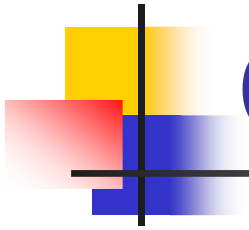
TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

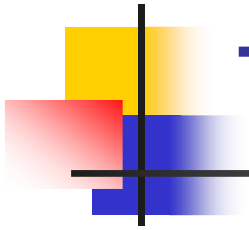
Introduction

Unit 0



Objetives

- Introduction to the field of Information Extraction
- Study of the techniques to transform unstructured information into structured information
- Introduction to the resources (POS tagger, EuroWordNet, etc.) that provide the necessary information to the information extraction system
- Introduction to some of the classical information extraction systems that have participated in different MUC conferences such as LaSIE and PROTEUS.



Theoretical/Practical Contents

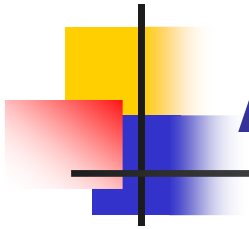
UNIT 1	Introduction to Information Extraction
UNIT 2	Information Extraction systems
UNIT 3	Architecture of Information Extraction Systems
UNIT 4	Entity Recognition
UNIT 5	Correference resolution
UNIT 6	IE Applications: Temporal Expressions
UNIT 7	Pattern inference
UNIT 8	PRACTICAL WORK: Python and Template filling task with IE

Sessions

	Wednesdays	**Thursdays**
Session 1	Introduction Unit 1, 2 and 3	Practical work: Python
Session 2	Unit 4 and 5.	Practical work Unit 6
Session 3	Unit 6 (I)	Practical work Unit 6
Session 4	Unit 6 (II)	Practical work Unit 8
Session 5	Unit 7	Practical work Unit 8
Session 6	Group session: Presentations of practical works	Final test

Basic bibliography (CV)

- *Information extraction.* SARAWAGI, Sunita. Now Publishers Inc (2008).
- *Named Entities : recognition, classification, and use. SEKINEN, Satoshi Sekinen (Ed).* RANCHHOD, Elisabeth. Published by: John Benjamins Publishing Company (2009).
- *Mining the social* web. RUSSELL, Matthew A. O`Reilly Media (2011). ACCESO ONLINE DESDE UA
- *Text mining handbook : advanced approaches in analyzing unstructured data.* Ronen Feldman and James Sanger Cambridge University Press, USA (2011).



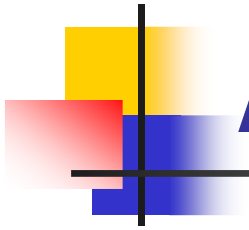
Additional bibliography

- *Information Extraction*. J. Cowie and W. Lehnert. Communications of ACM 1996

- *Information Extraction: Technique and Challenges.* R. Grishman. In Maria Teresa Pazienzia (Ed.). LNCS, 1299:10-27. Springer-Verlag 1997

- *Cross Lingual Information Extraction and Automated text Summarization.* E. Hovy. Chapter III of Multilingual Information Management: Current Levels and Future Abilities 1999

- *Introduction to Information Extraction technology.* D.E. Appelt and D.J. Israel. Tutorial for IJCAI 1999 *http://www.ai.sri.com/~appelt/ie-tutorial*



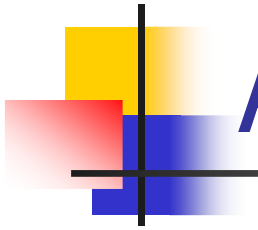
Additional bibliography

- *Empirical Methods in Information Extraction*. C. Cardie. AI Magazin Winter 1997

- *Information Extraction a User Guide*. H. Cunningham. Research Memo CS-97-02. Univ. Sheffield 1997

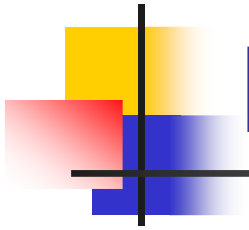
- *Information Extraction as a core language technology. What is IE?.* Y. Wilks. In Maria Teresa Pazienzia (Ed.). LNCS, 1299:10-27. Springer-Verlag 1997

- *Information Extraction*. J. Cowie, Y. Wilks. In R. Dale, H. Moisl and H. Somers (eds.) The Handbook of Natural Language Processing. 2000. http://www.dcs.shef.ac.uk/~yorick/papers/infoext.pdf



Additional bibliography

- *Customization of Information Extraction Systems*. R. Yangarber and R. Grishman. In Maria Teresa Pazienzia (Ed.). LNCS, 1299:10-27. Springer-Verlag 1997
- *Inferential Information Extraction*. M. Vilain. Proceeding of ACL 1999
- MUC-6. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
- MUC-7 http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

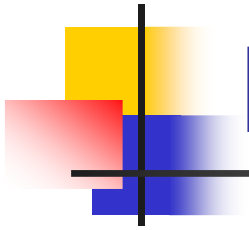


Recommended links (CV)

- Home page of Ralf Grishman, father of Information Extraction
  - http://cs.nyu.edu/grishman/
- Download JET, (Java Extraction Toolkit) a platform for building Information Extraction systems
  - http://cs.nyu.edu/grishman/jet/license.html
- Wikipedia
  - http://en.wikipedia.org/wiki/Information_extraction
- Temporal information annotation tool Callisto
  - http://fofoca.mitre.org/annotation_tools/timex2_annotation_tools.html
- GATE platform for building Information extraction systems (Sheffield University)
  - http://gate.ac.uk/ie/

Recommended links(CV)

- The Stanford Natural Language Processing Group. It includes a link for downloading Stanford Named Entity Recognizer.
  - http://nlp.stanford.edu/ner/index.shtml
- Link to Natural Language Toolkit that includes an IE subsystem.
  - http://nltk.org/
- TIDES annotation schema
  - http://projects.ldc.upenn.edu/ace/docs/English-TIMEX2-Guidelines_v0.1.pdf
- Webpage with information and temporal processing systems
  - http://timexportal.wikidot.com/start
- Timeml annotation schema
  - http://www.timeml.org/site/index.html

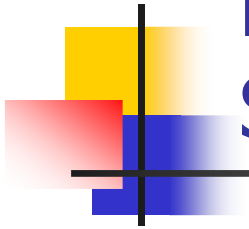


Evaluation

- Participation
- Practical works

Introduction to Information Extraction

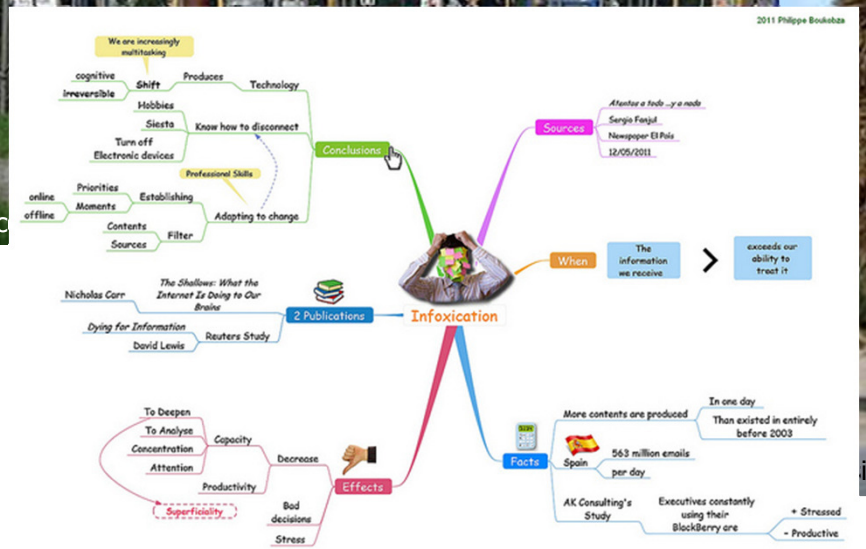
Unit 1



Motivation
State of the art TODAY

- Growth of the amount of electronic documents.
  - Use of paper is obsolete
- Information Society
  - Access to users without expertise
- Multilingual Society
  - Necessity of machine translation
- High maintenance costs for documental DB
- Impossible for human beings to process the huge amount of new digital information:
  - 500 million Tweets / day (data from October 2012)
  - 100 million more than 3 months before…
- Need for tools to treat texts:
  - HLT Human Language Technologies

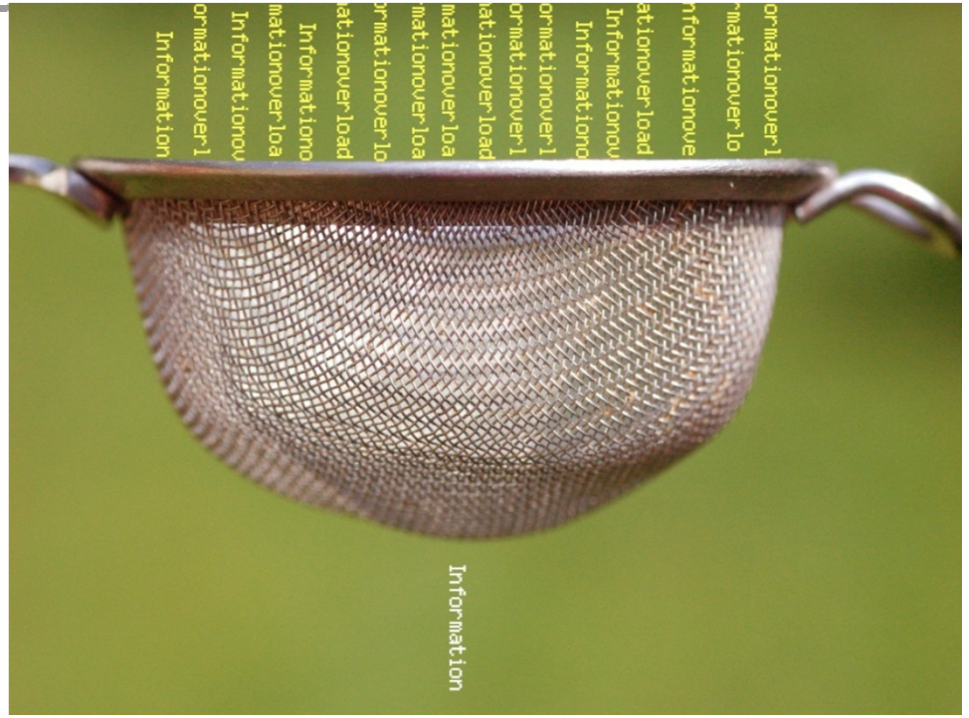
The problem: infoxication



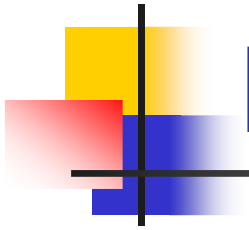
Excess of information

- Delays our actions…
- Prevents us from reacting on time..

The solution
Human Language Technologies

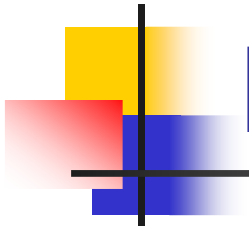


- What is this?
  - A technology to make the computer think
  - In order to understand the information written by human beings
  - Capable of FILTERING huge amounts of written information



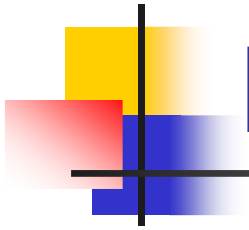
HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarization
- Data Mining
- Web Mining
- Knowledge Extraction



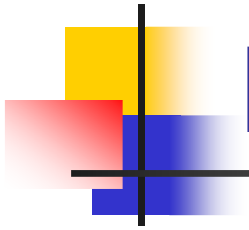
HLT Applications

- Information Retrieval
  - From a set of documents, it chooses the set of them that are related with a question entered by the user, by using a set of keywords.
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarisation
- Data Mining
- Web Mining
- Knowledge Extraction



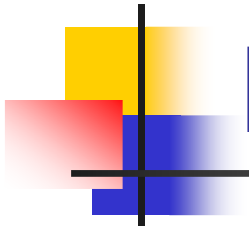
HLT Applications

- Information Retrieval
- Information Extraction
  - Obtaining the relevant information of a relevant document. What constitutes relevant information must be defined beforehand.
- Question Answering Systems
- Document Classification
- Summarisation
- Data Mining
- Web Mining
- Knowledge Extraction



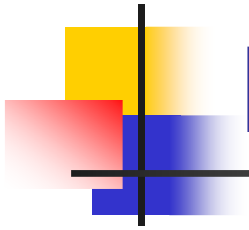
HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
  - Simplified version of IR. It only provides the text fragment where the answer to the user is found.
- Document Classification
- Summarisation
- Data Mining
- Web Mining
- Knowledge Extraction



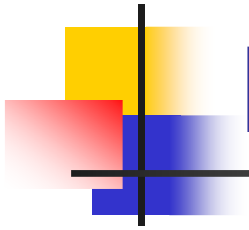
HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
  - Simplified version of IE. It assigns a category to the document. Previously, a set of categories must be defined.
- Summarisation
- Data Mining
- Web Mining
- Knowledge Extraction



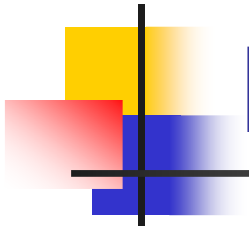
HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarisation
  - It creates automatic summaries of a document.
- Data Mining
- Web Mining
- Knowledge Extraction



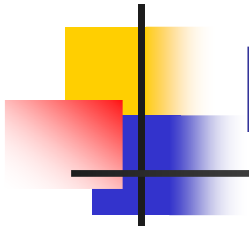
HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarisation
- Data mining
  - Within a given data structure, finding groups of them that are related by similar behaviour or properties.
- Web Mining
- Knowledge Extraction



HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarisation
- Data mining
- Web mining
  - It helps to navigate, search or visualise contents
- Knowledge Extraction

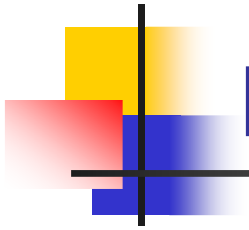


HLT Applications

- Information Retrieval
- Information Extraction
- Question Answering Systems
- Document Classification
- Summarisation
- Data Mining
- Web Mining
- Knowledge Extraction
  - Beyond Data Mining. It does not only extract data schemas but also tries to discover new properties by interpreting them and then using a model.

Information Extraction Systems

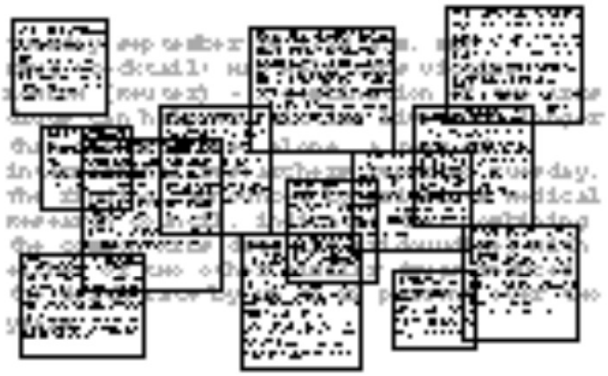
Unit 2

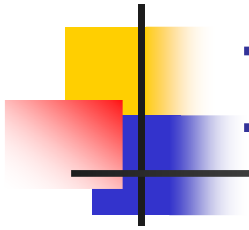


Definition of Information Extraction

- Cowie and Lehnert (1996). "Technique that provides certain information known to be relevant from a set of relevant texts"
- Gaizauskas and Wilks (1998). "It is the task of automatically extracting a type of prespecified information from text"

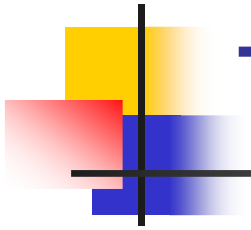
Information Extraction





Information Extraction

- Building systems that find and relate relevant information whereas the non relevant information is ignored.
- Relevant information is determined by a predefined guide of the domain, in which the type of information to be extracted must be as specific as possible
- From the NLP point of view, IE systems must work on different levels:
  - from word recognition to sentence analysis
  - From sentence-level understanding to the whole text



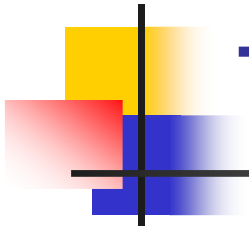
Templates

<Team-0001> :=
NAME: :
FIELD :
ADDRESS :
PRESIDENT :
COACH:

<Person-0001>:=
NAME: :
ALIAS:
POSITION :
TEAM :
GOALS:

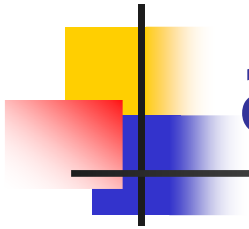
<Judge-0001> : =
NAME: :
SCHOOL :
INTERNATIONAL:
CATEGORY :

<MATCH-0001> : =
TEAMS :
REFEREES :
RESULTS :
GOAL_SCORERS:
CARDS:



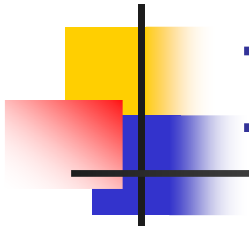
Template filling

<Team-0001> := NAME: : Levante FIELD : Nuevo Estadio ADDRESS : PRESIDENT : COACH: **<Person-0001>:=** NAME: : ALIAS: Veiga POSITION : Goalkeeper TEAM : <Team-0001> GOALS:	**<Team-0002>** := NAME: : Real Jaén FIELD : ADDRESS : PRESIDENT : COACH: **<Person-0017>:=** NAME: : ALIAS: Pepelu POSITION : Forward TEAM : <Team-0002> GOALS: 1	**<Judge-0001> : =** NAME: : Tristan Oliva SCHOOL : Murciano INTERNATIONAL: CATEGORY : **<MATCH-0001>** : = TEAMS : <Team-0001>, <Team-0002> REFEREES : <Judge-0001>, ... RESULTS : 0-1 GOAL_SCORERS: <Person-0017> CARDS: <Person-00XX>, ...



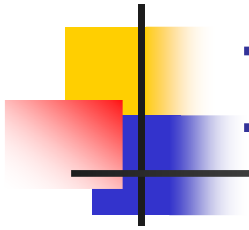
IE compared to other applications

- IE vs. information retrieval
- IE vs. full text understanding



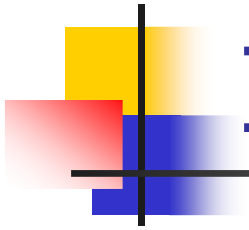
IE vs. information retrieval

- Information retrieval (IR)
  - From a query, IR selects a subset of relevant documents from a big collection
  - The users manually select the document that meets the requirements
- IE extracts the relevant information from the documents -> IR and IE are complementary technologies



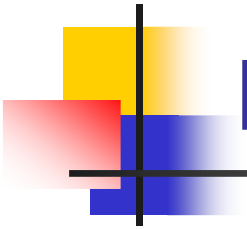
IE vs Text understanding

- IE
  - Generaly, only a part of the text is relevant
  - The information is mapped in a predefined representation, quite simple and invariable (structuring of the information)



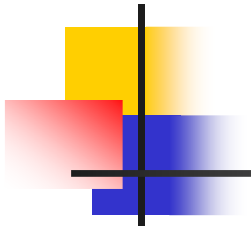
IE vs Text Understanding

- Text Understanding
  - The objective is to get the meaning of the whole text
  - The final representation must take into account all the complexities of the language



Brief histoy

- Before DARPA intervention
- Under DARPA guides.
- LRE projects



Brief history

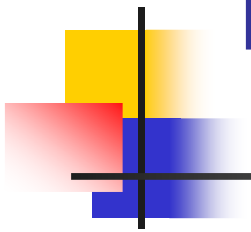
Before DARPA intervention

- SAGER (1981). Template filling from text of a sublanguage of the domain.
- DeJong (1982)
- Zarri (1983)
- JASPER (1986). First commercial system

Brief histoy
Under DARPA intervention

- MUC (Message Understandig Conference)
  - Established a quantitative evaluation system for all the IE systems, establishing the task to be performed by the systems and the set of text over the tasks that must be performed
- 7 MUC were developed

Conference	*Year*	*Text Source*	*Topic (Domain)*
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches



Brief history

Under DARPA intervention

- MUC-1 (May-87)
  - 6 systems participated
  - Domain: naval operations
  - The task wasn´t defined and no evaluation measures were established.

Brief history
Under DARPA intervention

- MUC-2 (May-89)
  - 8 systems participated
  - Domain: naval operation
  - A template was defined for the filling of its characteristics
  - Some evaluation criteria were defined but they were rejected because they were not optimal.

Brief history
Under DARPA intervention

- MUC-3 (May-91)
  - 15 systems participated
  - Domain: terrorism in Latin America
  - The evaluation criteria used in IR were adapted
    - Precision. It measures how good the system is
    - Recall. It measures how complete the system is

Brief history
Under DARPA intervention

- MUC-4 (June-92)
  - 17 systems participated
  - Domain: terrorism in Latin America
  - New evaluation criteria, independent from the texts, were defined:
    - F-measure. A combination of the measures of precision and recall. It measures how good and how complete the system is

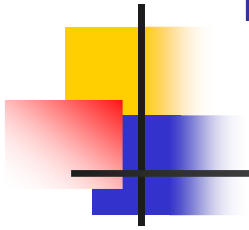
Brief history
Under DARPA intervention

- MUC-5 (August-93)
  - 17 systems participated
  - Domain: Company merging and microelectronic products advertisements
  - English and Japanese
  - A new evaluation measure is defined: error per response fill. It measures how much error the system generates:

    Non-relevant extractions+ Partial Extractions+ Relevant data that is not extracted

Brief history
Under DARPA intervention

- MUC-6 (November-95)
  - 17 systems participated
  - Domain: Company merging and microelectronic products advertisements
  - Objetive: Modularity and portability of the systems
  - Precision and recall are used again
  - Tasks definition
    - Entity recognition (ER)
    - Correference Resolution (CO)
    - Element Templates (ET)
    - Scenarios Templates (ST)



Brief history
Under DARPA intervention

- MUC-7 (April-98)
  - 17 systems participated
  - Domain: plane accidents and misile launching
  - Definition of a new task:
    - Template Relation (TR): *employee of*, *located at and product of*

Traditional features of the IES

Consequences of MUCS

- Basic process in two phases:
  - Recognition
  - Classification
- Importance of Knowledge Engineering
  - Modularity
- Use of local and weak knowledge
- NL Resources
  - Ontologies, Lexicons, Corpora …
- Heuristic / Stadistic techniques
  - Rule-based systems/ machine learning systems
- Output
  - Tagged texts / databases

Quality measures of IES

Consequences of MUCS

- Recall
  - Number of correct extracted items (Ec)
  - Number of existing items to extract in the text(Ee).

$$R = Ec / Ee$$

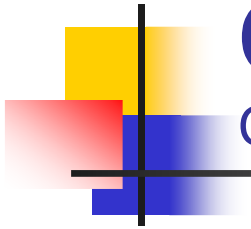
- Precision
  - Number of correct extracted items (Ec)
  - Number of extracted items (Et).

$$P = Ec / Et$$

- F-Measure

$$F = (\beta^2+1) * P * R / \beta^2*P + R$$

(β>1 more importance to R, β<1 more importance to P)



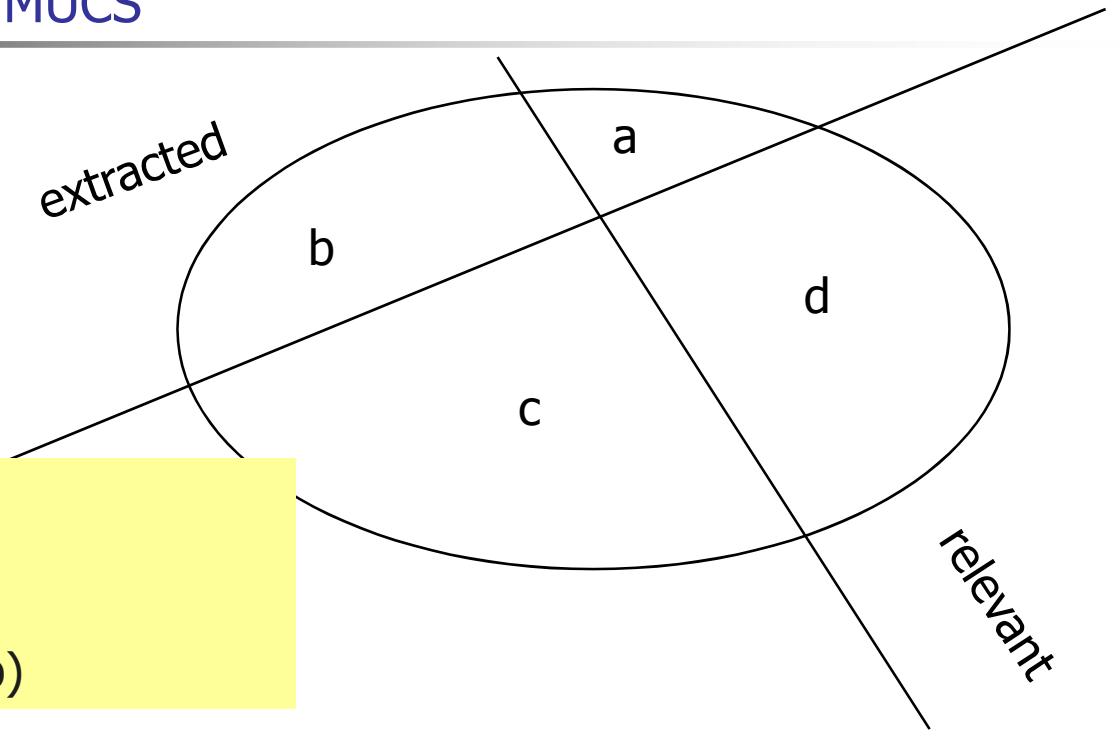
Quality measures of IES

Consequences of MUCS

- Precision and recall are independent measures
- Suppose we have a T and F test of 4 questions. If we answer only one questions correctly, we have R=25%
- Precision depends on:
  - If we have only answered one question P=100%
  - If we have answered two questions P=50%
  - If we have answered three questions P=33,33%
  - If we have answered four questions P=25%

Quality measures of IES

Consequences of MUCS



extracted = a + b
relevant = a + d
recall = a / (a + d)
precision = a / (a + b)

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

Extraction quality measures

Example

Entity person | Entity.organization | Temporal expression | Numeric expression

GOLDSTANDARD

El vicepresidente del Gobierno valenciano y conseller de Presidencia, José Císcar, ha asegurado hoy que la Generalitat saldará su deuda con todos los municipios de la Comunitat en 2013, ejercicio en el que su departamento reducirá en 4,5 millones (un 62 %) los servicios destinados a entidades locales.

Así lo ha avanzado Císcar en su comparecencia en la Comisión de Economía, Presupuestos y Hacienda de Les Corts para informar sobre las cuentas de Presidencia para 2013, que contará en 2013 con 34,3 millones de euros

Number of existing extractions (Ee) = 17
Number of correct extractions (Ec) = 13
Number of treated extractions(Et) = 15

IES

El vicepresidente del Gobierno valenciano y conseller de Presidencia, José Císcar, ha asegurado hoy que la Generalitat saldará su deuda con todos los municipios de la Comunitat en 2013, ejercicio en el que su departamento reducirá en 4,5 millones (un 62 %) los servicios destinados a entidades locales.

Así lo ha avanzado Císcar en su comparecencia en la Comisión de Economía, Presupuestos y Hacienda de Les Corts para informar sobre las cuentas de Presidencia para 2013, que contará en 2013 con 34,3 millones de euros

P = 13 / 15 = 0.8666 (86,66%)
R = 13 / 17 = 0.7647 (76,47%)
F (β=1) = 0,8124 (81,24%)

Resources and organizations

Resources and organizations

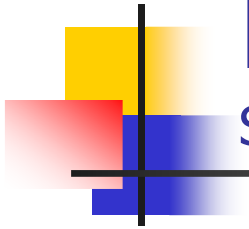
International organizations

- Consortium for Lexical Research (NMSU)
- Association for Computational Linguistics
- Oxford University Press, SRI, Oxford University, British Library and Lancaster and Cambrigde Universities that are part of a consortium to build the British National Corpus.
- Electronic Dictionary Research Project (EDR), in Japan

Resources and organizations

International organizations

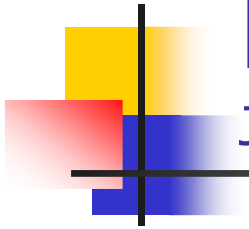
- Speech systems with DARPA
- Real Academia de la Lengua Española, RAE ("Royal Academy of the Spanish Language")
- Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), (Spanish Society for Natural Language Processing)



Resources and organizations

Specific conferences

- ACL (EACL, NAACL) - Annual Meeting of Association for Computational Linguistics.
- COLING - International Conference on Computational Linguistics
- CLNLP - Computational Logic for Natural Language Processing.
- DARPA Speech and Natural Language Workshop
- EUROSPEECH
- SEPLN – Conference of Spanish Society for Natural Language Processing
- Conference of natural and formal language
- DEXA
- TSD - Text, Speech and Dialogue



Resources and organizations

Journals

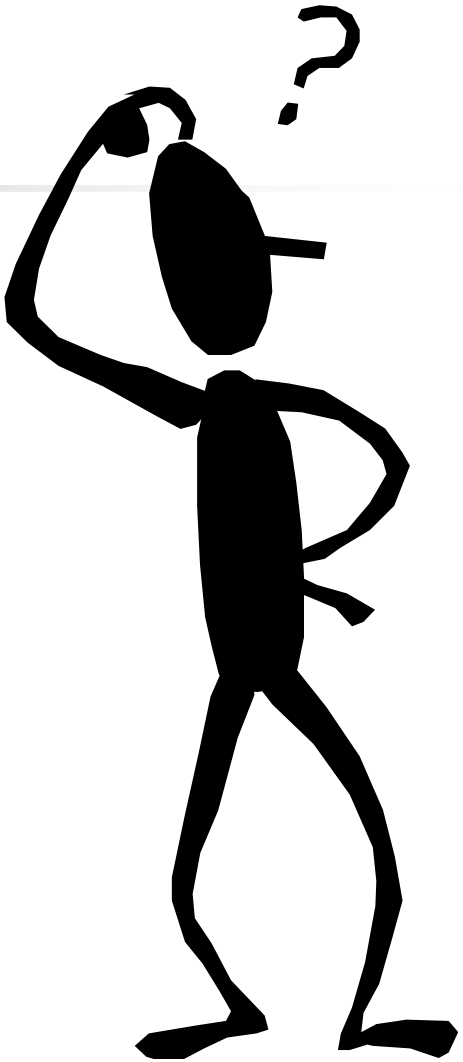
- Computational Linguistics (formerly American Journal of Computational Linguístics).
- Journal of Artificial Intelligence Research
- Artificial Intelligence
- Computing and Humanities
- Natural Language Engineering
- ACM of Communications
- Machine Translation
- Natural Language Processing
- Iberoamerican journal of Artificial Intelligence
- Novatica (Language technologies)

Resources and organizations

Links

- SEPLN- www.sepln.org
- ACL- www.aclweb.org
- COLING - www.coling.org
- CRL - crl.nmsu.edu
- CLG - www.wlv.ac.uk/sles/compling

Questions?



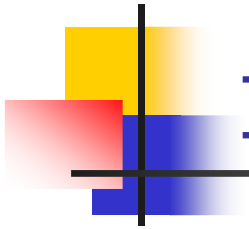
TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

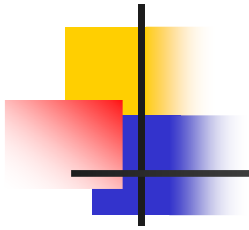
Information Extraction: Architecture of the IES

Unit 3



Index

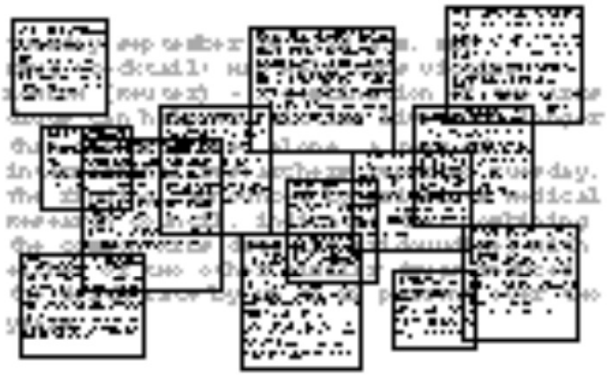
1. Introduction
2. Architecture
3. Pre-process
4. Resources

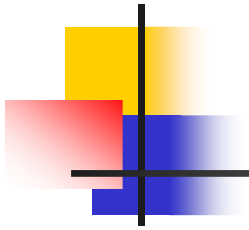


1. Introduction

- Information Extraction
  - Technique that provides certain information known as relevant from a set of relevant texts.
  - It is the task to automatically extracting a certain type of pre-specified information from texts

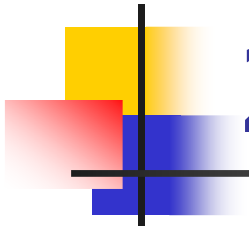
1. Introduction





1. Introduction

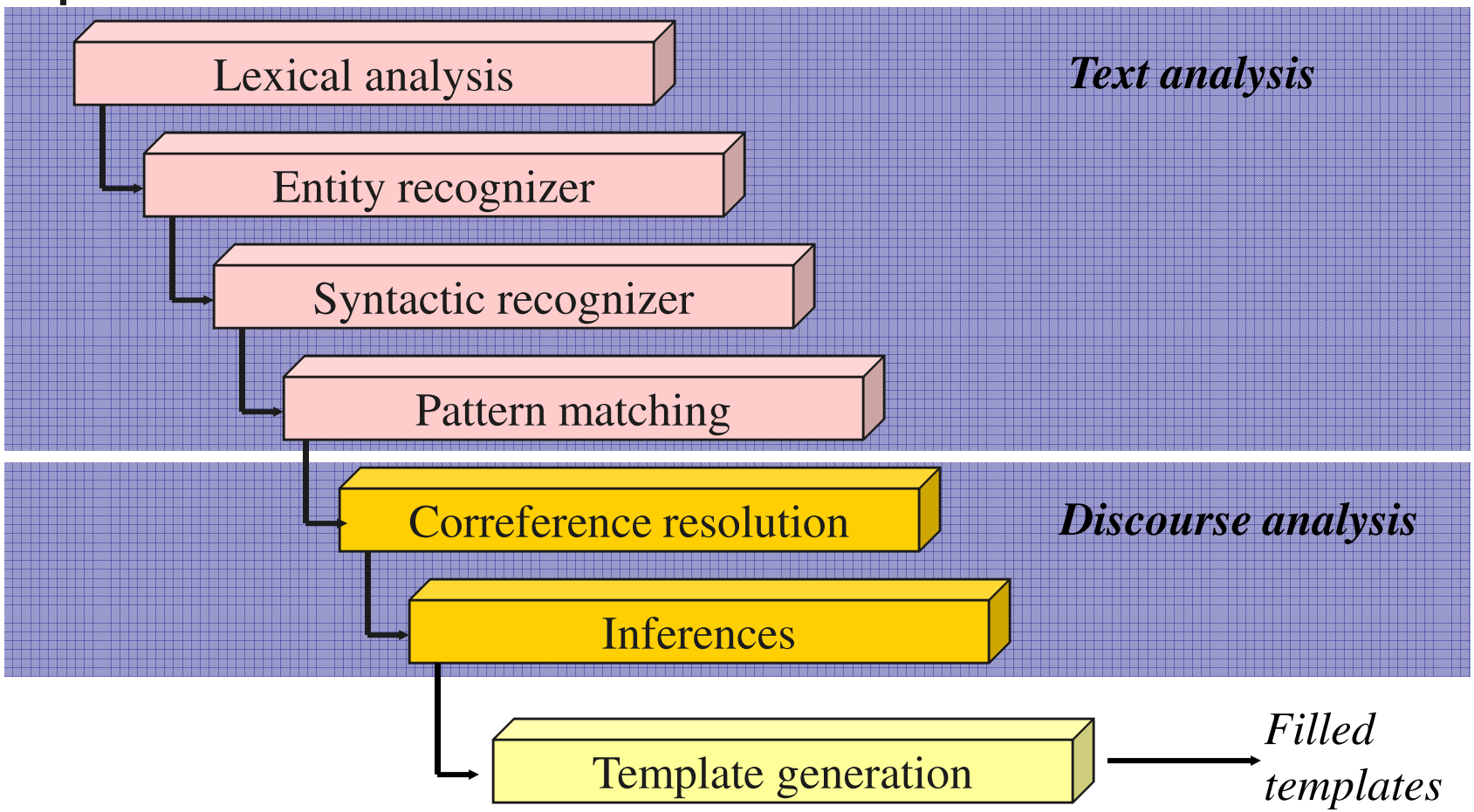
- IE.- Building systems that find and relate relevant information whereas non relevant information is ignored
- The relevant information is determined by predefined domain guides, in which the type of information to extract must be specified as exactly as possible
- From the NLP point of view, IE systems must work on different levels: from the recognition of words to the analysis of sentences and from the understanding of the sentence to the whole text



2. Architecture

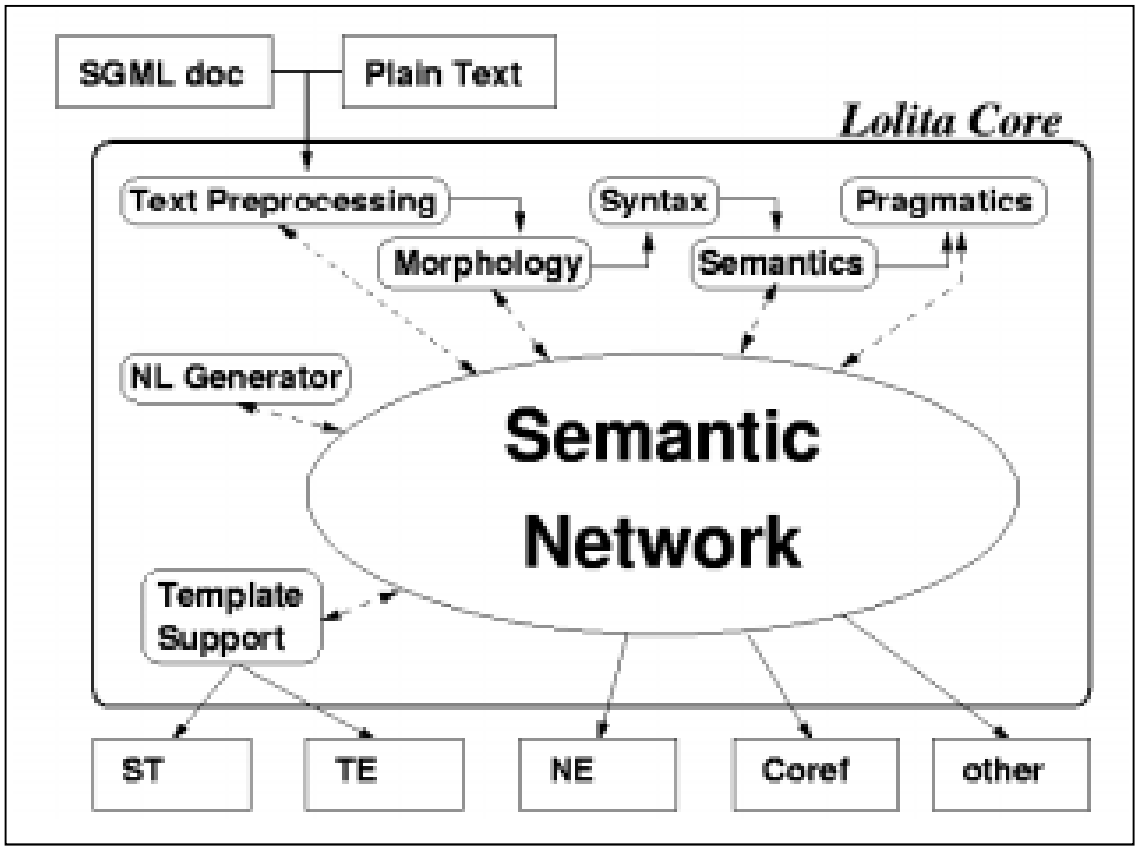
- Modular or Pipeline arch.
- Blackboard Arch.

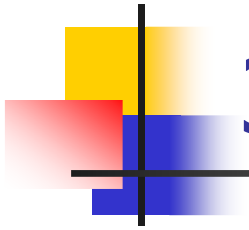
2. Modular Architecture



2. Blackboard architecture

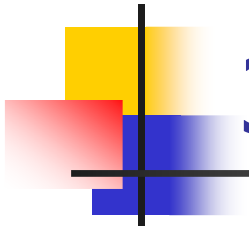
- Lolita´s system





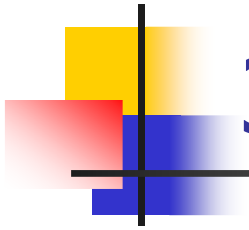
3. PRE-PROCESS

- Features of the text
  - Format
    - TXT
    - HTML
    - DOC
    - RTF
    - PDF
    - Etc.
  - Language
    - Spanish
    - English
    - French
    - Etc.



3. PRE-PROCESS

- Tasks to be performed
  - Converting to text (if original format is doc, rtf, pdf, etc.)
  - Avoiding unnecessary tags
  - Dividing the text into sentences
  - Tokenizing



3. PRE-PROCESS

- Simple tasks?
  - There are thousands of tools to convert into text
    - Xxx2txt
  - There are tools to eliminate tags or functions in programming languages to do so
  - Is there a simple problem then?

Butragueño: "La aportación de Robinho puede ser espectacular" - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://www.marca.es/edicion/noticia/0,2458,649341,00.html

Google Buscar en la Web Pop-ups del sitio permitidos Opciones Open PDF in Word

FÚTBOL
Equipos
1ª División
2ª División
2ª B, 3ª y más fútbol
Altas y bajas
Quiniela
Selección
Clas. Mundial '06
Internacional
Fútbol Sala

MADRID 2012
ALMERÍA 2005

BALONCESTO-3,05
Liga ACB
NBA
AJEDREZ
ATLETISMO
BALONMANO
CICLISMO
Tour de Francia
GOLF
MOTOR
Mundial motos
Fórmula 1
Mundial de rallies
Superbikes
Trial
NIEVE
PÁDEL
TENIS
Wimbledon

Atrás

01.07.2005 - 14:17

"TRABAJAMOS PARA FICHARLE LO ANTES POSIBLE, PERO HAY QUE ESPERAR"

Butragueño: "La aportación de Robinho puede ser espectacular"

Emilio Butragueño, vicepresidente del Real Madrid, declaró que Robinho, delantero del Santos por el que el conjunto blanco sigue negociando, ha sido la sensación del último año, aunque afirmó también que habrá que esperar a que todo concluya y considerarlo jugador del conjunto blanco.

"Es innegable que desde el club se está negociando con la intención de incorporarlo lo antes posible, pero de momento hay que esperar. Ha sido la sensación del último año. Nosotros debemos esperar y desear que el desenlace sea el que deseamos. Aunque es muy joven, su aportación al equipo puede ser espectacular. Luxemburgo tardó segundos en dar su opinión. Se encuentra encantado con su llegada", declaró a Realmadrid.com y Realmadrid TV.

Butragueño destacó la calidad de los dos últimos fichajes del equipo, los uruguayos Pablo García y Carlos Diogo. "Pablo García es un especialista que nos va a ayudar mucho con su calidad y experiencia y sentido táctico, que le va a venir muy bien al equipo. Diogo en un joven jugador con una gran proyección maravillosa, física y técnicamente es muy poderoso y puede jugar den varias posiciones".

Plantilla larga

"Nuestro objetivo es confeccionar una plantilla para ganar títulos para el año que viene y que duda cabe que estos dos jugadores nos van a ayudar mucho en esa faceta", añadió.

El dirigente madridista también se refirió a la llegada como entrenador de porteros de Pedro Jaro, ex guardameta del conjunto merengue.

"Estamos muy contentos porque es sin duda un buen refuerzo para el cuadro técnico. Pedro que fue jugador del Real Madrid, ya estuvo como técnico en las categorías inferiores y ahora ha sido contratado para entrenar a los porteros del primer equipo y además, nos ayudará en las categorías

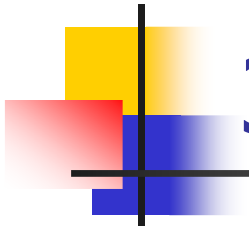
TODAS LAS NOTICIAS
RESULTADOS
LA PLANTILLA
EL CLUB
CALENDARIO
ESTADISTICAS
FORO

ARCHIVO
Todos los especiales

3. PRE-PROCESS

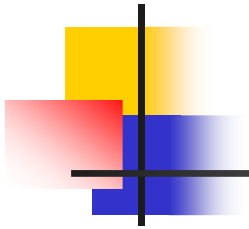
- Converting to text (if is doc, rtf, pdf, etc.)

```
<!-- Vignette StoryServer 5.0 Fri Jul 01 14:24:35 2005 -->
<!-- e-Scriba. Proyectos daVinci  -->
<!-- a -->



<META HTTP-EQUIV="Expires" CONTENT="Tue, 20 Jun 1995 04:13:09 GMT">


                                                <!-- Vignette StoryServer 5.0 Fri Jul 01 14:24:35 2005 -->
                <html>
<head> <!-- ____________________________________________________________________________ -->
<TITLE>Butragueño: "La aportación de Robinho puede ser espectacular"</TITLE> <!-- Vignette StoryServer 5.0 Fri Jul 01 14:24:36 2005 -->

 <!-- Vignette StoryServer 5.0 Sun Jun 19 01:54:09 2005 -->
 <style>
.Arial21az {font-family: Arial, Helvetica, sans-serif; font-size: 21px; font-weight: bold; text-decoration: none;color: #003C78}
.Arial14az {font-family: Arial, Helvetica, sans-serif; font-size: 14px; font-weight: bold; text-decoration: none;color: #003C78}
.Arial11az {font-family: Arial, Helvetica, sans-serif; font-size: 11px; text-decoration: none;color: #003C78}
.Arial11 {font-family: Arial, Helvetica, sans-serif; font-size: 11px; color: #000000}
.Arial21 {font-family: Arial, Helvetica, sans-serif; font-size: 21px; font-weight: bold; text-decoration: none;color: #000000}
.a12n {font-family: Arial, Helvetica, sans-serif; font-size: 12px; color: #000000}
.a11az {font-family: Arial, Helvetica, sans-serif; font-size: 11px; text-decoration: none;color: #003C78}
.bold {font-weight: bold}
.tr11azcl {font-family: "Trebuchet MS",Arial, Helvetica, sans-serif; font-size: 11px; text-decoration: none;color: #9EC9F0}
.a10n {font-family: Arial, Helvetica, sans-serif; font-size: 10px; color: #000000}</style>


                </head><!-- ____________________________________________________________________________ -->
<body bgcolor="#FFFFFF" leftmargin="0" topmargin="0" marginwidth="0" marginheight="0">


<table  width=750  CELLSPACING=0 CELLPADDING=0 border=0><!-- INiCIO TABLA ARRIBA -->
<tr><td>
                <!-- Vignette StoryServer 5.0 Tue Jun 28 01:42:18 2005 -->

                <!-- Vignette StoryServer 5.0 Tue Jun 28 01:35:01 2005 -->


                                                <link rel="stylesheet" href="/estilos/marca.css" type="text/css">
<SCRIPT LANGUAGE="JavaScript"> <!--
_version=10; //--> </script>
<SCRIPT LANGUAGE="JavaScript1.1"> <!--
_version=11;
if (navigator.userAgent.indexOf('Mozilla/3') != -1){
_version=10;} // --> </script>
<SCRIPT LANGUAGE="JavaScript"> <!--
var server = "http://ads.recoletos.es";
var sitepage = "marca.recoletos.es/rma_fut";
var position ="Bottom,Middle1,Middle2,Middle3,Top,Top1,x02!Top1";

if (! (RN)) {
var RN = new String (Math.random());
var RNS = RN.substring (2, 11);
}
var oas=server + '/RealMedia/ads/';
var oaspage= sitepage + '/1' + RNS + '@' + position;

if (_version < 11) {
  document.write ('<A HREF="' + oas + 'click_nx.ads/'+ oaspage + '" TARGET="_top" ><IMG SRC="' + oas + 'adstream_nx.ads/' + oaspage + '" BORDER=0 WIDTH=468 HEIGHT=60 ALT="Click!"></a>');

} else {
  document.write ('<SCR'+'IPT LANGUAGE="JavaScript1.1" SRC="' + oas + 'adstream_jx.ads/' + oaspage + '">');
  document.write ('\<!-- --\>');
  document.write ('\<\/SCRIPT\>');
  document.write ('\<!-- --\>');
}
// -->
</SCRIPT>
<table width="750" border="0" cellspacing="0" cellpadding="0">
  <tr>
    <td><img src="/imagenes/home/sup_interior_vodafone.gif" width="750" height="14"></td>
  </tr>
</table>
<table width=750 border=0 cellspacing=0 cellpadding=0>
  <tr>
    <td width=147 rowspan=2 align=center><img src="/img/home/mancheta.gif" width=122 height=144 vspace=8 usemap="#Map" border="0"><br>
    </td>
    <td height="76" align=center valign=top>
<div id="Layer1" style="position:relative; width:468px; height:60px; z-index:1">
  <div align="center">
    <SCRIPT LANGUAGE="JavaScript"> <!--
_version=10; //--> </script>
<SCRIPT LANGUAGE="JavaScript1.1"> <!--
_version=11;
if (navigator.userAgent.indexOf('Mozilla/3') != -1){
_version=10;} // --> </script>
<SCRIPT LANGUAGE="JavaScript"> <!--
var server = "http://ads.recoletos.es";
var sitepage = "marca.recoletos.es/rma_fut";
var position ="Bottom,Middle1,Middle2,Middle3,Top,Top1,x02!Top";

if (! (RN)) {
var RN = new String (Math.random());
var RNS = RN.substring (2, 11);
}
var oas=server + '/RealMedia/ads/';
```



3. PRE-PROCESS

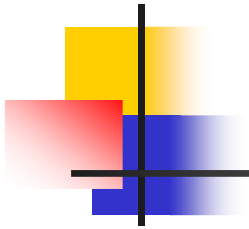
- PROBLEM: Is all the text necessary?
- There are parts of the document that are not part of the information itself
- Then, which is the valid information?
- Determining this is the problem



3. PRE-PROCESS

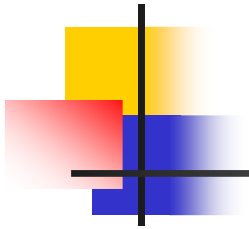
- Avoiding language directives

```
<SCRIPT LANGUAGE="JavaScript"> <!--
var server = "http://ads.recoletos.es";
var sitepage = "marca.recoletos.es/rma_fut";
var position ="Bottom,Middle1,Middle2,Middle3,Top,Top1,x02!Top1";

if (! (RN)) {
var RN = new String (Math.random());
var RNS = RN.substring (2, 11);
}
var oas=server + '/RealMedia/ads/';
var oaspage= sitepage + '/1' + RNS + '@' + position;

// -->
</SCRIPT>
```



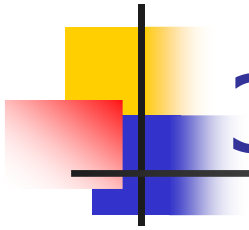
3. PRE-PROCESS

- Which is the problem when eliminating language directives?



3. PRE-PROCESS

- Which is the problem when eliminating language directives?
  - Dynamic pages



3. PRE-PROCESS

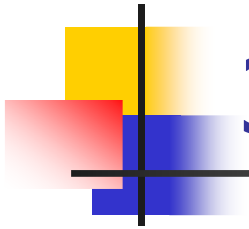
- Eliminating comments

```
<!-- Vignette StoryServer 5.0 Fri Jul 01 14:24:35 2005 -->
<!-- e-Scriba. Proyectos daVinci  -->
<!-- a -->
// -->
```

3. PRE-PROCESS

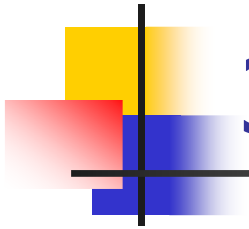
- Eliminating unnecessary tags

```
<html>
<META HTTP-EQUIV="Expires" CONTENT="Tue, 20 Jun 1995 04:13:09 GMT">
<head> <!--
    ____________________________________________________________________
    -->
<TITLE>Butragueño: "La aportación de Robinho puede ser espectacular"</TITLE>
<style>
.Arial21az {font-family: Arial, Helvetica, sans-serif; font-size: 21px; font-weight: bold; text-
    decoration: none;color: #003C78}
.Arial14az {font-family: Arial, Helvetica, sans-serif; font-size: 14px; font-weight: bold; text-
    decoration: none;color: #003C78}
.Arial11az {font-family: Arial, Helvetica, sans-serif; font-size: 11px; text-decoration: none;color:
    #003C78}
.a10n {font-family: Arial, Helvetica, sans-serif; font-size: 10px; color: #000000}</style>
</head>
```



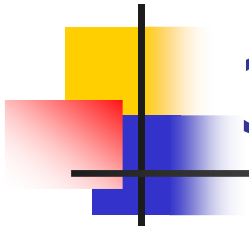
3. PRE-PROCESS

- Eliminating unnecessary tags
- (II)
  - <body>"TRABAJAMOS PARA FICHARLE LO ANTES POSIBLE, PERO HAY QUE ESPERAR"
    <h2>**Butragueño: "La aportación de Robinho puede ser espectacular"**<h2>
    Emilio Butragueño, vicepresidente del Real Madrid, declaró que Robinho, delantero del Santos por el que el conjunto blanco sigue negociando, ha sido la sensación del último año, aunque afirmó también que habrá que esperar a que todo concluya y considerarlo jugador del conjunto blanco.
    "Es innegable que desde el club se está negociando con la intención de incorporarlo lo antes posible, pero de momento hay que esperar. Ha sido la sensación del último año. Nosotros debemos esperar y desear que el desenlace sea el que deseamos. Aunque es muy joven, su aportación al equipo puede ser espectacular. Luxemburgo tardó segundos en dar su opinión. Se encuentra encantado con su llegada", declaró a Realmadrid.com y Realmadrid TV.
    Butragueño destacó la calidad de los dos últimos fichajes del equipo, los uruguayos Pablo García y Carlos Diogo. "Pablo García es un especialista que nos va a ayudar mucho con su calidad y experiencia y sentido táctico, que le va a venir muy bien al equipo. Diogo en un joven jugador con una gran proyección maravillosa, física y técnicamente es muy poderoso y puede jugar den varias posiciones".
</body>



3. PRE-PROCESS

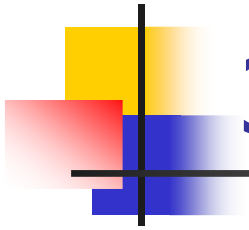
- Dividing into sentences
  - TRABAJAMOS PARA FICHARLE LO ANTES POSIBLE, PERO HAY QUE ESPERAR"
  - **Butragueño: "La aportación de Robinho puede ser espectacular"**
  - Emilio Butragueño, vicepresidente del Real Madrid, declaró que Robinho, delantero del Santos por el que el conjunto blanco sigue negociando, ha sido la sensación del último año, aunque afirmó también que habrá que esperar a que todo concluya y considerarlo jugador del conjunto blanco.
  - "Es innegable que desde el club se está negociando con la intención de incorporarlo lo antes posible, pero de momento hay que esperar.
  - Ha sido la sensación del último año.
  - Nosotros debemos esperar y desear que el desenlace sea el que deseamos.
  - Aunque es muy joven, su aportación al equipo puede ser espectacular.
  - Luxemburgo tardó segundos en dar su opinión.
  - Se encuentra encantado con su llegada", declaró a Realmadrid.com y Realmadrid TV.
  - Butragueño destacó la calidad de los dos últimos fichajes del equipo, los uruguayos Pablo García y Carlos Diogo.



3. PRE-PROCESS

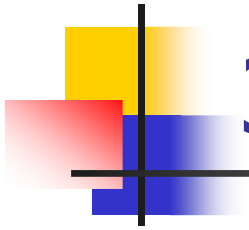
- Exercise: Divide the text into sentences

A shark has bitten Friday an Austrian tourist who was swimming in the Gulf of Mexico. It is the third attack of the same type registered in Florida in a week, as he has informed the FBI The injured man, 19, was taken to a hospital for surgery. A. Trojer, from Baden, Austria, was airlifted to a hospital in Fort Myers, where he is out of danger. It is confirmed that it was a shark attack, has indicated the police spokesman, Mr. LiMarzi. The spokesman added that the injured have bites over 31.5 cm. without disclosing further details. The beach where the accident happened remains open and no other sharks were spotted from helicopters that have flown over the area.



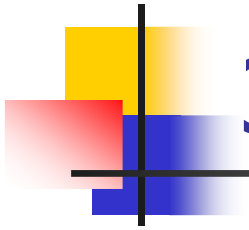
3. PRE-PROCESS

- Exercise: Dividing the text into sentences
  - How do you perform it?
  - Which rules have you used?



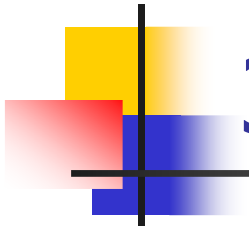
3. PRE-PROCESS

- Dividing the text into sentences
  - Simple task?
  - Look for “.”, “?”, “!” and then Capital letters
  - Sometimes “)”, “—” and then Capital letters



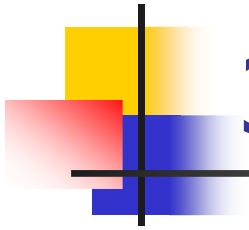
3. PRE-PROCESS

- PROBLEM: Dividing text into sentences
  - Decimals 31.5
  - Abbreviations cm., F.B.I.
  - Initials in names A. Trojer
  - Titles Mr. LiMarzi
  - E-mail addresses rafael@dlsi.ua.es
  - Webpages www.dlsi.ua.es



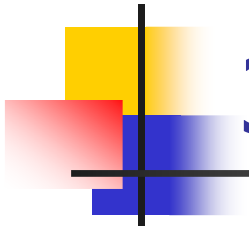
3. PRE-PROCESS

- Algorithms to divide sentences
  - Based on rules
  - Machine Learning



3. PRE-PROCESS

- Algorithms to divide sentences based on rules
  - Regular expressions
    - [Cutting1991]
    - Mark Wasson convert a grammar in a finite automat with 1419 states and 18002 transactions.
  - Ending words
    - [Müller1980] used a big list of words.



3. PRE-PROCESS

- Algorithms to divide sentences based on Machine Learning
  - [Riley1989] using regression trees
  - [Palmer and Hearst1997] using decision trees or neuronal networks

3. PRE-PROCESS

- Output after dividing sentences

<s num=1>"TRABAJAMOS PARA FICHARLE LO ANTES POSIBLE, PERO HAY QUE ESPERAR"</s>

<s num=1>**Butragueño: "La aportación de Robinho puede ser espectacular"**<s>

<s num=1>Emilio Butragueño, vicepresidente del Real Madrid, declaró que Robinho, delantero del Santos por el que el conjunto blanco sigue negociando, ha sido la sensación del último año, aunque afirmó también que habrá que esperar a que todo concluya y considerarlo jugador del conjunto blanco.</s>

<s num=1>"Es innegable que desde el club se está negociando con la intención de incorporarlo lo antes posible, pero de momento hay que esperar. </s>
<s num=2>Ha sido la sensación del último año. </s>
<s num=3>Nosotros debemos esperar y desear que el desenlace sea el que deseamos. </s>
<s num=4>Aunque es muy joven, su aportación al equipo puede ser espectacular. </s>
<s num=5>Luxemburgo tardó segundos en dar su opinión.</s>
<s num=6>Se encuentra encantado con su llegada", declaró a Realmadrid.com y Realmadrid TV. </s>

3. PRE-PROCESS

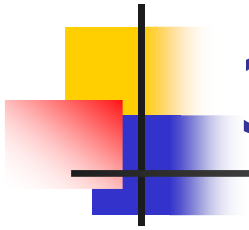
- Output after dividing sentences (II)

<s num=1>"TRABAJAMOS PARA FICHARLE LO ANTES POSIBLE, PERO HAY QUE ESPERAR"</s>

<s num=2>**Butragueño: "La aportación de Robinho puede ser espectacular"**<s>

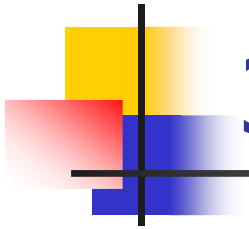
<s num=3>Emilio Butragueño, vicepresidente del Real Madrid, declaró que Robinho, delantero del Santos por el que el conjunto blanco sigue negociando, ha sido la sensación del último año, aunque afirmó también que habrá que esperar a que todo concluya y considerarlo jugador del conjunto blanco.</s>

<s num=4>"Es innegable que desde el club se está negociando con la intención de incorporarlo lo antes posible, pero de momento hay que esperar. </s>
<s num=5>Ha sido la sensación del último año. </s>
<s num=6>Nosotros debemos esperar y desear que el desenlace sea el que deseamos. </s>
<s num=7>Aunque es muy joven, su aportación al equipo puede ser espectacular. </s>
<s num=8>Luxemburgo tardó segundos en dar su opinión.</s>
<s num=9>Se encuentra encantado con su llegada", declaró a Realmadrid.com y Realmadrid TV. </s>



3. PRE-PROCESS

- Tokenizing
  - Dividing the text in the tokens that compose it
  - A token is the minimum unit of information

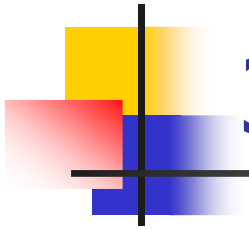


3. PRE-PROCESS

- Tokenizing

  WE
  WORK
  TO
  DO
  IT
  IN
  SPITE
  OF
  THE
  PROBLEMS
  …

Is there any problem?



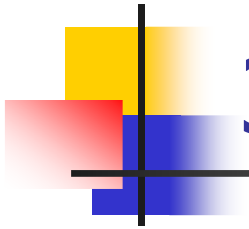
3. PRE-PROCESS

- Tokenizing

  WE
  WORK
  TO
  DO
  IT

  IN SPITE OF

  PROBLEMS
  ,



3. PRE-PROCESS

- POS tagger is in charge of
  - Assigning morphological categories
  - Tokenizing
  - Dividing into sentences

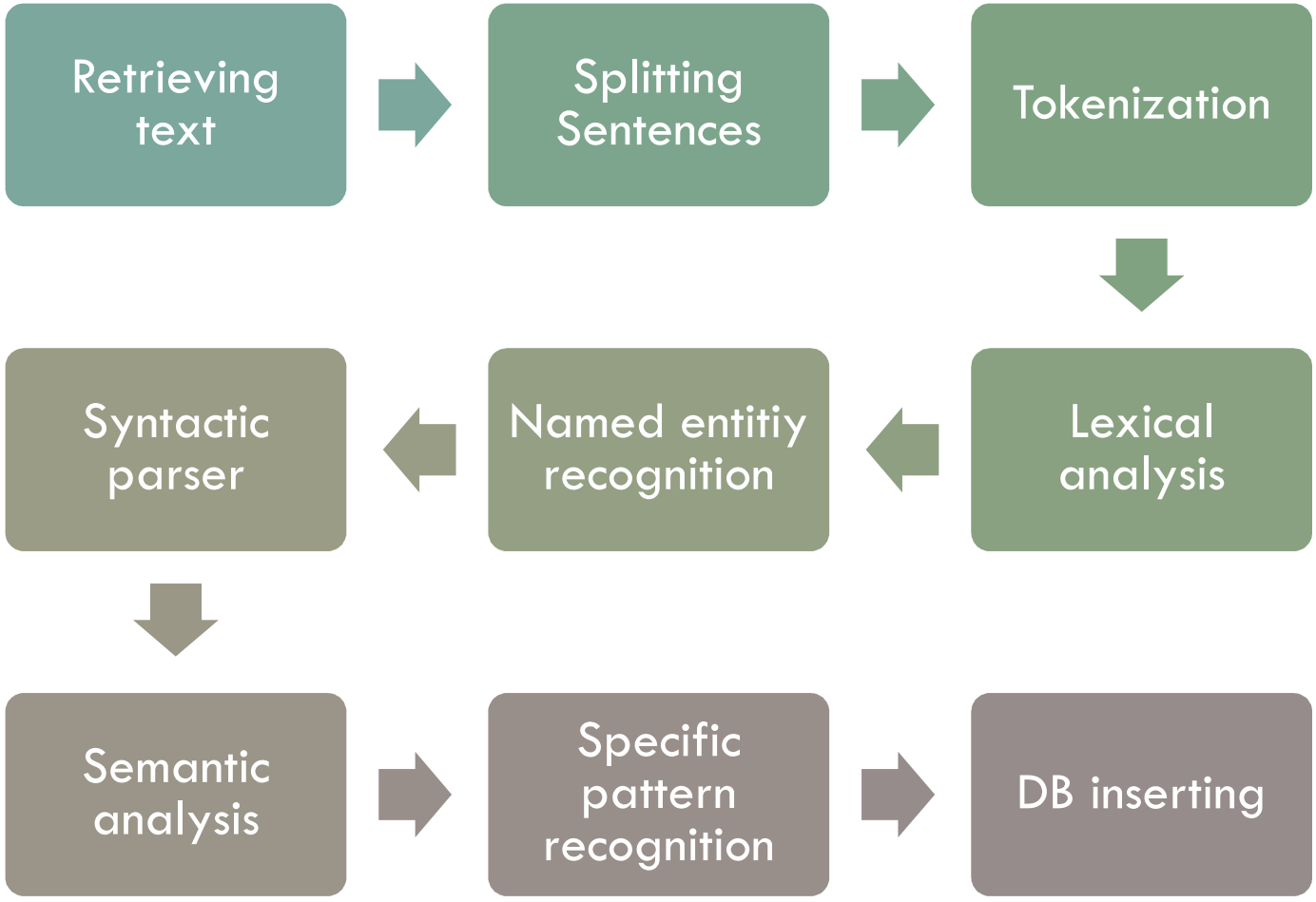
Questions

TEXTUAL INFORMATION EXTRACTION

Estela Saquete Boró
Departamento de Lenguajes y Sistemas Informáticos

Practical work

The process of textual information extraction



Jet (Java Extraction Toolkit)

- Jet (Java Extraction Toolkit)
  - Tool designed for the analysis of natural language, and especially for information extraction.
  - Jet can be downloaded from the web page of his author, Ralph Grishman (http://cs.nyu.edu/cs/faculty/grishman/jet/license.html) for educational use.

Jet. Installing

- Jet works over Java 1.4 or later programming environment.
- Check it with "java –version"
- Go to http://www.java.com/es/download/ and download lastest version.
- Download the package Jet.130.zip from the web and unzip it in a folder called Jet.
- Manual in Campus Virtual and on-line in http://cs.nyu.edu/cs/faculty/grishman/jet/doc/Jet.html

Contents of "Jet" folder

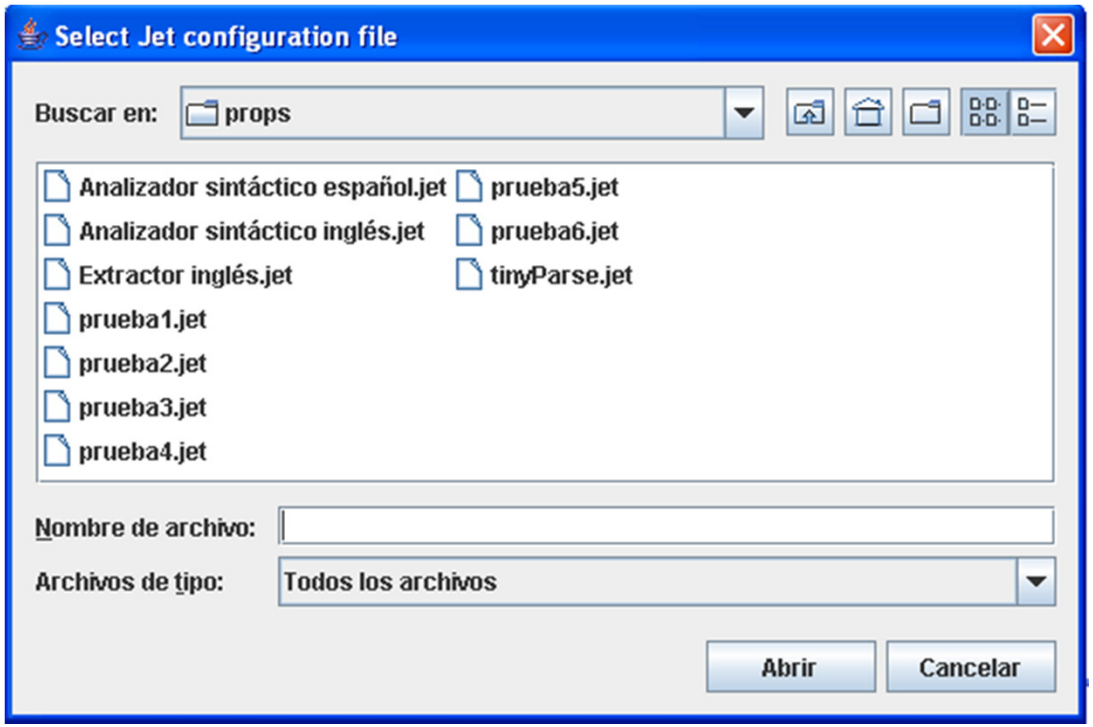
- jet.jar – the executable program of Jet
- lib – folder with the executables used by Jet
- doc – folder with the documentation files
- props – folder with the configuration files
- data – folder with the data files
- win – folder with the scripts for Windows execution
- unix – folder with the scripts for Linux (Unix) execution

Adding test files

- Download the following files from the Virtual Campus:
  - Props.rar
  - Data.rar
- Add them to the "props" and "data" folders, respectively

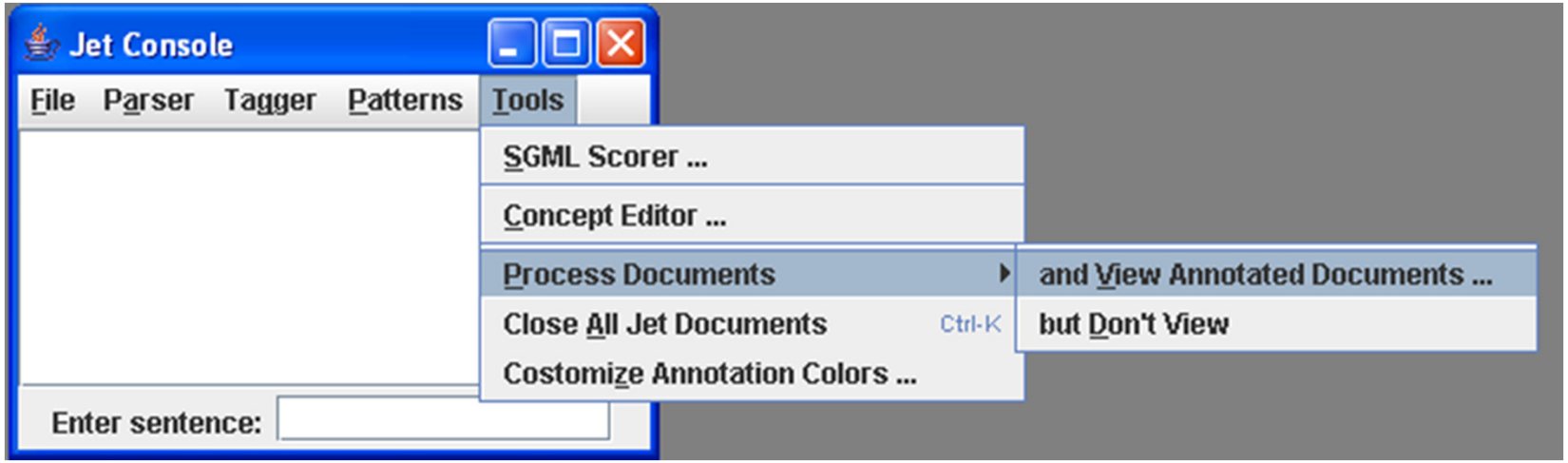
Executing Jet

- Execute Jet.jar and choose the configuration file
  - Each of them has a sequence of operations to execute



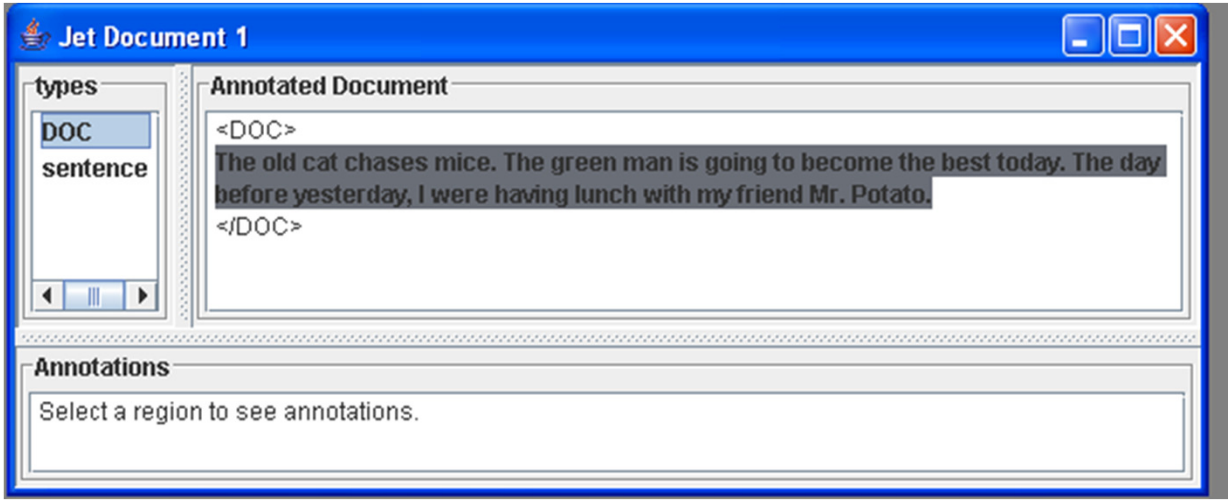
Jet console

- Ways of working:
  - Introduce a sentence and directly process it (in the cell "Enter sentence"),
  - Process complete documents.



Document processing

- The document must be between the XML tags: <DOC> … </DOC>
- The name of the document must be included in the configuration file
- The result window shows the processed text



Configuration file *.jet

- Jet\props*.jet
- Contents:
  - Data folder
  - Names of files
    - Grammar for parsing
    - Lexicon for lexical-morphological analysis
    - HMM model for POS tagger
    - HMM for NAME tagger
    - Concepts hierarchy
    - Patterns
    - Text to be processed
  - Execution sequence
  - Execution patterns

*.jet: folders and files

```
# JET properties file (class -- asgn #2)
# data folder:
Jet.dataPath        = data

# grammar for parsing
Grammar.fileName     = myGrammar.txt

# lexicon
EnglishLex.fileName1 = Jet4.dict

# HMM model for POS tagger
Tags.fileName = pos_hmm.txt

# HMM model for name tagger
NameTags.fileName = MUCnameHMM.txt

# file with concepts hierarchy
Concepts.fileName = myHierarchy.hrc

# files of patterns
Pattern.fileName = myPatterns.txt

# file to process
JetTest.fileName = texto_prueba.txt
```

- Grammar for parsing
- Lexicon for lexical-morphological analysis
- HMM model for POS tagger
- HMM for NAME tagger
- Concepts hierarchy
- Patterns
- Text to be processed

*.jet: execution sequence

```
#execution per document
processDocument = tag(TEXT), TEXT:processTextZone

#execution per document
processTextZone = sentenceSplit, sentence:processSentence

#execution per sentence:
processSentence = tokenize, lexLookup, pruneTags, tagNames, pat(vuelta1), pat(vuelta2)

# tags to be shown in the response file-
WriteSGML.type = PATTERN

#to execute in batch mode
#Jet.batch = 1
```

Jet Commads

- tag(tag): it recognizes fragments of tagged text between the tags <tag> … </tag> (pay attention to capital letters)
- tag:name_sequence : it executes the operations sequence for each recognized fragment
- sentenceSplit it adds <sentence>
- tokenize it adds <token>
- lexLookUp (Lexical searching) it adds <constit> for each entry in the lexicon. It includes lexical and morphological info
- pruneTags (POS Tagger) it leaves a unique <constit> according to the HMM model. It adds <tagger> to include the same in PennTreeBank format
- tagNames (Named Entity Tagger) it adds <ENAMEX> <TIMEX> <NUMEX> to NE, TE and numeric expressions respectively.
- pat(set_patterns) It recognizes patterns and executes the corresponding orders

Building patterns in Jet

- Namepattern := <pattern> it describes the pattern
- **pattern set** nameset it defines a set of patterns to detect in a single execution
- **when** namepattern <execution> it defines what to do with the pattern when found
  - **write** "message" it shows a message on the screen
  - **add** [tag arg1="xx" arg2="yy" …] it adds the tag and its arguments to the recognized text

Examples of patterns in Jet (myPatterns.txt)

- It looks for an specific token
  - patdet:= "the";
- It looks for a token in capital letters
  - patmays:=[token case=cap];
- It looks for a token whose core is “man”
  - pathombre:=[constit pa=[head=man]];
- It looks for tokens tagged as verbs by the POS tagger
  - patverbo := [tagger cat="VBZ"]; (PennTreeBank categories)
  - patverbo2:=[constit cat=v]; (Jet categories)
- It looks for a token that belongs to the semantic category “human_being”
  - patser:= [constit pa=[head ?isa(human_being)]];
- It looks for consecutive tokens: adjective and verb according to the POS tagger
  - patmultiple:=[constit cat=adj] [constit cat=n];

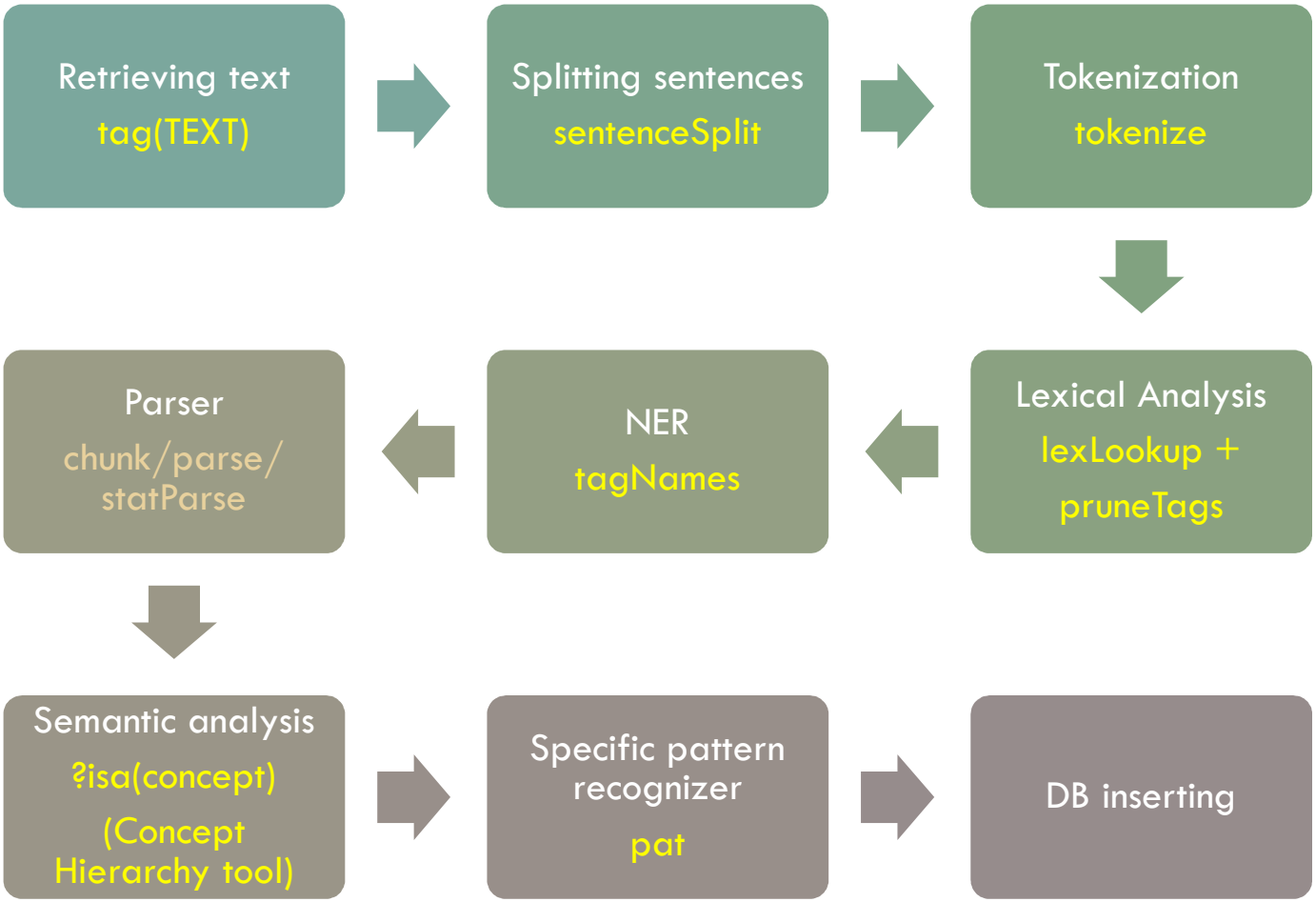
Examples of patterns in Jet (myPatterns.txt)

- It looks for 0 or n tokens tagged as “PATRON”
  - patbus:=[PATRON]*;
- It looks for 1 or n tokens tagged as “PATRON” with the argument“tipo=multiple”
  - patmarca:=[PATRON tipo=multiple]+;
- It looks for a token tagged as location by the Names tagger
  - patlugar:=[ENAMEX TYPE="LOCATION"];
- It looks for a token tagged as person by the Names tagger
  - patpersona:=[ENAMEX TYPE="PERSON"];
- It looks for a “no” after any token and stored with token in the variable “Negado”
  - patnegado:="no" ([token]): Negado;
- It looks for the relationship patron+verb+patron with any type of sequence between them
  - patrelacion:=[PATRON] [token]* [PATRON tipo=verbo] [token]* [PATRON];
- It looks for a relationship between people and locations;
  - patvive:=[person] [token]* [PATRON tipo=verb] [token]* [localtion];

Example of executions of patterns in Jet (myPatterns.txt)

```
pattern set round1;
#take care with the order of execution;
    when patverbo write “Verb found", add [PATRON tipo=verb];
    when patser write “ human_being concept found", add [PATRON tipo=being];
    when patlugar write “location pattern found", add [PATRON tipo=location], add
        [location];
    when patpersona write “Person pattern found", add [PATRON tipo=person], add
        [person];
    when patmays write “A capital letter word found”,add [PATRON tipo=capital];
    when patnegado write “tagging only the content of variable Negado", add
        [NEGADO] over Negado;

pattern set round2;
    when patmarca write “finding pattern in second turn", add [PATRON tipo=turn2];
    when patvive write “finding an special relationship", add [PATRON tipo=lives];
    when patrelacion write “findind a relationship", add [PATRON tipo=relationship];
```

Reviewing the process of textual information extraction with Jet



Textual information extraction

Practical work on temporal information

As previously explained, the first schema used for temporal annotation was TIDES. Nowadays, the systems are migrating to TIMEML because it is more expressive. Taking into account only the recognition and resolution of temporal expressions in both schemas and using the guidelines as a supporting material (http://www.timeml.org/site/index.html) and TIDES (http://www.cs.brandeis.edu/~im5/papers/MTRAnnotationGuide_v1_02.pdf) and the demo web application: http://gplsi.dlsi.ua.es/demos/T2T3 , you have to:

1) Describe the main similarities and differences between both annotation schemas

2) Determine the transformation rules between TIMEX2 and TIMEX3 regarding the extension of the expressions and attributes. As a final result we will obtain two sets of rules: a) Rules for the adaptation of the extension of the recognized expression, and b) Rules for the adaptation of the normalization attributes. Use these examples as a supporting material:

   a) Significant examples of the adaptation rules applied to the extension of the expression:

      i) The different groups will meet at 11 a.m. on Jan. 3, 2005

      ii) The bug will get fixed between now and Monday morning

      iii) I'm leaving on vacation two weeks from Tuesday

      iv) The firefighters came home five days after the fire

   b) Significant examples of the adaptation rules applied to normalization attributes:

      i) The id, type, val, anchorTimeId and mod attributes do not need examples to specify their rules, make a proposal if you consider it necessary.

      ii) For temporalFunction compare between:

         (1) Eleven in the morning

         (2) January, 31

         (3) Last year

         (4) Twelve o'clock January, 31, 2005

         (5) Summer of 1956

      iii) For beginPoint and endpoint attributes:

         (1) Six months until March 31, 2009

(2) Six months from March 31, 2009
(3) Three hours today

iv) For quant and freq attributes:
(1) Every October
(2) Annually
(3) Twice a week
(4) Three days each week

Textual information extraction

Practical work

Description of the problem

In the website http://www.usnewslink.com/uscasualties2007.htm there is a record of the victims of the military US operations during the year 2007, published by the Defense Department of this country. Here is an example:

> **12/31/2007 DoD Identifies Navy Casualty**
> The Department of Defense announced today the death of a sailor who was supporting Operation Iraqi Freedom.
> **Petty Officer 1st Class Victor W. Jeffries, 52, of Honolulu, Hawaii,** died Dec. 31, 2007 as a result of injuries suffered Dec. 24 in a vehicular accident in Kuwait. He was permanently assigned to the Navy Expeditionary Logistics Support Group, Kuwait.

From each line of this text, the following relevant data could be extracted:

1. Date of the announcement: "**12/31/2007**"
2. Name of the dead soldier: "**Victor W. Jeffries**"
3. Military squad: "sailor"
4. Military range: "**Petty Officer 1st Class**"
5. Age: "**52**"
6. Origins: "**Honolulu, Hawaii**"
7. Assigned military operation: "Operation Iraqi Freedom."
8. Date of decease: "Dec. 31, 2007"
9. Cause of decease: "injuries suffered Dec. 24 in a vehicular accident in Kuwait"
10. Permanent military unit: "Navy Expeditionary Logistics Support Group"
11. Permanent destination: "Kuwait"

Proposal technique

Build an Information extraction system based on JET to tag the text with XML tags <PATTERN Type=XXXX> trying to recognize each one of the 11 relevant types of data detected. The XML annotation will be used later to populate the DB.

Methodology

With the final aim of building a universal recognition model, it is important to ensure that the process to build the recognition model (training phase) is not affected by the concrete observed data. Due to this fact, the announcement corpus will be divided into two parts:

a) Training corpus: a subset formed by at least 30 announcements that will be taken from the data from December. This corpus will be used during the whole training phase to adjust the features of the recognition model. The errors made can be observed to improve the system.

b) Test corpus: We will reserve a subset of at least 20 announcements from the January data. This corpus will be used in the test phase and cannot be observed until the training phase is finished.

When evaluation/testing is finished we will calculate an estimation of PRECISION and RECALL.

PRECISION= No. correct data / No. extracted data

RECALL = No. correct data / No- existing data

Deliverables

You must deliver the following documents:

a) the .txt file with the definition of the built patterns.
b) the .txt with the training corpus.
c) the .txt with the test corpus.
d) the .jet file with the definition of the commands to execute the system.
e) The .hrc file with the ontology (if defined).
f) Report with the precision and recall results obtained

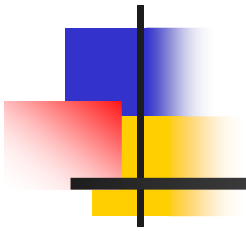
Delivery date: 15th December 2014

TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

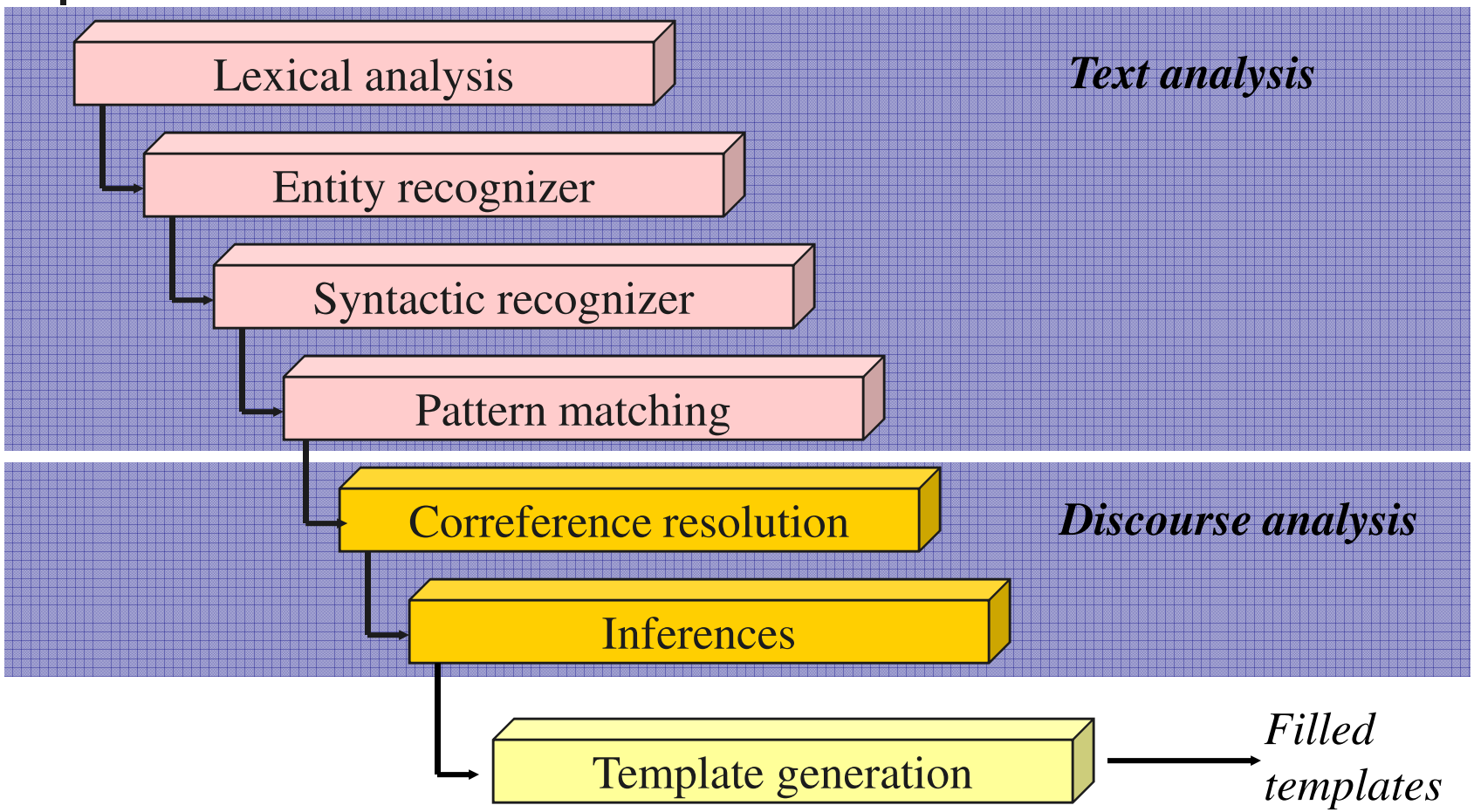
Universitat d'Alacant
Universidad de Alicante

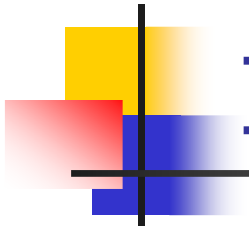


Information Extraction: Inference module

Unit 7 (b)

1. Introduction





Inference module

- Through inference rules information that appear implicit in the text is extracted

Inference module

- Through inference rules information that appear implicit in the text is extracted

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM :
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM :
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach +
Team.coach= Person1

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH : <Person-0002>

<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach +
Team.coach= Person1 +
Person2.team=<<NULL>>

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH : <Person-0002>

<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM :
GOALS:

<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Questions

TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

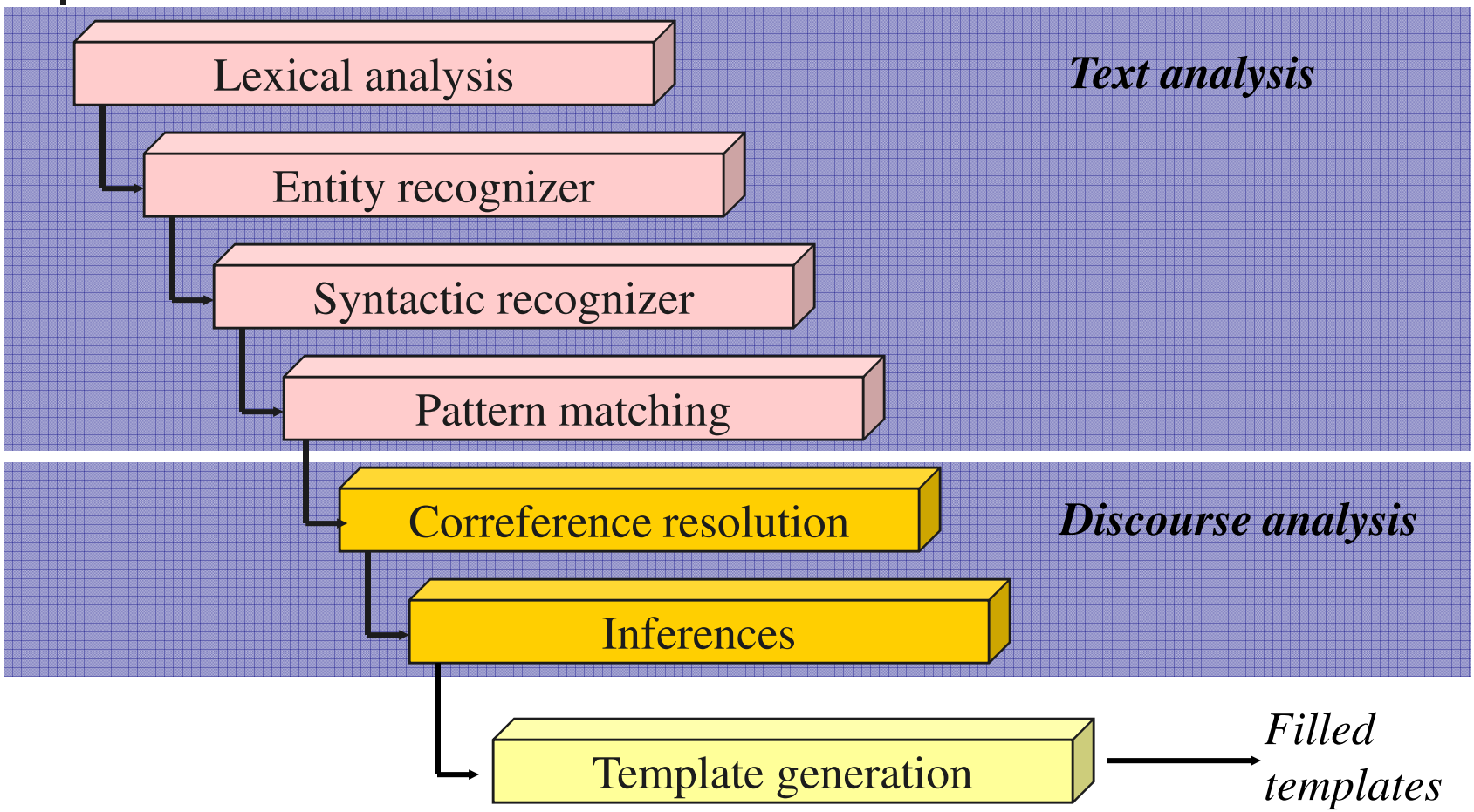
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

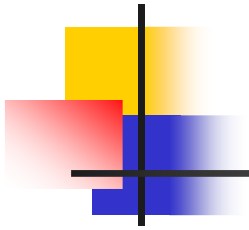
Universitat d'Alacant
Universidad de Alicante

Information Extraction: Pattern Matching

Unit 7 (a)

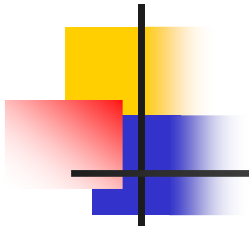
1. Introduction





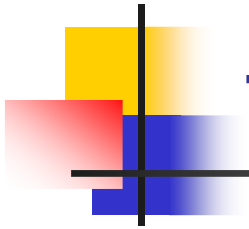
Extraction patterns

- It is usual to apply pattern matching techniques to extract information
- The patterns are learned from training corpus
- The application is similar to the algorithm of eliminating candidates
- The patterns must be applied from the general ones to the specific ones
- The general patterns can be applied for more than one case



Extraction patterns

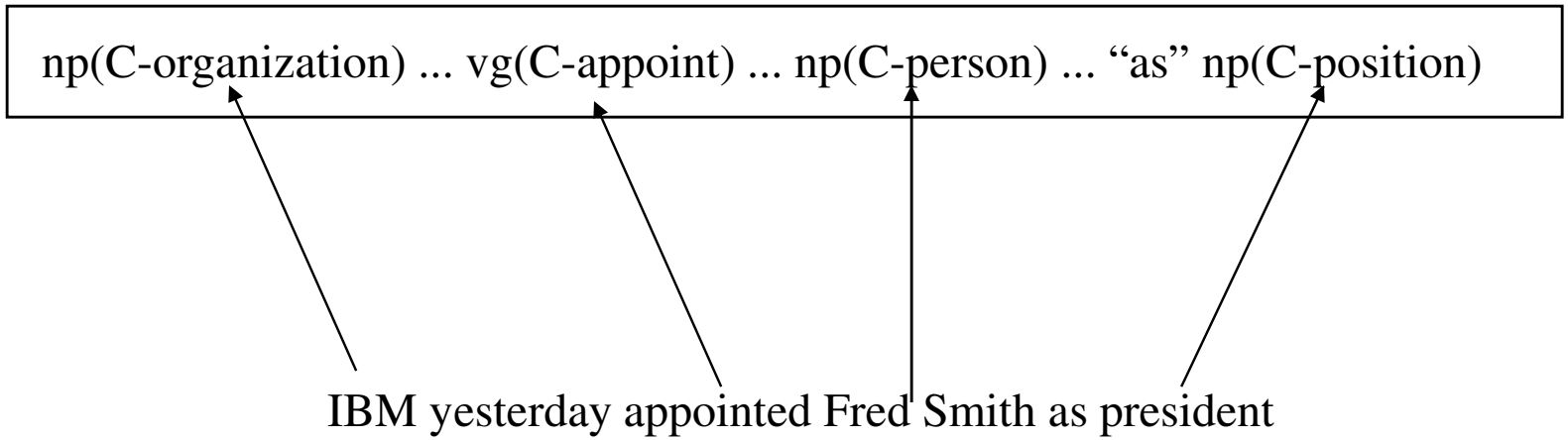
- The extraction rules can usually have
  - A pattern that must be applied over the structure (tagged text, analysis forest, logical forms). This pattern was previously obtained in the matching processing
  - One or more actions to be performed
    - Building examples of objects
    - Attributes´filling
    - Relationship establishment

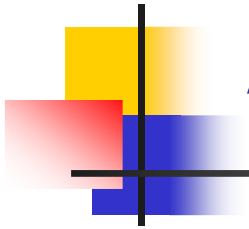


Type of patterns

- 3 levels
  - Low level: huge applicability (usually included in the system)
  - intermediate: patterns´libraries (applied to different domains)
    - ex. Entity extractors (person, company, location, organization)
    - Relationship extractors (person/organization, organization/location)
  - Specific of the domain

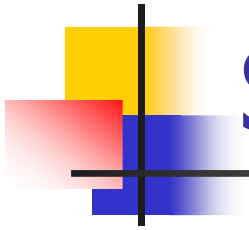
Example (PROTEUS)





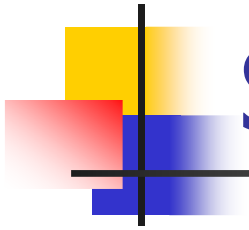
Acquisition of extraction patterns

- Usually manually built
- Limited used (but growing) of ML techniques
  - How to link templates with its NL expression
  - Until this moment Supervised learning
  - More often Unsupervised learning
    - Semi-supervised
    - Active Learning
    - Bootstrapping
    - Co-training



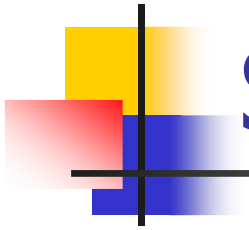
Syntactic analysis

- The first set of patterns uses basic components as the tagged words groups by means of a syntactic analyzer (parser) as NP (noun phrase) or VG (verb groups)



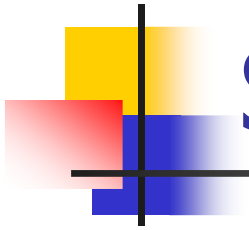
Syntactic analysis

- <np entity="e1"> Sam Schwartz </np> <vg>retired</vg> as <np entity="e2"> executive vice president</np> of <np entity="e3">the famous hot dog manufacturer</np>, <np entity="e4"> Hupplewhite Inc.</np>
- <np entity="e5">He</np> <vg>will be succeeded</vg> by <np entity="e6">Harry Himmelfarb</np>.



Syntactic analysis

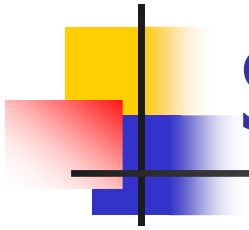
- Associated to each constituent there are some features that must be checked by the patterns in next steps
  - for VG: tense (past/present/future), voice (active/pasive), lemma/root
  - for NP: lemma/root, is it a name?, number (singular/plural)



Syntactic analysis

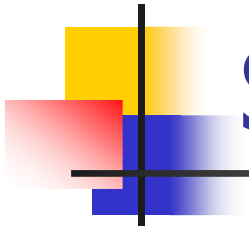
- For each NP, the system creates a semantic entity

entity e1	type: person	name: ”Sam Schwartz”
entity e2	type: position	value: ”executive vice president”
entity e3	type: manufacturer	
entity e4	type: company	name:”Hupplewhite Inc.”
entity e5	type: person	
entity e6	type: person	name: ”Harry Himmelfarb”



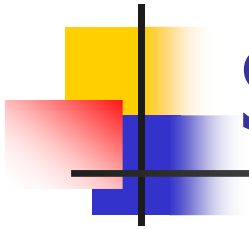
Syntactic analysis

- Semantic Restrictions
  - The next set of patterns builds a wide structure of NP adding modifiers
  - Duet to the syntactic ambiguity of the modifiers, the patterns must add some semantic restrictions (specific of the domain)



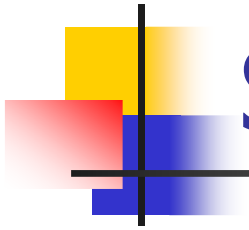
Syntactic analysis

- In our example, two patterns will recognize the apositive construction:
  - *company-description, company-name*,
- An the construction of the prepositional syntagm:
  - *position* of *company*
- and:
  - *position* matches any NP whose entity is of type ”position”
  - *company* respectively



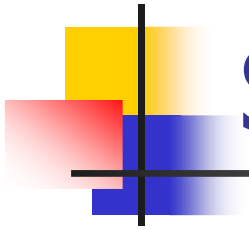
Syntactic analysis

- The system can include a small semantic hierarchy (*is-a* hierarchy)
  - e.g. manufacturer is-a company
  - Pattern matching uses the relationship *is-a*, this is a subtype of company (as manufacturer)



Syntactic analysis

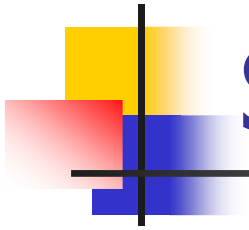
- In the pattern
  - *company-name*: NP with type ”company” whose core is a name
    - e.g. ”Hupplewhite Inc.”
  - *company-description*: NP with type ”company” whose core is a common name
    - e.g. ”the famous hot dog manufacturer”



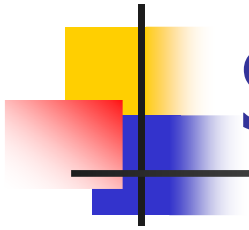
Syntactic analysis

- After applying the first pattern:
  - 2 NPs combined in one:
    - ”the famous hot dog manufacturer, Hupplewhite Inc”.
- Later, when applying the second pattern:
    - ”executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.”
  - A new NP + the relationship between "*the position*" and "*the company*"

Syntactic analysis



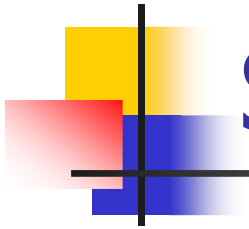
- <np entity="e1"> Sam Schwartz </np> <vg>retired</vg> as <np entity="e2"> executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.</np>
- <np entity="e5">He</np> <vg>will be succeeded</vg> by <np entity="e6"> Harry Himmelfarb</np>.



Syntactic analysis

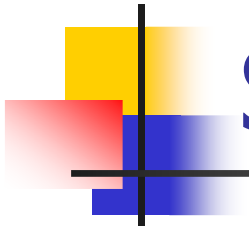
- The entities are updated as shown next:

entity e1	type: person	name: ”Sam Schwartz”
entity e2	type: position	value: ”executive vice president”
		company: e3
entity e3	type: manufacturer	name: ”Hupplewhite Inc.”
entity e5	type: person	
entity e6	type: person	name: ”Harry Himmelfarb”



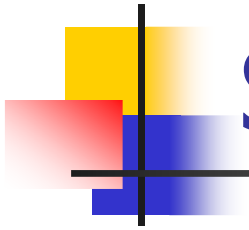
Scenario patterns

- The role of scenario patterns is extracting the events or the relationships relevant to the scenario
- In our example, there will be 2 patterns
  - *person* retires as *position*
  - *person* is succeeded by *person*
- *person* and *position* are items of patterns that match NPs with the associated type
- ”retires” and ”is succeeded” are items of patterns that match VG active and passive, respectively



Scenario patterns

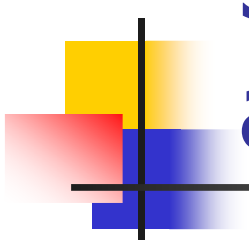
- *person* retires as *position*
  - Sam Schwartz **retired as** executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.
  - -> event leave-job (person, position)
- *person* is succeeded by *person*
  - He **will be succeeded by** Harry Himmelfarb
  - -> event suceed (person, person)



Scenario patterns

entity e1	type: person	name: ”Sam Schwartz”
entity e2	type: position	value: ”executive vice president”
		company: e3
entity e3	type: manufacturer	name:”Hupplewhite Inc.”
entity e5	type: person	
entity e6	type: person	name: ”Harry Himmelfarb”
event e7	type: leave-job	person: e1 position: e2
event e8	type: succeed	person1: e6 person2: e5

Scenario patterns for terrorist attacks



- For example, in Fastus there are 95 scenario patterns
  - killing of <HumanTarget>
  - <GovOfficial> accused <PerpOrg>
  - bomb was placed by <Perp> on <PhysicalTarget>
  - <Perp> attacked <HumanTarget>'s <PhysicalTarget> with <Device>
  - <HumanTarget> was injured

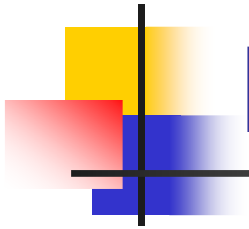
Linguistic patterns

- <subj> passive verb
- <subj> active verb
- <subj> infinitive verb

- <subj> name aux
- Passive verb<dobj>
- Active verb <dobj>
- infinitive <dobj>

- <victim> was dead
- <perpetrator> attacks
- <perpetrator> tries to attack

- <victim> was victim
- It was killed <victim>
- Attacked <target>
- kill <victim>



Exercises

- Look for patterns in the following domain:
  - Sports

Questions

TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco

Dr. Rafael Muñoz Guillena

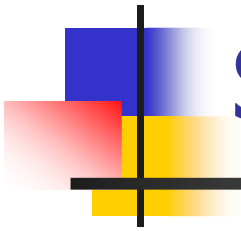
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

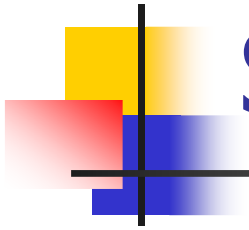
Universitat d'Alacant
Universidad de Alicante

Temporal Representation and Reasoning

Unit 6

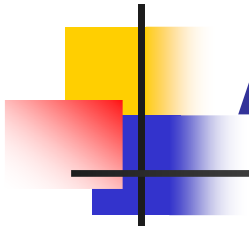


State of the art



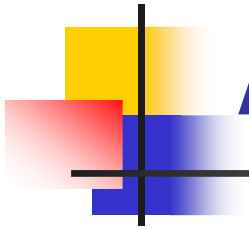
State of the art

- To locate and resolve information:
  - Knowledge based approaches
  - Corpus based approaches
  - Hybrid systems



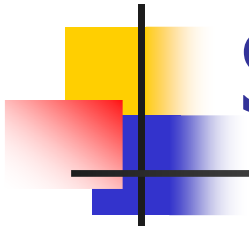
Approaches

- Based on knowledge→ systems that resolve the problem in an specific domain using a symbolic representation developed by a human.
- Based on corpus→ They use ML and corpus techniques to deal with temporal information
- Advantages and drawbacks



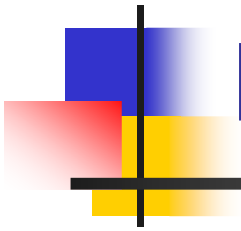
Approaches

- To locate temporal expressions, both approaches perform well
- For resolution, until this moment, only knowledge approaches are used.
- Most systems use one approach to locate and the other to resolve → hybrid systems



State of the art

Approach	System	
Knowledge	TERSEO HeidelTime TARSQI	*[Saquete2004]* *[Strotgen2010]* *[Verhagen2005]*
Corpus	*[Boguraev&Ando2005]* *[Bethard2006,2007]*	
Hybrid	TRIOS TIPSEM	*[Uzzaman&Allen2010]* *[Llorens2011]*



Proposal of a temporal model

Definition of a temporal model

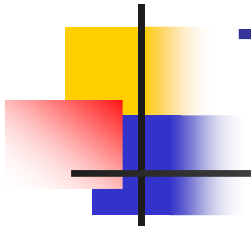
- It has two parts
  - Temporal ontology: temporal concepts and temporal relationships between the concepts
  - A set of two types rules:
    - Syntactic part of the model: identification rules (dependent of the language)
    - Semantic part of the model: resolution or normalization rules (independent of the language)

Temporal ontology

- It defines the relevant concepts within the temporal domain and the defined relationship between those concepts.
- Temporal concepts:
  - Resolution type
  - Referent date
  - Temporal Unit
  - Temporal caracter
  - Temporal value

Temporal ontology

- According to the resolution type:
  - Explicit TE: It denotes a date or period.No need to solve it:
    - Complete date with or without time: 11/11/2005
    - Expressions that refer to known events: Christmas
  - Implicit TE: They need to be solved in order to know the exact date or period they refer to. We need resolution rules for these expressions: yesterday, the day before,..



Temporal ontology

- According to the referent date:
  - TE that refers to the date of the document: yesterday
  - TE that refers to a previously named TE: a month later

Temporal ontology

- According to the temporal uni:
  - DAY: TE as yesterday, tomorrow
  - MONTH: TE as a month later
  - YEAR: TE as a year before
  - WEEK: TE as next week
  - WEEKEND: TE as this weekend
  - WEEKDAY: TE as last Monday
  - EVENT: TE as last Easter

Temporal ontology

- According to the temporal caracter:
  - PAST: referring to a previous time to the referent date last week
  - PRESENT: present moment this year
  - FUTURE: referring to a future time to the referent date next week
  - PAST/PRESENT: Period that starts in the past and finish in the present with respect to the referent date since five days
  - PRESENT/FUTURE: Period that starts in the present and finish in the future with respect to the referent date during the following five months

Temporal ontology

- According to the temporal value:
- Based on the type of definition of the expression:
  - CONCRETE: concrete date→ yesterday
  - FUZZY: fuzzy expression→ some days ago
- Based on the wideness of the expression:
  - SIMPLE: it is referring a simple value→tomorrow
  - PERIOD: it is referring periods of dates →during weeks

Temporal relationships

- Temporal relationships between concepts:
  - IS A: It is the relation defined by the different taxonomies, for example yesterday is a DAY or some days ago is a Fuzzy
  - Precedence relation: It is the relation that defines the temporal ordering between the concepts in the ontology and it can be used to establish the chronological ordering between the events, for example yesterday precedes today and today precedes tomorrow. This relation defines a horizontal relation between concepts of the ontology

Temporal rules of the model

- Two types of rules:
  - Identification rules: They allow to identify the temporal expressions withing the concepts defined in the ontology. They are language dependent
  - Resolution rules: They allow to resolve the concepts of the ontology. They are implemented in first order logic and based on Hobbs defined functions (2002)

Example of resolution rule

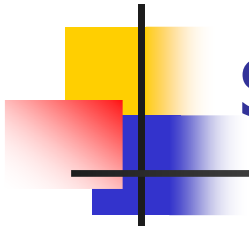
- Example of resolution for DAY (PAST):

yesterday(x):
∃ y,z[documentDate(y)
∧
before(x,y)
∧
interval-between (z,x,y)
∧
duration(z,*Day*)=1
∧
day(x)]

Proposal of a system based on knowledge

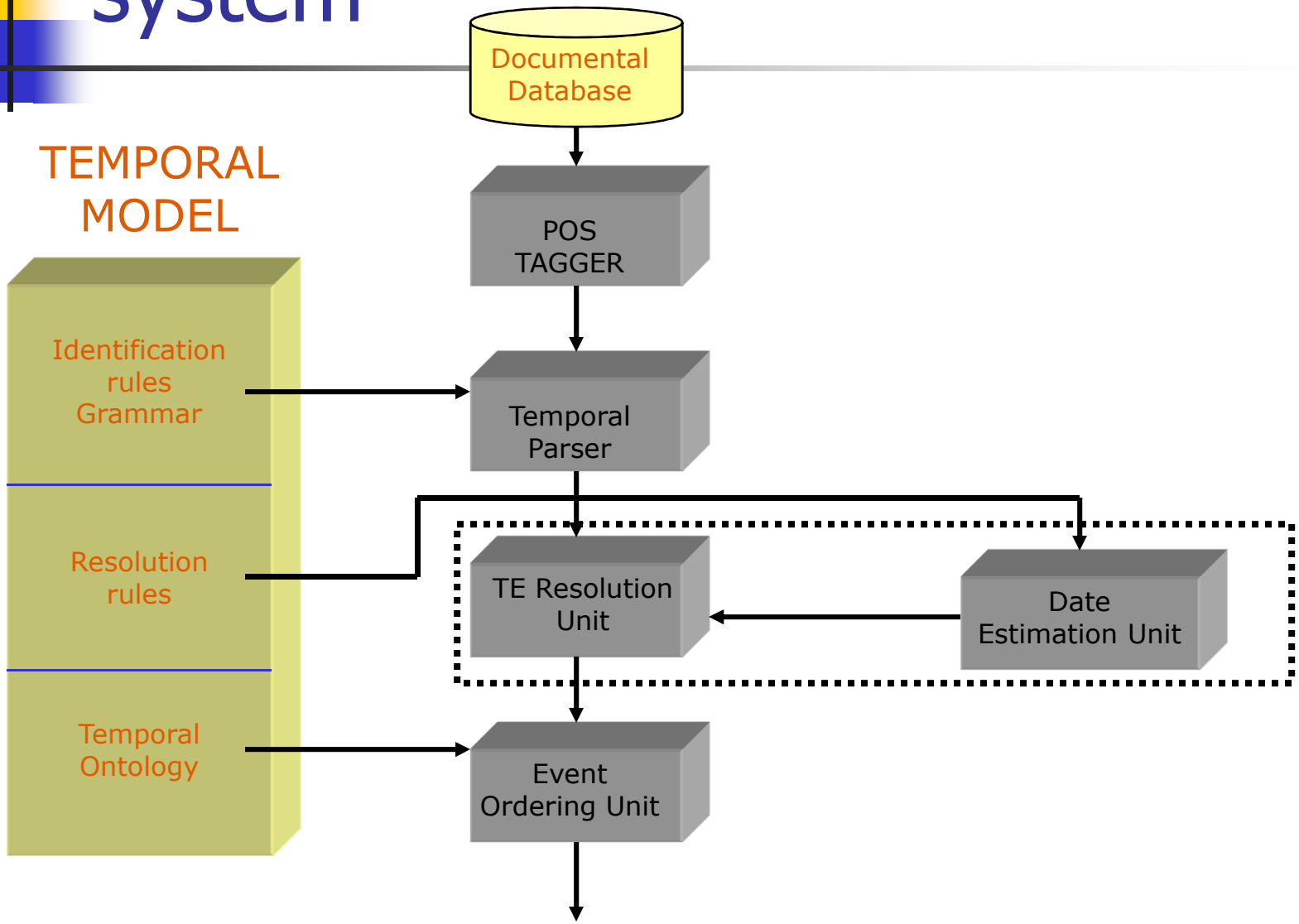
TERSEO system

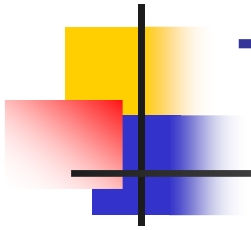
TE identification and resolution system



- A proposal of a system allows to:
  - Identify temporal expressions
  - Resolve those temporal expressions
  - Annotate them using an annotation schema
  - Order the expressions in the text
- System based on knowledge
- Developed for Spanish

TE identification and resolution system





Temporal Parser

- It is based on the identification rules of the temporal model and uses a temporal grammar
- It performs a shallow parsing, so the grammar rules are only recognizing TE:
  - Rules for explicit dates
  - Rules for implicit dates

Identification rules

Implicit dates Document Date Concrete	Reference→'mañana' *(tomorrow)* Reference→'ayer' *(yesterday)*
Implicit Dates Previous Date Period	Reference→'un mes después' (*a month later*) Reference→num+'años después' (*num years later*)
Implicit Dates Previous Date Concrete	Reference→'un día antes' (*a day before*)
Implicit Dates Previous Date Fuzzy	Reference→ 'días después' (*some days later*) Reference→ 'días antes' (*some days before*)

TE resolution unit

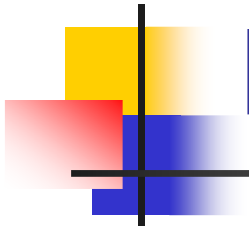
- For TE resolution→ resolution rules of the temporal model
- Two possible referent dates:
  - Document date(DateP)
  - Date previously named in the text (DateA)

- Proceso para buscar la fecha de referencia:
  - Por defecto se usa la fecha del documento
  - Cuando una ET explicita es encontrada, su valor se guarda en DateA. El valor de DateA es actualizado con cada nueva TE encontrada.

Resolution rules

REFERENCE	RESOLUTION RULES
'ayer' (yesterday) 'mañana' (tomorrow)	DateP – 1 DateP +1
'durante el mes siguiente' (during the following month)	[DayI/Month(DateA) +1/Year(DateA) -- DayF/Month(DateA)+1/Year(DateA)]
'un día antes' (a day before)	DateA-1
'días después' (some days later)	>>>>>DateA

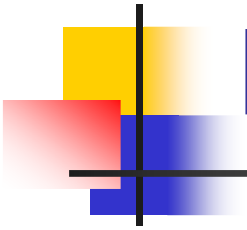
Event ordering unit



- We built a table with the complete information extracted from the XML tags
  - The table includes a column for each tag ID, VALDATE1, VALTIME1, VALDATE2, VALTIME2 y VALORDER.
- Ordering rules:
  - TE1 is previous to TE2 if the range associated with TE1 is previous and disjoint to the one associated with Te2.
  - TE1 is concurrent with TE2 if the range associated with TE1 overlaps the one associated with TE2

Multilingual extension

- Monolinguility problem:
  - Language barrier
  - Difficulty to compare systems performing in other languages
- Current tendency of the systems:
  - Multilinguality. CLEF →tasks related with multilingual IR
- Getting knowledge for a new language → Solution???



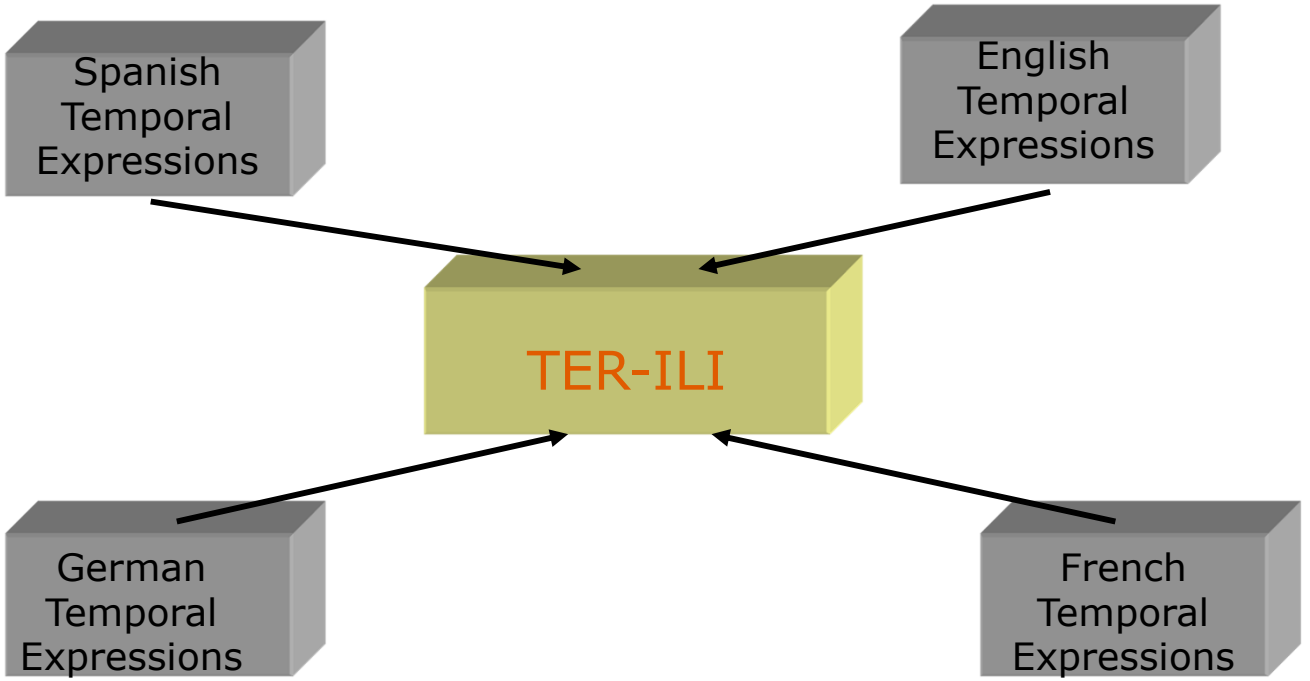
Multilingual extension

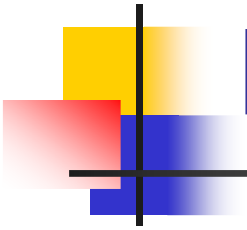
- For ML systems: they need an annotated corpus in the new language
- For knowledge based systems: they need to manually determine the new rules for identification in the new language.
- Very costly solutions

Multilingual extension

- Novel proposal: Developing a new platform that allows to automatically obtain the new knowledge required to identify and resolve the TE in the new language using the initial language as a base
- Similar architecture to EuroWordNet (Vossen, 2000)

Multilingual extension





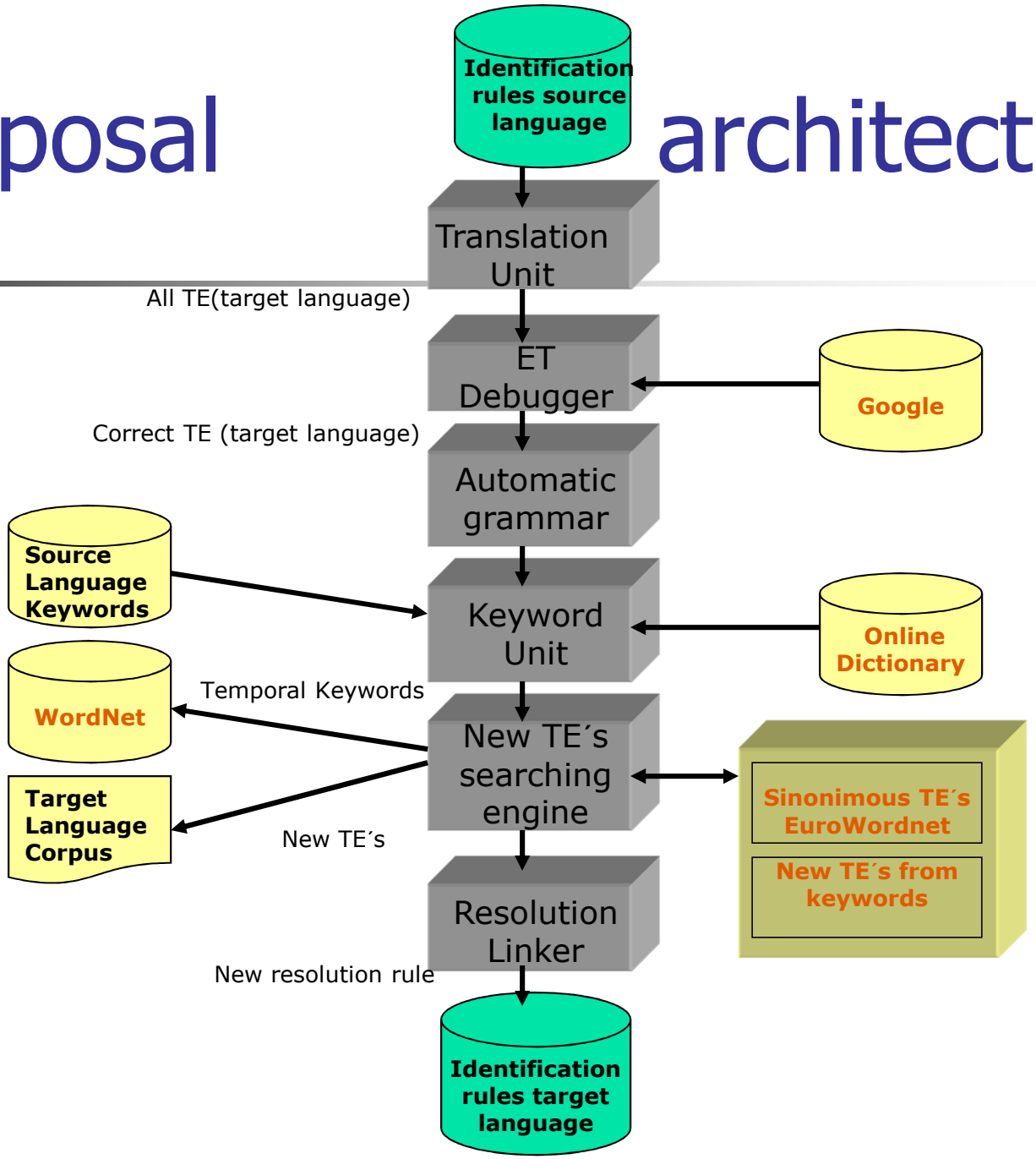
Multilingual extension

- In order to perform this extension we will use automatic translator to translate the initial expressions to the new language
- The resolution rule is the same as the one of the original expression
- Google is used to filter the translated expressions

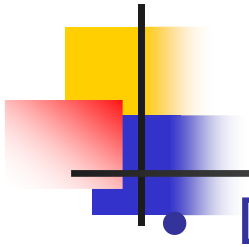
Multilingual extension

- Another automatic extension options:
  - Getting new expressions from corpus annotated with temporal expressions
  - Using parallel corpora to obtain the equivalence between a temporal expression in one language and another. The obtained expression is always correct whereas the automatic translated can have some errors.

Proposal architecture



Translation Unit



- Direct translation of the temporal expressions stored in TERSEO system to the destination language (English, French, Italian, …)
- Three machine translation systems were used: BabelFish, FreeTranslator, Power Translator
- Each expression in the destination language will be linked with the resolution rule of this expression in the original language

Temporal expression debugger

- Objective: avoiding possible errors in the translation
- All the translated expressions are filtered using Google as a resource
- Each exactly expression is searched in Google and it will be considered wrong if Google do not return any results

Automatic grammar generation

- We use two resources to generate the partial grammar in the destination language in an automatic way:
  - MACO POS TAGGER in English
  - All the translated expressions to English
- It generates a set of rules that will conform the new grammar
- It is a partial grammar, so, it only implements a shallow parsing that recognize TE in the text

Keyword unit

- The set of keywords is used to look for new TE in the destination language.
- Each keyword has a set of temporal features.
- In order to obtain them we use:
  - All the expressions previously translated to the destination language
  - Wordnet. Wordnet is a lexical resource used to get synonyms of the keywords

Keyword unit

- Two types of temporal keywords:
  - — High temporality keywords: words that have only a temporal sense and only this one, for instance *"January", "day",..*
  - — Low temporality keywords: words with high probability of being part of a temporal expression but not always, for instance *"next", "last",..*
- When looking for new TE´s the keywords with high temporality are going to be looked for in the text first

New Temporal Expression Searching Engine

- Using the temporal keywords, this unit access to a non-annotated corpus in the destination language, returning possible new TE.
- If the temporality keywords set (high and low temporality) are found in the text, this set of words is considered a possible TE.
- The expression is filtered using Google.
- The expression is not linked to any resolution rule in a first moment

New Temporal Expression Searching Engine (Ejemplo)

- *"There were two accidents days ago."*
  - If our keyword sets is form by: day, days, ago, before,…
  - The searching engine found the high temporality keyword: "days" and looks for more temporal keywords (high and low) backwards and forward of this word in the text.
  - The searching will finish when no more adjacent temporal words are found.

Resolution Linker

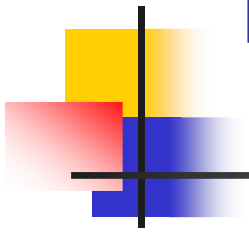
- It assigns a resolution rule for each expression using the temporal features assigned to each word in an expression
- For example, the resolution rule:

  Month (Date) / Day (Date)-1 / Year (Date)

  - Expressions in the origin language related with this rule:

  ayer, el pasado día, el último día,…

  - Expressions in the destination language related to this rule: yesterday, the past day, the last day, the day before …



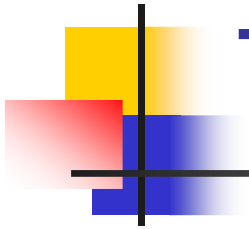
Exercise

- Access to the demo web of TERSEO (only English):

http://gplsi.dlsi.ua.es/demos/TERSEO/

Proposal of a system based on corpus

TIPSEM system



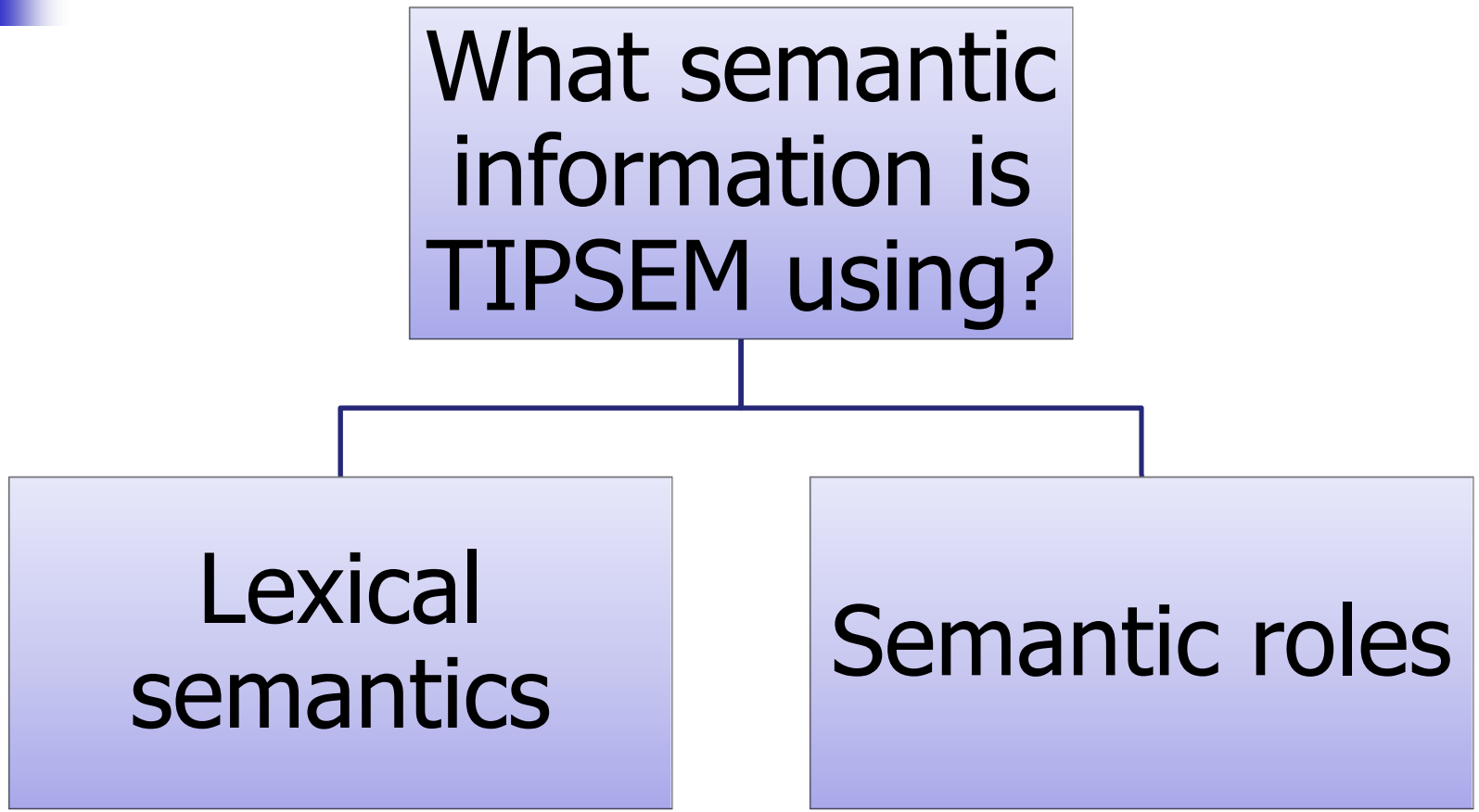
TIPSEM

- TIPSEM: **T**emporal **I**nformation **P**rocessing using **Sem**antics
- TIPSEM is in fact a hybrid system:
  - Machine learning to locate the expressions
  - Rules for the resolution of the expressions

TIPSEM

- The difference with other systems is that they only use morphosyntactic information (words, lemmas, syntax,…) and this system propose using also:
  - Semantic information
- It was implemented for English but it has been adapted to Spanish, Italian and Chinese

TIPSEM


What semantic information is TIPSEM using?
Lexical semantics
Semantic roles

TIPSEM

Lexical semantics

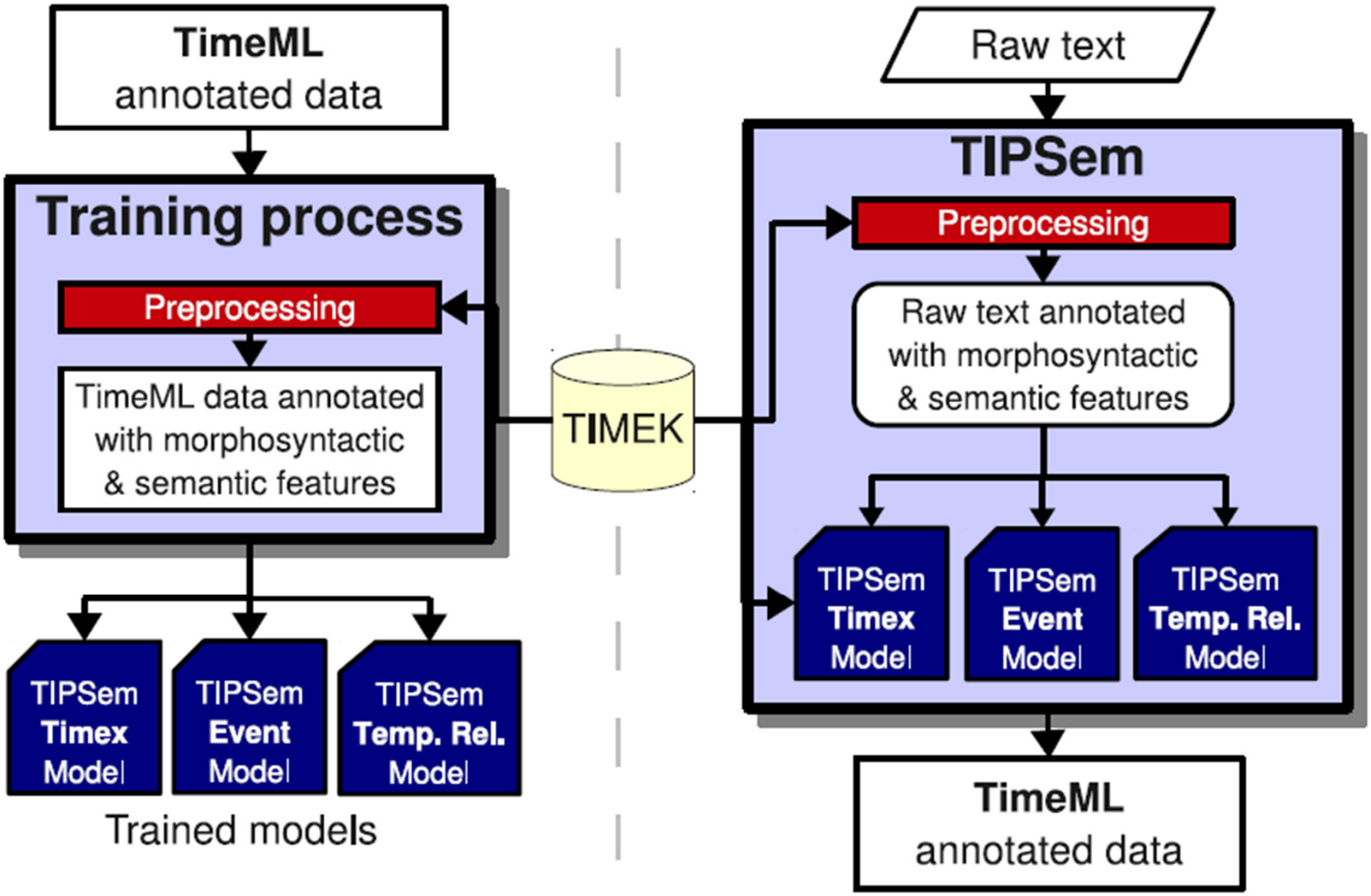
- Semantic information at word level
- It includes synonyms, hyperonyms, hiponyms,…
- We use a semantic resource called **Wordnet**:
  - "December" is semantically related with "time"
  - "war" is related with "event"

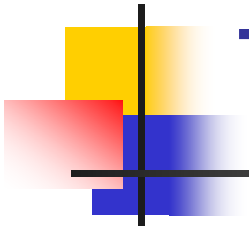
TIPSEM

Semantic roles

- Semantic information in a sentence level
- It specifies the role of the predicate arguments
- We use the **Propbank set of roles**:
  - *[He A0] visited [Alicante A1] [in April AM– TMP].*
  - *[He A0] likes [April A1]*

Arquitectura TIPSEM



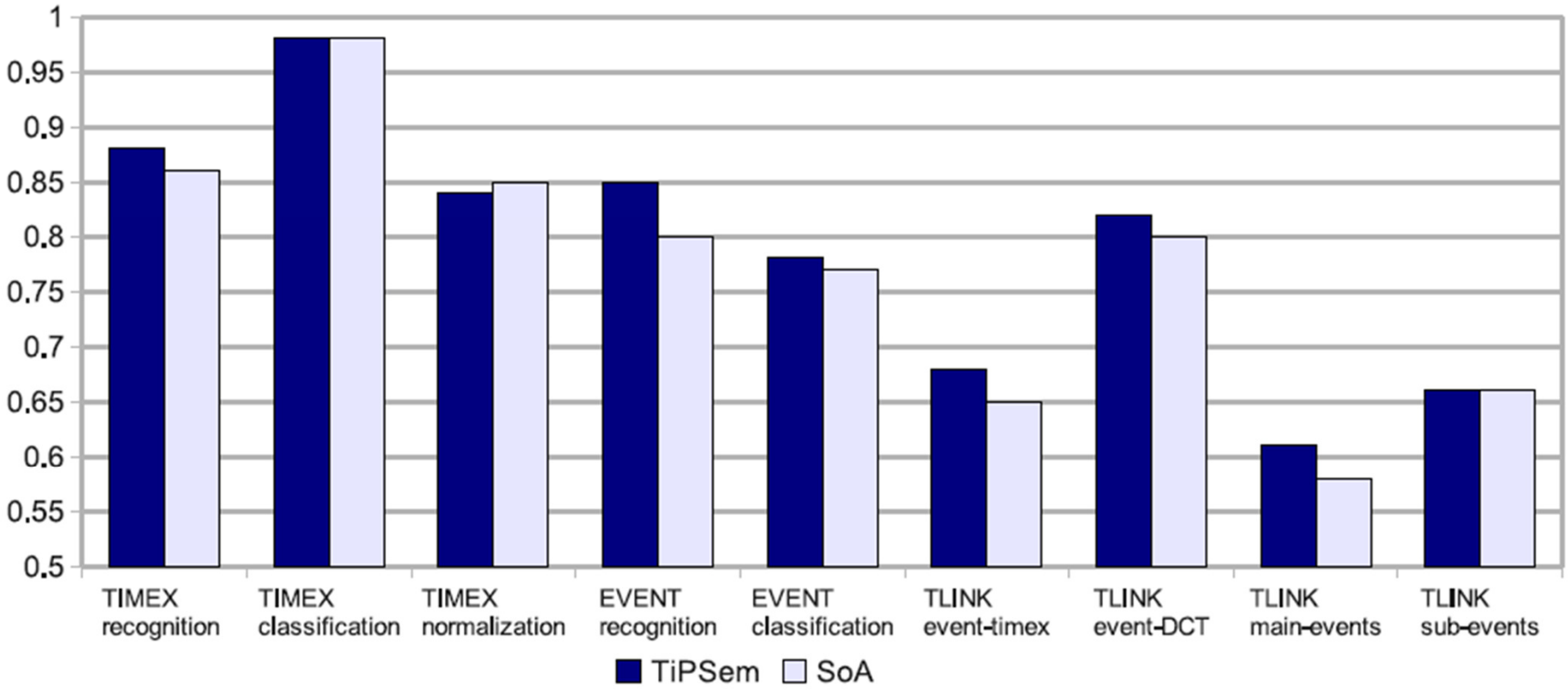


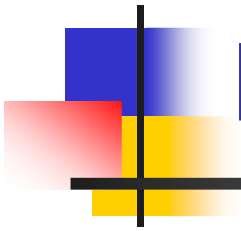
TIPSEM

Two machine learning techniques

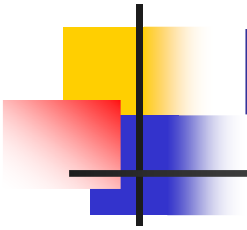
- CRFs: very appropriate to problems related with sequences (recognition of timex and event, and categorization of temporal relations)
- SVMs: very appropriate to problems when sequence is not relevant (classification of timex and event, types of normalization of timex)

TIPSEM evaluation



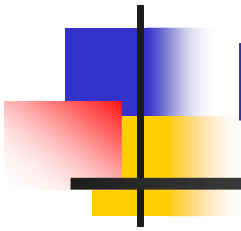


Possible applications



Possible applications

- Temporal representation and reasoning can be applied to many final applications in NLP, for example:
  - Chronological representation
  - Temporal QA
  - Multi-document Summarization



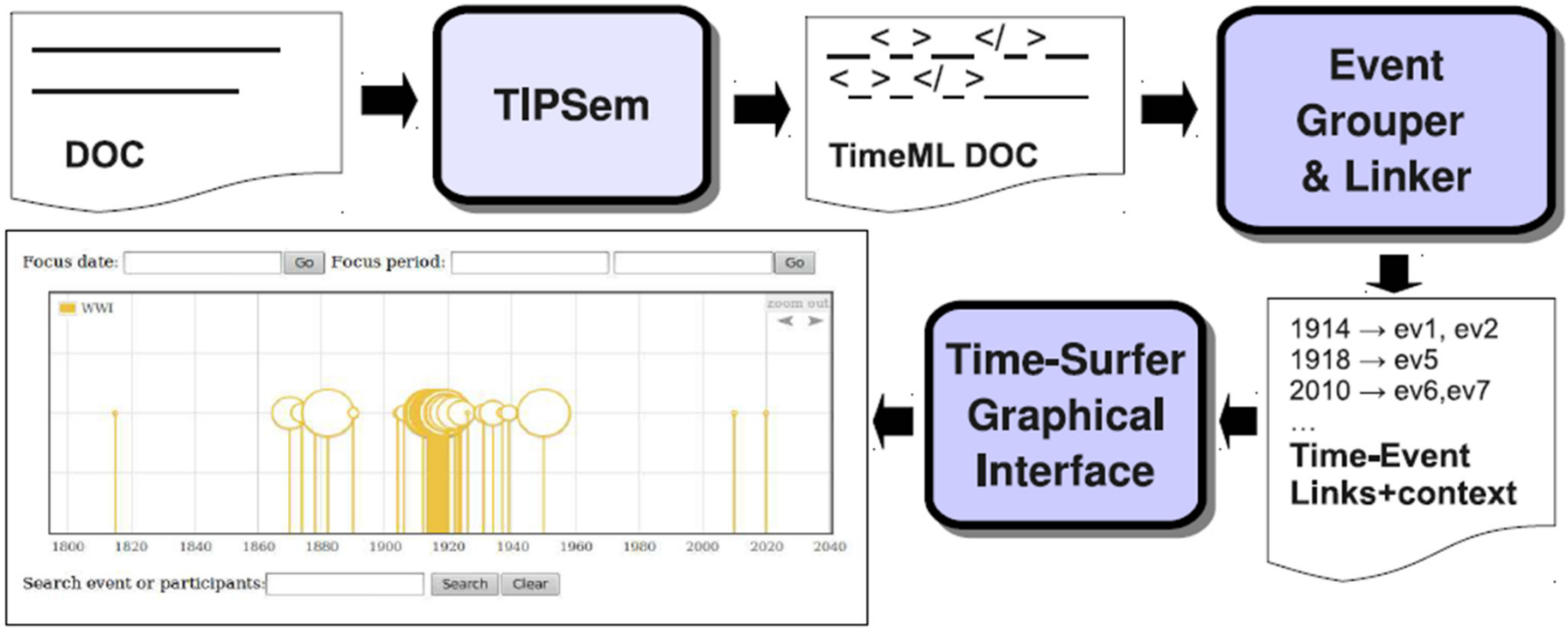
Possible applications

Chronological representation
TIMESURFER

Representación cronológica

- TIPSEM has been applied to build a graphical interface of the content of the document → TIMESURFER
- TIMESURFER allows the user to search and explore the information related to time.
  - It draws a timeline with the events of the text.

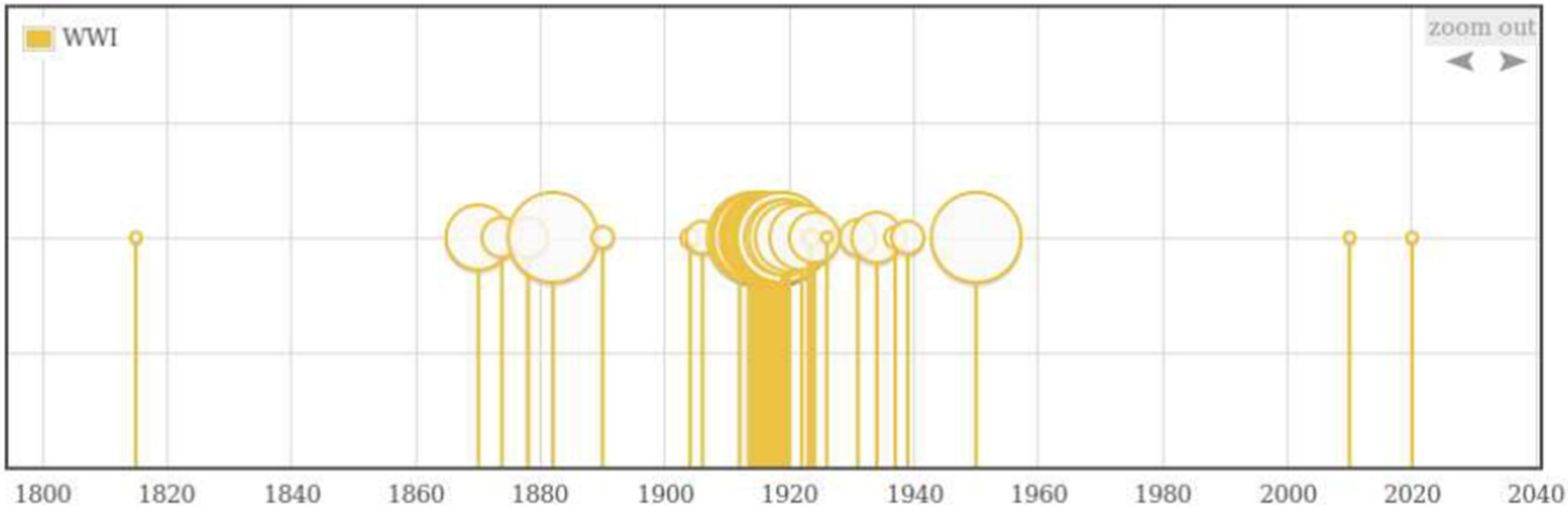
TIMESURFER


DOC
TIPSem
TimeML DOC
Event Grouper & Linker
1914 → ev1, ev2
1918 → ev5
2010 → ev6,ev7
…
Time-Event Links+context
Time-Surfer Graphical Interface
Focus date:
Go
Focus period:
Go
WWI
zoom out
1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000 2020 2040
Search event or participants:
Search
Clear

Time-Surfer: history/WWI

Time representation of history/WWI article in Wikipedia.

Focus date: Go Focus period: Go



Earliest year: *1815* - **Latest year**: *2020* - **Total event groups**: *70*

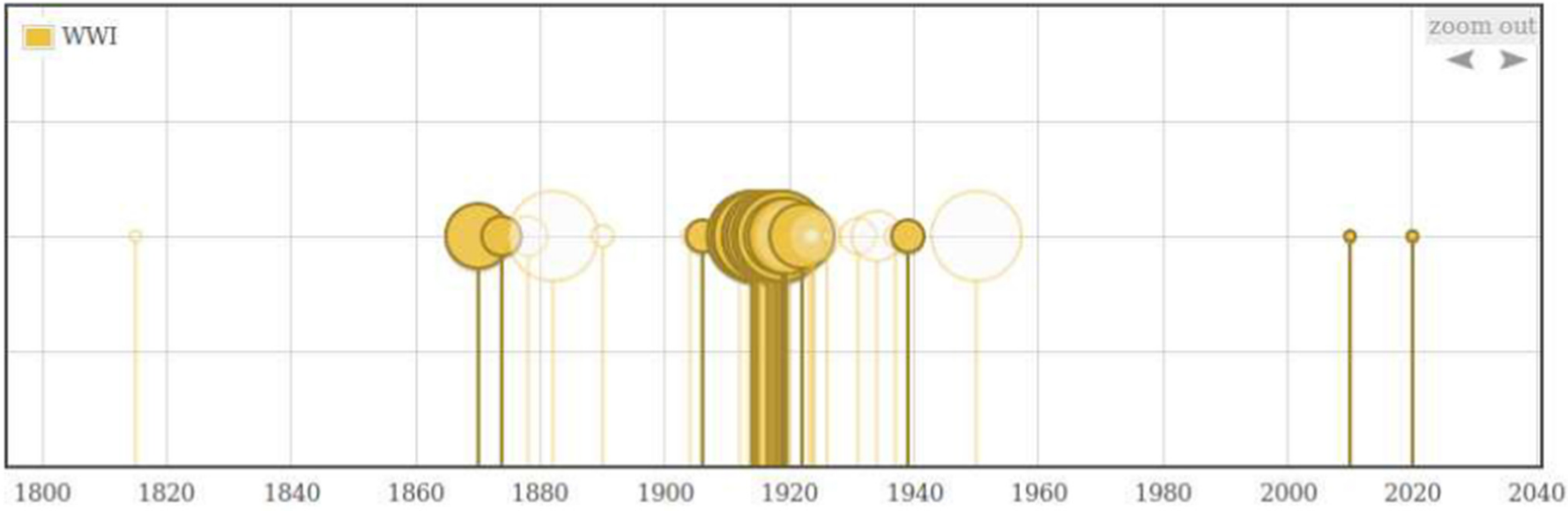
Search event or participants: Search Search&Focus Clear

Search entity relations - E1: E2:

Time-Surfer: history/WWI

Time representation of history/WWI article in Wikipedia.

Focus date: Go Focus period: Go



Earliest year: *1870* - **Latest year**: *2020* - **Total event groups**: *28*

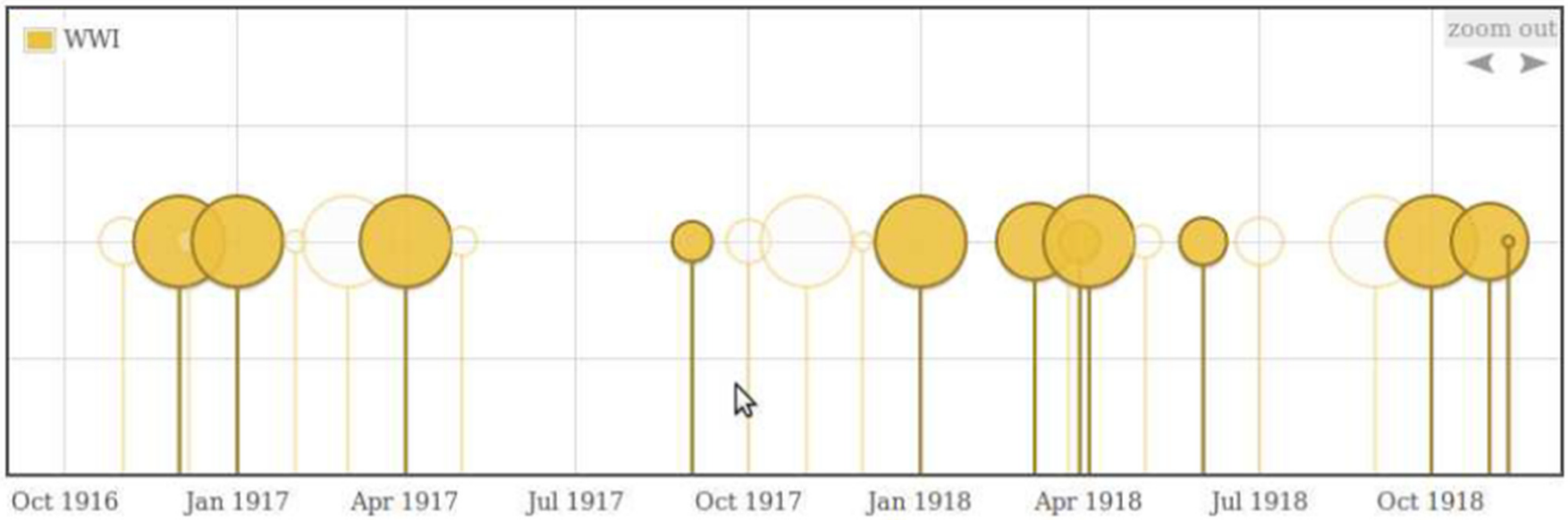
Search event or participants: Germany Search Search&Focus Clear

Search entity relations - E1: E2:

Time-Surfer: history/WWI

Time representation of history/WWI article in Wikipedia.

Focus date: Go Focus period: Go



Earliest year: *1870* - **Latest year**: *2020* - **Total event groups**: *28*

Search event or participants: Germany Search Search&Focus Clear

Search entity relations - E1: E2:

Time-Surfer: history/WWI

Time representation of history/WWI article in Wikipedia.

Focus date: Go Focus period: Go

Earliest year: *1870* - **Lates**

Search event or participants: Germany

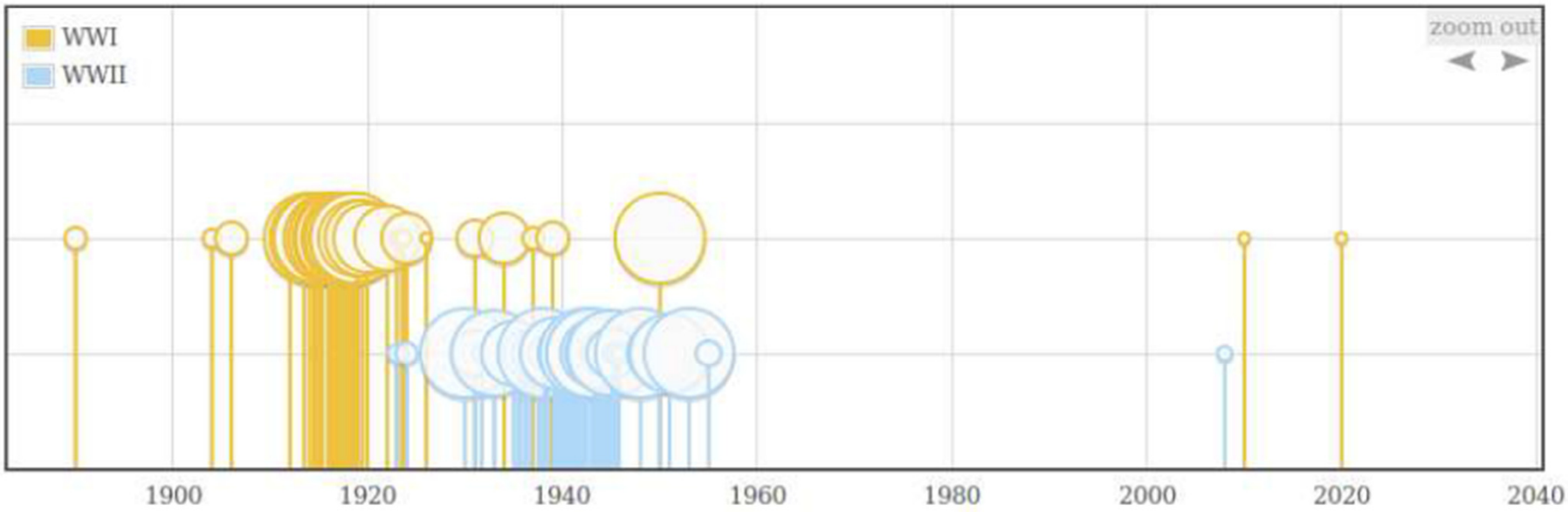
Search entity relations - E1: E2:

Time-Surfer: history/WWI and history/WWII

Time representation of history/WWI article in Wikipedia.

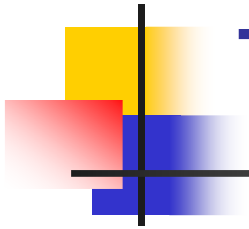
Time representation of history/WWII article in Wikipedia.

Focus date: Go Focus period: Go



Earliest year: *1815* - **Latest year**: *2020* - **Total event groups**: *139*

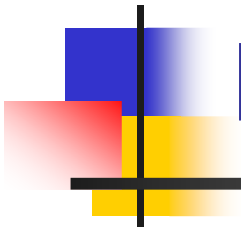
Search event or participants: Search Search&Focus Clear



TIMESURFER

- Exercise: Read the paper about TIMESurfer and try the demo:

http://gplsi.dlsi.ua.es/demos/TIMEE/Time-Surfer/index.php



Possible applications

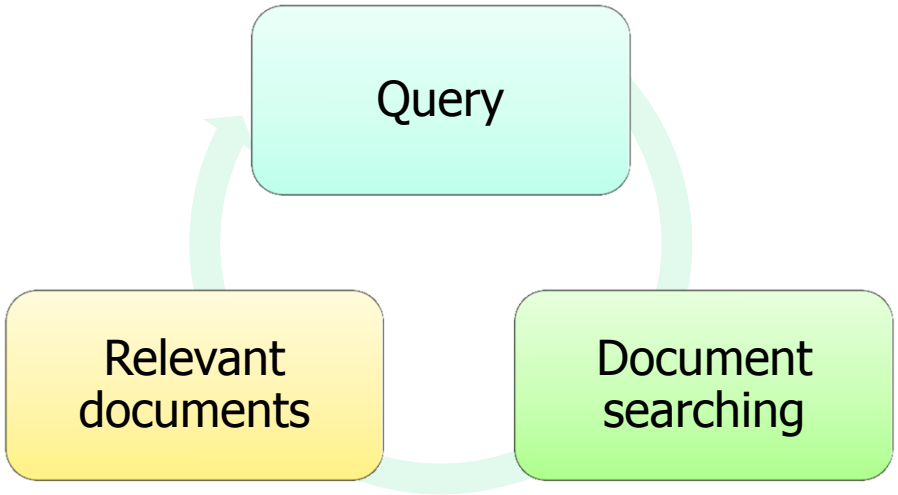
Temporal Question Answering

Temporal QA

- Question Answering?: it is the process to automatically look for the answer of questions of the user
- The questions will be done in natural language as well as the answers
- Nowadays, the QA systems only deal with factual questions → *"What is the capital of Brazil?"*

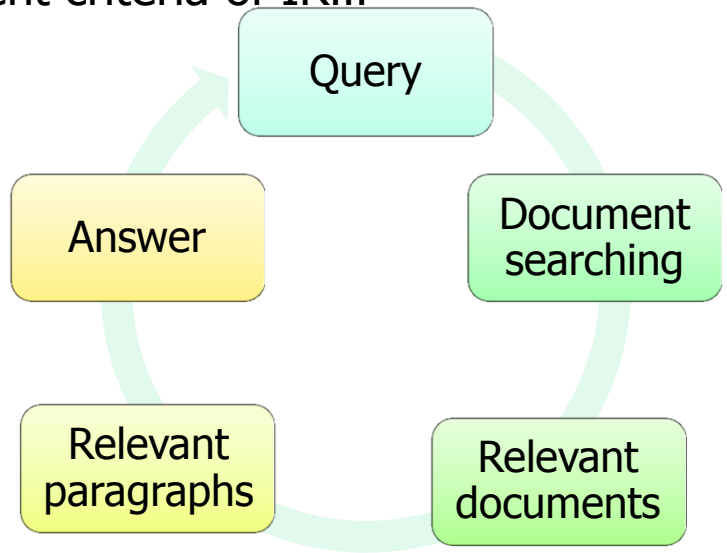
QA vs. IR

- Information retrieval
  - The user makes a query
  - It returns the most relevant documents related to this query
  - It is usually based on statistical information

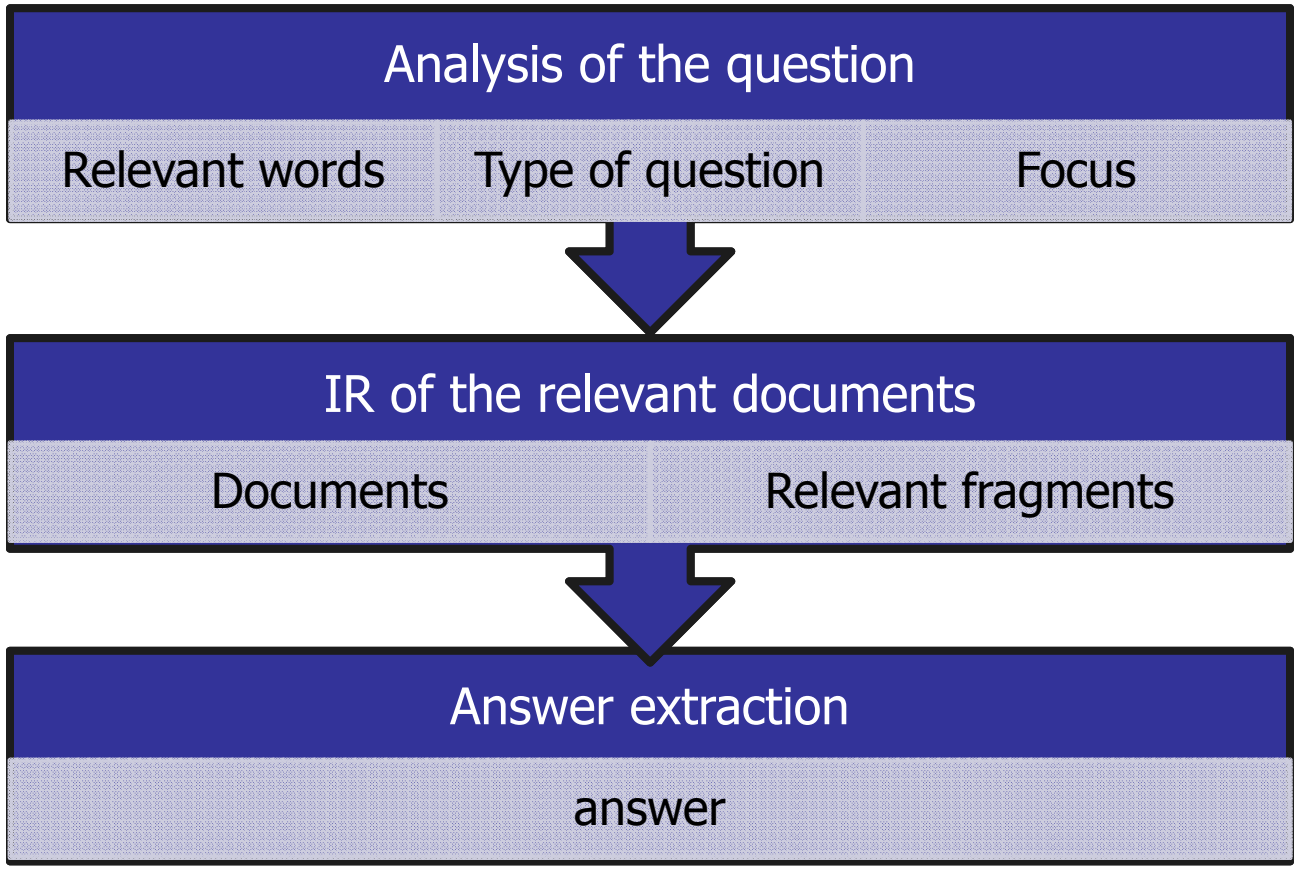


QA vs. IR

- QA systems
  - Same origin: query of the user
  - It gets the most relevant documents (and paragraphs) according to the query
    - It is possible that use different criteria of IR…
  - It returns a concrete answer
  - It uses more NLP



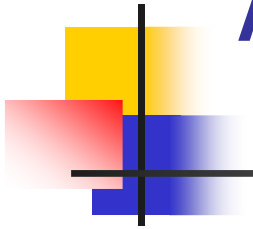
Basic QA architecture



Basic QA architecture

Analysis of the question

- Objetives
  - Determining the type of question
    - What?
    - Who ? ...
  - Determining the type of answer
    - Object
    - Person …
  - Determining the focus of the question
    - The most important term in the query, the one about we are asking
  - Extracting the terms of the query for the IR
  - Determining the semantic context of the answer



Analysis techniques

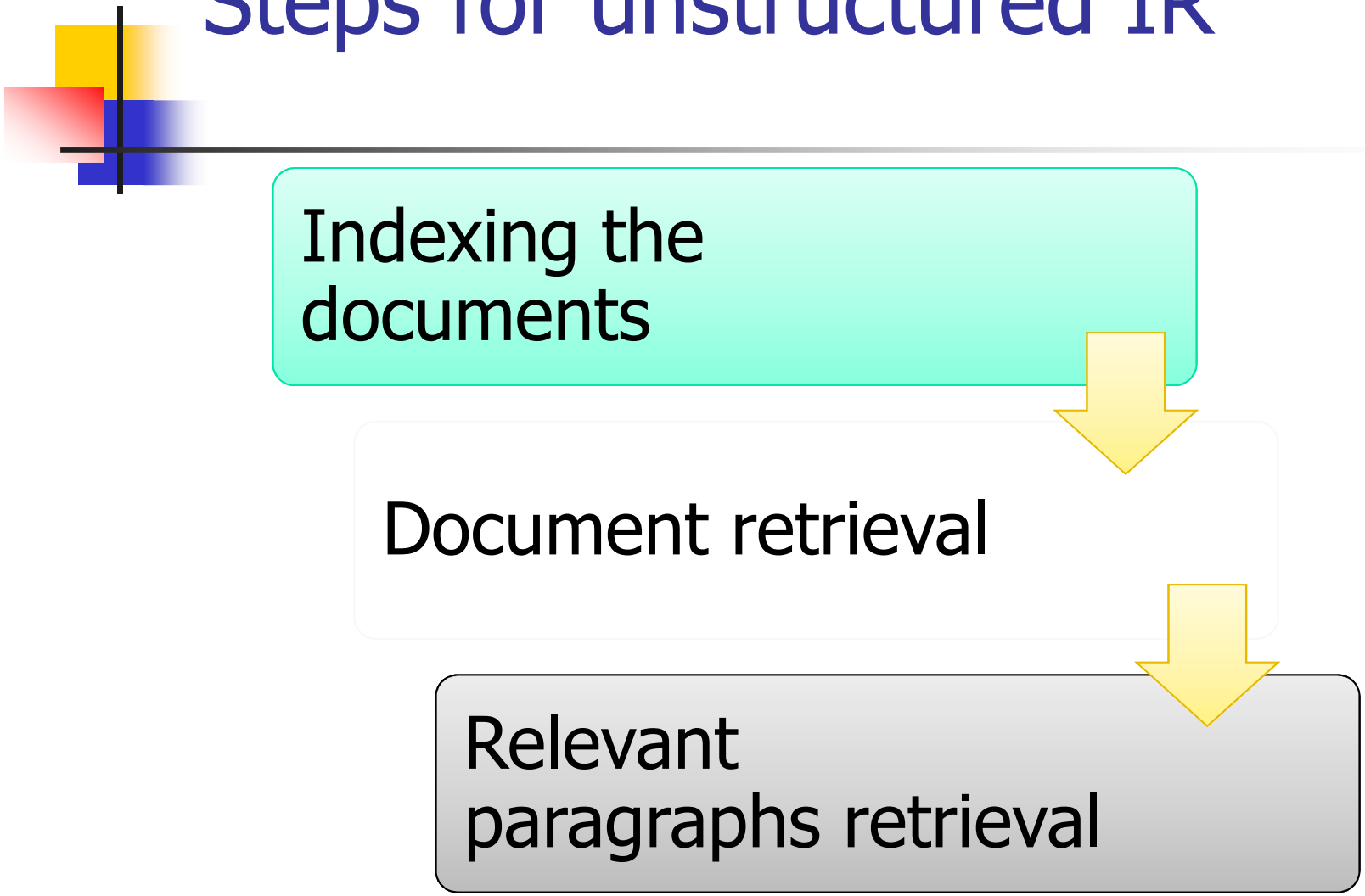
- Limited NLP
  - Bag of words
  - Stemming
  - Lemmatization
  - Avoiding stopwords
- Advanced NLP
  - Shallow syntactic parsing
  - Dependency syntactic parsing
  - Word disambiguation
  - Semantic roles detection
  - Semantic relations
  - Question classification system
  - Getting the focus of the question
  - Extracting of the keywords of the question
  - NER, multiword recognition, abbreviations, dates,…
  - Wikipedia support

Basic QA architecture

Information retrieval

- Relevant documents to the question
  - Between them the answer is found
  - IR depends on the type of access for the data
    - Structured data
      - IR = access to the DB through its query language
      - Previous mapping of the DB schema → ontology
      - conversions FL → SQL
    - Unstructured and semi-structured data
      - The information that could have the answer must be analised and related to the question

Steps for unstructured IR

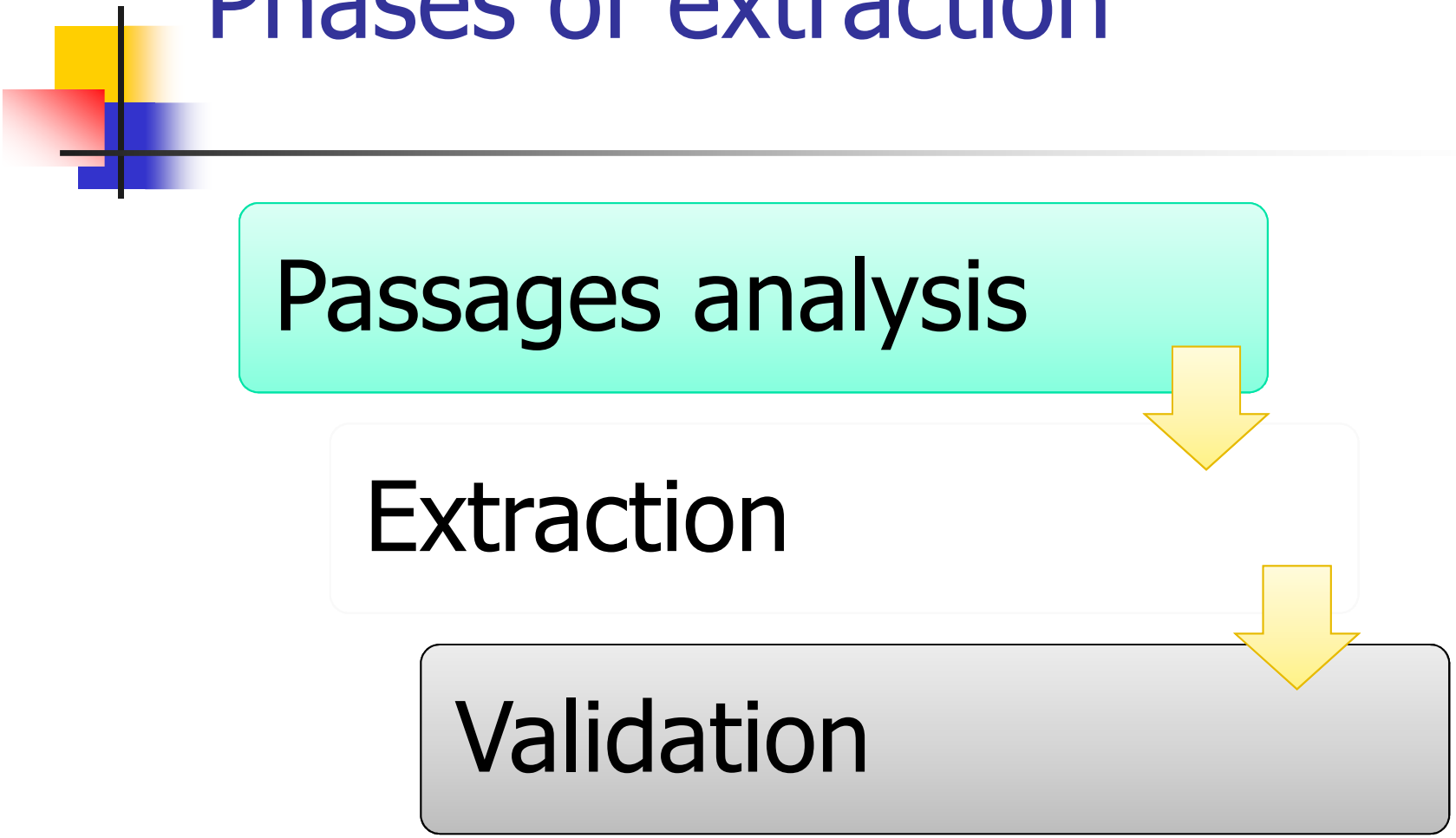


Basic QA architecture

Answer extraction

- Locate the answer in the relevant Localización de la respuesta en los fragmentos relevantes
- La dificultad de esta tarea depende del tipo de respuesta esperada
- Objetivo fundamental
  - inferir información sobre la respuesta desde la pregunta
  - diversas técnicas de proyección pregunta-respuesta

Phases of extraction



Extraction of the answer

- Projection models
  - Automatic translation models
    - From the language of the questions to the language of the answers (Berger et al., 2000)
  - Machine learning techniques from a corpus of question/answer pairs
    - From the FAQ files it will learn the terms, sentences and patterns that appear in the document to be retrieved (Agichtein et al, 2001)
  - Lexical-semantic patterns
    - SiteQ (Lee et al., 2001) relies on 361 lexical-semantic patterns to locate the answer

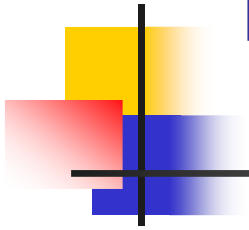
Approaches for validation

- Based on nearness (Magnini et al. 2002)
  - It is using the redundancy of the Web
  - It will check if the selected answer appears near to the terms of the question
- Based on patterns (Subbotin y Subbotin, 2001)
  - It will built validation patterns
  - The question will be rewritten as an affirmation that includes the candidate answer
  - It will check if this sentence appears in the Web
- Statistical (Turney, 2001)
  - It will check if the question and the answer are usually together in the web
  - It does not need to download the whole documents to check nearness, it will only count them

Examples of validation

- Approach based on nearness
  - **Question**: Who invented the telegraph?
  - **Candidate answer**: Samuel Morse
  - Query to the web (snippets): "Samuel" "Morse" "invent" "telegraph"
    - The person who **invented** the **telegraph**, was **Samuel Morse**.
    - Until in 1854 the US Supreme Court passed the **Morse** was who **invented** the first **telegraph**.
    - **Samuel** Finley Breese **Morse** (27th april 1791, Charlestown, Massachusetts – died the 2nd of April 1872, Nueva York), **inventor** of the **telegraph**.
    - **Samuel Morse** had made history because he has **invented** the electric **telegrap** and the **Morse** alphabet.
  - Validation:
    - High nearness between terms:: Samuel Morse, invent, telepgrph.
    - Correct answer.

Examples of validation

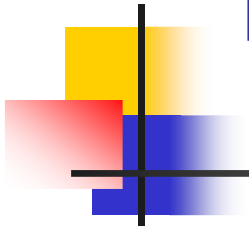


- Approach based on patterns (simple approach)
  - **Question**: Who invented the telegraph?
  - **Candidate answer**: Samuel Morse
  - Query to the web: "Samuel Morse invented the telegraph"
    - 23 ocurrences
  - Validation:
    - The question rewritten as an affirmation appears many times in the web
    - Correct answer.
  - Problem:
    - Not always the question appear written in the same way

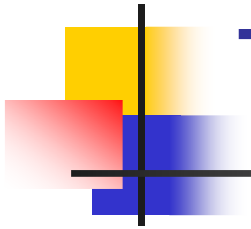
Examples of validation

- Approach based on patterns (AVE approach)
  - **Question**: Who invented the telegraph?
  - **Candidate answer**: Samuel Morse
  - Text (snippets):
    - **Samuel Morse** had made history because he has **invented** the electric **telegrap** and the **Morse** alphabet.
  - Hipothesis (question+answer rewritten in affirmative):
    - "Samuel Morse invented the telegraph"
  - Validation:
    - Using textual entailment techniques we have to guess if the snippet (or document) satisfies the hipothesis.
      - Answer Validation Exercise (CLEF)
    - Correct answer

Examples of validation



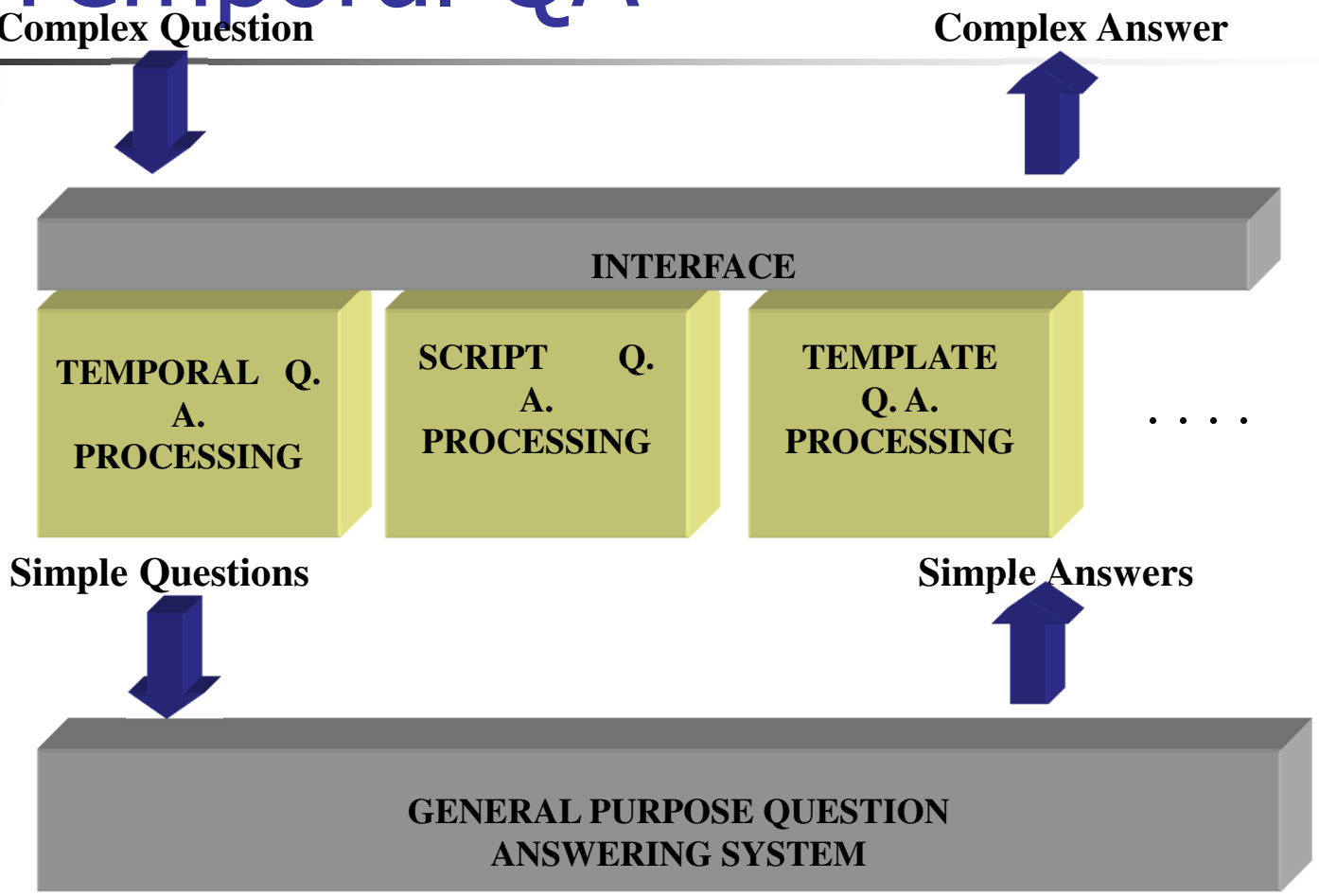
- Statistical information
  - **Question**: Who invented the telegraph?
  - **Candidate answer**: Samuel Morse
  - Query to the web: “Samuel” “Morse” “invent” “telegraph”
    - 9590 ocurrences
  - Validation:
    - The query occurs frequently
    - No need to download the document and check
    - Correct answer



Temporal QA

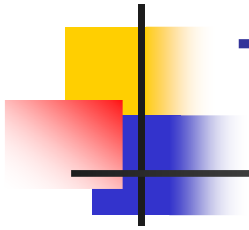
- There are two types of problems that are not being treated by QA systems at this moment:
  - Very complex questions, whose answer must be obtained by gathering information from different documents and ordered in time to get the correct answer: "Who was spokesman of the Soviet Embassy in Baghdad ***during*** the invasion of Kuwait?"
  - Questions that contain implicit temporal expressions that must be resolved in order to answer the questions. For example, comparing these two questions:
    - "Who was the President of USA ***five years ago***?"
    - "Who is the President of USA?"

Temporal QA


Complex Question
Complex Answer
INTERFACE
TEMPORAL Q. A. PROCESSING
SCRIPT Q. A. PROCESSING
TEMPLATE Q. A. PROCESSING
. . . .
Simple Questions
Simple Answers
GENERAL PURPOSE QUESTION ANSWERING SYSTEM

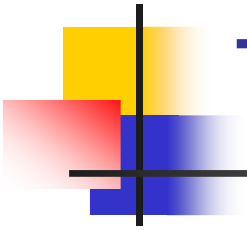
Temporal QA

- Advantages of multilayer architecture:
  - It allows to use any existent QA system →only adapting inputs and outputs
  - The QA system does not need to be modified because the processing of the complex question is in a superior layer
  - Each layer is independent of the other



Taxonomy temporal questions

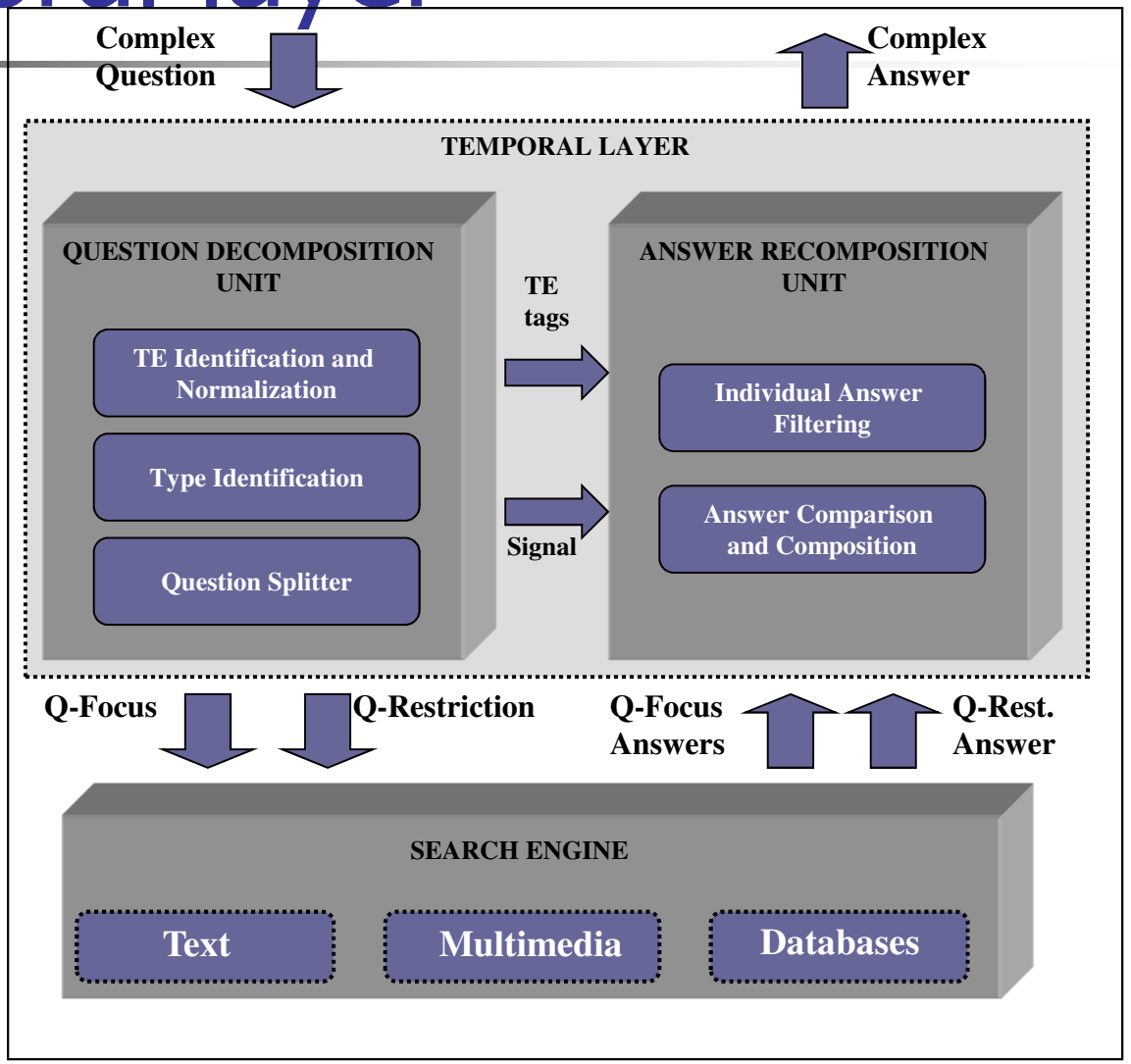
- Two types of questions
  - Simple temporal questions
  - Complex temporal questions
- Simple temporal questions:
  - Type 1: "When did Jordan close the port of Aqaba to Kuwait?"
  - Type 2: "Who won the *1988* New Hampshire republican primary?



Taxonomy temporal questions

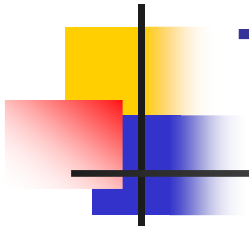
- Complex Temporal Questions:
  - Type 3: "What did George Bush do *after* the U.N. Security Council ordered a global embargo on trade with Iraq *in August 90*?"
  - Type 4: "What happened to world oil prices *after* the Iraqi annexation of Kuwait?"

Temporal layer



Temporal layer

- Question Decomposition Unit:
  - The types 3 and 4 are divided.
  - *"Where did Bill Clinton study before going to Oxford University?"* is divided into:
    - Q-Focus: *Where did Bill Clinton study?*
    - Q-Restriction: *When did Bill Clinton go to Oxford University?*

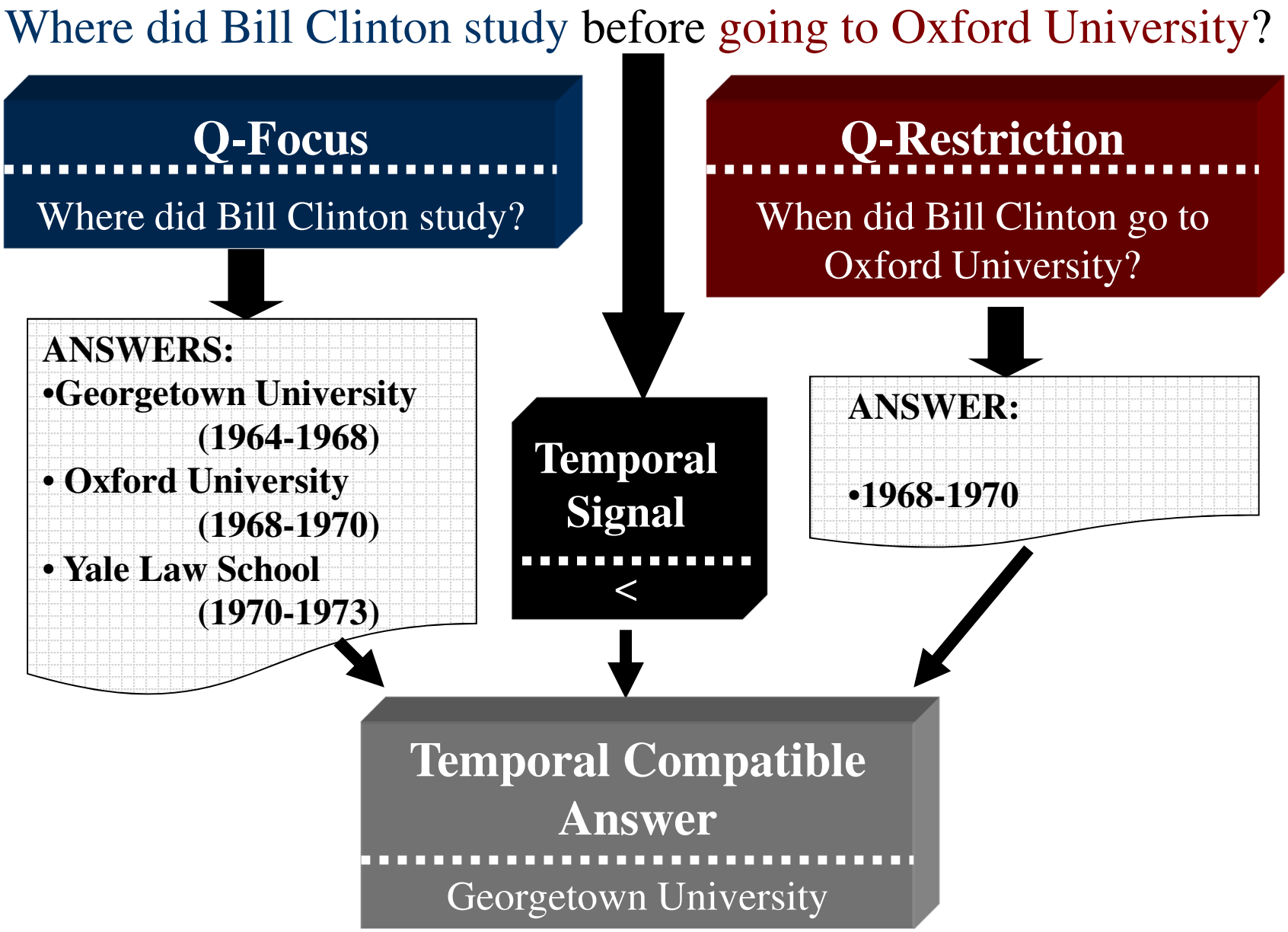


Temporal layer

- General Purpose QA *"Where did Bill Clinton study before going to Oxford University?"* :
  - Answer to the simple questions:
    - Q-Focus answers:
      - *Georgetown University (1964-68)*
      - *Oxford University (1968-70)*
      - *Yale Law School (1970-73)*
    - Q-Restriction answer: *1968*

Temporal layer

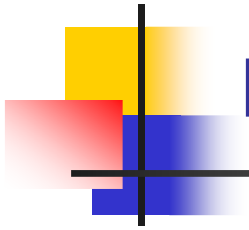
- Answer Recomposition Unit *"Where did Bill Clinton study before going to Oxford University?"* :
  - It builds the answer to the original question using the answers of the subquestions:
    - Respuesta: *Georgetown University (1964-68)*
  - Temporal compatibility: The only temporal compatible answer with the Q-Restriction answer is that one taking into account that the temporal signal is *"before"*


Where did Bill Clinton study before going to Oxford University?
Q-Focus
Where did Bill Clinton study?
Q-Restriction
When did Bill Clinton go to Oxford University?
ANSWERS:
•Georgetown University (1964-1968)
• Oxford University (1968-1970)
• Yale Law School (1970-1973)
Temporal Signal
<
ANSWER:
•1968-1970
Temporal Compatible Answer
Georgetown University

Decomposition of the question

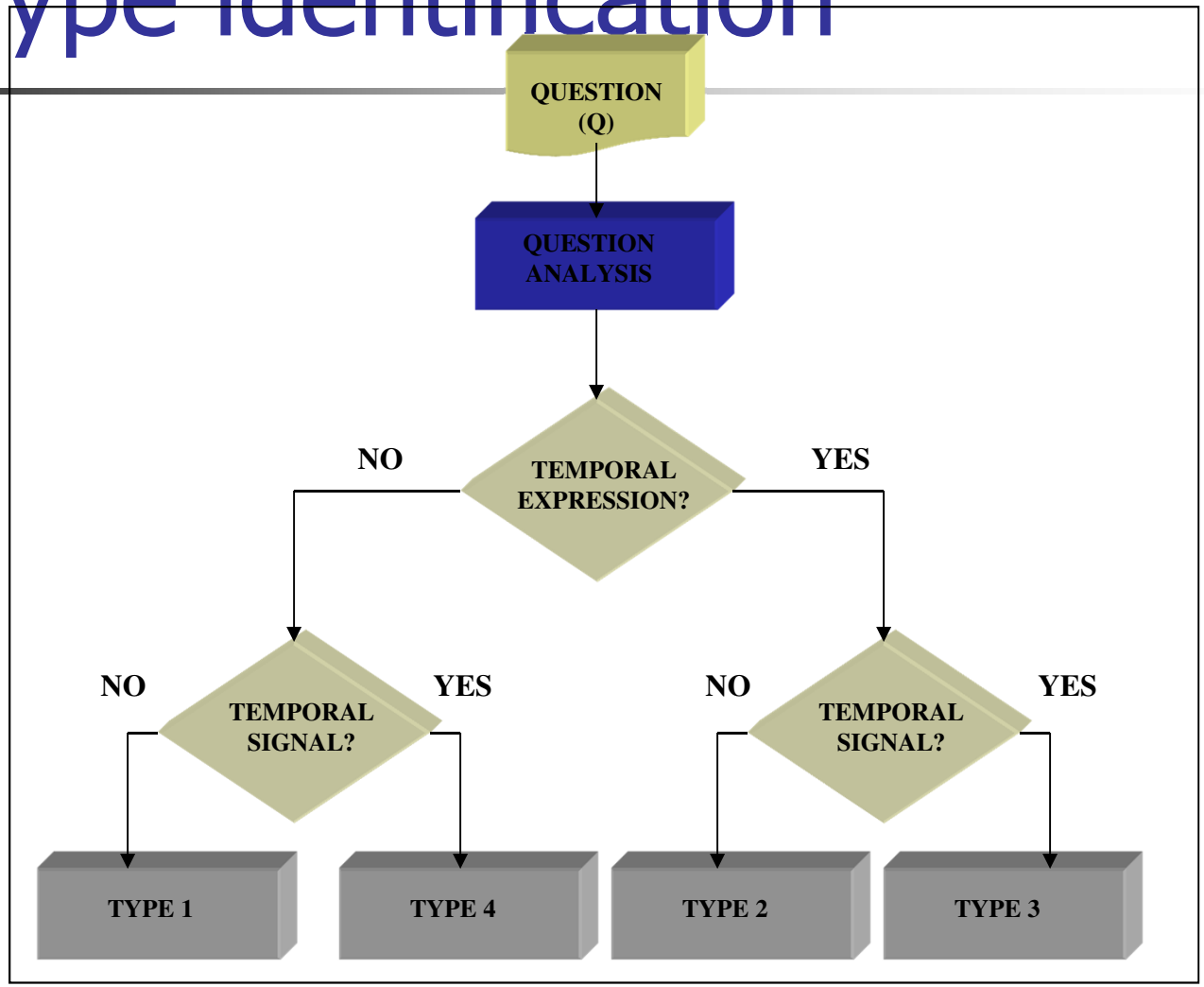
Modules in depth

TE identification and normalization



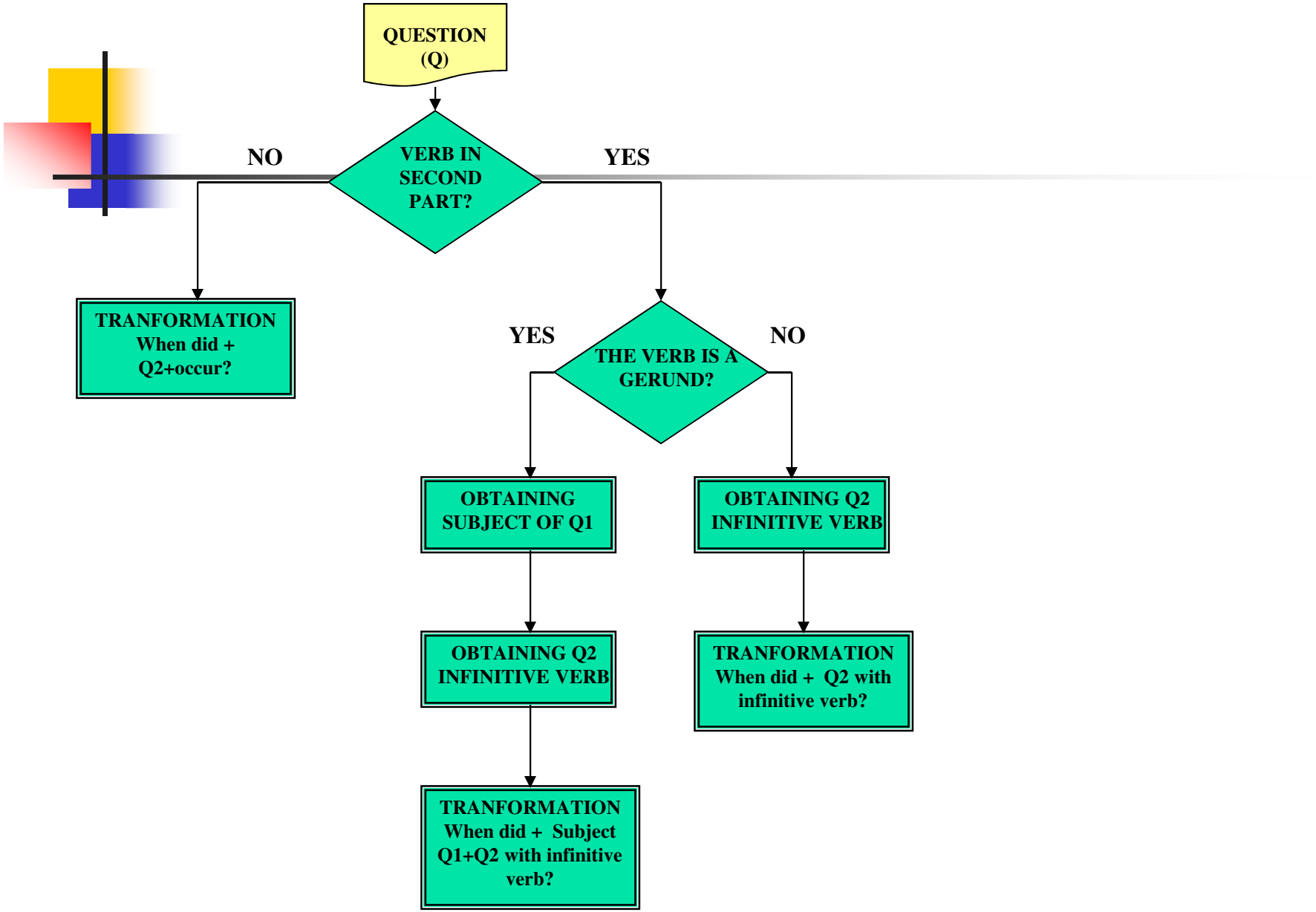
- This module is using TERSEO system to identify and resolve temporal expressions. The system is able to do these two task and in more than one language if necessary

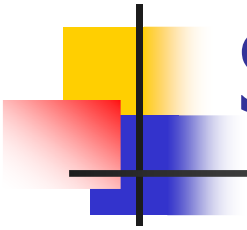
Type identification



Question Splitter

- It is based on the identification of the temporal signal that relates the two events and establishes a temporal order between them
- The complex questions are divided in two parts:
  - Q-Focus: what the user needs to find. We will get a set of answers from the system
  - Q-Restriction: the temporal restriction that must be fullfilled by the answer obtained for the Q-Focus. The answer must be unique


QUESTION (Q)
VERB IN SECOND PART?
NO
YES
TRANFORMATION When did + Q2+occur?
YES
THE VERB IS A GERUND?
NO
OBTAINING SUBJECT OF Q1
OBTAINING Q2 INFINITIVE VERB
OBTAINING Q2 INFINITIVE VERB
TRANFORMATION When did + Q2 with infinitive verb?
TRANFORMATION When did + Subject Q1+Q2 with infinitive verb?

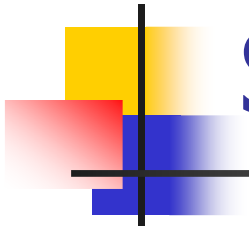


Splitting patterns

- First case
  - The questions that follows the temporal signal has no verb:
  - *"What happened to the world oil prices after the Iraqi invasion of Kuwait?":*
    - *Q-Restriction: When did the Iraqi invasion of Kuwait occur?*

Splitting patterns

- Second case
  - The questions that follows the temporal signal contains a gerund verb
  - *"Where did Bill Clinton study before going to Oxford University?":*
    - We extract the subject of the first part
    - We transform the verb of the second part to an infinitive
    - Q-Restriction: *"When did Bill Clinton go to Oxford University?"*



Splitting patterns

- Third case
  - The second part of the question contains a conjugated verb and its own subject
  - *"What did George Bush do after the U.N. Security Council ordered a global embargo on trade with Iraq?":*
    - We conjugate the verb of the Q-Restriction using the appropriate auxiliar verbs
    - Q-Restriction: *"When did the U.N. Security Council order a global embargo on trade with Iraq?"*

Temporal signals

SIGNAL	ORDERING KEY
After	F1 > F2
When	F1 = F2
Before	F1 < F2
During	F2i <= F1 <= F2f
Previously	F1 > F2
On/in	F1 = F2
While	F2i <= F1 <= F2f
For	F2i <= F1 <= F2f

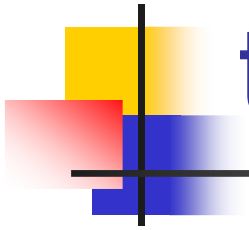
Temporal signals

SIGNAL	ORDERING KEY
After	F1 > F2
Before	F1 < F2
During	F2i <= F1 <= F2f
Previously	F1 > F2
On/in	F1 = F2
While	F2i <= F1 <= F2f
For	F2i <= F1 <= F2f

Reconstruction of the answer

Modules in depth

Preprocessing of the output of the QA system



- Te recomposition unit is waiting for answers with their related dates
- In order to obtain these dates, TERSEO system will be applied to the passages of the documents were the answer is found
- The system looks for the event in the document and obtains the near date to this event

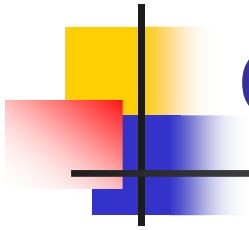
Individual Answer filtering

- For the subquestions with temporal expressions, only the answers that fullfill the temporal restrictions are obtained by the Recognition and Resolution TE unit
- Those answers that fullfill the temporal restrictions will go to the next module.

Individual Answer filtering

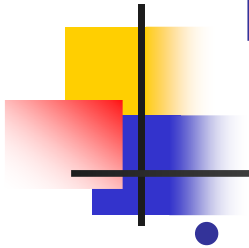
- example: *"Where did Bill Clinton study before going to Oxford University in the sixties?"*:
  - Answers of the QA system:
    - Answers Q1: *Georgetown University (1964-68) // Oxford University (1968-70) // Yale Law School (1970-73)*
    - Answers Q2: *1968*
  - Temporal restrictions:
    - **<DATETIMEREF valdate1="01/01/1960" valdate2="31/12/1969"> in the sixties </DATETIMEREF>**
  - The first two answers → possible answers, they will go to the next module

Answer Comparison and composition



- Using the ordering key, imposed by the temporal signal, the answer to the complex question will be obtained from the list of possible answers

Exercise 1 Temporal QA



- Annotate the question corpus following the previous approach

Exercise 2 Temporal QA

- Download the paper and prepare it:

- *http://www.timeml.org/site/terqas/documentation/TimeML-use-in-qa-v1.0.pdf*

- Group discussion

TEXTUAL INFORMATION EXTRACTION

Dr. Patricio Martínez Barco

Dr. Rafael Muñoz Guillena

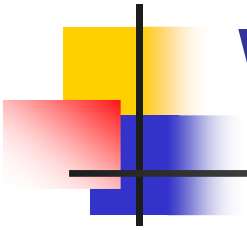
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Temporal representation and reasoning

Unit 5



What is temporal information?

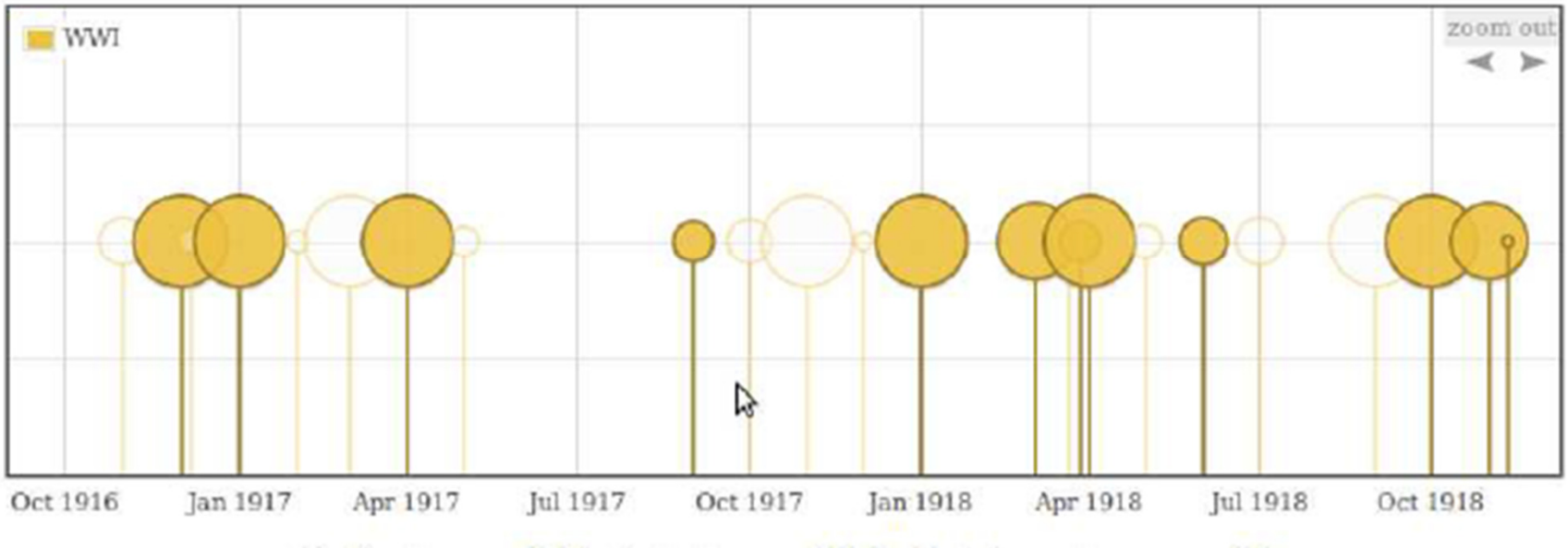
- Why to detect it?
- Temporal information in written texts:
  - Concrete dates: *10/10/2005*
  - Temporal expressions: *yesterday, the day before*
  - Verbal tenses: *do, did, will do*
  - Events: the meeting, crash
  - Temporal relationships

Objective of processing them

Locate the events of a text in a timeline

Cronological order

Objective of processing them



Example of TIP

He arrived in the bar at 8 p.m.

However, she had already left.

Therefore, he went back home,

after talking with some friends.

Example of TIP

He arrived in the bar at 8 p.m.

However, she had already left.

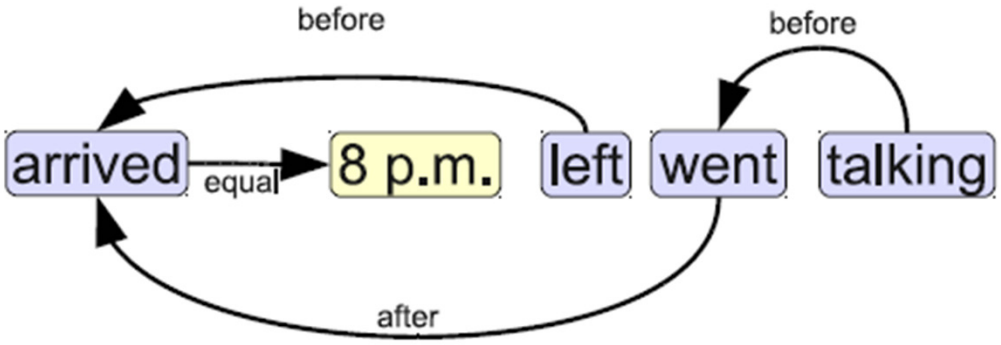
Therefore, he went back home,

after talking with some friends.

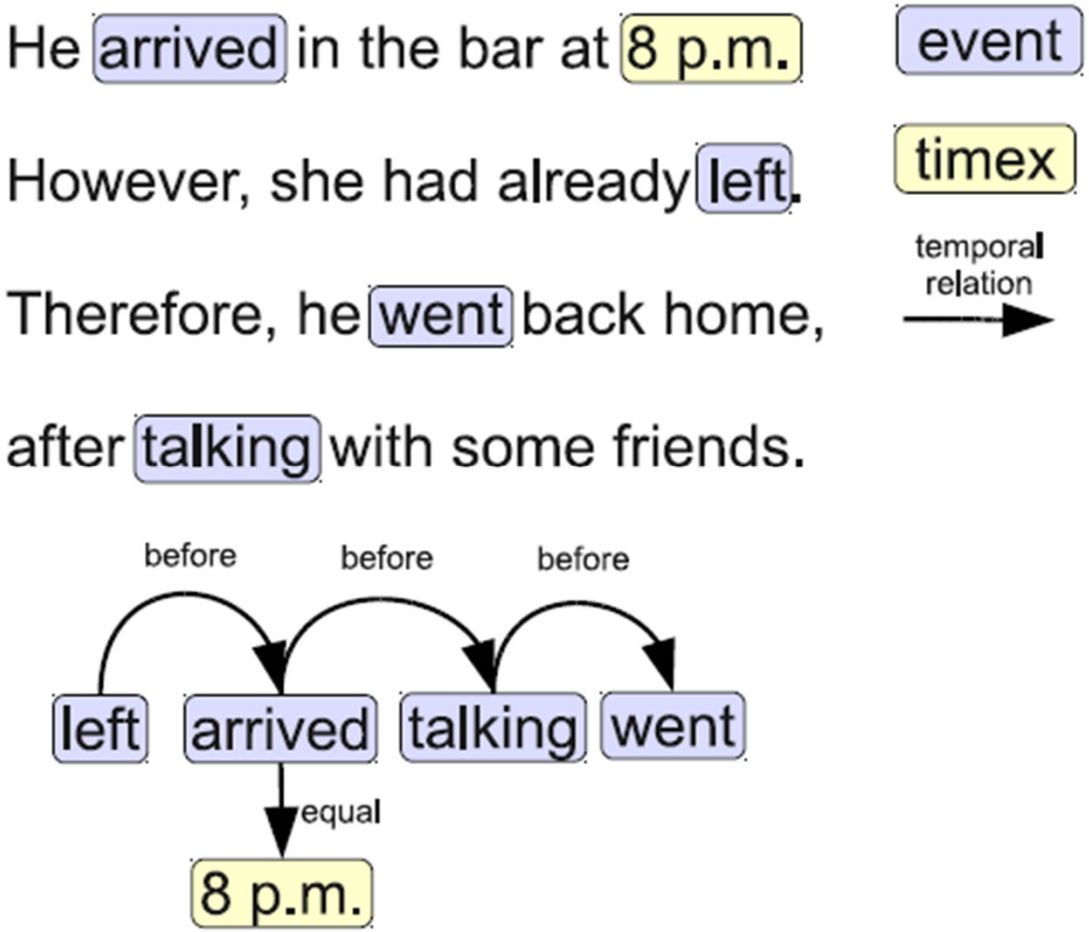
event

timex

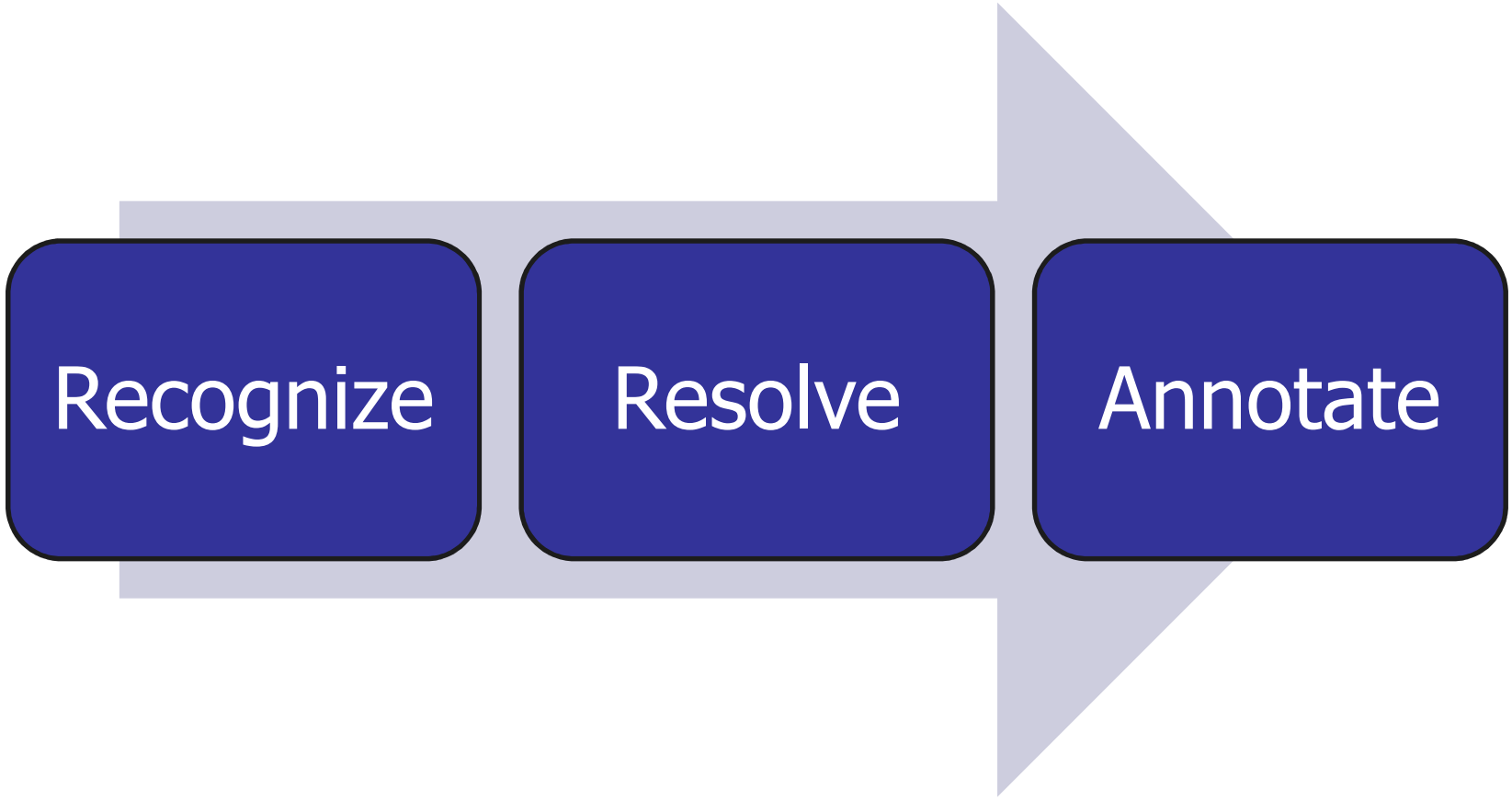
temporal relation

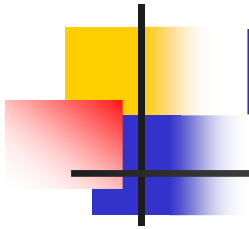


Example of TIP



Tasks to be performed



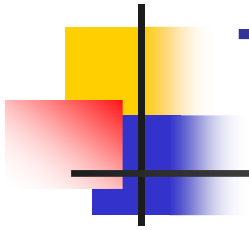


Practical example

- Given the following text:
  - Recognize temporal information
  - Resolve temporal information
  - Classify the temporal information in types
  - Define a possible annotation schema
- Individual work first
- Group discussion

Temporal annotation

- There are different approaches for the temporal problem, but all of the system must have a comparable output → standard of the output using annotation schemas

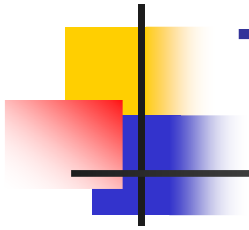


Temporal annotation

- Evolution of temporal schemas:
  - TIMEX: conferences MUC 6 and 7. First approaches to absolute and relative temporal expressions.
  - TIMEX2: produced by the TIDES program and tested in the TERN2004 conference. The aim of this schema was the normalization of expressions

Temporal annotation

- Evolution of the temporal schema:
  - TIMEML: standard ISO that covers TE, events and relationships between them. This is the most complete schema until this → TIMEBANK corpus



TIDES o TIMEX2

DOWNLOAD SCHEMA

http://www.timeml.org/site/terqas/readings/MTRAnnotationGuide_v1_02.pdf

Temporal annotation (TIDES)

- TIDES: developed by DARPA for:
  - Annotating the temporal expressions of a document→ TIMEX2 tag
  - Identifying the temporal value that the expression is representing:
    - VAL
    - MOD
    - ANCHOR_VAL
    - ANCHOR_DIR
    - SET
    - COMMENT

Temporal annotation(TIDES)

- VAL attribute can be classified in three types:
  - Time points: Exp. That are answers to When? questions
  - Durations: Exp. That are answers to How long? questions
  - Frequences: Exp. That are answers to How often? questions
- Take care with the granularity of the expression

Temporal annotation(TIDES)

- VAL: it contains a date and/or time
  - VAL="1965-10-16"
  - VAL="1965-10-16T08:00"
  - VAL="P5Y"
- Format:
  - Concrete dates:

  YYYY-MM-DDThh:mm:ss
  - Periods:

  PnYnMnDTHnMnS --- PnW (for weeks)

Temporal annotation(TIDES)

- MOD: it captures the temporla modifiers→ MOD=“APPROX”/”END”/”START”
- SET: it identifies expressions that denote sets→ YES/NO

Temporal annotation(TIDES)

- ANCHOR_VAL: it contains a normalized form of an anchor date/time → ANCHOR_VAL="1964". Same format as VAL
- ANCHOR_DIR: It captures the direction or relative orientation between VAL and ANCHOR_VAL

Temporal annotation(TIDES)

- ANCHOR_DIR: Possible values
  - WITHIN: the referent date is within the period indicated by the duration
  - STARTING: the period indicated in the duration starts with the referent date
  - ENDING: the period indicated in the duration ends with the referent date
  - AS_OF: the period of duration of the expression is present but indefined
  - BEFORE: the TE does not indicated when the duration starts/ends but the verbal tens (past) indicates that it was before the referent
  - AFTER: the TE does not indicated when the duration starts/ends but the verbal tens (future) indicates that it was after the referent

Temporal annotation(TIDES)

- Besides, there are fuzzy expressions whose values for VAL and ANCHOR_DIR as fixed:
  - val="PRESENT_REF" y anchor_dir="AS_OF" → now, today, presently, nowadays,…
  - val="FUTURE_REF" y anchor_dir="AFTER" →future, later, ahead, in a few years,..
  - Val="PAST_REF" y anchor_dir="BEFORE" → past, former, recently, long ago,…

Temporal annotation(TIDES)

- Examples of annotated expressions:
  - Dates and times:

<TIMEX2 val="2000-10-01T18:46:13.59">10/01/2000 18:46:13.59</TIMEX2>

- Durations:

<TIMEX2 val="P30Y" anchor_val="2008" anchor_dir="ENDING">the last 30 years</TIMEX2>

Temporal annotation(TIDES)

- Expressions with modifiers

  <TIMEX2 val="2000-W40" mod="END">late last week</TIMEX2>

- Expressions that use tokens to certain periods of time:

  <TIMEX2 val="2008-01-15TNI">Tuesday night</TIMEX2>

Temporal annotation(TIDES)

- Expressions with undefined values

<TIMEX2 val="PXY" anchor_val="2008" anchor_dir="ENDING">last years</TIMEX2>

- Sets

<TIMEX2 val="XXXX-XX-XX" set="YES">each day</TIMEX2>

Temporal annotation(TIDES)

<DOC><DATE_TIME>
<TIMEX2 val="2000-10-01T18:46:13.59">10/01/2000 18:46:13.59</TIMEX2> </DATE_TIME>
<BODY><TEXT>
it began when a hasidic jewish family bought one of the town's two meat-packing plants
<TIMEX2 val="1987">13 years ago</TIMEX2>. first they brought in other hasidic jews, then mexicans, palestinians, ukrainians.
postville **<TIMEX2 val="PRESENT_REF" anchor_val="2000-10-01T18:46:13.59" anchor_dir="AS_OF" comment="">now</TIMEX2>** has 22 different nationalities, but some resident haven't welcomed the newcomers.
</TEXT></BODY>
<END_TIME> **<TIMEX2 val="2000-10-01T18:48:39.30" >10/01/2000 18:48:39.30</TIMEX2>**
</END_TIME>
</DOC>

Temporal annotation(TIDES)

- Exercises:
  - Yesterday
  - The next three years
  - The past four days
  - Daily

Temporal annotation(TIDES)

- Solution:
  - Yesterday: **<TIMEX2 VAL="2007-02-04">**
  - The next three years: **<TIMEX2 VAL="P3Y" ANCHOR_VAL="2007" ANCHOR_DIR="STARTING">**
  - The past four days **<TIMEX2 VAL="P4D" ANCHOR_VAL="2007-02-05" ANCHOR_DIR="ENDING">**
  - Daily: **<TIMEX2 val="XXXX-XX-XX" mod="" set="YES" non_specific="" anchor_val="" anchor_dir="" comment="">**

Temporal annotation(TIDES)

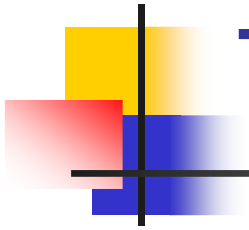
- Complex expressions without embedding:
  1. **Conjoined Expressions: <TIMEX2 VAL="2009-01-19">today</TIMEX2> and <TIMEX2 val="2009-01-20">tomorrow</TIMEX2>**
  2. **Exprerssions that are treated as a unit: <TIMEX2 VAL="XXXX-12-02>the second of December</TIMEX2>**

Temporal annotation(TIDES)

- Complex expressions with embedding:
  1. **Time-Anchored Expressions: <TIMEX2 VAL="2009-02-03">two weeks from <TIMEX2 val="2009-01-20">next Tuesday</TIMEX2></TIMEX2>**
  2. **Possessive Constructions: <TIMEX2 VAL="2009-SU"><TIMEX2 VAL="2009">this year</TIMEX2>'s summer</TIMEX2>**
  3. **Pre-Modifier and head triggers: <TIMEX2 VAL="P1D">a <TIMEX2 VAL="PT12H">12 hour</TIMEX2>day</TIMEX2>**

Temporal annotation(TIDES)

- Exceptionss:
  - **<TIMEX2 VAL="1998-WXX-4" SET="YES">Some Thursdays</TIMEX2> in <TIMEX2 VAL="1998">1998</TIMEX2>**
  - **There is a value dependency but are annotated as independent expression in their extension**

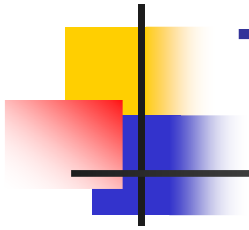


Temporal annotation(TIDES)

- exercise: Using annotation guidelines of TIDES, determine the problems associated to:
  - **Cultural expressions with a possesive: "next year's Christmas"**
  - **Two expressions joined by a conjunction and anchored to a third one: "six months or a year from now"**
  - **Temporal expressions anchored to events: "two days after the disaster"**

Temporal annotation(STAG)

- STAG: part of the thesis of Setzer 2001
- It includes events and temporal relationships, apart from temporal expressions:
  - Event
  - TIMEX
  - Signal

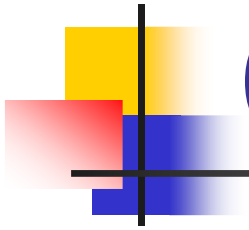


TIMEML

DOWNLOAD SCHEMA

http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf

Temporal annotation (TIMEML)



- This is the most complete schema dn the most standarized:
  - Temporal expressions: TIMEX3
  - Events: EVENT
  - Signals: SIGNAL (“on”,”during”)
  - Temporal relationships: TLINK (temporal), ALINK (aspectual), SLINK (subordinated)

Temporal annotation(TIMEML)

- TLINK: relationship between two temporal expressions or and event and a temporal expression
- ALINK: aspectual relationship between two events: “began to sink”, “stop looking”. They are initiation, culmination,etc relationships.
- SLINK: relationship established byt verbs like want, deny, believe plus another verb.

Temporal annotation(TIMEML)

- Atributos TIMEX3
  - ID: compulsory and automatic attribute
  - Type:
    - DATE: “October of 1963”
    - TIME: “half past noon”
    - DURATION: “2 months”
    - SET: “every two days”

Temporal annotation(TIMEML)

- Atributos TIMEX3
  - Value: equivalent to VAL attribute of TIMEX2
  - Mod: equivalent to MOD in TIMEX2
  - temporalFunction: binary attribute that express if the value of a expression must be obtained through a temporal function: “last week”→true, “January 31, 1999” → false

Temporal annotation(TIMEML)

- Atributos TIMEX3
  - anchorTimeID: optional and denotes when a expression is anchored to another.
    - Example: <TIMEX3 tid="t11" type="DATE" value="2008-11-01" temporalFunction="true" anchorTimeID="t0">today</TIMEX3>

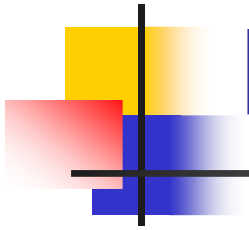
Temporal annotation(TIMEML)

- Atributos TIMEX3
  - beginPoint y endPoint: optional and denotes a duration:
    - Example: <TIMEX3 tid="t1" type="DURATION" value="P6M" beginPoint="t3" endPoint="t2">six months</TIMEX3> until <TIMEX3 tid="t2" type="DATE" value="2009-03-31">March 31, 2009</TIMEX3>

      <TIMEX3 tid="t3" type="DATE" value="2009-09-30" anchorTimeID="t6"/>

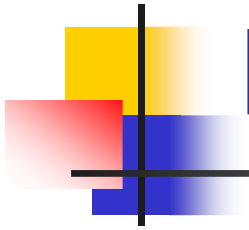
Temporal annotation(TIMEML)

- Atributos TIMEX3
  - Quant y freq: optional, it is used only en set expressions:
    - <TIMEX3 tid="t1" type="SET" value="P1W" freq="2x"> twice a week </TIMEX3>
    - <TIMEX3 tid="t1" type="SET" value="P2D" quant="EVERY"> every 2 days </TIMEX3>



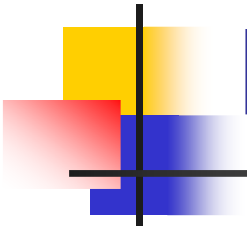
Events

- We distinguish between events and states (dictionary definition) :
  - Event: “a thing that happens”
  - State: “the condition or position existent of a person or a thing”
- As can be observed, defining exactly what is an event or an state is a difficult task



Events

- State (Setzer, 2000)
  - Entity or relation between entities that is able to change but in the moment of the observation is keeping the same form during a time, and very often without starting or ending point
  - Typically, a change of state denotes and event
  - Example: *"Elvis está vivo"*



Events

- Event (Setzer, 2000)
  - Something that happens, with a defined starting and ending point
  - Example: *"a plane has crashed in the Atlantic Ocean"*

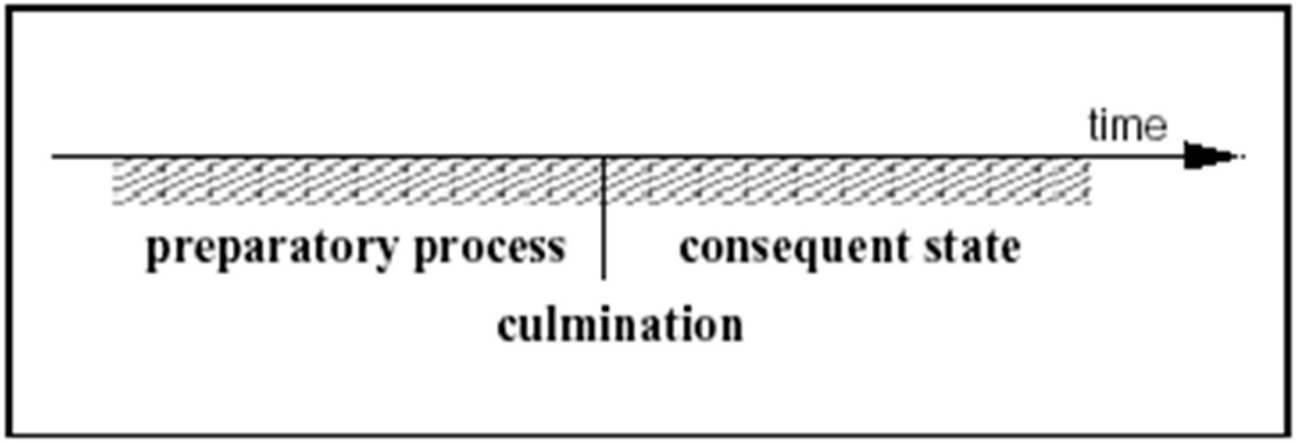
Events

- Classification of event according to Moens and Steedman, 1988

	events		states
	atomic	extended	
+conseq	**CULMINATION**	**CULMINATED PROCESS**	
	recognise, spot, win the race	build a house, eat a sandwich	understand,love, know, resemble
-conseq	**POINT**	**PROCESS**	
	hiccup, tap, wink	run, swim, walk, play the piano	

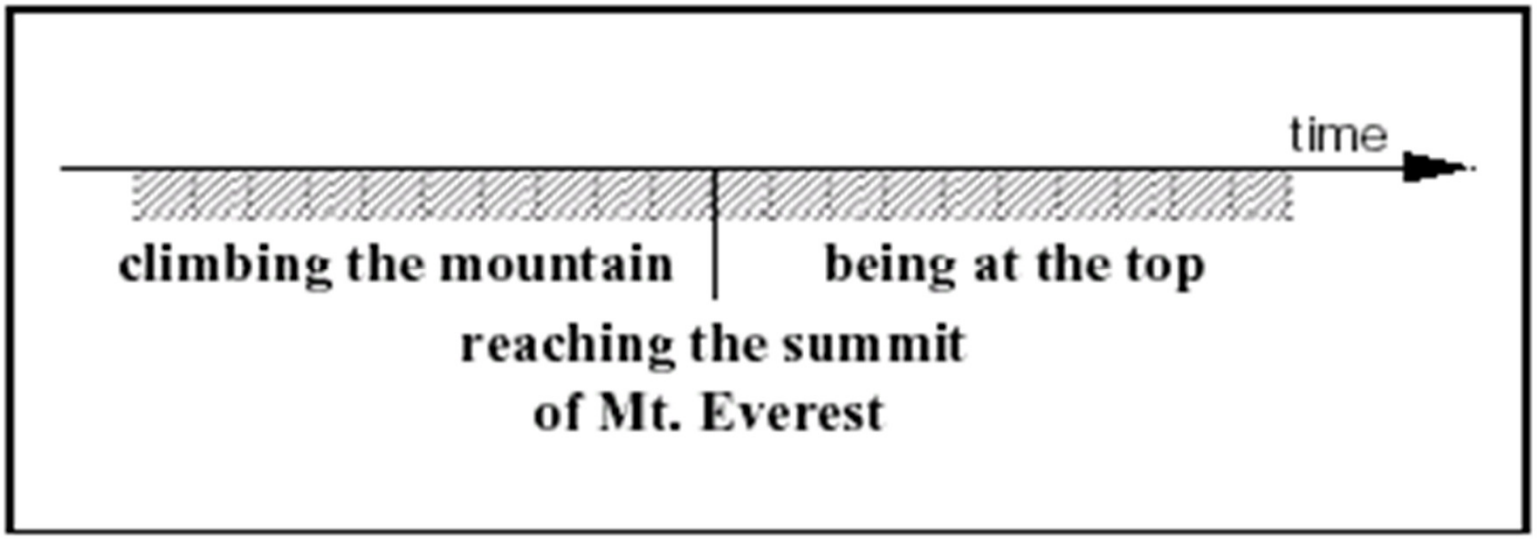
Events

- Moens and Steedman defined also the concept of event core, with the following parts:

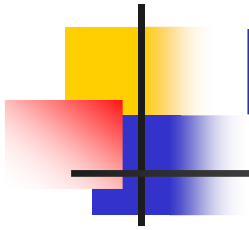


Events

- Here is an example of event:



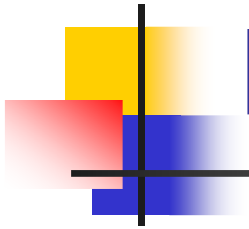
Expressing events in NL



- The events can be expressed as verbs:
  - *The plane was introduced in the hangar.*
- Nominalizations that express verbs:
  - *The cause of the crash …*
- However, the presence of these items are not always an event. Due to this fact is difficult of detect it automatically:
  - *The house is 55 miles from the island*

Expressing events in NL

- Therefore, the states can be expressed with verbs
- Besides, what in one context is considered an event, in another can change:
  - *For example:*
    - *Rajoy is the president of Spain now:* State
    - *Rajoy was elected president of Spain in 2011:* Event



Expressing events in NL

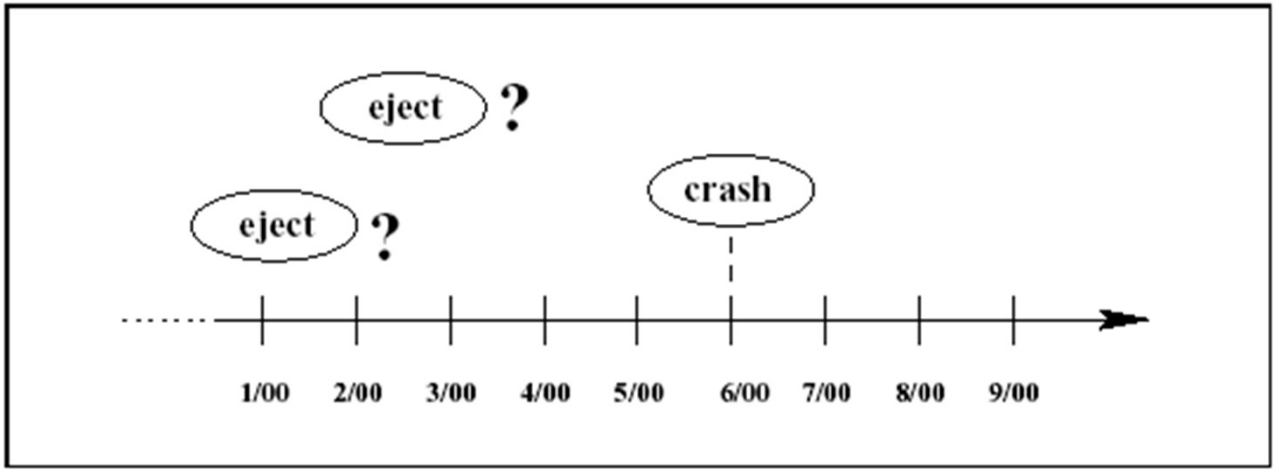
- This means that it is not possible to distinguish events only with syntatic information, becuase the Verb category can be for events or states
- It is also necessary some semantic information about the verb. Not all the classes of verbs are marked as events

Expressing events in NL

- Besides verbs can be events but also:
  - Reporting events: The government said yesterday that…
  - Reported events: the prices will raise this year 15%
- Our purpose is to annotate reported events that are the actions that happen and they can later be used in QA systems.

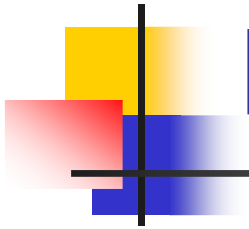
Relationship between events

- Apart from determining the events that can be found, it is also important to establish the cronological order between them and locating them in the timeline



Relationship between events

- The problem is that sometimes we don´t have the precise information in order to locate the events in the timeline exactly.
- If the event is associated to a temporal expression it can be located without problems
- But, if we find expressions like: "*The accident was December 25th. Sometime later there was an accident"*

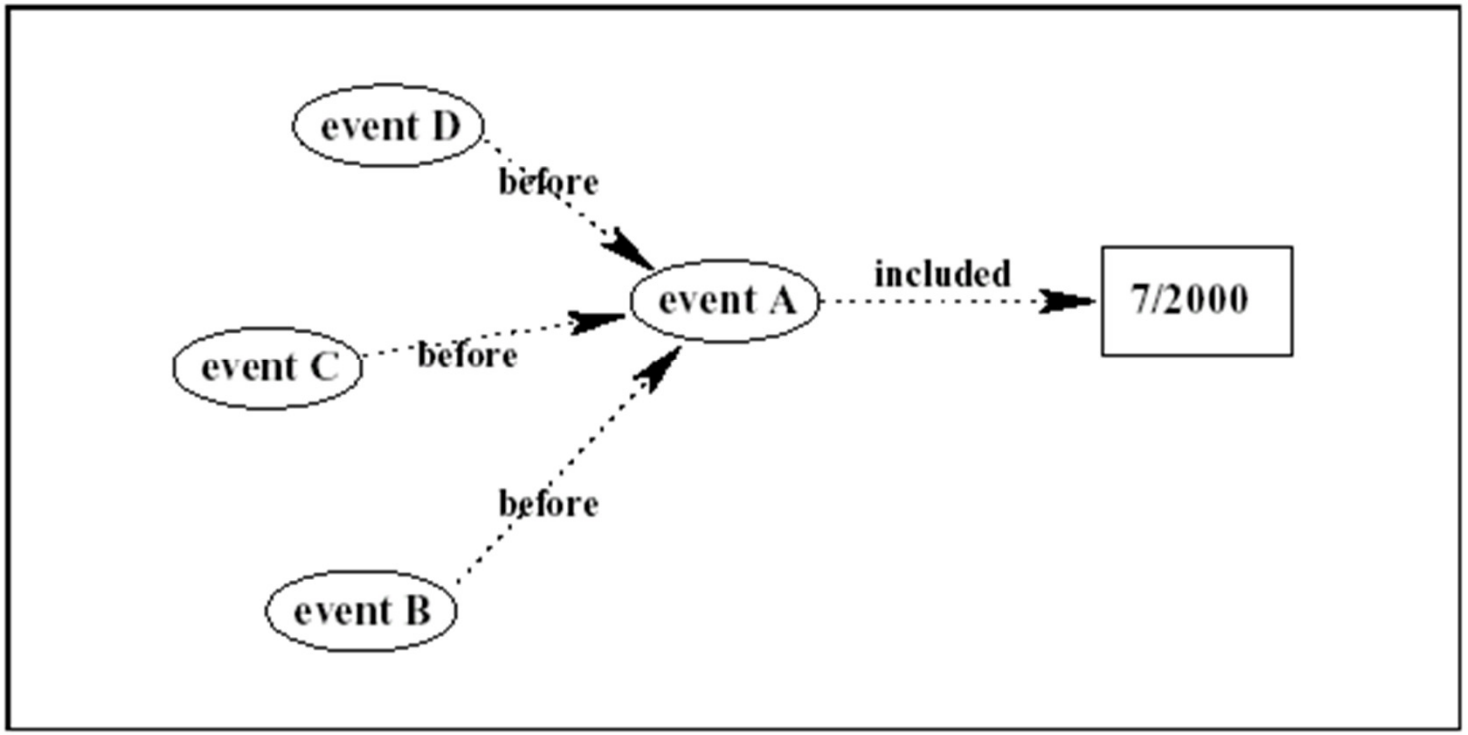


Relationship between events

- *"The accident was December 25th. Sometime later there was an accident"*
- In this example we have two events:
  - Event 1: First accident (12/25)
  - Event 2: Second accident(>>> 12/25)
- It is not possible to exactly locate the second event in the timeline
- Possible solution?

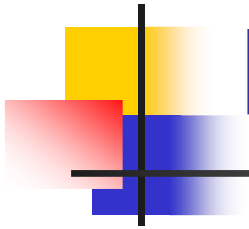
Relationship between events

- A possible solution is using a graph:



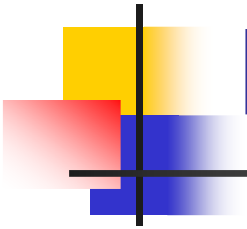
Relationship between events

- Relationships betweem events are usually express uisng temporal conjuctions such as:
  - before
  - after
  - during
  - Included/excluded
- These conjunctions are denoting the cronological order between the events → temporal relationships between events



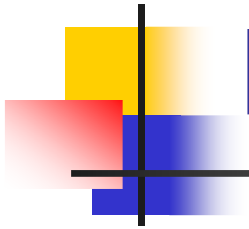
Possible solutions

- In order to detect events, if we have a annotated corpus with these events, ML systems can be trained in order to recognize new events.
- This is Setzer´s solution



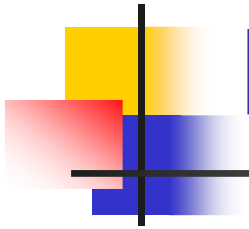
Event annotation

- In order to annotate events there are different approaches
- The simplest one is to annotate the complete sentence where the event is found
- However, the standard annotation in TIMEML proposed a more detailed events annotation.
- At first, with TIMEX2, events were not annotated



Event annotation

- TIMEML proposes the following advances from TIMEX2:
  - It introduces the EVENT concept
  - It introduces the SIGNAL concept, in order to tag the signals that goes with a temporal expression and that are not part of it, but can modify the final meaning of the expression
  - It introduces the MAKEINSTANCE concept, to events that are repeated or happen frequently in time



Event annotation TIMEML

- EVENT tag:
  - Situations that happen are considered events
  - Events can be points of time or periods
    - Points: Magallanes **got** the islands
    - Period: Companies **are growing** on the Internet

Event annotation TIMEML

- How we annotate events? Generally events are expressed as:
  - Infinitive, gerund and participle verbs:
    - A fresh flow of lava, gas and debris **erupted** there Saturday.
    - Prime Minister Benjamin Netanyahu called the prime minister of the Netherlands **to thank** him for thousands of gas masks his country has already contributed.

Event annotation TIMEML

- How we annotate events?
  - Nominalization
    - Israel will ask the United States to delay a military **strike** against Iraq until the Jewish state is fully prepared for a possible Iraqi **attack**.
  - Adjectives:
    - **…** after a Philippine volcano **dormant** for six centuries began exploding with searing gases, thick ash and deadly debris.
  - Predicative clauses:
    - "There is no reason why we would not **be prepared**," Mordechai told the Yediot Ahronot daily.

Event annotation TIMEML

- How we annotate events?
  - Prepositional sentences:
    - *All 75 people **on board** the Aeroflot Airbus died.*
- The annotation for simple events as verbs, names and adjectives is:

  A fresh flow of lava, gas and debris

  <EVENT eid=1> erupted </EVENT>

  there Saturday.

Event annotation TIMEML

- However, in the case of more complex events, it presentation is text can also be discontinuous. For example:
  - There is no reason why we would not **be prepared.**
    There is no reason why we would not **be** fully **prepared.**
  - They will definitely **take into consideration** our readiness

    They will definitely **take** it **into consideration**.

Event annotation TIMEML

- In order to avoid problems when tagging the following strategies can be used:
  1. Si el evento está formado por un cluster verbal, sólo se anotaría la cabeza de dicho cluster: *Israel has been **scrambling** to **buy** more masks abroad.*
  2. Si tenemos un phrasal verb, sólo se anota el verbo y no la preposición: *Additional distribution centers would be **set** up next week.*

Event annotation TIMEML

3. If the event is a name, but appears with more elements in the noun phrase, only the name that denotes the event is tagged: *The young industry's rapid **growth** also is attracting regulators eager to police its many facets.*
4. If the event is a predicative clause only the predicative element is tagged: *There is no reason why we would not be **prepared**.*
5. If the event is expressed with a prepositional sentence only the nominal part is tagged: *All 75 people on **board** the Aeroflot Airbus died.*

Event annotation TIMEML

- EVENT tag attributes:
  - Id of the event(eid): Compulsory. Number that identifies the event. It is automatically assign to each recognized event
  - Clase (class): Compulsory. It indicates the type of the event:
    - STATE
    - REPORTING
    - ASPECTUAL
    - I_ACTION I_STATE
    - OCURRENCE
    - PERCEPTION

Event annotation TIMEML

- CLASS attribute:
  - STATE: We tagged as states those circumstances that are not persistent, so, they can change and not being states anymore:
    - No **injuries** were reported over the weekend.
  - REPORTING: Describe actions of a person or organization declaring something, explaining an event, etc.
    - No injuries were **reported** over the weekend.

Event annotation TIMEML

- CLASS attribute:
  - ASPECTUAL: aspects of the verb:
    a) Initiation: *begin*, *start*, *commence, set out, set about, lead off, originate, initiate*
    b) Termination: *stop*, *end*, *halt*, *terminate*, *cease*, *discontinue*, *interrupt*, *quit*, *give up*, *abandon*, *block*, *break*, *lay off*.
    c) Completion: *finish*, *complete*.
    d) Continuation: *continue*, *keep, go on, proceed, go along, carry on, uphold, bear on, persist, persevere.*

Event annotation TIMEML

- CLASS attribute:
  - I_ACTION I_STATE: It is denoted by verbs like *believe, think, expect, suspect, fear, want, hope, appear, seem* , *promise, agree, offer, ask*, …I_ACTION for dynamic and I_STATE for static
  - OCURRENCE: In this class, we have all the types of events that describe things that happen
    - *Two moderate **eruptions** shortly before 3 p.m. Sunday appeared to **signal** a larger **explosion**.*

Event annotation TIMEML

- CLASS attribute:
  - PERCEPTION: It refers to events defined by verbs like *see, watch, glimpse, behold, view, hear,listen,…*

Event annotation TIMEML

- EVENT attributes:
  - Tense: Compulsory. Possible values are:
    - PAST: *No injuries were **reported** over the weekend.*
    - PRESENT: *The young industry's rapid growth also is **attracting** regulators eager to police its many facets.*
    - FUTURE: *"Anything along its path will be **destroyed**."*
    - IRREALIS: Switzerland offered to **lend** Israel another 25,000 masks
    - NONE: Se usa especialmente para nominalizaciones, adjetivos o infinitivos, participios y gerundios: *The **evacuation** was to take four hours, he said.*

Event annotation TIMEML

- EVENT attributes:
  - Aspect : Optional. It will be marked when possible and it has two values:
    - PERFECTIVE: *Philippines officials earlier had **ordered** the evacuation of more than 11,000 people.*
    - PROGRESSIVE: *The volcano began **showing** signs of activity in April for the first time in 600 years,...*

Event annotation TIMEML

- EVENT attributes:
  - Polarity: Compulsory. It has two values:
    - POSITIVE: *Villagers from a 12-mile radius of the mountain **fled** the area on foot.*
      *All 75 people on board the Aeroflot Airbus **died**.*
    - NEGATIVE: *"There is no reason why we would not be **prepared**," Mordechai told the Yediot Ahronot daily.*
      *No **injuries** were reported over the weekend.*

Event annotation TIMEML

- Complete syntax of EVENT tag:

attributes ::= eid class tense [aspect] polarity

eid ::= <integer>

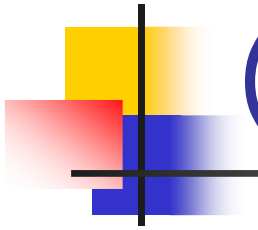
class ::= 'OCCURRENCE' |'STATE' | 'REPORTING' |
'ASPECTUAL' | 'I_ACTION‘ | ‘I_STATE’ |
PERCEPTION’

tense ::= 'PAST' | 'PRESENT' | 'FUTURE' |
'IRREALIS' | 'NONE'

aspect ::= 'PROGRESSIVE' | 'PERFECTIVE'

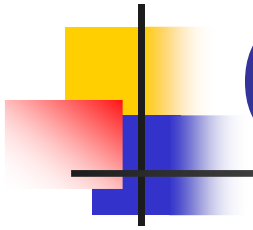
polarity ::= 'POSITIVE' | 'NEGATIVE'

Event annotation TIMEML (exercises)



- *The young industry's rapid growth also is attracting regulators eager to police its many facets.*

Event annotation TIMEML (exercises)



The young industry's rapid

<EVENT eid=1 class=OCCURRENCE tense=NONE polarity=POSITIVE> growth </EVENT>

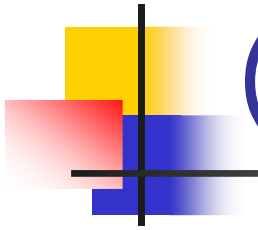
also is

<EVENT eid=2 class=OCCURRENCE tense=PRESENT aspect=PROGRESSIVE polarity=POSITIVE> attracting </EVENT> regulators

<EVENT eid=4 class=STATE polarity=POSITIVE> eager </EVENT> to

<EVENT eid=5 class=OCCURRENCE tense=IRREALIS polarity=POSITIVE> police </EVENT> its many facets.

Event annotation TIMEML (exercises)



- *Several pro-Iraq demonstrations have taken place in the last week.*

Event annotation TIMEML (exercises)

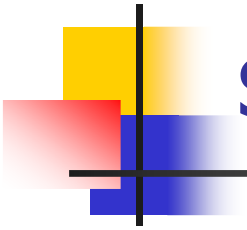
Several pro-Iraq

<EVENT eid=1 class=OCCURRENCE tense=NONE polarity=POSITIVE> demonstrations </EVENT> have

<EVENT eid=2 class=OCCURRENCE tense=PAST aspect=PERFECTIVE polarity=POSITIVE> taken </EVENT>

place in the last week.

Temporal Annotation (another schemas)



- TERSEO annotation schema: It allows to annotate expressions with its specific values for the dates that the expression refers to.
- It defines two tags with the same type of attributes:
  - DATETIME: for explicit expressions
  - DATETIMEREF: for implicit expressions

Temporal Annotation (another schemas)

- Explicit expressions

<DATE_TIME ID="value"

TYPE="value" (Concrete,Period,Fuzzy)

VALDATE1="value"

VALTIME1="value"

VALDATE2="value"

VALTIME2="value"

VALORDER="value">
Expression</DATE_TIME>

Temporal Annotation (another schemas)

- Implicit expressions

<DATE_TIME_REF ID="value"

TYPE="value" (Concrete,Period,Fuzzy)

VALDATE1="value"

VALTIME1="value"

VALDATE2="value"

VALTIME2="value"

VALORDER="value">
Expression</DATE_TIME_REF>

Temporal Annotation (another schemas)

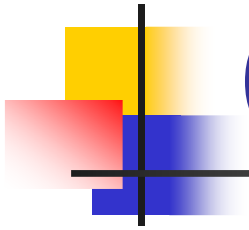
- *Explicit expressions*

  <**DATE_TIME ID=**1 **TYPE=**"CONCRETE"
  VALDATE1="06/14/2005"
  VALTIME1="19:00">
  June 14th ,2005 at 7 in the evening
  </**DATE_TIME**>

- *Implicit expressions DateP=06/14/2005*

  <DATE_TIME_REF ID=2 **TYPE=**"CONCRETE"
  VALDATE1="06/13/2005**">**
  yesterday
  </DATE_TIME_REF>

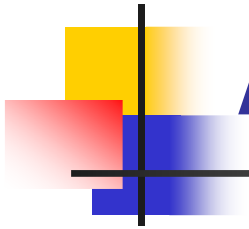
Temporal annotation (TERSEO)



- Exercises:
  - Yesterday
  - The next three years
  - The past four days
  - The last years

Temporal annotation (TERSEO)

- Solution:
  - Yesterday: **<DATETIMEREF type="C" valdate1="2011-12-21">**
  - The next three years: **<DATETIMEREF type="R" valdate1="2012" valdate2="2015">**
  - The past four days **<DATETIMEREF type="R" valdate1="2011-12-18" valdate2="2011-12-21">**
  - The last years **<DATETIMEREF type="F" valdate1=">>>2001" valdate2="<<<2011">**



Annotation exercise

- Annotate the delivered text with:
  - TIMEX2
  - TIMEML
- Correction in group

Textual information extraction

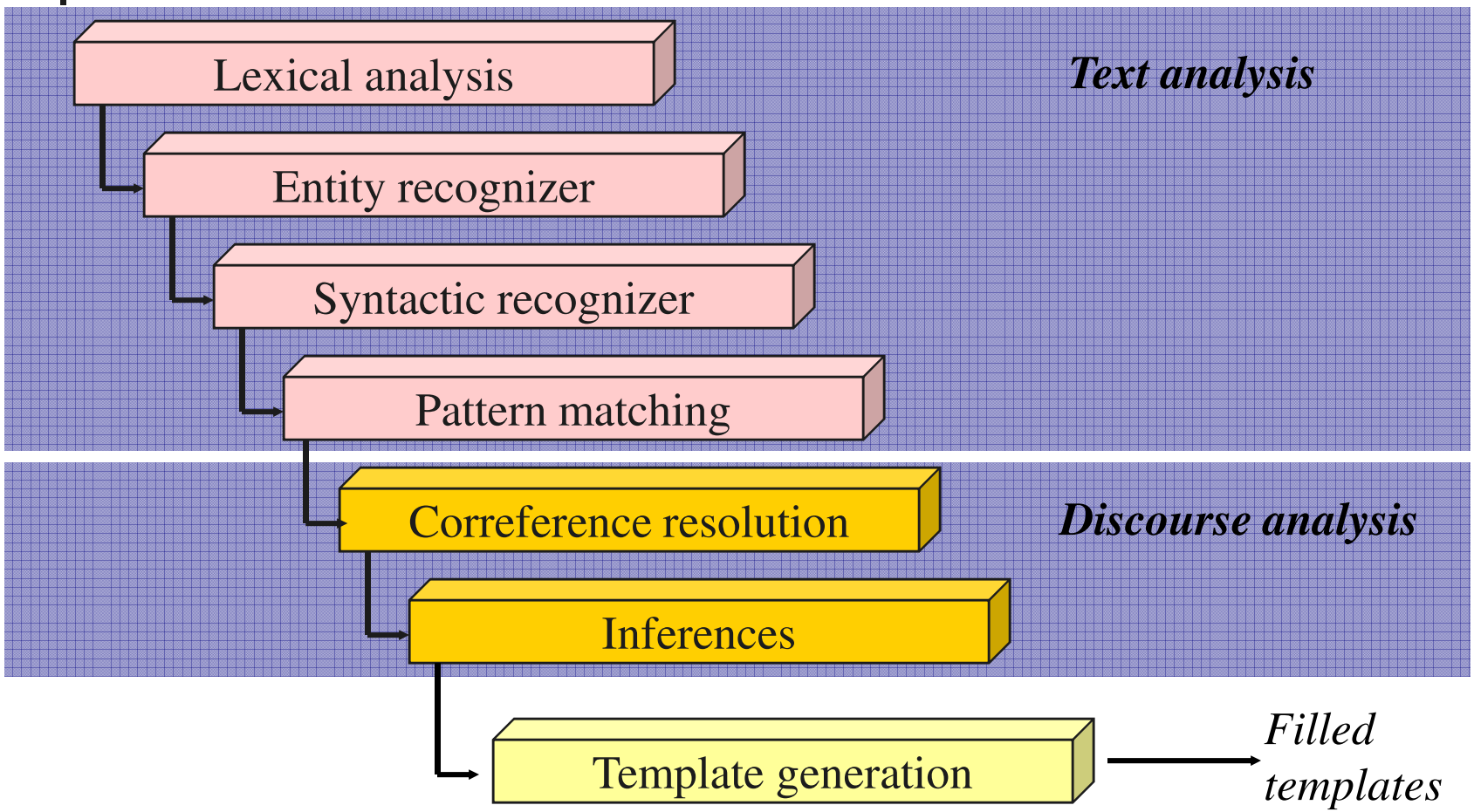
Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

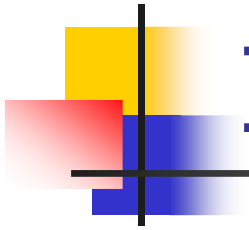
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Information Extraction: Correference Resolution

Unit 5

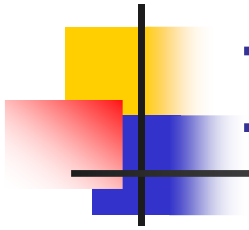
Introduction





Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution

What is the anaphora?

DEFINITION (Hirst, 1981)

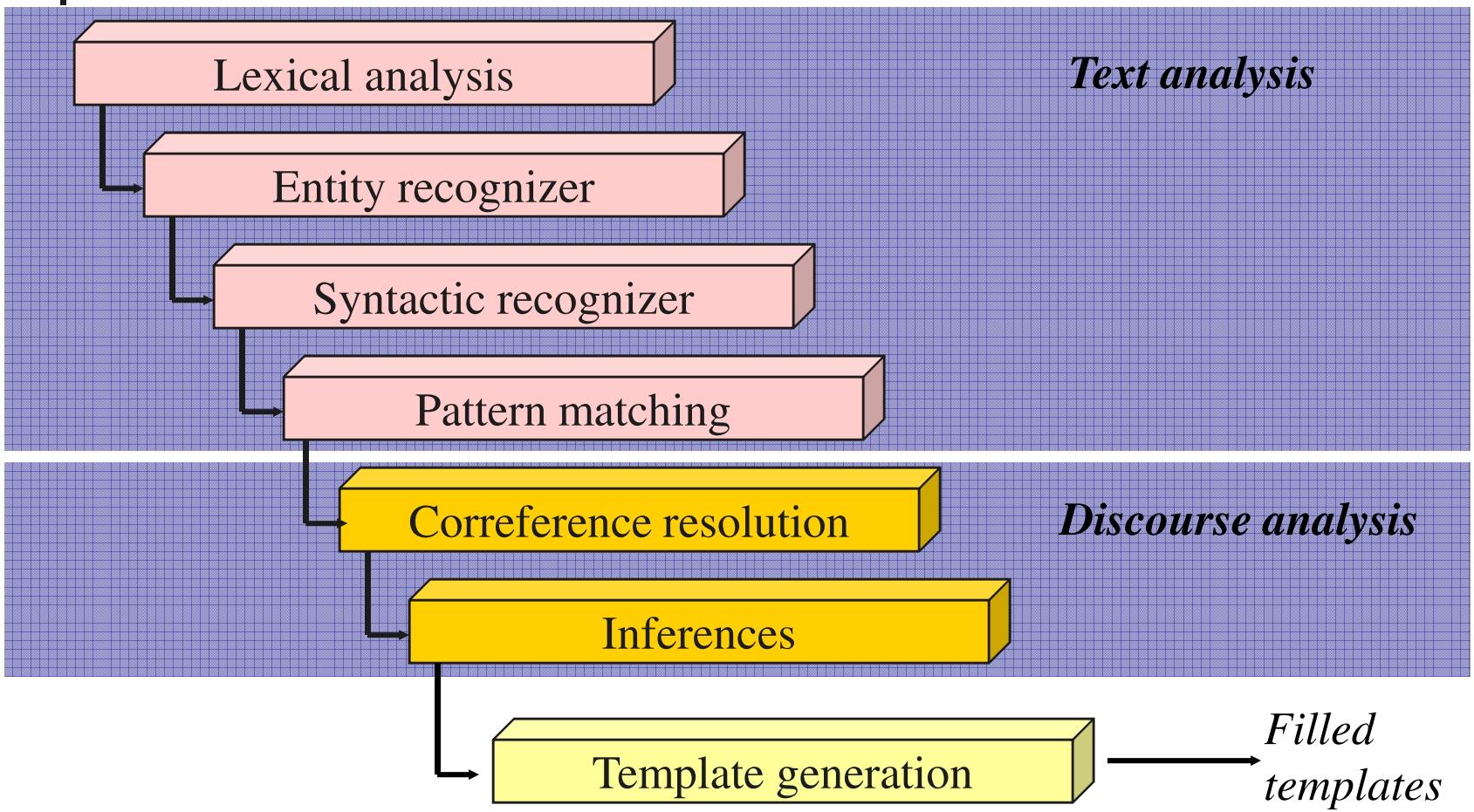
"Mechanism that allows to make in a discourse an abbreviated reference to any entity or entities with the confidence that the receiver of the discourse is able to interpret the reference and determines the entity that is referring to."

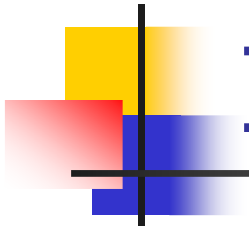
The girl listened the explanations of the lecturer.

She was always interested in his words

What is the anaphora?

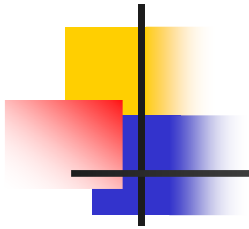
The anaphora in the Information Extraction





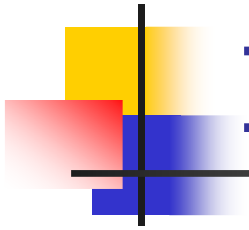
Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



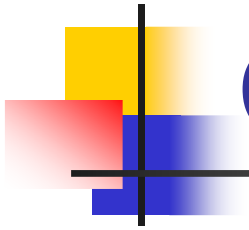
Why is it important to resolve it

- Discourse phenomena ⟹ coherence in the text.
- It is important to resolve it in order to extract implicit information from the text.
- It avoids the overgeneration of templates



Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



General concepts

- Referent
- Correferent
- Endophora
- Exophora
- Correference strings

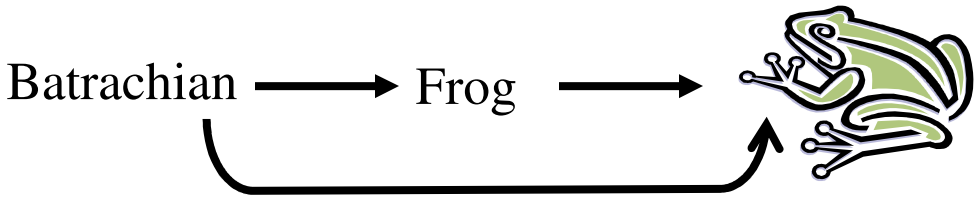
General concepts

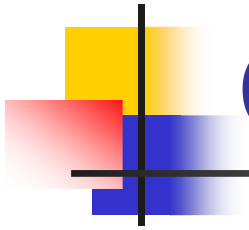
- Referent
  - Symbolic relationship existent between a linguistic expression and the specific or abstract object the it represent.

Frog

General concepts

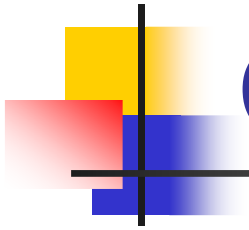
- Correferent
  - When a linguistic expression aludes to another linguistic entity that has the same referent.





General concepts

- Endophora
  - Correference of a linguistic expression with another that is located or before or after, but always within the linguistic discourse
  - Anaphora, when the aluded expression appears previously
  - Cataphora, when the aluded expression appears after

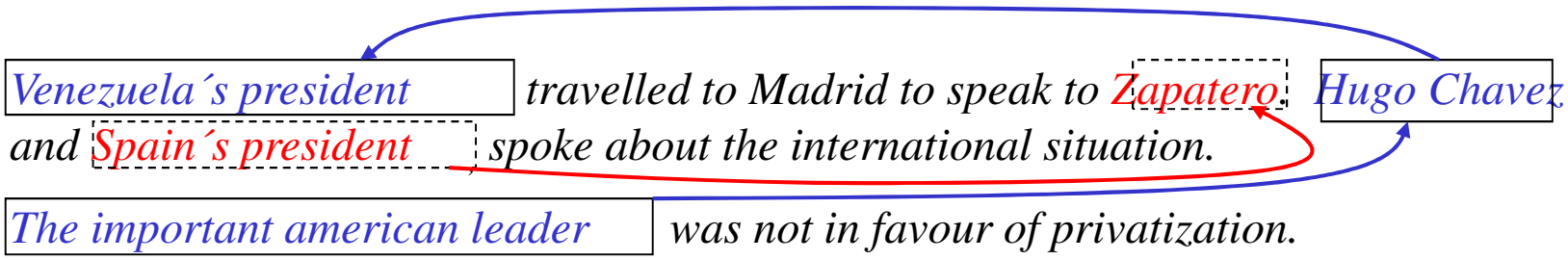


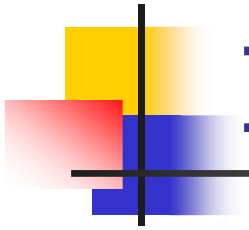
General concepts

- Exophora
  - When a reference is done. That means, we are aluding to an extra-linguistic object.
  - Deixis, when the information is in the physical environment of the dialogue
  - Homophora, when general knowledge is requiered to resolve it

General concepts

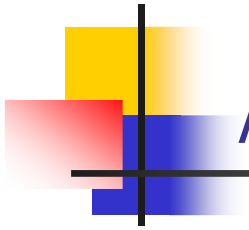
- Correference strings
  - When the antecedent of an anaphora aludes also to another previous linguistic expression.





Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



Types of anaphora

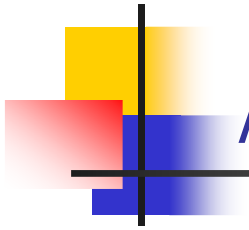
Anaphoric expressions

- Pronouns
- Definite noun phrase anaphora
- Adverbs

Types of anaphora

Anaphoric expressions

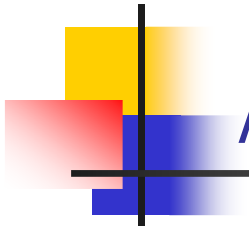
- Pronouns
  - The most widespread type of anaphora is the pronominal anaphora which is realised by anaphoric pronouns.
  - Example:
    - Computational Linguists from many different countries attended the tutorial.
    - They took extensive notes.
  - It should be pointed out that not all pronouns in English are anaphoric. For instance, "it‟ can often be non-anaphoric such as in the case of the previous sentence. Other examples of non-anaphoric “it” include expressions such as "It is important", "It is necessary", "It has to be taken into account". A non-anaphoric "it" is termed pleonastic (Lappin & Leass 1994).



Types of anaphora

Anaphoric expressions

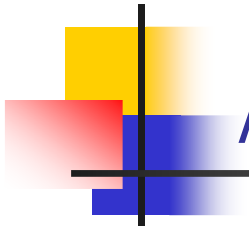
- Pronouns
  - Syntactic role
    - Subject
      - You must pass this course
    - Complement



Types of anaphora

Anaphoric expressions

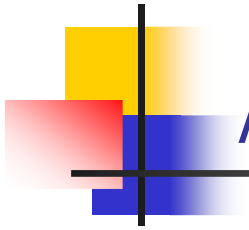
- Pronouns
  - Syntactic role
    - Subject
    - Complement
      - I don´t have it



Types of anaphora

Anaphoric expressions

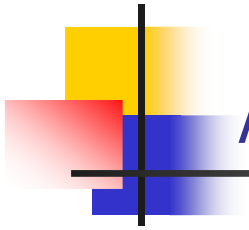
- Definite noun phrase anaphora
  - With definite article
  - With demonstrative
  - With possesive



Types of anaphora

Anaphoric expressions

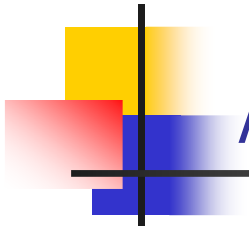
- Definite noun phrase anaphora
  - With definite article
    - It, they,..
  - With demonstrative
  - With possesive



Types of anaphora

Anaphoric expressions

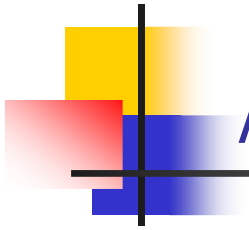
- Definite noun phrase anaphora
  - With definite article
  - With demonstrative
    - This, these, that, those, ..
  - With possesive



Types of anaphora

Anaphoric expressions

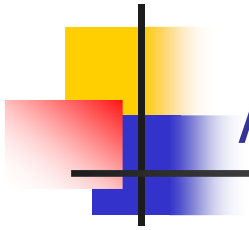
- Definite noun phrase anaphora
  - With definite article
  - With demonstrative
  - With possesive
    - Mine, yours, ours,…



Types of anaphora

Anaphoric expressions

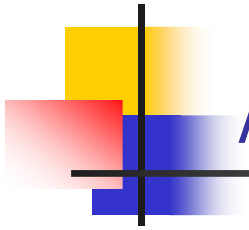
- Definite noun phrase anaphora
  - Non-anaphoric
  - Anaphoric
    - Direct
      - Same core
      - Alias
    - Indirect (bridge references)
      - Core semantically related
      - Thematic role
      - Proper nouns
      - Discourse topics
      - Inferences



Types of anaphora

Anaphoric expressions

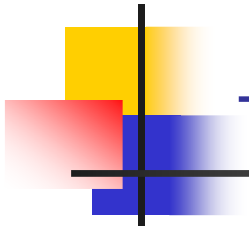
- Adverbs
  - Time
    - Before, after, tomorrow, etc.
  - Location
  - Etc.



Types of anaphora

Anaphoric expressions

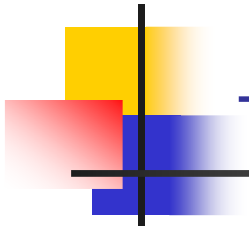
- Adverbs
  - Time
  - Location
    - Here, there, etc.
  - Etc.



Types of anaphora

Types of antecedents

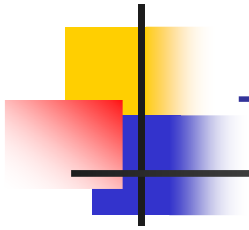
- Noun phrases
- Events or actions
- Sentences
- Paragraphs
- Documents



Types of anaphora

Types of antecedents

- Noun phrases
  - Complete noun phrases
  - Simple noun phrases
  - Pronouns
  - Proper nouns
- Events or actions
- Sentences
- Paragraphs
- Documents



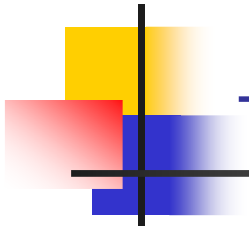
Types of anaphora

Types of antecedents

- Noun Phrases
  - Complete noun phrases

    The wood house that was sold ...

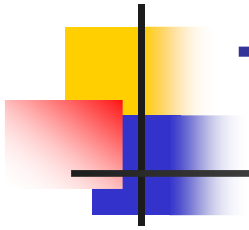
  - Simple noun phrases
  - Pronouns
  - Proper nouns
- Events or actions
- Sentences
- Paragraphs
- Documents



Types of anaphora

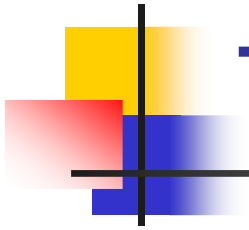
Types of antecedents

- Noun Phrases
  - Complete noun phrases
  - Simple noun phrases
  - Pronouns
  - Proper nouns:
    - Juan went to the party and I saw him
- Events or actions
- Sentences
- Paragraphs
- Documents



TYPES OF ANAPHORA

- Pronominal anaphora
- Definite noun phrase anaphona
- Numeric shallow anaphora
- Temporal anaphora
- Verbal anaphora
- Adverbial anaphora



TYPES OF ANAPHORA

- Pronominal anaphora

  Personal (subject)

  Juan knows the password. **He** was travelling.

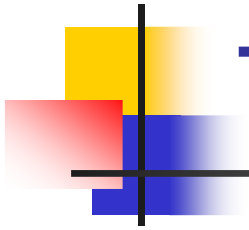
TYPES OF ANAPHORA

- Pronominal anaphora

  Personal (complement)

  I haven´t heard any news about **Luis**. I haven´t seen **him** since October.

  The television was on when Luisa came. She turns **it** off when she goes to sleep

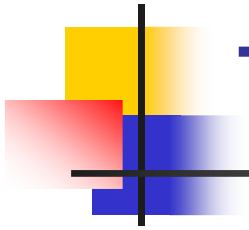


TYPES OF ANAPHORA

- Pronominal anaphora

  Demonstrative

  Between the audience I could see **a table**. **This** was red.



TYPES OF ANAPHORA

- Pronominal anaphora

  Posessive

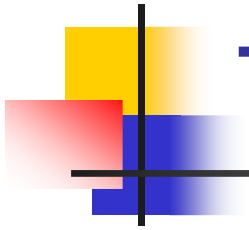
  Your eyes are blue. **Mine** are green.

TYPES OF ANAPHORA

- Pronominal anaphora

  Reflexive

  Luis goes to the park every day. He buys **himself** an ice-cream.

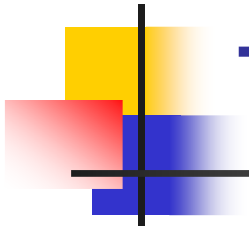


TYPES OF ANAPHORA

- Pronominal anaphora

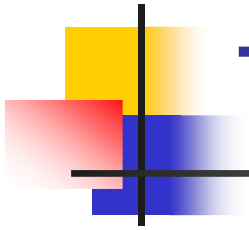
  Relative

  The CD´s that I lent you were very expensive



TYPES OF ANAPHORA

- Pronominal anaphora
- Definite noun phrase anaphora
- Numeric shallow anaphora
- Temporal anaphora
- Verbal anaphora
- Adverbial anaphora



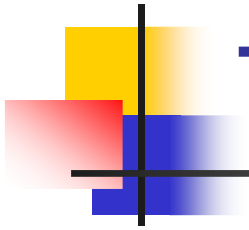
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Direct anaphora: Same core

  Luis has **an important company**. **The company** has 200 employees.

  Luis has **an important company**. **This company** has 200 employees.



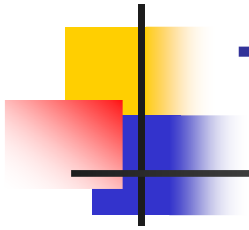
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Direct anaphora: Alias

  Rafael Muñoz has **an important company**. M**r. Muñoz** has 200 employees.

  International Bussines Machines presented new computer models. **IBM is an important company**.



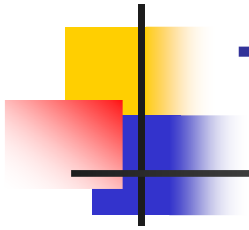
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Indirect anaphora: Related cores

  The company is in bankrupt. **The company** pretends to fire 100 employees.

  Bamboo is the base of our products. The East provides us **this plant**.



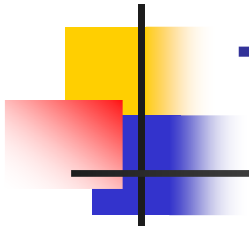
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Indirect anaphora: Thematic role

  Juan sold his house to Pedro. **The seller** has to pay the taxes.

  ESPACIO S.A. is building houses in San Juan. **The property development company** has three thousand square meters to build.



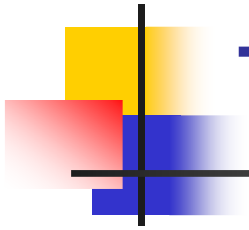
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Indirect anaphora: Proper nouns

  Juan A. Samaranch traveled to USA to see the Olympic Games. **The ex president** was invited by the Committee.

  Denilson played his 100 match at first division. **The Brazilian** scored three goals.

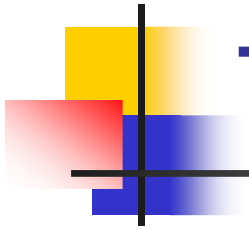


TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Indirect anaphora: Discourse topics

  The Spanish basketball is happy because all **the teams** go to the next round.



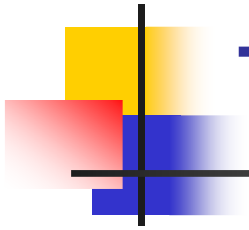
TYPES OF ANAPHORA

- Definite noun phrase anaphora

  Indirect anaphora: Inferences

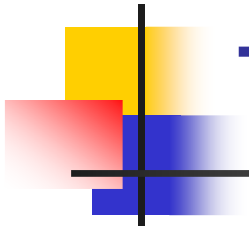
  The terrorist attacks were usual. **Victims** appeared anywhere.

  The earthquake destroy Turkey. **The huge destruction** cause many casualties.



TYPES OF ANAPHORA

- Pronominal anaphora
- Definite noun phrase anaphona
- Numeric shallow anaphora
- Temporal anaphora
- Verbal anaphora
- Adverbial anaphora

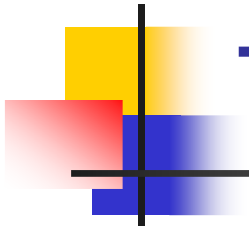


TYPES OF ANAPHORA

- Numeric shallow anaphora

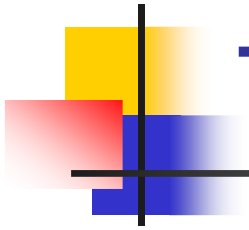
Luis and Mariano have a shop. **The first one** works only in the mornings.

Rome, Milan, Madrid, **Barcelona** and Paris present their Autumn collections. **The second of the Spanish cities** has more designers this year.



TYPES OF ANAPHORA

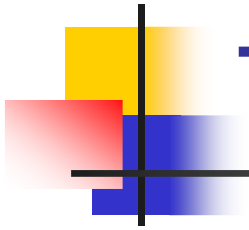
- Temporal anaphora
  - 31st of May will finish the classes. Two days later the exams will start
  - The Real Madrid played the final the 5th of March. The following week the goalkeeper had an accident.



TYPES OF ANAPHORA

- Verbal anaphora

It is prohibited to **smoke** in this area, so do not do **it**.



TYPES OF ANAPHORA

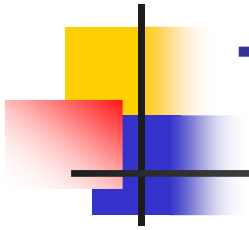
- Pronominal anaphora
- Definite noun phrase anaphora
- Numeric shallow anaphora
- Temporal anaphora
- Verbal anaphora
- Adverbial anaphora

TYPES OF ANAPHORA

- Adverbial anaphora

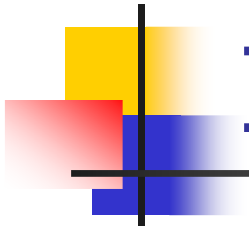
I will not finish my studies until **next year**. **Then** I will work at a company.

In front of the office there is **a car repair shop**. **There** you will find all you need for your car.



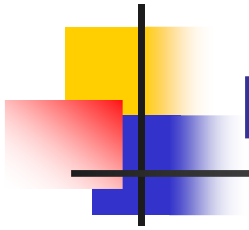
TYPES OF ANAPHORA

- Pronominal anaphora
- Definite noun phrase anaphora
- Numeric shallow anaphora
- Temporal anaphora
- Verbal anaphora
- Adverbial anaphora



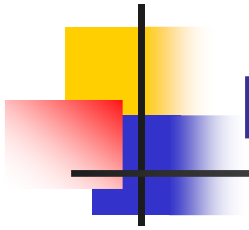
Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



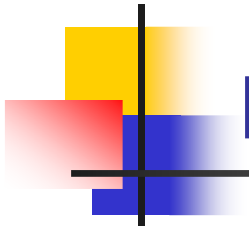
Resolution strategies

- Based on linguistic knowledge
- Based on corpus



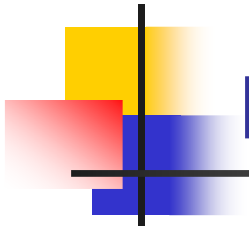
Resolution strategies

- Based on linguistic knowledge
  - They imitate human knowledge sources
  - Consultives
    - A unique source of information
  - Democratics
    - Combine various information sources
    - Mechanisms of restrictions and preferences
      - Rules to discard candidates
      - Rules to order candidates
- Based on corpus



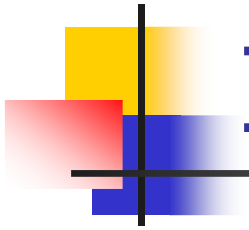
Resolution strategies

- Based on linguistic knowledge
- Based on corpus



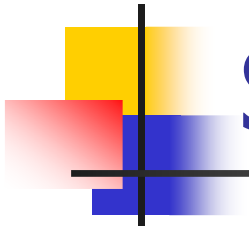
Resolution strategies

- Based on linguistic knowledge
- Based on corpus
  - They study corpus through statistical tools
  - They propose probabilistic models



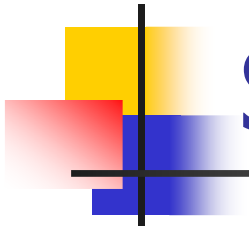
Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



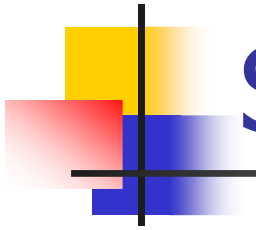
SOURCES OF INFORMATION

- Morphological
- Syntactic
- Semantic
- World knowledge



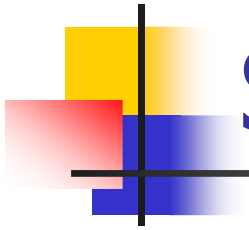
SOURCES OF INFORMATION

- Morphological
  - Morphological tagger (POS tagger)
    - Set of tags
    - Information of the tags:
      - Grammatical category
      - Gender
      - Number
      - Tense
      - Person, etc.
    - Examples: NCMS000, VI3PS00, DD3MS00



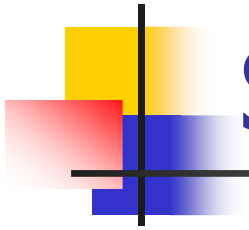
SOURCES OF INFORMATION

- Syntactic
  - Syntactic role
    - Role: Subject, direct complement, indirect complement
    - Position regarding the verb
    - Composition of the noun phrases
  - C-command, Non correferencial rules
    - It discards candidates in the same sentence of the anaphoric expression
    - For complete analysis (Reinhart 1983; Lappin & Leass 1994)
    - For shallow analysis (Palomar et al. 2001)



SOURCES OF INFORMATION

- Semantic
  - Semantic Relationships (synonyms, hiperonyms, meronyms, troponyms, etc..)
  - Semantic Resources: WordNet, EuroWordNet,
  - Ontologies: Mikrocosmos, WordNet, EuroWordNet

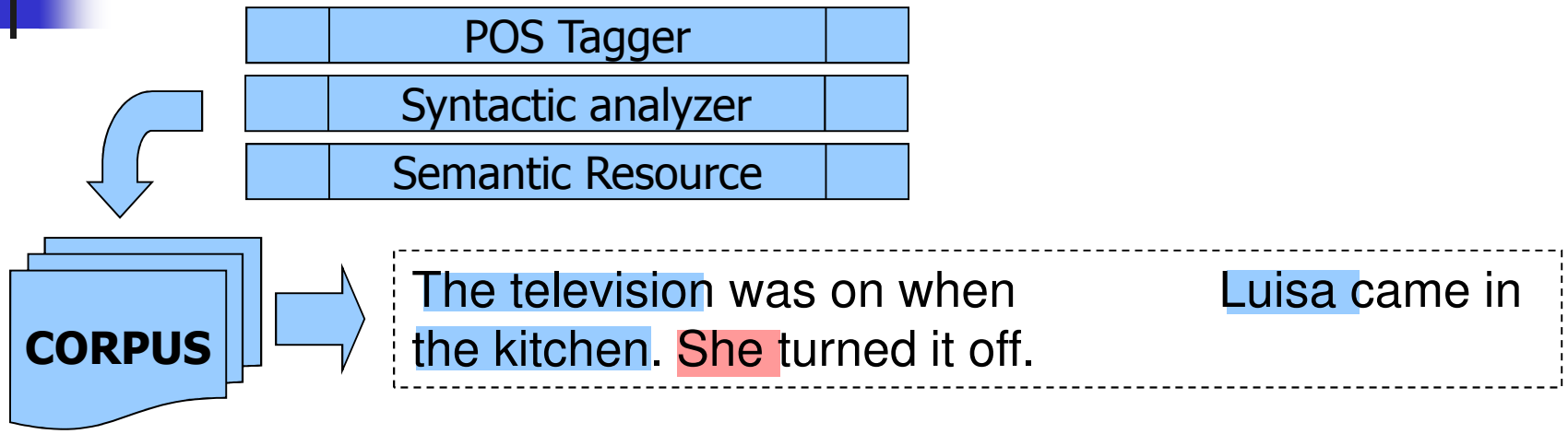


SOURCES OF INFORMATION

- World knowledge
  - Resources that are quite difficult to maintain
  - They contain different types of information
    - Fixed (Eiffiel Tower is in Paris)
    - Variable (The president of USA is Barack Obama)

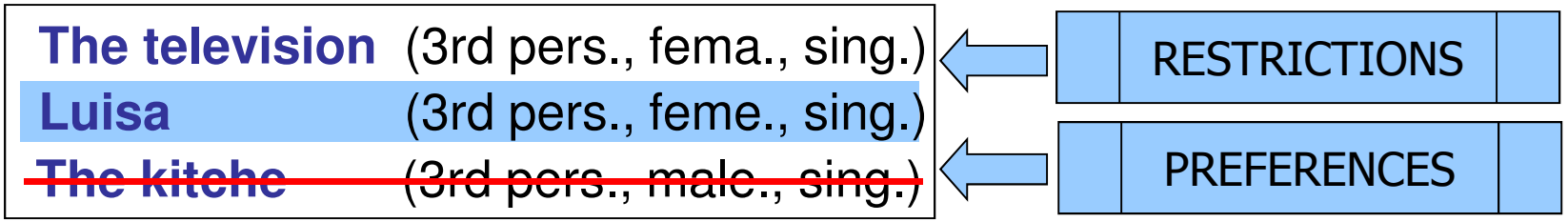
SOURCES OF INFORMATION

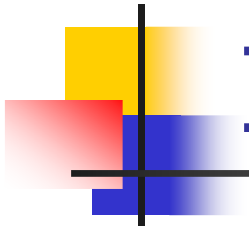
Application to the resolution: Schema



Pronoun: **She** (3rd pers., female., sing.)

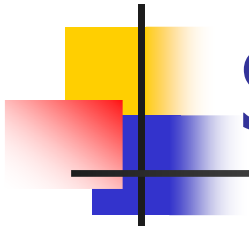
List candidates:





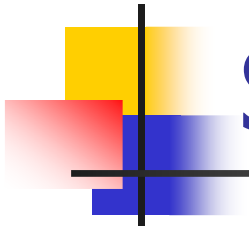
Index

- What is the anaphora?
- Why is it important to resolve it?
- General Concepts
- Types of anaphora
- Resolution strategies
- Sources of information
- System of resolution



Systems of resolution

- Pronominal Anaphora
- Definite Noun Phrases anaphora



Systems of resolution

- Pronominal Anaphora
  - Based on Knowledge
    - Hobbs (1978)
    - Lappin & Leass (1984)
    - GPLSI
  - Based on Machine Learning
    - Ge, Hale & Charniak (1998)

Systems of resolution

Pronominal anaphora: Methods based on knowledge

- Hobbs (1978)
  - It uses trees of shallow syntactic analysis
  - It applies a simple searching on trees algorithm
  - The algorithm selects the first candidate that founds in a left-right way that satisfies a set of morpho-syntactic restrictions
    - Gender coherence
    - Application of c-command

Systems of resolution

Pronominal anaphora: Methods based on knowledge

- Lappin & Leass (1994)
  - Resolution of 3rd person personal pronouns and lexical anaphora (reflexives and reciprocal)
  - It uses morphological and syntactic information
  - It discards incompatibles candidates (morphological info.)
  - It uses non correference rules to discard candidates within the sentence
  - Identification of pleonastic (non anaphoric)
  - Assigning valued that shows the relevance of a candidate
  - Getting the candidate with highest relevance

Systems of resolution

Pronominal anaphora: Methods based on knowledge

- GPLSI
  - Application to all types of pronouns
  - More relevant papers
    - Palomar et al. (2001) Computational Linguisitics
    - Martínez and Palomar (2001) Journal of Artificial Intelligence Research
    - Saíz-Noeda et al. (2001) CONLL workshop of ACL
    - Peral and Ferrández (2000) workshop of ACL
    - Ferrández et al (1998) COLING

Systems of resolution

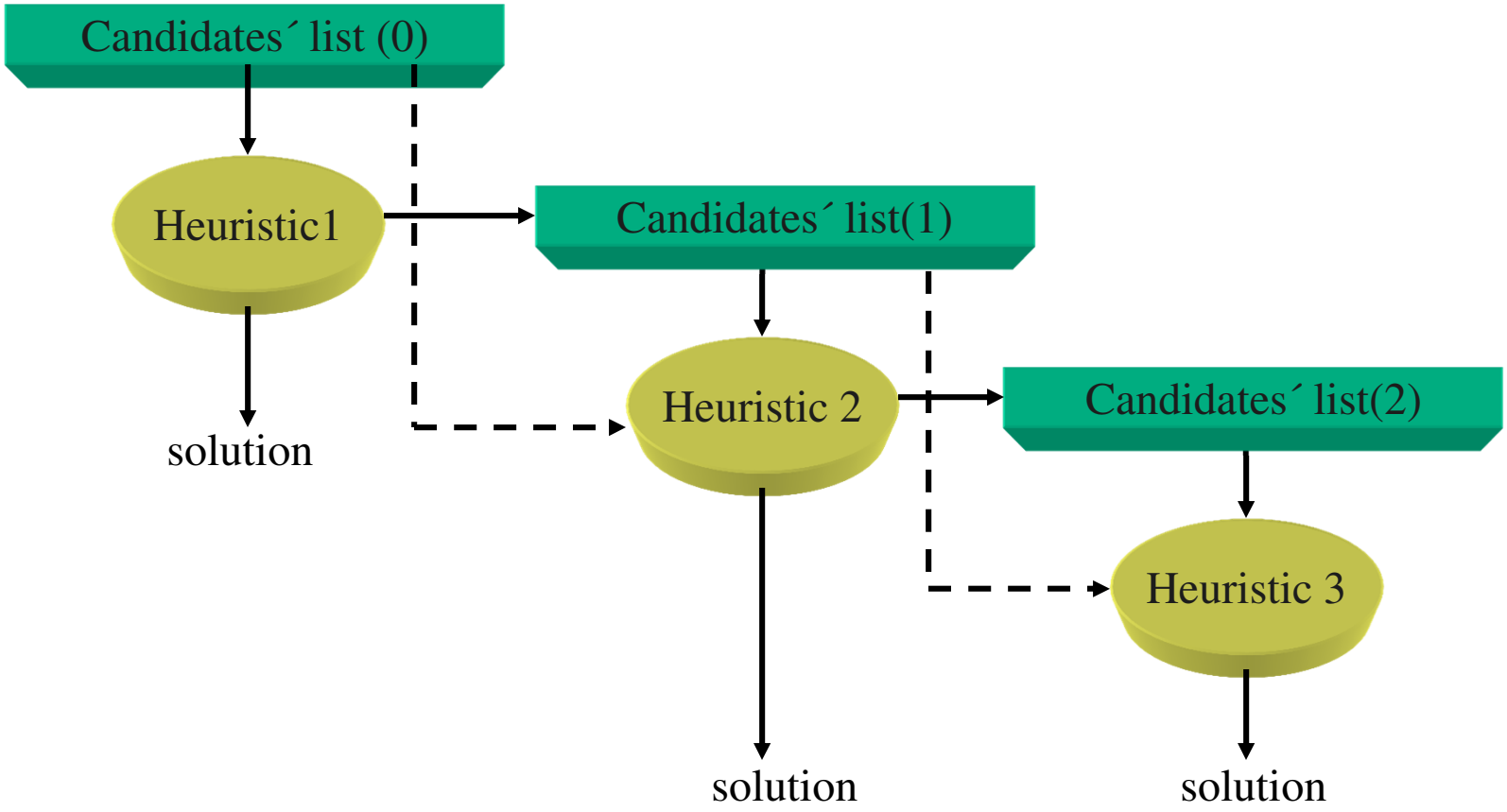
Pronominal anaphora: Methods based on knowledge

- GPLSI
  - It uses different types of information sources: morphological, syntactical and semantical
  - Application of non correference rules using shallow syntactic parsing
  - Set of restrictions and preferences specific for each type of pronoun
  - Application of restrictions and preferences with weight systems and filtering systems

Systems of resolution

Pronominal anaphora: Methods based on knowledge

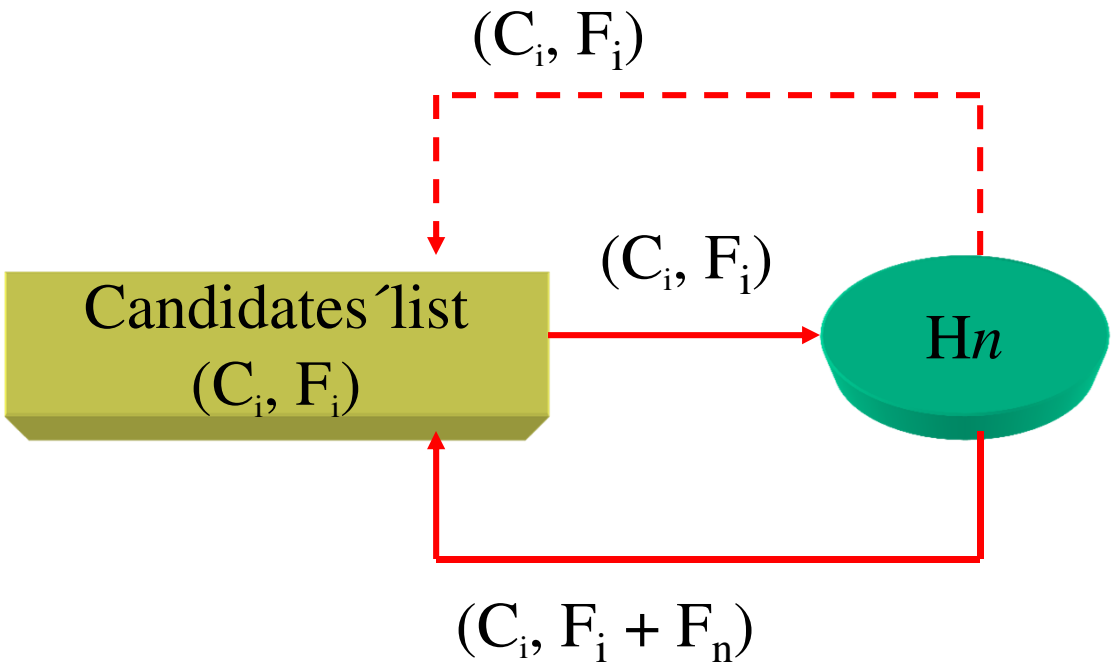
- GPLSI: Applications of rules with filtering system



Systems of resolution

Pronominal anaphora: Methods based on knowledge

- GPLSI: Application of rules with weight system



Systems of resolution

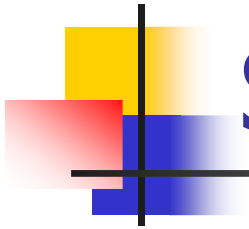
Pronominal anaphora: Methods based on knowledge

- GPLSI
  - Application in different systems:
    - Monologues
    - Dialogues
    - Information extraction
    - Information retrieval
    - Question Answering
    - Automatic translation

Systems of resolution

Pronominal anaphora: Methods based on Machine Learning

- Ge, Hale & Charniak (1998)
  - Resolution of pronominal anaphora
  - Unsupervised method
  - It used attribute-value features:
    - Distance between anaphora and antecedent
    - Features of movement, gender and number
    - It uses Hobbs´algorithm to obtain the shallow parser trees (Initial Phase)



Systems of resolution

- Definite Noun Phrases anaphora
  - Methods based on knowledge
    - Poesio & Vieira (1999)
    - GPLSI
      - Muñoz & Palomar (1999,2000,2001)
      - Muñoz et al. (2000)
  - Methods based on Machine Learning
    - Cardie & Wagstaff (1999)
    - Bean & Rillof (1999)

Systems of resolution

Definite Noun Phrase: Methods based on knowledge

- Poesio & Vieira (1999)
  - Step 1: Identification of some of the non anaphoric DNP (study of restrictive modifiers)
  - Step 2: Searching the antecedent with the SAME CORE of the DNP
  - Step 3: Application of a set of rules to identify non anaphoric cases
  - Step 4: Application of semantic information to resolve some of the non anaphoric DNP

Systems of resolution

Definite Noun Phrase: Methods based on knowledge

- GPLSI
  - Identification of the non anaphoric DNP
  - Resolution of the anaphoric DNP
  - Resolution to the correferences type identity or part-of
  - More relevant papers
    - Muñoz et al. (2002) PorTAL
    - Muñoz y Palomar (2001) TSD, RANLP, NLP, etc.
    - Palomar y Muñoz (2000) IBERAMIA,
    - Muñoz et al. (2000) MICAI, ACIDCA

Systems of resolution

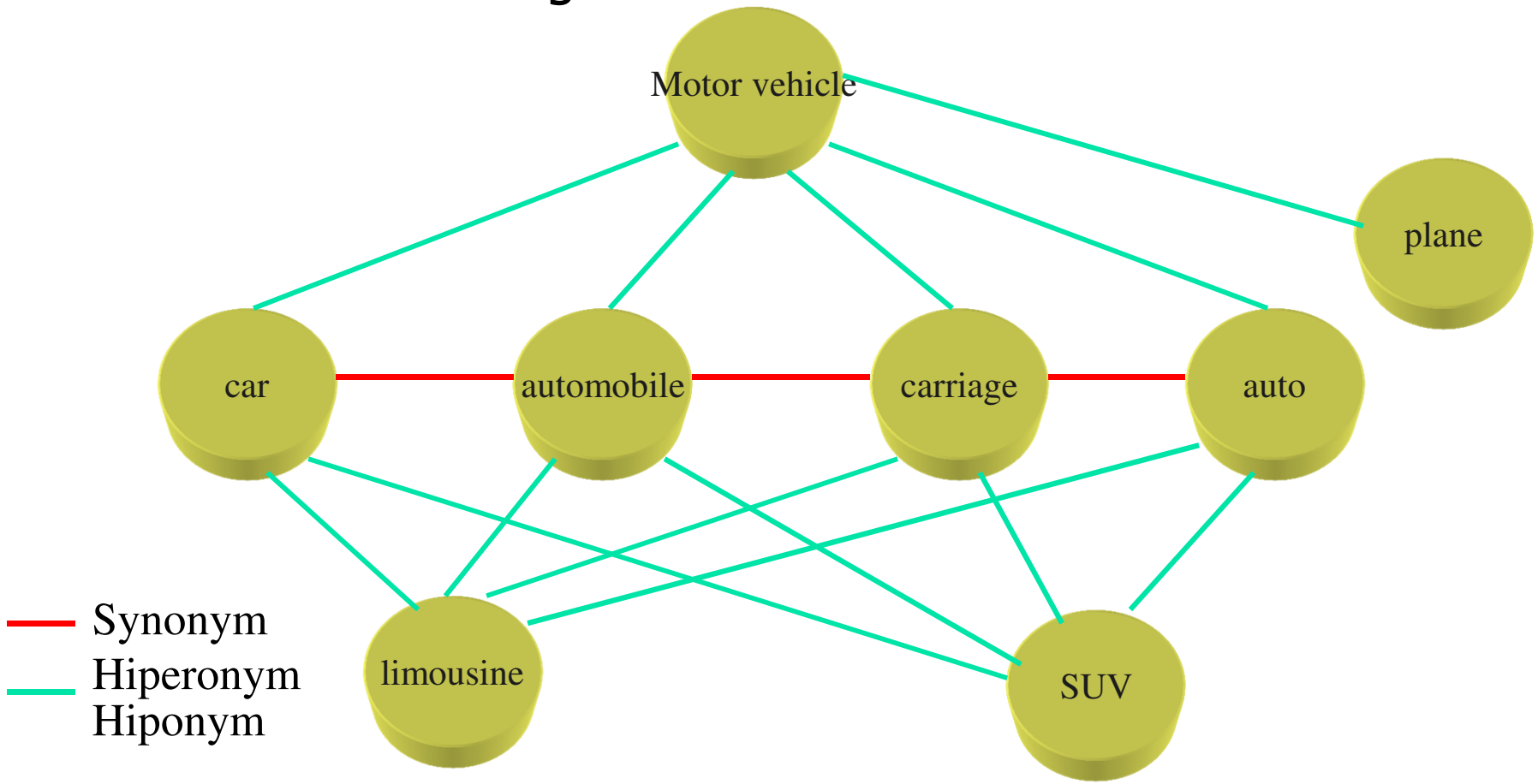
Definite Noun Phrase: Methods based on knowledge

- GPLSI
  - Building a semantic net
    - Identification of part of the non anaphoric DNP
    - Reduction of the number of candidates to compare, preserving the space of accessibility
    - Using semantic information extracted fromWordNet (synonym, hypernym relations, etc.) and from an onology
    - Using Magnini domains to make easy the resolution of part-of correferences
    - Application of rules using for systems: filtering and weight

Systems of resolution

Definite Noun Phrase: Methods based on knowledge

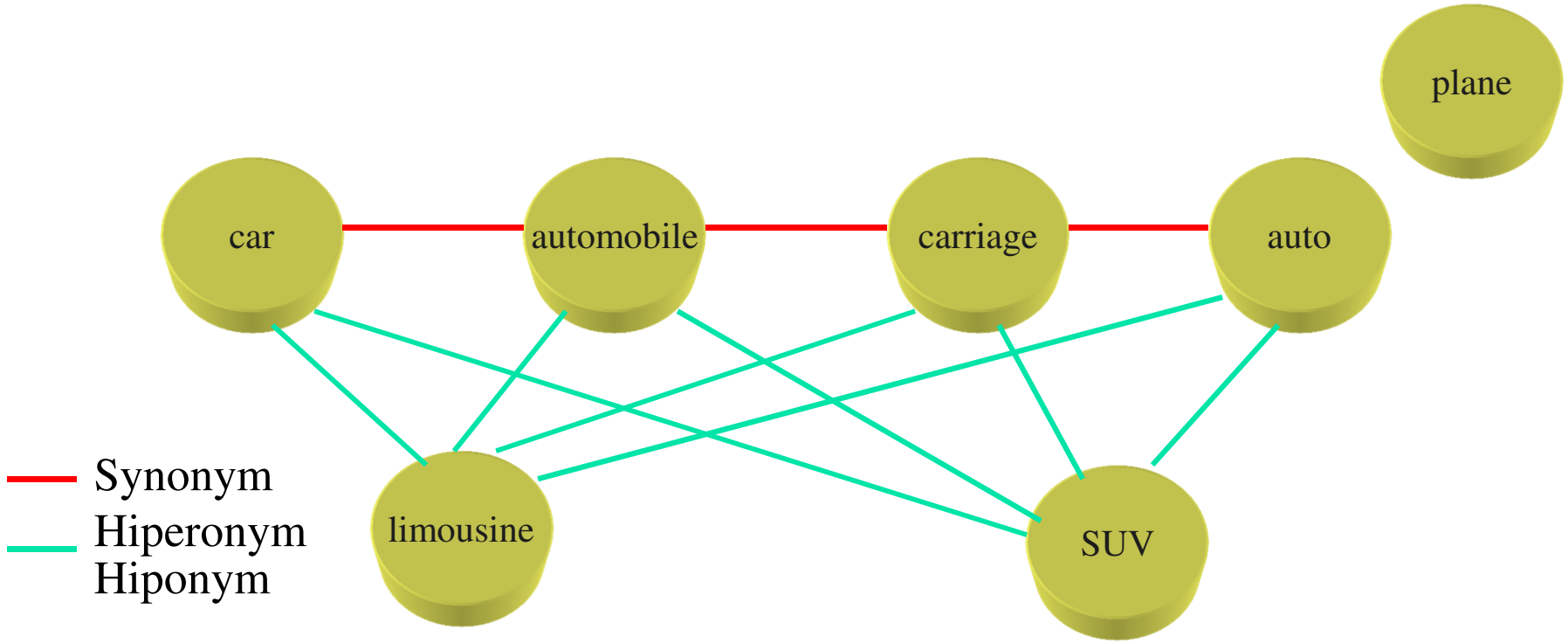
- GPLSI: Building a semantic net



Systems of resolution

Definite Noun Phrase: Methods based on knowledge

- GPLSI: Building a semantic net



Systems of resolution

Definite Noun Phrase: Methods based on Machine Learning

- Cardie & Wagstaff (1999)
  - Unsupervised algorithm
  - Resolution of noun phrases
  - It deals the correference problem as a clustering problem
  - It uses attribute-value feature vector
    - Words that form the NP
    - Core
    - Position
    - Article
    - Proper name
    - Etc.

Systems of resolution

Definite Noun Phrase: Methods based on Machine Learning

- Bean & Rillof (1999)
  - Statistical method for the identification of non-anaphoric DNP
  - Steps:
    - Application of syntactic heuristics: study of modifiers. (The president of USA --- referential)
    - The first NP is not anaphoric
    - Building patterns using the extracted as first NP from the training corpora
    - Using definite noun phrases lists

Questions

TEXTUAL INFORMATION EXTRACTION

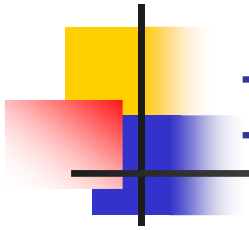
Dr. Patricio Martínez Barco
Dra. Estela Saquete Boró

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

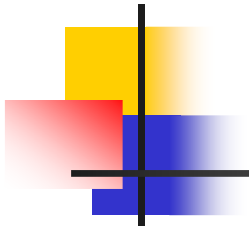
Information Extraction: Recogniton of Entities

Unit 4



Index

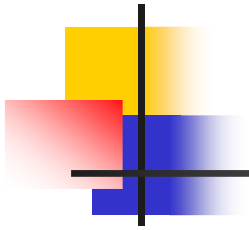
1. Introduction
2. General features
3. Tagging
4. Resources
5. NE algorithms
6. Demos



1. Introduction

- Information extraction
  - Technique that provides certain information known as relevant from a set of relevant texts
  - It is the task for automatically extracting a type of pre-specified information from texts

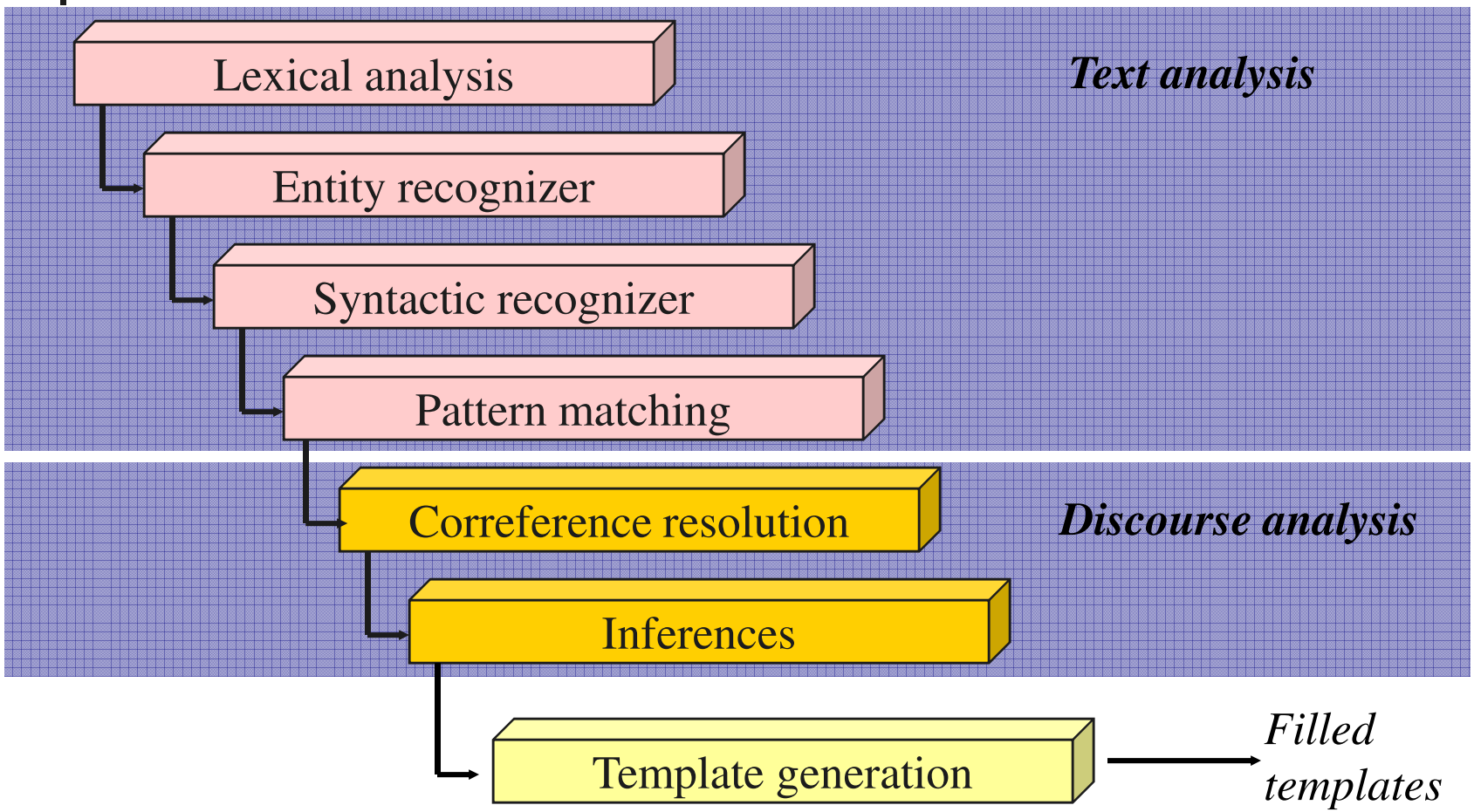
1. Introduction



1. Introduction

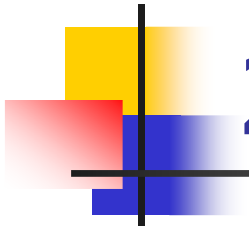
- IE.- Building systems that find and relate relevant information while at the same time ignoring non-relevant information
- Relevant information is determined by predefined domain guides, where the type of information to extract must be specified as exactly as possible
- From the NLP point of view, IE systems must work on different levels: from the recognition of words to the analysis of sentences and from the understanding of the sentence to the whole text

1. Introduction



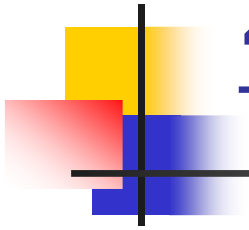
2. General features

- This task consists on the identification of proper nouns and some numeric data and its classification in a set of predefined categories of interest.
- Example:
  - The city of Madrid waits for an answer of the members of the Olimpic Commitee about the candidature to organize the next Olympic Games



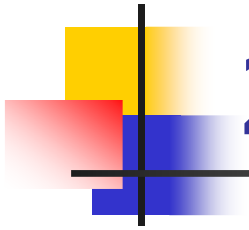
2. General features

- In general, there are three categories: Person, Organization and Location.
- Other usual categories: temporal expressions, money, weight, telephones, e-mail addresses, etc.
- Other categories typical of domains (name of laws, bibliographic references, etc.)



2. General features
Type of entities

- Type of entities
  - Strong
    - Madrid, Comité Olímpico, Rafael Muñoz, etc.
  - Weak
    - The president of the USA, Bill Gates´company, etc.



2. General features

- Used techniques:
  - Based on knowledge
    - Using dictionaries or lists
    - Using rules
  - Based on machine learning (statistical)
    - Supervised
    - Unsupervised

2. General features

Techniques based on rules

- Steps:
  - Identification of the entity.- Identification of the starting and ending of the entity.
  - Classification of the entity.- Classification of the type of entity (person, organization, location, etc.)

2. General features

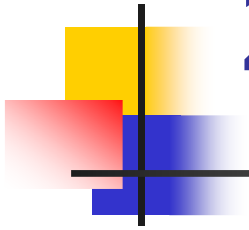
Based on knowlede

- The identification of strong entities is quite simple. The identification of weak entities has more difficulties.
- Classifying strong and weak entities is extremely difficult

2. General features

Recognition techniques based on knowledge

- Approaches based on list search
  - Systems that recognize ONLY the entities that are stored in their lists (geographical encyclopedias, databases of names, lists of names, etc.
  - Strong points: Simple, fast, language independent, easy to redirect.
  - Weak points: Gathering and maintenance of the lists, name variants are not detected, ambiguity is not solved (Africa, April)



2. General features

Recognition techniques based on knowledge

- Types of list used:
  - General lists: locations + organizations + names
  - Machine learned lists from a training corpus from MUC7 (100 articles of NYT)
  - Specific lists: names of medicines, names of laws, etc.

2. General features

Recognition techniques based on knowledge

- Approaches based on rules
  - Systems that recognized the entities that match the defined rules, supported by some types of lists or dictionaries (abbreviations, names, etc).
  - Weak points: Dependent on the language.
  - Strong points: They can solve name variants and ambiguities.

2. General features

Recognition techniques based on knowledge

- Evidences (McDonald 1995)
  - External.- Words or names that go with certain entities in certain contexts.
  - Internal.- The entities usually have a certain internal structure. Some components of this structure can be specified in lists or dictionaries, and others can be implied.

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of strong entities.-
  - Capital words (PalMay)
  - Problems:
    - Presence of prepositions, defined articles and demonstrative pronouns between capital words (Joaquín de la Cierva)
    - Starting of sentences(The Coca-cola is the most ...)
    - Coordination of entities through conjunctions or prepositions (Construcciones and Contratas S.A., IBM and Microsoft)

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of weak entities.-
  - Keywords
    - Mr. Ms.
    - S.A., Ltd.
    - Avenue, Beach
  - Words that belong to a specific ontology within a certain category
    - President (human)
    - company (organization)
    - airport (location)

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of entities. Internal evidences.
  - Person

Title (Mr.) + {PalMay}	Mr. Gómez
Title (Mr.) + Name + PalMay	Mr. Rafael Muñoz
Name + CapitalLetter + ”.” + {PalMay}	Juan A. García
Title (Mr.) + PREP + {PalMay}	Mr. De Pedro
Title (Mr.) + PREP + ART + {PalMay}	Mr. De la Fuente
Title (Mr.) + PalMay + CONJ + PalMay	Mr. Ramón y Cajal

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of entities. Internal evidences.
  - Organization
    - {PalMay} + Trigger (S.A., S.L.,) — Coca-Cola S.A.
    - Trigger (Coop.) + {PalMay} — Coop. Algodón
    - Acronym + Trigger (S.A., S.L.,) — IBM S.A.
    - Activity + {PalMay} — Cafeteria Pedro
    - {PalMay}+PREP+{PalMay} + Trigger (S.A.,) — Casa de Cultura
    - {PalMay} + NUM + Trigger (S.A., S.L.,) — Tien 21 S.A.
    - PalMay+CONJ + PalMay+Trigger (S.A., S.L.,) — Construcciones y Contratas S.A.

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of entities. Internal evidences.
  - Location
    - PalGeo + {PalMay} — Cape San Antonio
    - Trigger (Str.,Sq.) + {PalMay} — Sq. Manila
    - Trigger (Str.,Sq.)) + {PalMay} + “,” + NUM +”,” + NameCity
      Sq. Manila, 2, Alicante
    - NameCity — Alicante
    - NameCountry — Spain

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of entities. External evidences.
  - Person
    - Job + {PalMay} — the player Alfaro
    - {PalMay} + "," + Job — Alfaro, player...
    - {PalMay} + ... + Verbs — Alfaro scored...

2. General features

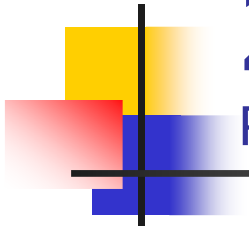
Recognition techniques based on knowledge : PHASES

- Identification of entities. External evidences.
  - Organization
    - Job + PREP + {PalMay} — manager of Coca-Cola SA
    - "Company" + {PalMay} — Company Dragados

2. General features

Recognition techniques based on knowledge : PHASES

- Identification of entities. External evidences.
  - Location
    - ART + PNTOCARD + {PalMay} — the south of Alicante
    - {PalMay} + "is" + ART+ ADJ + PalGeo — Ness is a big lake



2. General features

Recognition techniques based on knowledge : PHASES

- Classification of the entities.-
  - Presence of triggers (Mr, Co., street, etc)
    - Application of rules
  - Problems:
    - Structural ambiguity
      - IBM and Microsoft S.A. (one or two entities??)
    - Semantic ambiguity:
      - Mr. Donut (person or company)
      - John F. Kennedy (person or location)

2. General features

Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem: Structural ambiguity (words that are part of it)
    - Starting of sentences
      - The Real Jaén played.... (THE REAL JAÉN or REAL JAÉN)

2. General features

Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem: Structural ambiguity (words that are part of it)
    - Starting of sentences
      - The Real Jaén played.... (THE REAL JAÉN or REAL JAÉN)
    - Solution: Compare with the presence of similar entities. Usually if it is an article IS NOT part of the entity.

2. General features

Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem: Structural ambiguity:
    - A conjunction appears between two names
      - Construcciones and Contratas (ONE or TWO COMPANIES)

2. General features

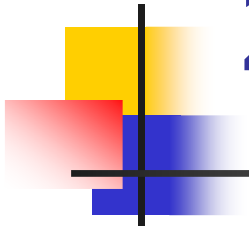
Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem: Structural ambiguity:
    - A conjunction appears between two names
      - Construcciones and Contratas (ONE or TWO COMPANIES)
    - Solution:
      - Studying the number (Sing, Plural) of the verb
      - Study of the triggers appearing before and after the conjunction

2. General features

Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem: Structural ambiguity : Study of the elements appearing at both sides of the conjuction
    - Both substrings contain triggers, TWO ENTITIES (IBM S.A. and Microsoft S.A.)
    - The left substring has a trigger at the end, TWO ENTITIES (IBM S.A. and Microsoft)
    - The right substring has a trigger and the left substring an activity, TWO ENTITIES (Hospital San Carlos and Hércules C.F.)
    - Acronym in one of the substrings, TWO ENTITIES (CAM and Hércules)
    - Otherwise, ONE ENTITY or verb disambiguation (Construcciones and Contratas S.A. acquires...)



2. General features

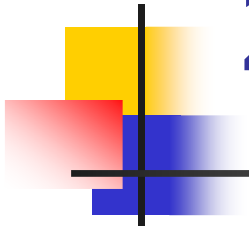
Recognition techniques based on knowledge: PHASES

- Classification of entities.-
  - Problem Semantic ambiguity: Classification of the type of entity
    - John F. Kennedy (AIRPORT or PERSON)
    - Adolfo Domínguez (ORGANIZATION or PERSON)

2. General features

Recognition techniques based on knowledge : PHASES

- Classification of entities.-
  - Problem Semantic ambiguity: Classification of the type of entity
    - John F. Kennedy (AIRPORT or PERSON)
    - Adolfo Domínguez (ORGANIZATION or PERSON)
  - Solution: Adding semantic information associated to the verb or an ontology
    - Landing at John F. Kennedy (LOCATION)
    - Adolfo Domínguez visits ... (PERSON)



2. General features

Machine Learning

- Based on Machine Learning.-
  - What is machine learning?
    - Getting the description of a CONCEPT in a representation language that explains done observations and helps to predict NEW instances of the same distribution
  - Representation language
    - Feature´s vectors

2. General features

Machine Learning

- Based on Machine Learning.-
  - An **instance** is a vector: $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$ whose components, called **features** (or attributes), are discrete or real values.
  - where X is the space of all the possible instaces.
  - $Y = \{y_1, \ldots, y_m\}$ is the set of categories or classes.
  - The objective is learning the function , $f : X \rightarrow Y$
  - An example is one instance $\mathbf{x}$ *belonging to X,* tagged as a correct value $f(\mathbf{x})$, for example, a pair $\langle \mathbf{x}, f(\mathbf{x}) \rangle$
  - where D is the set of all the examples.
  - The space of hypothesis, H, is the set of functions $h: X \rightarrow Y$ that can be considered as possible definitions
- The objective is finding a function h belonging to H that for each pair $\langle \mathbf{x}, f(\mathbf{x}) \rangle$ belonging to D, $h(\mathbf{x}) = f(\mathbf{x})$

2. General features

Machine Learning

- Based on Machine Learning.-
  - Example

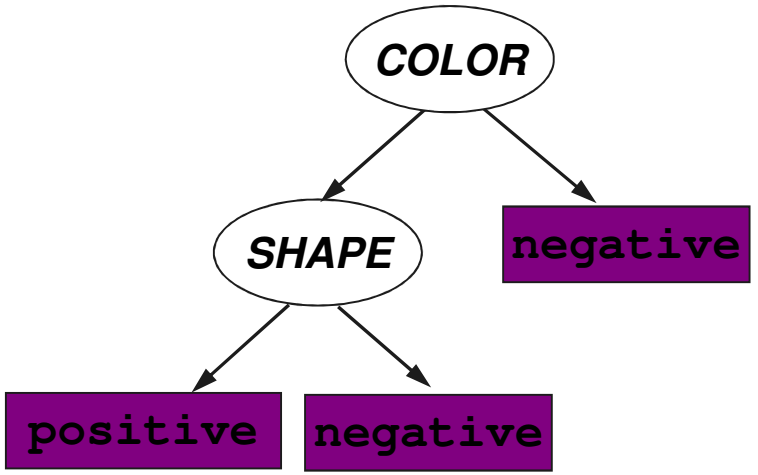
Ejemplo	TAMAÑO	COLOR	FORMA	CLASE
1	pequeño	rojo	círculo	Positiva
2	grande	rojo	círculo	Positiva
3	pequeño	rojo	triangulo	Negativa
4	grande	azul	círculo	negativa

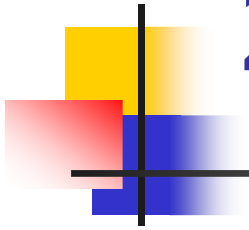
Rules

(COLOR=red) $\wedge$

(SHAPE=circle) $\Rightarrow$ positive

Decision tree





2. General features

Machine Learning

- Based on Machine Learning.-
  - It is basically a categorization problem.
  - It needs positive and negative examples to learn
  - It needs a set of features that must be observed and learned to be able to classify.

2. General features

Machine Learning

- Based on Machine Learning.-
  - It can be done in one or two steps:
    - 2-steps
      - Identification. Consists on identifying the category of each word (B=starting of the entity, I=inside the entity, O=outside the entity)
      - Classification. The words identified as B or I must be classified in one of the searched categories (PER,ORG,LOC, etc)
      - Usually a post-processing is needed to solve sequences as OIIO or OBBIO and transform them into OBIO or OBIIO, respectively

2. General features

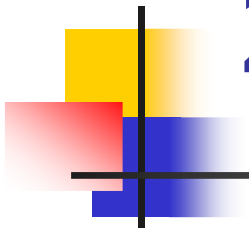
Machine Learning

- Based on Machine Learning.-
  - It can be done in one or two steps:
    - 1-step
      - Identification and classification. Consists on identifying the category of each word
        - Person (B-PER=starting of the entity, I-PER=within the entity)
        - Organization (B-ORG, I-ORG)
        - Location (B-LOC, I-LOC)
        - XXX (B-XXX, I-XXX)
        - O=outside the entity

2. General features

Machine Learning

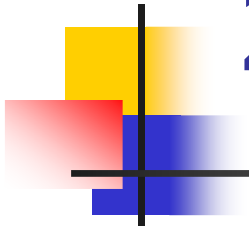
- Based on Machine Learning.-
  - Advantages and drawbacks:
    - 2-steps
      - Less number of classes(BIO)
      - All the examples of the different entities contributed
    - 1-step
      - The classification of unambiguous words can help next ones
      - O,B-PER,I-xxx,I-xxx → O,B-PER,I-PER,I-PER



2. General features

Machine Learning

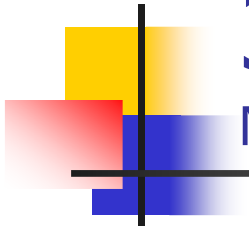
- Methods of Machine Learning.
  - HMM
  - ME
  - TimBL
  - C4.5
  - AdaBoost
  - etc



2. General features

Machine Learning

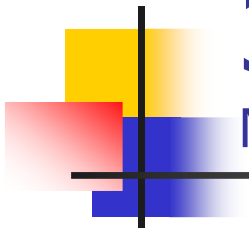
- Features
  - Domain or environment(previous and following words)
  - Belonging to dictionaries
  - Triggers
  - Morphology (categories, sufixes, etc.)
  - Syntax (sentences)
  - Semantic (ontologies, domains, rol, etc)
  - Etc.



3. Tagging of entities

MUC tagging

- Tagging used in MUC-6 (1995) and MUC-7 (1998).
- According to the specifications 7 types of entities were defined and their correspondant tags:
  - PERSON (ENAMEX)
  - ORGANIZATION (ENAMEX)
  - LOCATION (ENAMEX)
  - DATE (TIMEX)
  - TIME (TIMEX)
  - MONEY (NUMEX)
  - PERCENT (NUMEX)

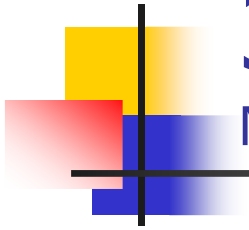


3. Tagging of entities

MUC tagging

- The three subtasks ENAMEX, TIMEX and NUMEX were subclassified with SGML tags using the TYPE attribute whose possible values were: PERSON, ORGANIZATION, LOCATION, DATE, TIME, MONEY y PERCENT.
- The structure of the tag was:

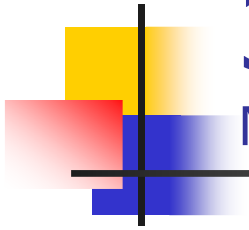
  <ENAMEX TYPE="value"> .... </ENAMEX>



3. Tagging of entities

MUC tagging

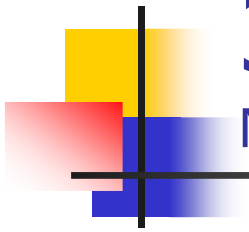
- Tagging used at MUC.
- "Flavel Donne is an analyst with General Trends, which has been based in Little Spring since July 1998."



3. Tagging of entities

MUC tagging

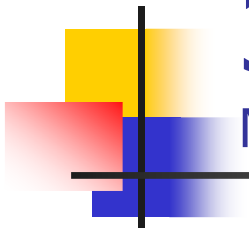
- "<ENAMEX TYPE='PERSON'>Flavel Donne<\ENAMEX> is an analyst with General Trends which has been based in Little Spring since July 1998."



3. Tagging of entities

MUC tagging

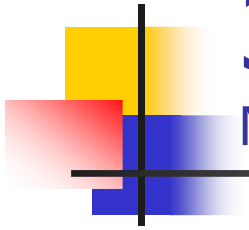
- "<ENAMEX TYPE='PERSON'>Flavel Donne<\ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends<\ENAMEX>, which has been based in Little Spring since July 1998."



3. Tagging of entities

MUC tagging

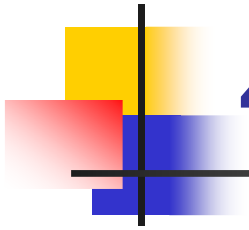
- "<ENAMEX TYPE='PERSON'>Flavel Donne<\ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends<\ENAMEX>, wich has been based in <ENAMEX TYPE='LOCATION'>Little Spring<\ENAMEX> since July 1998 ."



3. Tagging of entities

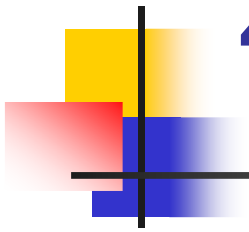
MUC tagging

- "<ENAMEX TYPE='PERSON'>Flavel Donne<\ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends<\ENAMEX>, wich has been based in <ENAMEX TYPE='LOCATION'>Little Spring<\ENAMEX> since <ENAMEX TYPE='DATE'>July 1998 <\ENAMEX>."



4. Resources

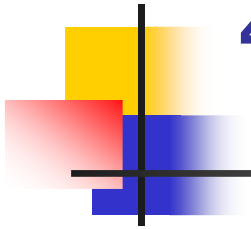
- Resources.
  - POS-tagger
  - Lists of names and acronyms
  - Geographical enciclopedia
  - Annotated corpora
  - WordNet



4. Resources

POS-tagger

- It provides a tag for each type of word (Name, Verb, adjective, etc.)
- Most of them include more information apart from the category (number, gender, etc.)
- Different set of tags
  - PAROLE
  - CRATER



4. Resources

POS-tagger

- EAGLES: Adjective (AQ) (AQ0FS00)
  - 1 Category: Adjetive A
  - 2 Type: Qualificative Q
  - 3 Grade:

    Positive P Comparative C Superlative S

    Intensive I Apreciative A
  - 4 Gender:

    Male M Female F Common C
  - 5 Number:

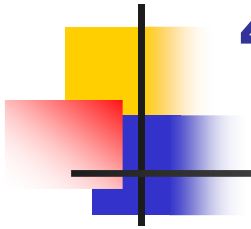
    Singular S Pural P Invariable I
  - 6 Case: 0
  - 7 Function:

    Modificator M Especificator S

4. Resources

POS-tagger

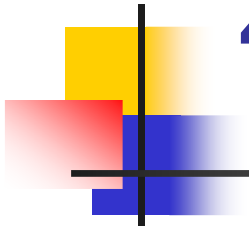
- EAGLES: Name (NC) (NCFS000)
  - 1 Category: Name N
  - 2 Type:
    Common C Proper P
  - 3 Gender:
    Male M Female F Common C
  - 4 Number:
    SingularS Plural P Invariable N
  - 5 Case: 0
  - 6 Semantic gender 0
  - 7 Grade: apreciative A



4. Resources

POS-tagger

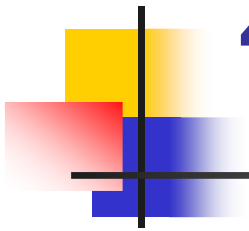
- EAGLES: Verb (VMI) (VMIP1SM)
  - 1 Category: Verb V
  - 2 Type:
    Main M Auxiliar A
  - 3 Mode:
    Indicative I Subjuntive S Modal M
    Conditional C Infinitive N Gerund G
    Participle P
  - 4 Tense:
    Present P Imperfect I Future F Past S
  - 5 Person:
    First 1 Second 2 Third 3
  - 6 Number: Singular S Plural P
  - 7 Gender: Male M Female F



4. Resources

Lists of names and acronyms

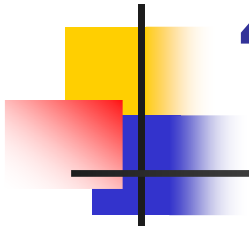
- Manually created
  - People
  - Locations
  - Companies
- Learned from annotated corpora
- Sometimes are part of the POS-tagger



4. Resources

Geographical enciclopedia

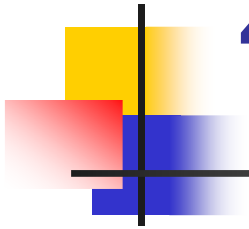
- The same as the lists of names or acronyms
- Manually generated (commercial enciclopedias)
- Generated from annotated corpora



4. Resources

Annotated corpora

- Corpora that identify (tags) certain types of information (Entities, anaphora, senses, etc.)
- They are necessary in methods based on machine learning or statistical ones
- There are corpora that are:
  - Supervised
  - Unsupervised



4. Resources

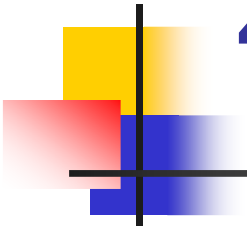
WordNet

- Tool used in different tasks of NLP
- It is within the set of dictionaries and electroninc corpora that support different types of information such as lexical, syntactic, semantic, etc.
- It is a database built over an extensive set of English words, but recent works are in other languages too (Spanish, Italian, Dutch, etc.)

4. Resources

WordNet

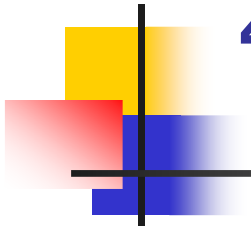
- Developed at the Princenton University under the supervision of Prof. G. Miller
- It structures the information in set of synonyms called *synset*.
- Obviously the polysemic words belongs to more than one set
- For each word it provides all the senses



4. Resources

WordNet

- For each synset provides the different semantic relationships with another synsets
  - Mero/Holonims
  - Hiper/Hiponims
  - Antonims
  - Troponims
- Lexical categories
  - Names
  - Verbs
  - Adjetives
  - Adverbs

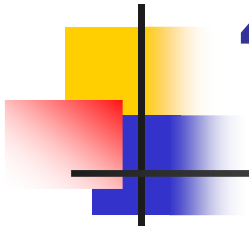


4. Resources

WordNet

- For the names there 25 main roots

1. Act
2. Animal
3. Artifact
4. Attribute
5. Body
6. Cognition
7. Communication
8. Event
9. Feeling
10. Food
11. Group
12. Location
13. Motive
14. Object
15. Person
16. Phenomenon
17. Plant
18. Possesion
19. Process
20. Quantity
21. Relation
22. Shape
23. State
24. Substance
25. Time

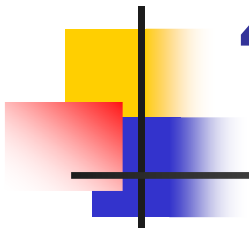


4. Resources

WordNet

- For verbs

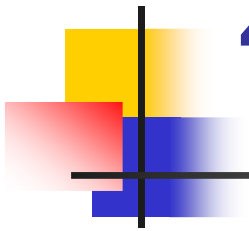
1. Body
2. Change
3. Cognition
4. Communication
5. Competition
6. Consumption
7. Contac
8. Creation
9. Emotion
10. Motion
11. Perception
12. Posession
13. Social
14. Static
15. Weather



4. Resources

EuroWordNet

- European Project that integrates Wordnets in 8 European languages
  - English
  - Dutch
  - Italian
  - Spanish
  - French
  - German
  - Czech
  - Estonian



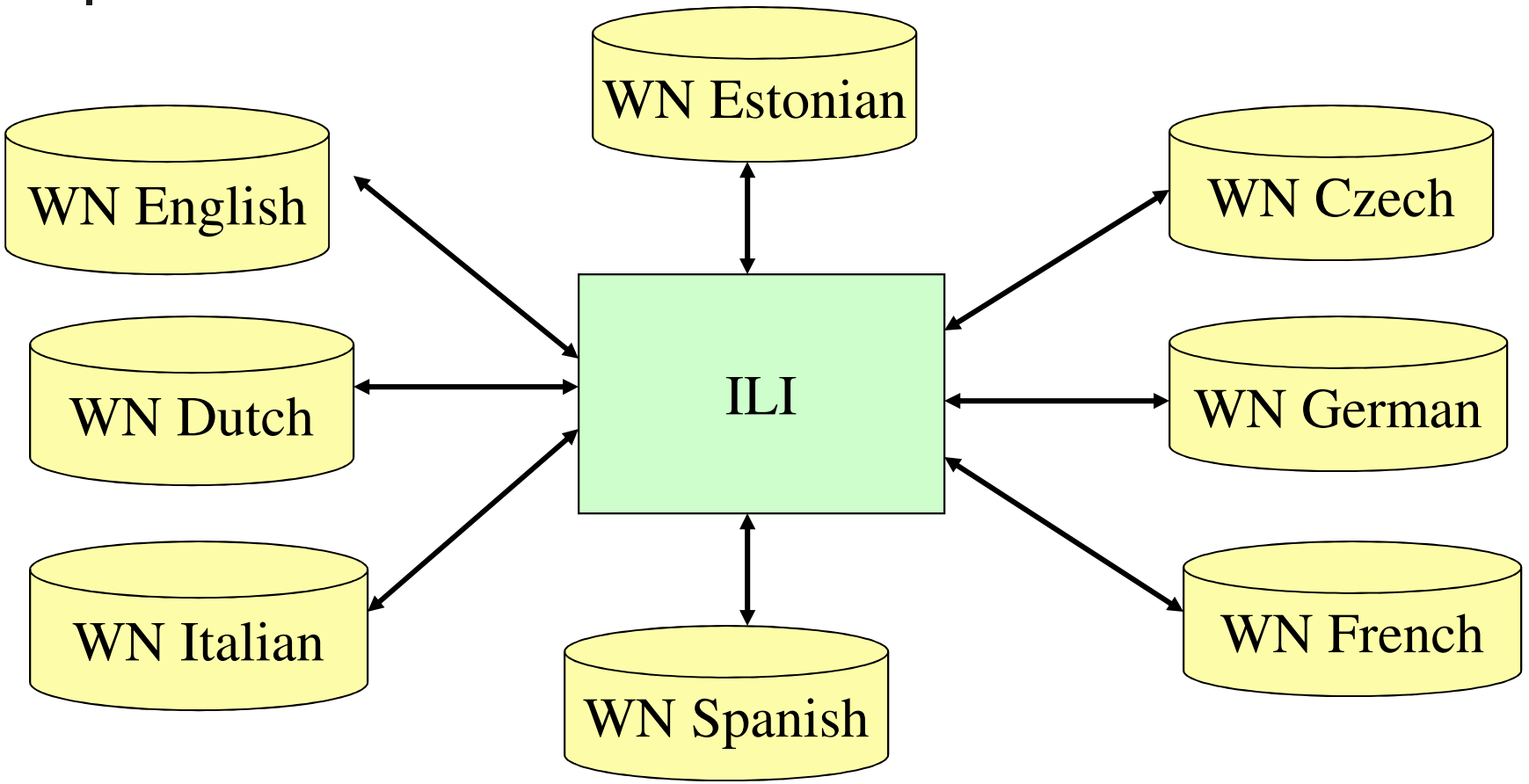
4. Resources

EuroWordNet

- Module called ILI
- Spanish Wordnet
- Ontology used in EuroWordNet

4. Resources

EuroWordNet



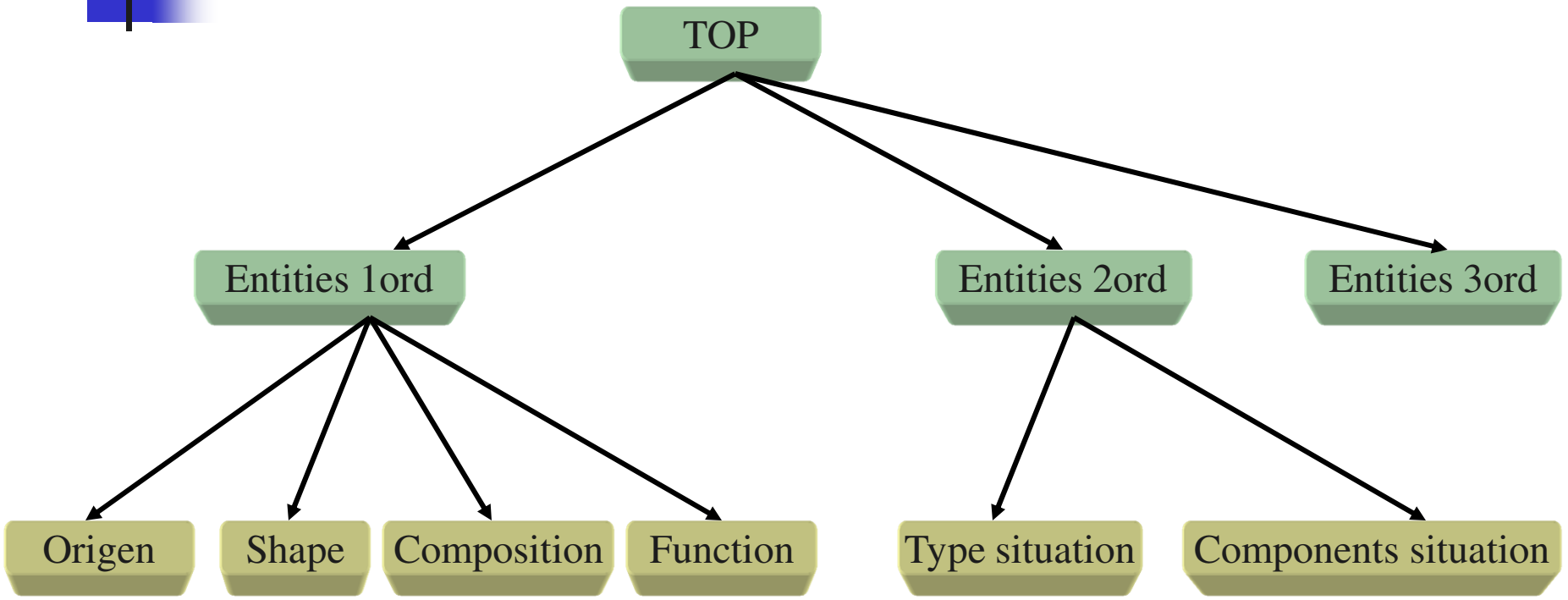
4. Resources

EuroWordNet

- WN Spanish
  - Lexical resource with semantic information
  - It is part of the EuroWordnet
  - It is formed by a set of grammatical categories: names, verbs and adjectives
  - Synsets:
    - Synonyms
    - Related through hipernym, hiponym, meronym, holonym, troponym, cause-effect, agent-action, etc.
    - There are not only relationship between the words of the same grammatical category

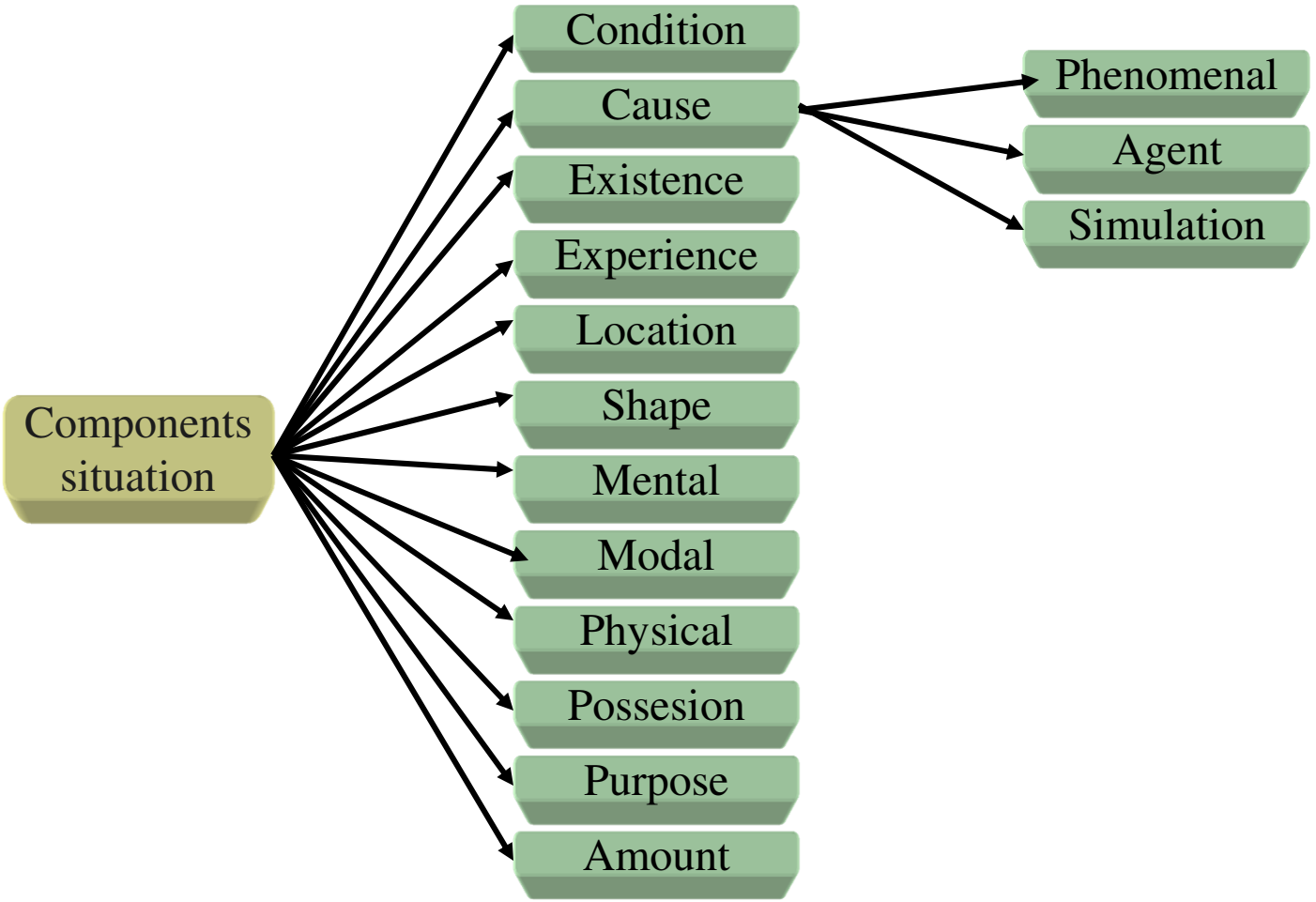
4. Resources

EuroWordNet : Ontology



4. Resources

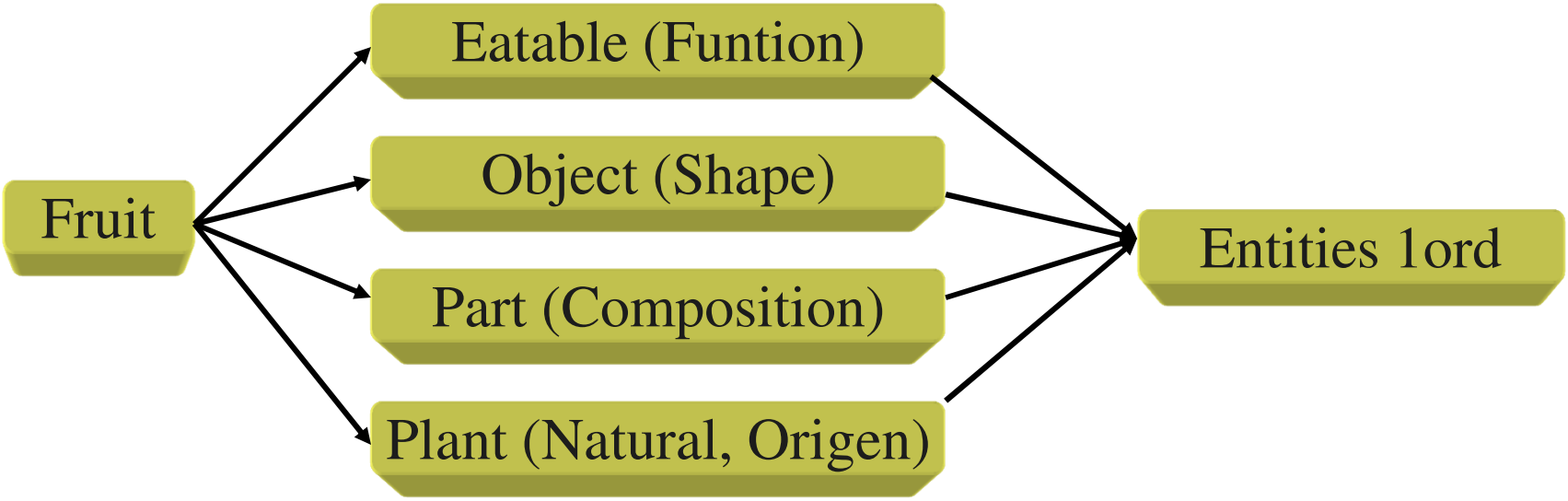
EuroWordNet : Ontology



4. Resources

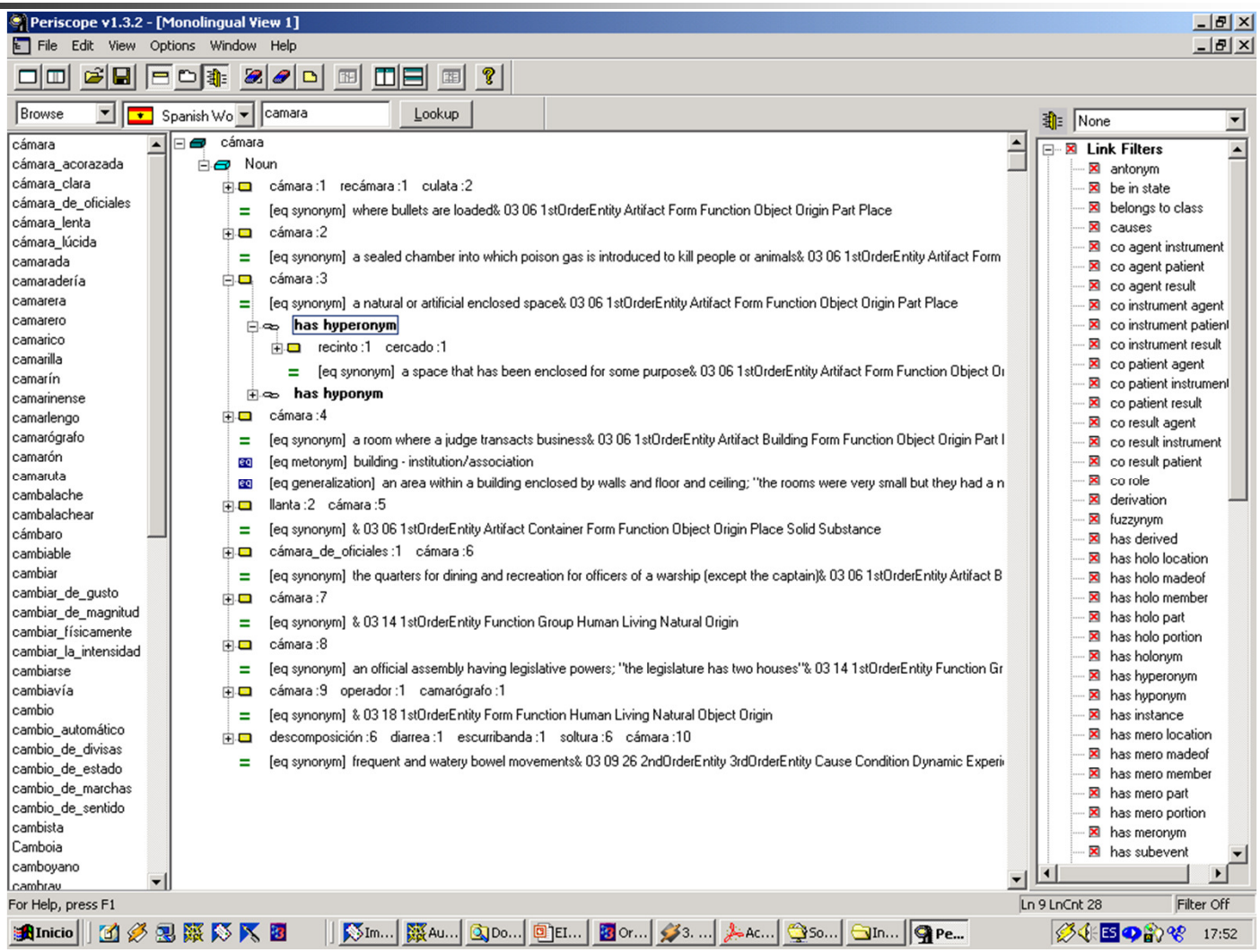
EuroWordNet : Ontology

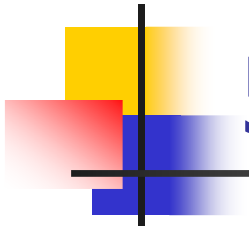
- Disjoint ontology only in the first level of it (Entities 1ord, Entities 2ord, Entities 3ord)



4. Resources

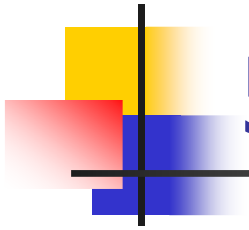
EuroWordNet





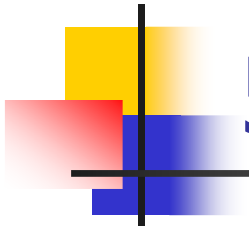
5. Algorithms

- Methods based on rules
- Statistical/Supervised methods
- Hybrid methods



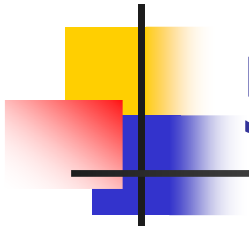
5. Algorithms

- Methods.
  - Methods based on rules
    - LaSIE (Wakao et al 1996)
    - FACILE (Black et al. 1998)
  - Statistical/Supervised methods
  - Hybrid methods



5. Algorithms

- Methods.
  - Methods based on rules
  - Statistical/Supervised methods
    - Decision trees: (Sekine et al. 1998),
    - Maximum entropy: MENE system (Borthwick et al. 1998)
  - Hybrid methods



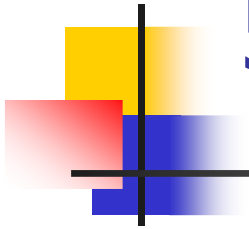
5. Algorithms

- Métodos.
  - Methods based on rules
  - Statistical/supervised methods
  - Hybrid methods
    - Mikheev et al 1999
    - Lin 1998

5. Algorithms

Based on rules

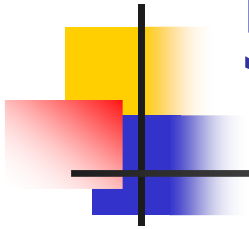
- LaSIE (Wakao et al 1996)
  - Recognized PERSON, ORGANIZATION, LOCATION and TIME
  - Uses
    - List of organization (2600 inputs)
    - List of locations (2200 names of countries, cities and provinces. Apart from 150000 places)
    - List of people (500 names)
    - List of people´s title (160)
    - List of company triggers (94)



5. Algorithms

Based on rules

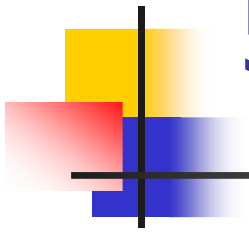
- LaSIE (Wakao et al 1996) Steps:
  - Lexical processing
  - Application of grammatical rules
  - Discourse interpretation



5. Algorithms

Based on rules

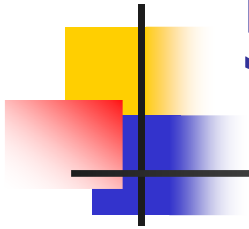
- LaSIE (Wakao et al 1996) Steps :
  - Lexical processing
    - POS Tagger: It is tagged as Proper Name all the capital words that are not in dictionaries or that appear in it like that.
    - Searching over lists(name and triggers)
  - Application of grammatical rules
  - Discourse interpretation



5. Algorithms

Based on rules

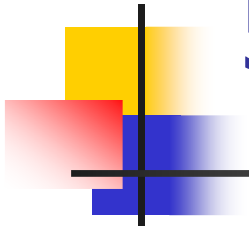
- LaSIE (Wakao et al 1996) Steps :
  - Lexical processing
  - Application of grammatical rules
    - It is integrated in the analyzer as rules of the PN, in a first step the proper names are identified, and in the second step, the rest of it is grammatically analised.
    - 177 rules (94 for ORG, 11 for LOC and 54 for PERS)
    - The analizer realises a semantic representation
  - Discourse intrepretation



5. Algorithms

Based on rules

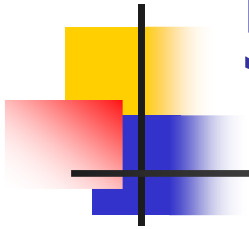
- LaSIE (Wakao et al 1996) Steps :
  - Lexical processing
  - Application of grammatical rules
  - Discourse interpretation
    - Correference of proper names
    - Inference of semantic types



5. Algorithms

Based on rules

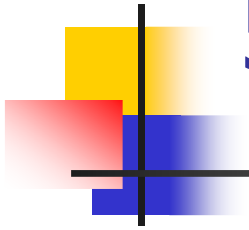
- LaSIE (Wakao et al 1996) Steps:
  - Lexical processing
  - Application of grammatical rules
  - Discourse interpretation
    - Correference of proper names
      - It recognizes alternatives forms of an entity (ALIAS)
      - It uses 31 heuristics to solve correference
    - Inference of semantic types



5. Algorithms

Based on rules

- LaSIE (Wakao et al 1996) Steps:
  - Lexical processing
  - Application of grammatical rules
  - Discourse interpretation
    - Correference of proper names
    - Inference of semantic types
      - It is used to obtain inferences that helps in the classification
      - The semantic relationships of the words that go with the entities influence in the classification of the entity (Erickson stocks, stocks related with organization then ERICKSON is organization)



5. Algorithms

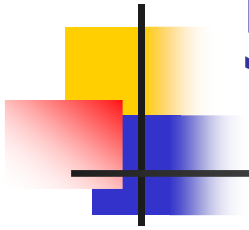
Based on rules

- FACILE (Black et al. 1998)
  - Its rule´s formalism support partial analysis
  - The rules use iteration operators like pattern-matching ones.
  - Its notation is simpler that the classical pattern-matching languages.
  - The rules are using weights to compete between similar rules
  - It has a module to detect correference of names
  - The system is not using machine learning

5. Algorithms

Based on rules

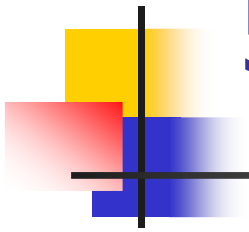
- FACILE (Black et al. 1998)
  - Notation of the rules
    - A ⇒ B \ C / D
    - *[syn=NP, sem=ORG] (0.9) ⇒*
      | [norm="university"],
      [token="of"],
      [sem=REGION | COUNTRY | CITY] /;



5. Algorithms

Statistical/Supervised methods

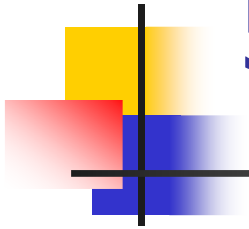
- MENE system (Borthwick et al. 1998)
  http://www.choicemaker.com/borthwick.htm
  - It is using maximum entropy
  - It used a huge amount of knowledge resources in the tagging decisions (capitalization feature, lexical features, text, etc)
  - It is using names, companies and sufixes dictionaries
  - It needs an annotated corpus to learn



5. Algorithms

Statistical/Supervised methods

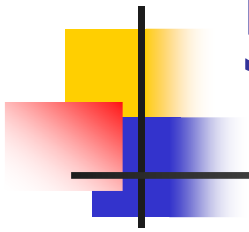
- Decision trees: (Sekine et al. 1998)
  - Algorithm for NE in Japanese
  - It is using a supervised method
  - Algorithm:
    - Phase 1: Building the decision tree (training)
    - Phase 2: Application of the tree to new texts (testing)



5. Algorithms

Statistical/Supervised methods

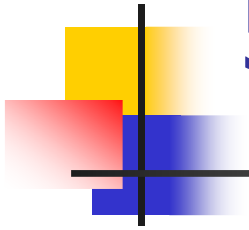
- Decision trees: (Sekine et al. 1998)
  - It uses three sets of different features:
    - POS tags
    - Information about letter type
    - Special dictionaries



5. Algorithms

Hybrid methods

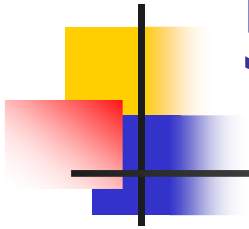
- LTG system Mikheev et al 1999
  - A string like "Adam Kluver" has an internal structure that suggests a person
  - But it can be an alias or acronym of "Adam Kluver Ltd." (organization) or "Adam Kluver Country Park" (location)
  - The searching over a list cannot help: maybe it is not in a list, or it is in more than one, or at a wrong list.
  - Sometimes, in the text, it appears what it is called contextual information that clarifies the type of entity
  - This system makes a decision after studying the contextual information.



5. Algorithms

Hybrid methods

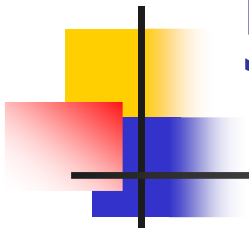
- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
  - Partial Match 1
  - Rule Relaxation
  - Partial Match 2
  - Title Assignment



5. Algorithms

Hybrid methods

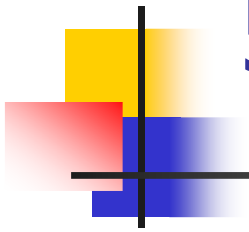
- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
    - They combine internal and external evidences
    - They are applied a POS and a simple semantic tagging
    - They only tag the things that appearing in lists are also appearing in the proper context.
  - Partial Match 1
  - Rule Relaxation
  - Partial Match 2
  - Title Assignment



5. Algorithms

Hybrid methods

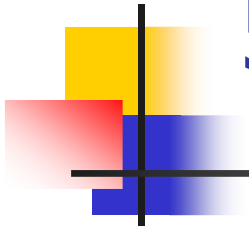
- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
  - Partial Match 1
    - It generates substrings but preserving the order of the identified entities (Adam Ltd., Adam Kluver, Kluver Ltd.)
    - Searching of these substrings in the text, pre-tagging them.
    - Application of a maximum entropy model with the contextual information (position in the sentence, etc.). The admitted in the probabilistic model preserve the tag.
  - Rule Relaxation
  - Partial Match 2
  - Title Assignment



5. Algorithms

Hybrid methods

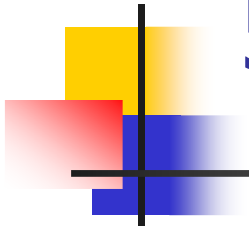
- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
  - Partial Match 1
  - Rule Relaxation
    - Application of contextual restriction rules much more relaxed (it prevails the internal structure Adam Kluver is PERSON, if it was not previously decided)
    - Conjuctions in ORGANIZATIONS are resolved (looking if each part appears more times in the text by itself)
    - Resolution of the problem of starting of the sentence and genitive looking for another appearances in the text.
  - Partial Match 2
  - Title Assignment



5. Algorithms

Hybrid methods

- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
  - Partial Match 1
  - Rule Relaxation
  - Partial Match 2
    - Recognition of the alias "White" when John White is tagged
    - Application of entropy model
    - Conjunction used to imply entities of the same type
  - Title Assignment



5. Algorithms

Hybrid methods

- LTG system Mikheev et al. 1999. Steps:
  - Sure-fire Rules
  - Partial Match 1
  - Rule Relaxation
  - Partial Match 2
  - Title Assignment
    - Identification of entities in titles (Everything in capital letters)
    - Comparison with text´s entities
    - Application of maximum entropy model

5. Algorithms

Hybrid methods

- Lin 1998.
  - It is based on searching of combined words that usually appear together (collocation) Example: “in fact”
  - Using a DB of rules for “collocations” in order to recognize NE in texts
  - It is using Naive-Bayes classifier to classify proper names not recognized by the “collocations”

5. Algorithms

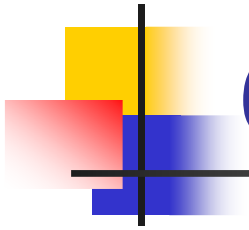
- Results

System	Prec.	Rec.	F-mes.	Type
LaSIE	93	91	91.75	Rules
FACILE	93	92		Rules
Sekine et al.	85	75	79.49	Statistical
MENE	91	78	84.22	Statistical
LTG				Hybrid
Lin	87	85	86.37	Hybrid

5. Algorithms

- Results of the LTG system (Mikheev et al. 1999)

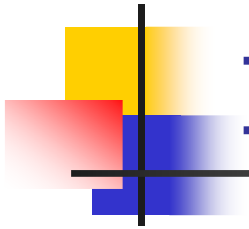
Type of entity	Complete Dict.		Limited dicti.		Some loc.		No Dictionary	
	P	R	P	R	P	R	P	R
organization	93	90	90	87	89	87	85	86
person	98	96	97	92	97	90	95	90
location	94	95	92	91	90	85	59	46



6. Demos

- Demos.
  - POSNER (Korean)
  - SweNam (Swedish)
  - GPLSI (Spanish)

Inference module



Inference module

- Through inference rules information that appear implicit in the text is extracted

Inference module

- Through inference rules information that appear implicit in the text is extracted

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM :
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM :
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION :
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH: <Person-0001
<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:
<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach +
Team.coach= Person1

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH : <Person-0002>

<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS:

Inference module

- Through inference rules information that appear implicit in the text is extracted
  - Sergio Kresic replaces Bernd Krauss in Mallorca´s

Person1 + replace (active) + *Person2* *(position=trainer)* ->
Person1.team=Person2.team +
Person1.position=coach +
Team.coach= Person1 +
Person2.team=<<NULL>>

<Team-0001> :=
NAME : Mallorca
FIELD : Son Mox
ADDRESS :
PRESIDENT :
COACH : <Person-0002>

<Person-0001>:=
NAME :Bernd Krauss
ALIAS:
POSITION : Coach
TEAM :
GOALS:

<Person-0002>:=
NAME :Sergio Kresic
ALIAS:
POSITION : Coach
TEAM : <Team-0001>
GOALS: