Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Extracción de Información

Introducción

- ¿Qué es la Información?
- ¿Qué es un Sistema de Información?
- ¿Qué es la Recuperación de Información?
- ¿Qué es la Extracción de Información?

Introducción

Información:

(1) ***5.*** *f. Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada.* (RAE)

(2) Es un conjunto organizado de datos procesados, que constituyen un mensaje que cambia el estado de conocimiento del sujeto o sistema que recibe dicho mensaje. (Wikipedia)

Introducción

Sistema de Información:

conjunto de funciones o componentes interrelacionados que forman un todo, es decir, obtiene, almacena, procesa y distribuye información para apoyar la toma de decisiones y el control en una organización. (Wikipedia)

Recuperación de Información:

es la disciplina encargada de la representación, almacenamiento y organización, y su posterior acceso y recuperación para responder a las necesidades de un usuario (Salton)

Introducción

Extracción de Información:

es un tipo de recuperación de la información cuyo objetivo es extraer automáticamente información estructurada o semiestructurada desde documentos legibles por una computadora. (Wikipedia)

Tarea de la EI

- Objetivo: responder consultas sobre textos
- Analizar texto sin restricciones para extraer cierto tipo de información contenida en él
- Intenta convertir información no estructurada según un “esquema de BD”

Tarea de la EI

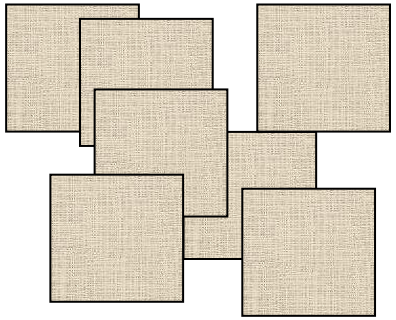
- Proceso
  - Identificar porciones de información (entidades y relaciones entre ellas) en un documento de texto
  - Pasar de texto a relaciones estructuradas

- Hipótesis
  - Existe información que podría representarse según una estructura más “manejable”

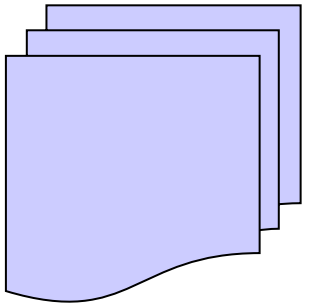
EI & RI

La idea es que un SRI y un SEI se complementen …

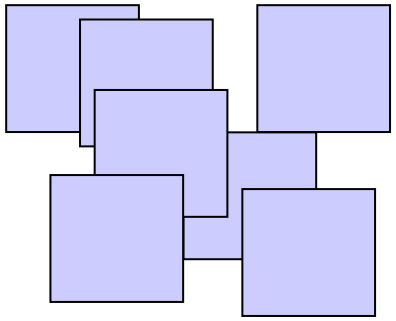
En los SRI



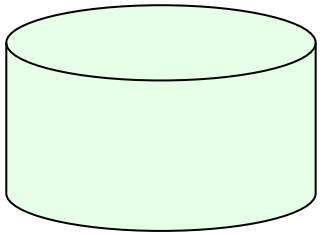
Se obtienen documentos



En los SEI



Se obtienen datos ó información útil



Extracción de Información

- Típicamente extrae:
  - Entidades
  - Relaciones
  - Eventos
- A partir de los documentos existentes en un dominio restringido
- Los sistemas por lo general "conocen" el dominio de actuación
- La idea es tener por ejemplo plantillas <***atributo,valor*** > a efectos de poder evaluar mejor el proceso de extracción

Extracción de Información

Ejemplos:

"Se concede licencia reglamentaria al docente Juan Pérez por un mes"

<Licencia>
<Nombre, Juan Pérez>
<Duración,1 mes>
<Motivo, Licencia reglamentaria>
</Licencia>

"Se resuelve otorgar una licencia de 1 semana a María Gómez para asistir al encuentro"

<Licencia>
<Nombre, María Gómez>
<Duración,1 semana>
<Motivo, Licencia especial>
</Licencia>

Extracción de Información

Ejemplo:

Sistema de extracción de información sobre el dominio de las inscripciones de inmuebles o terrenos a expropiar por un municipio para realizar obras.

Extracción de Información

Montevideo, 29/05/2006 Ref: 55847

Señor Director del Registro de la Propiedad Inmueble de Montevideo:

El suscrito Esc. José Somoza, cédula de identidad 1:373.017, integrante del cuerpo de escribanos del Servicio de Escribanía de la Intendencia Municipal de Montevideo, solicita a Ud. se proceda a la expropiación del bien empadronado con el Nro. 1019, ubicado en la calle Solano Antuña 1524 de esta ciudad con el fin de ensanchar la citada calle.

El mismo fue debidamente inscripto en la Sección respectiva del Registro que Ud. dirige con el Nro: 40607, Folio 30012 Libro 124 a los 26 días del mes de Mayo del 2004 teniendo como última reinscripción Nro 12556 Fo 8774 Lo 45 con fecha 23/04/2005.

Extracción de Información

<PLANTILLA>
NroReferencia,
Fecha
<INSCRIPCIÓN>
<ESCRIBANO_ACTUANTE>
Nombre, CI
</ESCRIBANO_ACTUANTE>
<MUNICIPIO> Nombre </MUNICIPIO>
<OBJETO_A_EXPROPIAR>
<UBICACIÓN> Padron, Direccion </UBICACIÓN>
<INSCRIPCIÓN> Nro, Folio, Libro, Fecha </INSCRIPCION>
<ULTIMA_INSCRIPCIÓN> Nro, Folio, Libro, Fecha
</ULTIMA_INSCRIPCION>
</OBJETO_A_EXPROPIAR>
<DESTINO> TipoDestino </DESTINO>
</INSCRIPCION>
</PLANTILLA>

Extracción de Información

Montevideo, **29/05/2006** Ref: **55847**

Señor Director del Registro de la Propiedad Inmueble de Montevideo:

El suscrito Esc. **José Somoza**, cédula de identidad **1:373.017**, integrante del cuerpo de escribanos del Servicio de Escribanía de la Intendencia Municipal de **Montevideo**, solicita a Ud. se proceda a la expropiación del bien empadronado con el Nro. **1019**, ubicado en la calle **Solano Antuña 1524** de esta ciudad con el fin de **ensanchar la citada calle**.

El mismo fue debidamente inscripto en la Sección respectiva del Registro que Ud. dirige con el Nro: **40607**, Folio **30012** Libro **124** a los **26 dias del mes de Mayo del 2004** teniendo como última reinscripción Nro **12556** Fo **8774** Lo **45** con fecha **23/04/2005**.

Sin otro particular, saluda a Ud. muy atentamente,

Extracción de Información

<PLANTILLA>
NroReferencia: 55847 , Fecha: 21/07/2006
<INSCRIPCIÓN>
<ESCRIBANO_ACTUANTE>
Nombre: José Somoza, CI:1:373.017
</ESCRIBANO_ACTUANTE>
<MUNICIPIO> Nombre: Montevideo </MUNICIPIO>
<OBJETO_A_EXPROPIAR>
<UBICACIÓN> Padron: 1019, Direccion: Solano Antuña 1524 </UBICACIÓN>
<INSCRIPCIÓN> Nro: 40607, Folio: 30012, Libro: 124,
Fecha: 26 dias del mes de mayo del 2004
</INSCRIPCION>
<ULTIMA_INSCRIPCIÓN> Nro: 12556, Folio: 8774, Libro: 45,Fecha: 23/04/2005
</ULTIMA_INSCRIPCION>
</OBJETO_A_EXPROPIAR>
<DESTINO> TipoDestino: ensanche de calle </DESTINO>
</INSCRIPCION>
</PLANTILLA>

Extracción de Información

Regla de Whisk (S. Soderland, 1998)

DOCUMENT:	EXTRACTED DATA:
Capitol Hill- 1 br twnhme. D/W W/D. Pkg incl $675. 3BR upper flr no gar. $995. (206) 999-9999 	<Bedrooms: 1 Price: 675> <Bedrooms: 3 Price: 995>

Extraction rule: * (<Digit>) 'BR' * '$' (<Nmb>)
Output: Rental {Bedrooms @1} {Price @2}

Extracción de Información

Regla de Whisk (S. Soderland, 1998)

DOCUMENT:
Capitol Hill- 1 br twnhme.
D/W W/D. Pkg incl $675.
3BR upper flr no gar. $995.
(206) 999-9999

EXTRACTED DATA:
<Bedrooms: 1
Price: 675>
<Bedrooms: 3
Price: 995>

Extraction rule: * (<Digit>) 'BR' * '$' (<Nmb>)
Output: Rental {Bedrooms @1} {Price @2}

Historia

Conferencias MUC (Messagge Understanding Conferences)

- celebradas entre 1987 y 1997; patrocinadas por el gobierno de EEUU
- comenzaron siendo sobre mensajes cortos para trabajar luego con artículos sobre dominios específicos
- actuaban como lugar para intercambio entre investigadores y tendencias
- para evaluar se usaban *precision*, *recall* y *F*

Historia

Conferencias MUC

- MUC-1 (1987) y MUC-2 (1989) : análisis de mensajes en operaciones navales
- MUC-3 (1991) y MUC-4 (1992) : actividades terroristas en Latinoamérica
- MUC-5 (1993) : extracción de eventos en noticias
- MUC-6 (1995) : extracción de eventos de actividad de empresas (fusiones, compras)
- MUC-7 (1997) : artículos sobre vehículos espaciales y lanzamiento de misiles y satélites

Historia

- La serie MUC es continuada por la serie ACE (*Automatic Content Extraction*)
- De 2000 a 2008
- Cobertura más general de noticias de prensa y extracción de contenidos
- En las últimas ediciones se incorporan otras lenguas (chino, árabe, español)

Historia

- Actualmente y desde el 2009 se evalúan sistemas de EI en KBP (*Knowledge Base Population*), que forma parte de la TAC (*Text Analysis Conference*)
- También en CONLL (*Conference on Natural Language Learning*)
- Formato:

  Se da una entidad con nombre, un artículo de referencia y alrededor de 2: de textos para realizar extracción sobre la entidad, definiendo atributos y poblando la KB

Corpus

¿Qué es un Corpus ?

- Es una colección de material lingüístico de ejemplos reales de uso de la lengua
- Es de utilidad en diferentes áreas, principalmente en lingüística computacional y lingüística teórica

Corpus

¿Cómo se construye un corpus?

- Recopilación de un conjunto de documentos
- Hay que definir las características deseadas:
  - escrito / oral
  - idioma (un idioma o multilingüe)
  - tipo de texto (prensa, literario, científico, ...)
  - dominio (arte, lingüística, bio-informática, ...)
  - anotado / no anotado (conjunto de etiquetas)

Brown Corpus

- 1961
- inglés americano
- 1.000.000 palabras
- sólo material escrito (500 textos de aprox. 2000 palabras)
- anotado (*POS tags*)
- disponible pagando licencia
- contiene material clasificado en 15 categorías:
  - prensa: reportajes, editoriales, …
  - literarios: misterio, ciencia ficción, romance, …
  - religión
  - humor
  - …

Penn Treebank Corpus

- 1989
- inglés americano
- 4:500.000 palabras
- banco de árboles lingüísticos
- anotado (*POS tags* + información sintáctica)
- disponible pagando licencia
- contiene:
  - artículos científicos
  - noticias
  - capítulos de obras literarias
  - oraciones de manuales de computación
  - el corpus Brown completo re-etiquetado

Corpus CREA

Corpus de Referencia del Español Actual (Real Academia Española)

- inicio 1993
- español actual
- 160:000.000 palabras
- material escrito y oral
- no anotado
- disponible para consultas en línea (sin suscripción)

http://corpus.rae.es/creanet.html

Corpus CORDE

Corpus Diacrónico del Español (Real Academia Española)

- inicio 1993
- español anterior a 1975, todas las variedades
- 250:000.000 palabras
- material escrito
- no anotado
- disponible para consultas en línea (sin suscripción)

http://corpus.rae.es/cordenet.html

Corpus del Español

(Mark Davies, Brigham Young University)

- 2001
- español histórico y actual
- 100:000.000 palabras
- material oral y escrito
- anotado (*POS-tags*, *lemas*)
- disponible para consultas en línea

http://www.corpusdelespanol.org/

Corpus ANCORA

Universidad de Barcelona

- 2008 (aprox.)
- español (500.000 palabras) y catalán (500.000 palabras)
- principalmente textos periodísticos
- anotaciones:
    - categoría morfológica
    - constituyentes y funciones sintácticas
    - estructura argumental y papeles temáticos
    - clase semántica verbal
    - sentidos de WordNet nominales
    - entidades nombradas
    - correferencia

http://clic.ub.edu/es/ancora-es/

Corpus

Algunos Corpus “nuestros”

- Corpus de noticias extraídas de la web
  - grande
  - se sigue generando
- Corin
  - proyecto desarrollado por la Licenciatura en Lingüística
  - conjunto de etiquetas muy completo
  - corpus pequeño
- Surgidos en proyectos particulares…
  - Opiniones
  - Eventos
  - Expresiones Temporales
  - Wikipedia
  - Tweets de humor

Tesauro

- Un tesauro es una lista de palabras o *términos* controlados y estructurados empleados para representar *conceptos* ordenados en forma alfabética, temática y jerárquica
- Se pueden construir
  - manualmente
  - en forma automática
- Los términos aparecen con:
  - una definición
  - una lista de descriptores
  - una lista de *relaciones* semánticas:
    - de equivalencia: *sinonimia*
    - asociativas: *hiperonimia*, *homonimia*, *antonimia*
    - jerárquicas: *es_un*, *parte_todo*

Extracción de Información

- Tareas
  - Reconocimiento de entidades con nombre (NER)
  - Resolución de correferencias
  - Extracción de relaciones semánticas entre entidades
  - Resolución y reconocimiento de expresiones temporales
  - Asignación de roles semánticos

Ejemplo

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Ejemplo

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Ejemplo

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Ejemplo

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Ejemplo

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Ejemplo

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Ejemplo

Barcelona autorizó a **Luis Suárez** a viajar el lunes a **Montevideo** para estar a la orden de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Ejemplo

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Barcelona autorizó a Luis Suárez a viajar el lunes a Montevideo para estar a la orden de la selección para los partidos de las eliminatorias ante Argentina y Paraguay, pese a que el jugador no fue incluido en la convocatoria del club catalán para el partido de este sábado ante Alavés, por la segunda fecha de la Liga Española.

Extracción de Relaciones

La mayor parte de los trabajos extraen:

- relaciones entre entidades mencionadas en la misma oración
- relaciones predeterminadas (dirección de una empresa, club donde juega un jugador, etc.)
- relaciones binarias; se habla de extracción de eventos cuando hay más de 2 argumentos

Extracción de Relaciones

Barcelona autorizó a **Luis Suárez** a **viajar** el lunes a **Montevideo** para **estar a la orden** de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el **jugador** no fue **incluido** en la convocatoria del **club catalán** para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

1) Relación : *autorizar*
2) 2 argumentos: A *autoriza a* B
3) Podríamos agregar CUÁNDO, A QUÉ, PARA QUÉ lo autorizó y entonces no sería solo extracción de relaciones sino también de **eventos**

Extracción de Relaciones

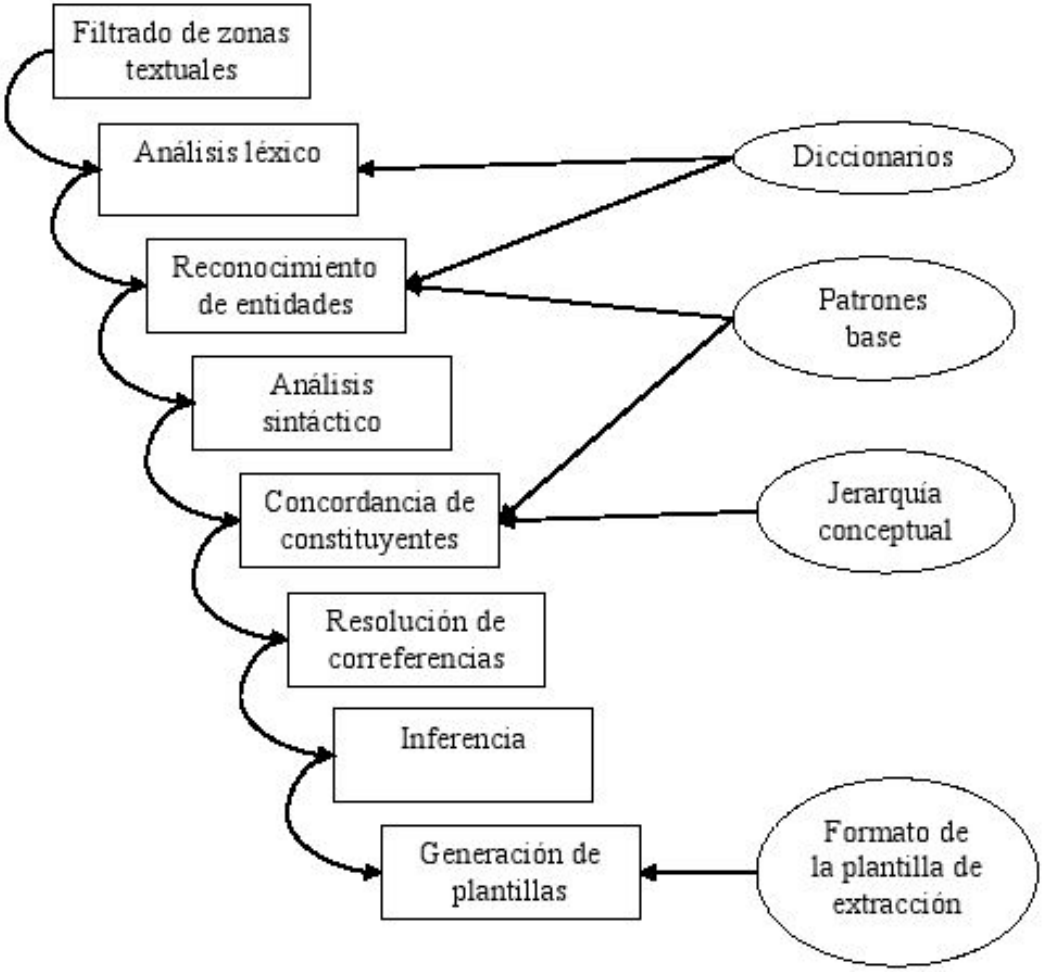
Barcelona **autorizó** a **Luis Suárez** a **viajar** el lunes a **Montevideo** para **estar a la orden** de la selección para los partidos de las eliminatorias ante **Argentina** y **Paraguay**, pese a que el **jugador** no fue **incluido** en la convocatoria del **club catalán** para el partido de este sábado ante **Alavés**, por la segunda fecha de la **Liga Española**.

Otras relaciones que se informan:

- viajar
- estar a la orden
- incluido en la convocatoria

En general se procede por etapas, primero las entidades y luego las relaciones

Arquitectura genérica de un SEI



Enfoques para la construcción de un SEI

- Métodos manuales
- Sistemas entrenados
  - aprendizaje supervisado
  - aprendizaje semi-supervisado

Enfoques para la construcción de un SEI

- **Manuales:** basados en conocimiento lingüístico
  - Patrones descubiertos por un experto humano mediante la inspección del corpus

    Ejemplo: Título ([\b] mayúscula minúscula+) → persona;
    siendo Título = {Sr.,Sra.,Ing., ...}

  - Es posible también manejar una lista de nombres de pila
  - Criterios similares pueden encontrarse para organizaciones, lugares, etc.
  - Llevan un importante trabajo de ajuste

Enfoques para la construcción de un SEI

- **Manuales:** basados en conocimiento lingüístico
  - Ventajas
    - Complejidad conceptual reducida y sencillos de desarrollar
    - Son sistemas que obtienen muy buen rendimiento
    - Posibilidad de identificar patrones no presentes en el corpus
    - Es común tener largas listas de personas, lugares geográficos, etc.
  - Algunos trabajos
    - Shrank (1972); Minsky (1975); Fillmore & Baker (2001): Hobbs (1996)

Enfoques para la construcción de un SEI

- Manuales: basados en conocimiento lingüístico
  - Desventajas
    - Recursos lingüísticos y tiempos de escritura de reglas
    - Dificultades para adaptarse a distintos dominios
    - Proceso de desarrollo trabajoso (1500 hs/hombre)
    - Las listas funcionan bien, pero deben ser actualizadas periódicamente
    - Hay problemas que requieren desambiguar (p.ej. Uruguay). Difícil solo con listas y patrones

Enfoques para la construcción de un SEI

- Sistemas entrenados
  - Aprendizaje supervisado
    - Uso de métodos estadísticos (HMM, CRF)
    - Aprendizaje de reglas a partir de:
      - corpus anotados manualmente
      - interacción con el usuario
    - Es un problema secuencial y es usual modelarlo con etiquetado BIO
      - B: 1er token del segmento (o único)
      - I: tokens siguientes del segmento
      - O: tokens externos

Enfoques para la construcción de un SEI

- Sistemas entrenados
  - Aprendizaje semi-supervisado
    - Uso de métodos estadísticos (Clustering)
    - Nos permite reducir la cantidad de ejemplos anotados manualmente
    - Hipótesis: Distintas ocurrencias de un mismo nombre tienen preferentemente la misma clase

Enfoques para la construcción de un SEI

- Sistemas entrenados
  - Ventajas
    - La portabilidad a otros dominios sencilla
    - La generación de reglas "*data driven*" asegura la cobertura completa de ejemplos
  - Desventajas
    - Requieren de grandes volúmenes de datos de entrenamiento
    - Datos de entrenamiento caros o inexistentes
    - Cambios en especificaciones pueden requerir re-anotaciones
    - Si tenemos “semillas” de un tipo predominante va a absorber todos los casos → habría que aprender todas las clases a la vez

Extracción de Información

Criterios de elección de un enfoque:

- Disponibilidad de recursos lingüísticos
- Posibilidad de escritura de reglas
- Disponibilidad de datos de entrenamiento
- Cambios posibles en la especificación
- Performance requerida