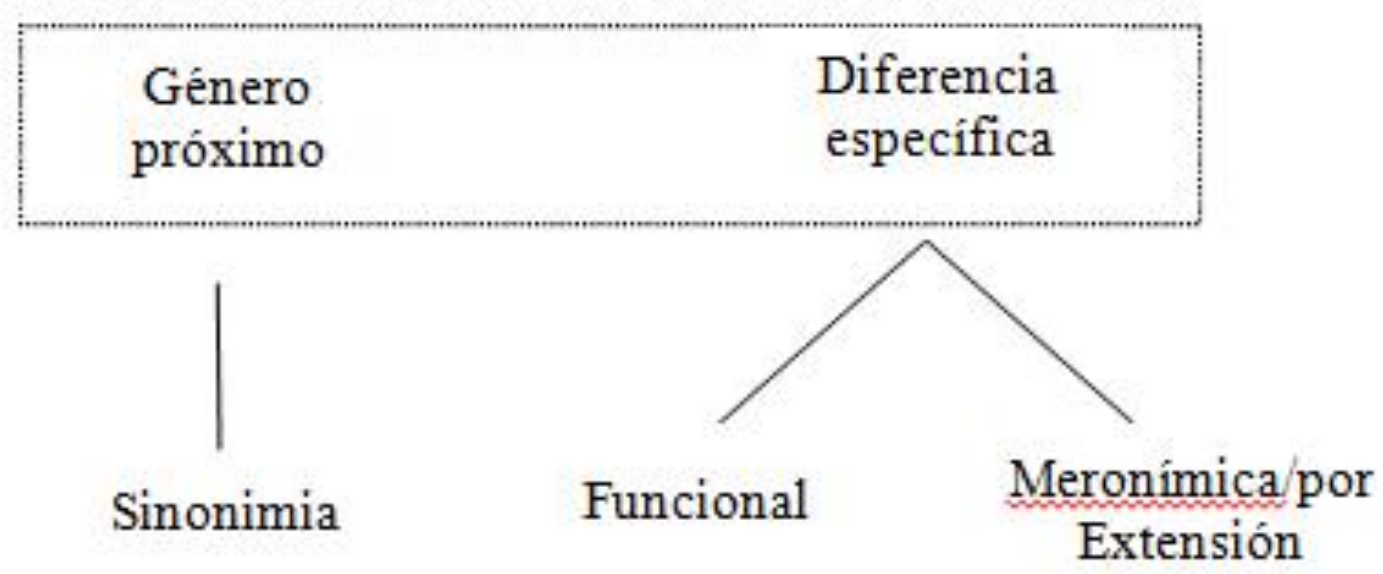
EXTRACCIÓN AUTOMÁTICA DE CONTEXTOS DEFINITORIOS EN A PARTIR DE APRENDIZAJE DE MÁQUINA

Víctor Germán Mijangos de la Cruz
Grupo de Ingeniería Lingüística
Posgrado en Lingüística

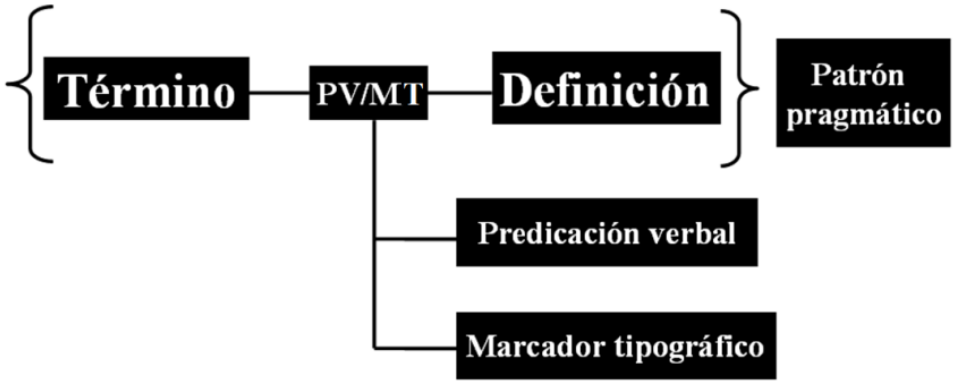
Contextos Definitorios

- Fragmento textual donde se encuentra un término y su definición
- Ayudan a entender un término
- Usos diversos en terminología, terminótica o PLN
- Estructura particular

Clasificación de CDs



Estructura de CDs



Patrón pragmático | Término | Predicación verbal | Marcador tipográfico

<Matemáticamente>, <la Teoría Lineal > <puede ser considerada como> <:> <una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje.>

CD

Definición

Reto

- ¿Con el conocimiento existente de los CDs es posible crear un sistema que los extraiga automáticamente?

- ¿Cómo debe ser este sistema?

Problemas

- Encontrar los patrones verbales
- Determinar cuándo los patrones introducen CDs
- Encontrar CDs en documentos
- Hacer todo esto automáticamente

ECODE

¿Qué es el eCode?

- Extractor de Contextos Definitorios

- Sistema computacional que extraes CDs en textos especializados

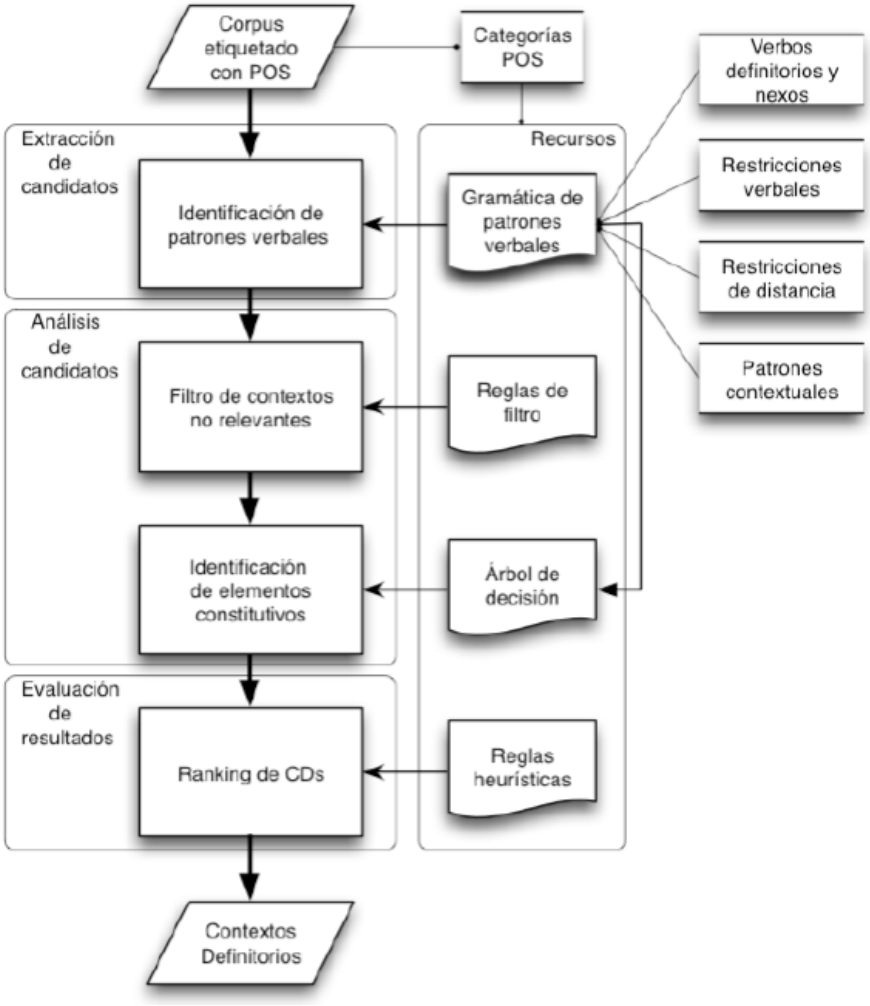
¿Cómo funciona?

- Expresiones regulares
- Patrones verbales

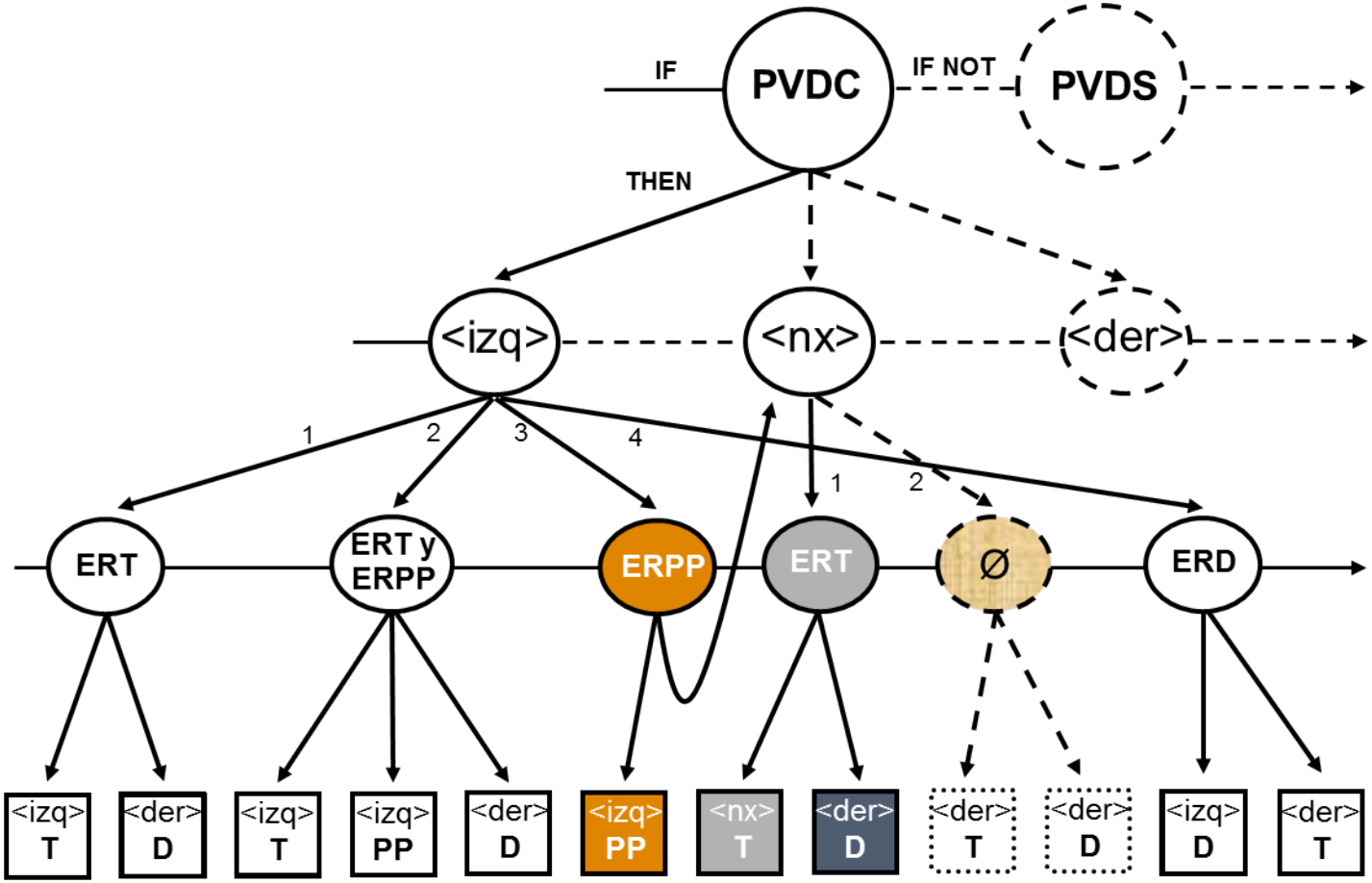
Definición	Verbos	Partículas asociadas
Analítica (Predicación primaria)	*Referir, Representar, Ser, Significar*	*a*
Analítica (predicación secundaria)	*Caracterizar,Comprender, Concebir, Conocer, Considerar, Definir, Describir, Entender, Identificar Visualizar*	*como por*
Sinonímica	*Equivaler, Llamar, Nombrar Ser_igual, Ser_similar*	*también a igual a similar a*
Funcional (Predicación primaria)	*Emplearse, Encargar, Funcionar Ocupar, Permitir, Servir, Usar Utilizar*	*de para*
Extensional (Predicación primaria)	*Componer, Comprender, Consistir Constar, Contar, Constituir, Contener, Incluir, Integrar*	*de por con*

Variable	Expresión regular	Ejemplos
Generales		
$token	[^ /]*	gen
$tag	/[A-Z][^]*	/N5-MS
$palabra	$token.$tag;	gen/ N5-MS
$inicioLinea	<doc_codi [^]*?>:	<doc_codi m00449>:
Clases de palabras		
$adj	$token./J[^]*	rojo/JQ--MS
$adv	$token./D[^]*	completamente/D
$conj	$token./C	y/C
$det	$token./(A|E|J(N|6))[^]*	los/AMP
$prep	$token./P	de/P
$pron	$token./R[^]*	se/REE6366
$signo	$token./Z	./Z
$sus	$token./N[^]*	cromosoma/N5-MS
$vbo	$token./V[^]*	entender/VI----
Tipos de verbos		
$vTag	/V[^G][^]*	/VDR3P-
$verboCon	$token./V(D|J|R)[^]*	define/VDR3S-
$verboInf	$token./VI[^]*	definir/VI----
$verboPar	$token./VC[^]*	definidos/VC--PM
$verboGer	$token./VG[^]*	entendiendo/VG----

Algoritmo


Corpus etiquetado con POS
Categorías POS
Verbos definitorios y nexos
Restricciones verbales
Restricciones de distancia
Patrones contextuales
Extracción de candidatos
Identificación de patrones verbales
Recursos
Gramática de patrones verbales
Análisis de candidatos
Filtro de contextos no relevantes
Reglas de filtro
Identificación de elementos constitutivos
Árbol de decisión
Evaluación de resultados
Ranking de CDs
Reglas heurísticas
Contextos Definitorios

Árbol de decisiones



Problemas con el eCode

- Expresiones regulares
  - Pueden no funcionar adecuadamente
  - Se basan en POST
- Necesita conocimientos lingüísticos previos
- Costo computacional

El mayor problema del eCode

- ¿Dónde está?

Reprogramar el eCode

- Necesidad de extraer CDs
- Trabajo exhaustivo
- Existencia de información previa

Aprendizaje de máquina

- Que la computadora aprenda
- Bayes ingenuo
- Máxima entropía
- Árboles de decisión

Corpus

- Corpus de Contextos Definitorios (Corcode)
- 1311 CDs
- Corpus de entrenamiento: 998 CDs (Aprox. 75%)
- Corpus de evaluación: 313 CDs

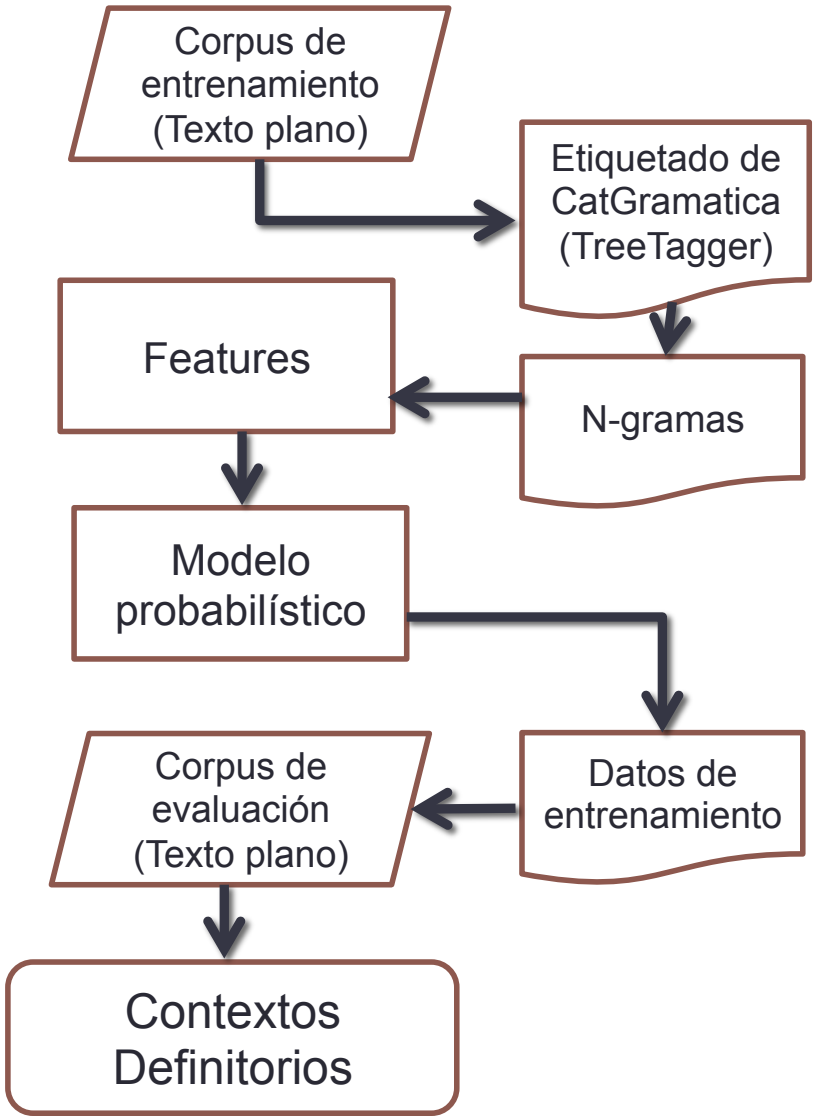
Features

- ¿Qué elementos de los CDs tomar?
- Elementos sintácticos
- Relaciones gramaticales
- N-gramas de categoría gramatical

N-gramas de categoría gramatical

- 3-gramas:
  - La AI se define como la técnica de software que los programas utilizan para dar solución a algún problema.
  - Art. + Sust. + Clit. + Verb. + Prep. + Art. + Sust. + Prep. + Sust. + Conj. + Art. + Sust. + Verb. + Prep. + Verb. + Sust. + Prep. + Adj. + Sust.
  - (Art. + Sust. + Clit); (Sust. + Clit. + Verb); (Clit. + Verb. + Prep.); (Verb. + Prep. + Art); (Prep. + Art. + Sust)……

Sistema


Corpus de entrenamiento (Texto plano)
Etiquetado de CatGramatica (TreeTagger)
Features
N-gramas
Modelo probabilístico
Corpus de evaluación (Texto plano)
Datos de entrenamiento
Contextos Definitorios

RESULTADOS

Árboles de decisión

N-Gramas	Precisión	Cobertura	Medida-F1
Bigramas	0.925	0.948	0.936
Trigramas	0.89	0.808	0.847

Máxima entropia

N-Gramas	Precisión	Cobertura	Medida-F1
Bigramas	1	0.884	0.938
Trigramas	0.943	0.968	0.955

Bayes

N-Gramas	Precisión	Cobertura	Medida-F1
Bigramas	0.993	0.952	0.972
Trigramas	0.993	0.948	0.970

Comparativa

Sistema	Precisión	Cobertura	Medida - F1
eCode (Alarcón)	0.53	0.79	0.634
eCode (Vieyra)	0.389	0.881	0.539
Mod. Aprendizaje	0.993	0.952	0.972

Comentarios finales

- Los métodos de aprendizaje de máquina han mostrado tener mejores resultados en la tarea de extracción de CDs.
- ¿Introducir métodos híbridos puede hacer al sistema más completo?
- El sistema tiene carencias

GRACIAS

VMijangosC@iingen.unam.mx