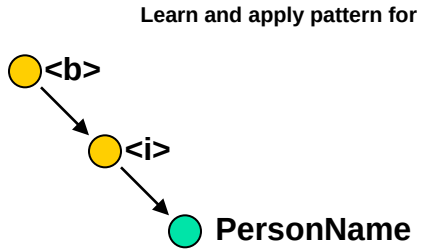
Finite State Machines

Abraham Lincoln was born in Kentucky.



Wrapper Induction

Abraham Lincoln was born in Kentucky.



Web Information Extraction

- I sistemi di IE non-web analizzano documenti non strutturati, utilizzando prevalentemente metodi di analisi del linguaggio naturale e tecniche di machine learning o reasoning (AI)
- I sistemi web analizzano **documenti semi-strutturati** (offerte di lavoro, avvisi commerciali, ...) e utilizzano prevalentemente pattern matching, comunque in generale metodi più semplici
- traditional IE usually takes advantage of NLP techniques such as lexicons and grammars, whereas Web IE usually applies machine learning and pattern mining techniques to exploit the syntactical patterns or layout structures of the template based documents

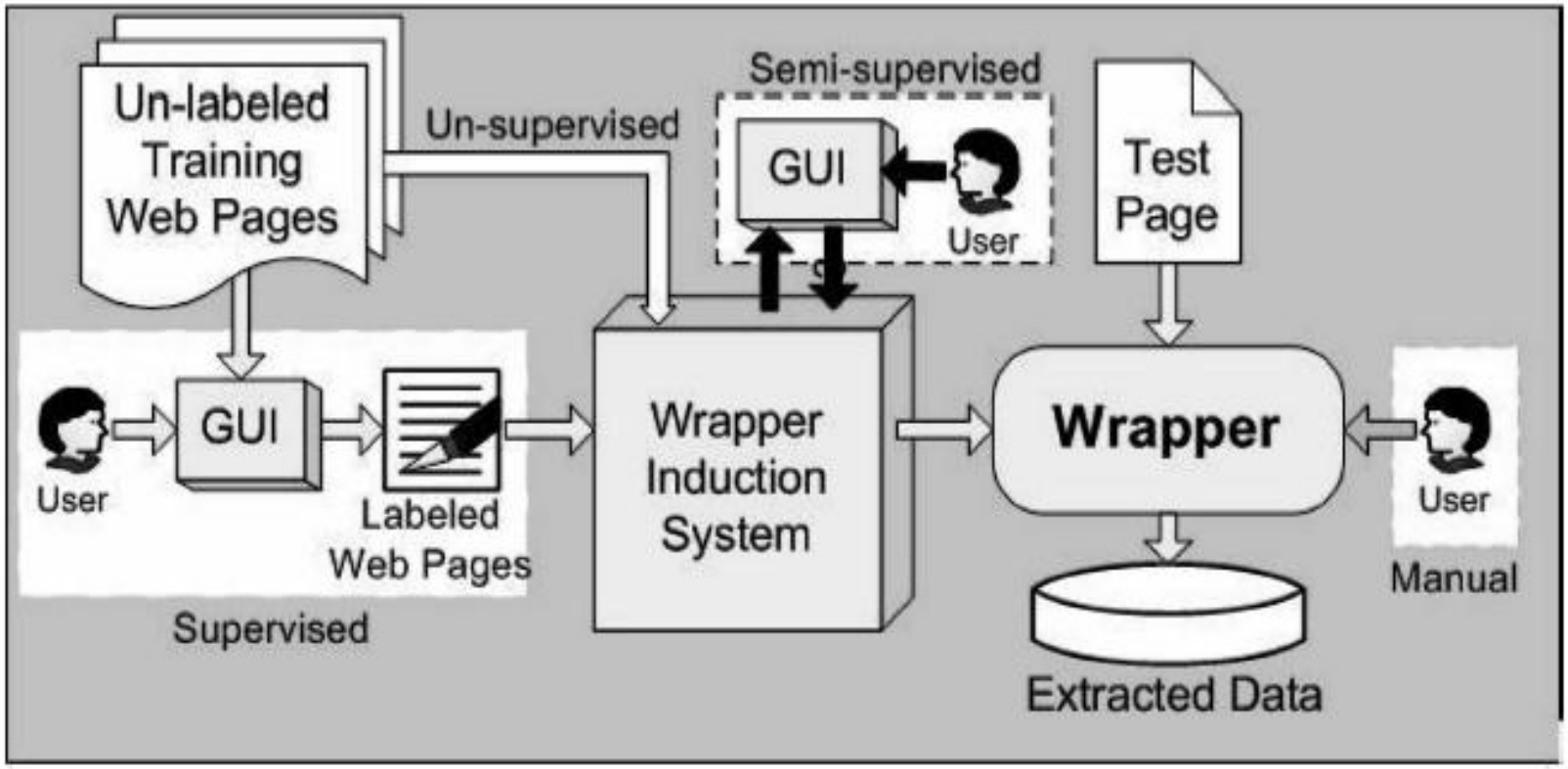
Web Information Extraction

- Sul web ci sono molte sorgenti di informazione strutturata, ad es: elenchi telefonici (pagine gialle), cataloghi di prodotti (es. cataloghi di libri, come AMAZON), previsioni del tempo, stock quotes..
- Queste risorse sono formattate per persone, non per la manipolazione da parte di computer
- Un "**Wrapper**" è una procedura per estrarre uno specifico contenuto (ad es. un template)
- I Wrapper possono essere codificati a mano, o usando tecniche di apprendimento
- Essendo i sorgenti parzialmente strutturati, questo compito è per certi versi più semplice che non estrarre template fillers da testi liberi.

Web Information Extraction

- A wrapper was originally defined as a component in an information integration system which aims at providing a single uniform query interface to access multiple information sources.
- In an information integration system, a wrapper is generally a program that "wraps" an information source (e.g., a database server or a Web server) such that the information integration system can access that information source without changing its core query answering mechanism.
- In the case where the information source is a Web server, a wrapper must query the Web server to collect the resulting pages via HTTP protocols, perform information extraction to extract the contents in the HTML documents, and finally integrate with other data sources.

Web Information Extraction



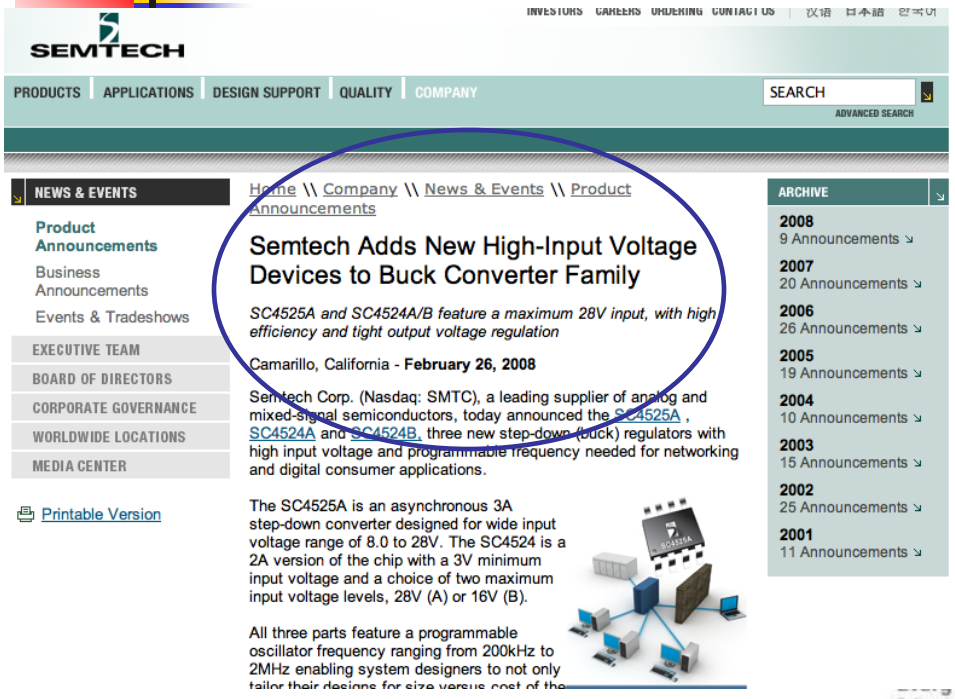
- Architettura di un sistema per Wrapper Induction (an IE system designed to generate wrappers)

Web Information Extraction

- Se il testo ha una struttura ripetitiva, ad esempio tag html o xml che siano informative, questo facilita l'analisi

```
<h2>Key Features of the SC4525A and SC4524A/B</h2>
<ul>
   <li>Wide input range of 8V to 28V (SC4525A);
from 3V to 28V (SC4524A); and from 3V to 16V (SC4524B)</li>
   <li>3A output current  (SC4525A)</li>
   <li>2A output current  (SC4524A/B)</li>
   <li>200kHz to 2MHz programmable frequency</li>
   <li>Peak current-mode control</li>
   <li>Cycle-by-cycle current limiting</li>
   <li>Hiccup overload protection with frequency foldback</li>
   <li>Programmable soft-start and enable</li>
   <li>Thermal shutdown</li>
</ul>
```

Web Information Extraction



I sistemi web-IE devono analizzare la struttura html delle pagine per identificare le porzioni rilevanti

Web Information Extraction

- La valutazione dell'accuratezza di un sistema di Web - IE va fatta su testi sui quali non sia stato fatto alcun apprendimento. Misura per ogni documento:
  - Numero totale di estrazioni corrette : *N*
  - Numero totale di coppie slot-valore estratte dal sistema : *E*
  - Numero totale di coppie slot-valore estratte dal sistema che sono corrette (rispetto al template-soluzione): *C*
- Misure di prestazione (simili a quelle per IR standard):
  - Recall = *C*/*N*
  - Precision = *C*/*E*
  - F-Measure = media armonica fra recall e precision

Web Information Extraction

- Se i documenti sono annotati (dal semplice XML fino alle annotazioni semantiche mediante ontologia) le tecniche di IE sarebbero banali. Ma è difficile annotare manualmente archivi documentali in xml o altri linguaggi di annotazione. Alcune industrie commerciali potrebbero essere riluttanti a fornire dati in formati cosè accessibili.
- In realtà, un'altra applicazione di IE è proprio quella di trasformare documenti non strutturati in files annotati in xml.

"Mr. John Smith è stato nominato Presidente della ACE Spa il 25 dicembre 2222".

<management_change>

<new_manager>Mr. John Smith</new_manager> è stato nominato <title>Presidente</title> della <company>ACE Spa</company> il <date_of_event>25 dicembre 2222 </date_of_event> </ management_change>