Procesamiento de Lenguaje Natural para la Memoria, la Verdad y la Justicia

Franco M. Luque

Grupo de PLN
FaMAF, UNC & CONICET
Córdoba, Argentina

Semana de la Memoria
1 de abril de 2015

Resumen

El Grupo de Procesamiento de Lenguaje Natural

`http://www.pln.famaf.unc.edu.ar/`

- Creado en 2005 por Gabriel Infante-Lopez y Laura Alonso Alemany.
- Cinco investigadores:
  - Laura Alonso Alemany
  - Martín Domínguez
  - Paula Estrella
  - Franco Luque
  - Jorge Sánchez
- Miembros afines y colaboradores estables.
- Dos estudiantes de doctorado.
- Algunos estudiantes de grado.

Procesamiento de Lenguaje Natural (PLN)

- El campo de las Ciencias de la Computación que estudia el procesamiento automático del lenguaje humano.
- Tareas relevantes:
  - Segmentación de palabras (tokenización) y oraciones,
  - etiquetado de tipos de palabras (categorías léxicas o part-of-speech),
  - análisis sintáctico,
  - extracción de información,
  - etc.
- También estudia modelos lingüísticos de base matemática y computacional.
- Por eso, también llamado Lingüística Computacional.
- Nada que ver con PNL!

El Archivo Provincial de la Memoria

`http://www.apm.gov.ar/`

- Creado en el 2006, conjuntamente con la Comisión Provincial de la Memoria, por la Legislatura de Córdoba.
- Actividades:
  - Rescate y conservación de documentos de la represión,
  - investigación, búsqueda y sistematización de información,
  - digitalización y gestión del software,
  - gestión de los Espacios para la Memoria,
  - etc. etc. etc.
- Área de Informática y Digitalización:
  - Marcelo Yornet
  - Marcos Kary
  - Paula de la Fuente
  - Laura Iturburu

Los Boletines del Ejército

- Documentos militares de 1975 a 1983.
- Periodicidad irregular, de acuerdo a las necesidades.
- Información:
  - Nombramientos y pases: movimientos del personal militar.
  - Cursos y aptitudes: realización y aprobación de cursos. Obtención y péridad de aptitudes especiales.
  - Altas, bajas, promociones (ascensos) y retiros.
  - Comisiones en el extranjero, causas de Justicia Militar, etc.

Los Boletines del Ejército

- Hechos a máquina.
- Buen estado de conservación.
- Escaneados y pasados por OCR.
- Texto semi-estructurado: mucha regularidad en la escritura.
- Información de interés:
  - Todas las menciones de personas de conocida participación en delitos de lesa humanidad.
  - Pases a unidades militares relacionados (e.g. Destacamentos de Inteligencia).
  - Realización de cursos relacionados (e.g. inteligencia o interrogatorio).

Extracción de Información

- Trata el problema del análisis de texto no estructurado para encontrar y estructurar determinada información de interés.
- Sub-tareas:
  - Reconocimiento de Entidades Nombradas: Etiquetado de menciones de entidades de diferentes tipos.
  - Extracción de Relaciones: Identificación de menciones de relaciones entre entidades.

Reconocimiento de Entidades Nombradas

(o *Named Entity Recognition (NER)*)

- Etiquetado de menciones de entidades de diferentes tipos (personas, lugares, fechas, etc.) en textos de lenguaje natural.
- Dos acercamientos al problema:
  - Sistemas basados en Reglas.
  - Sistemas basados en Aprendizaje por Computadora.

Ejemplo

El 23 de marzo de 1976 (FECHA) un grupo de personas del

Ejército Argentino (ORGANIZACIÓN) privaron de su libertad a Graciela Olivella (PERSONA).

Sistemas Basados en Reglas

- Características:
  - Reglas: Patrones que se pueden evaluar sobre porciones de texto.
  - El cumplimiento de un patrón indica la presencia de una entidad.
  - Los patrones se elaboran manualmente a partir del conocimiento de los datos.
- Ventajas:
  - Buenas para textos altamente estructurados.
  - Bajo costo de elaboración.
- Desventajas:
  - Malas para textos muy "naturales".
  - Rigidez: intolerancia a ruido o variación en los datos.
  - Muy específicos: no adaptables a otros dominios.

Reglas: Expresiones Regulares

Ejemplo

- Reconocimiento de personas:
  1. Grado militar: 17 posibilidades.
  2. Complemento del grado: palabras empezadas en mayúsculas, o "de".
  3. Condición opcional: "en comisión" o "(R - Art. 62)".
  4. "D" ó "Da" opcional.
  5. La **persona** propiamente dicha: palabras en mayúsculas.
  6. Identificador opcional: seis dígitos entre paréntesis (y con punto).

Teniente	Médica	"en comisión"	Da	DIANA ALICIA MONTES
1	2	3	4	5 (PERSONA)

Reglas: Expresiones Regulares

Ejemplo

- Reconocimiento de personas:
  1. Grado militar: 17 posibilidades.
  2. Complemento del grado: palabras empezadas en mayúsculas, o "de".
  3. Condición opcional: "en comisión" o "(R - Art. 62)".
  4. "D" ó "Da" opcional.
  5. La **persona** propiamente dicha: palabras en mayúsculas.
  6. Identificador opcional: seis dígitos entre paréntesis (y con punto).

Cabo 1ro	WALTER OMAR LOPEZ	(255.340)
1	5 (PERSONA)	6

Reglas: Expresiones Regulares

Ejemplo

- Reconocimiento de personas:
  1. Grado militar: 17 posibilidades.
  2. Complemento del grado: palabras empezadas en mayúsculas, o "de".
  3. Condición opcional: "en comisión" o "(R - Art. 62)".
  4. "D" ó "Da" opcional.
  5. La **persona** propiamente dicha: palabras en mayúsculas.
  6. Identificador opcional: seis dígitos entre paréntesis (y con punto).

Cabo	de Hornos	WAKA WAKA
1	2	5 (PERSONA)

Aprendizaje por Computadora

(o *Machine Learning*)

- Características:
  - Basado en datos (corpus): se requieren ejemplos de lo que se quiere obtener.
  - Se define un conjunto de características relevantes que el modelo debe observar.
  - Se entrena un modelo a partir de los datos y las características relevantes.
- Ventajas:
  - Toleran ruido y varaciones en los datos.
  - Son reusables: independientes del idioma o dominio.
  - Permiten una evaluación sistemática.
- Desventajas:
  - Son costosos de elaborar.
  - Requieren de datos anotados (a veces muchos).

Aprendizaje por Computadora

- Se reduce el problema de NER a un problema de etiquetado de palabras.
- Codificación BIO:
  - `O`: no es parte de una entidad.
  - `B-<E>`: comienzo de entidad de tipo `<E>`.
  - `I-<E>`: interior de entidad de tipo `<E>`.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	O

Ejército	Argentino	privaron	de	su	libertad	a	Graciela	Olivella	.
B-ORG	I-ORG	O	O	O	O	O	B-PER	I-PER	O

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El 23 de marzo de 1976 un grupo de personas del
O ???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El 23 de marzo de 1976 un grupo de personas del
O B-F ???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El 23 de marzo de 1976 un grupo de personas del
O B-F I-F ???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El 23 de marzo de 1976 un grupo de personas del
O B-F I-F I-F ???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El 23 de marzo de 1976 un grupo de personas del
O B-F I-F I-F I-F ???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	???				

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	???			

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	???		

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	???	

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	???

Ejército Argentino privaron de su libertad a Graciela Olivella .

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	O

Ejército Argentino privaron de su libertad a Graciela Olivella .
???

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	O

Ejército	Argentino	privaron	de	su	libertad	a	Graciela	Olivella	.
B-ORG		???							

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	O

Ejército	Argentino	privaron	de	su	libertad	a	Graciela	Olivella	.
B-ORG	I-ORG	???							

Aprendizaje por Computadora: Clasificación Ordenada

- Se etiqueta cada palabra de izquierda a derecha.
- Cada decisión se toma en base a la información disponible en el momento.

Ejemplo

El	23	de	marzo	de	1976	un	grupo	de	personas	del
O	B-F	I-F	I-F	I-F	I-F	O	O	O	O	O

Ejército	Argentino	privaron	de	su	libertad	a	Graciela	Olivella	.
B-ORG	I-ORG	O	O	O	O	O	B-PER	I-PER	O

Características Relevantes (*Features*)

- *Ingeniería de features:* seleccionar el conjunto de *features* que mejor funcione (expermiental, etc.).
- Algunos features que nos pueden servir:
  - La palabra a etiquetar (`word`),
  - si la palabra está en mayúsculas (`isupper`),
  - si la palabra tiene dígitos (`isdigits`),
  - si la palabra no tiene símbolos (`isalphanum`),
  - si la palabra es un nombre propio del castellano (`isname`),
  - el largo de la palabra (`wordlen`),
  - todos los anteriores sobre la palabra anterior (`prev`) y la siguiente (`next`),
  - la etiqueta de la palabra anterior (`prev(label)`).
- Obs: algunos features son palabras, otros son booleanos, otros son numéricos.
- Se computa un *vector de features*.

Características Relevantes (*Features*)

Ejemplo

... Teniente 1ro D ERNESTO HANSE LARRAMENDI ...
O O O ???

Vector de características para la palabra ERNESTO:

```
{'word=ERNESTO': 1, ...,
 'isupper': 1, 'isdigits': 0, 'isalphanum': 1,
 'isname': 1,  // !!!
 'wordlen': 7,
 'prev(word)=D': 1, ...,  // !!!
 'prev(isupper)': 1, ...,
 'next(word)=HANSE': 1, ...,
 'next(isupper)': 1, ...,
 'prev(label)=O': // etiqueta anterior
}
```

Clasificadores

- Un clasificador es una función de vectores de features en un conjunto finito de clases.
- La función se aprende a través de un algoritmo que toma como input vectores etiquetados con su clase.
- Pipeline de preprocesamiento de los vectores:
  - Escalado y ajuste de media.
  - Selección de features relevantes.
  - Reducción de dimensionalidad.
- Algunos clasificadores:
  - Árboles de decisión.
  - Naive y Multinomial Bayes.
  - Support Vector Machines (SVMs).
  - Métodos combinados.

Clasificadores: Árboles de decisión

- Un árbol cuyos nodos internos son condiciones sobre los features y cuyas hojas son etiquetas de salida.
- Ventajas:
  - Se pueden escribir manualmente o usar algoritmos de entrenamiento.
  - Soporta naturalmente clasificación multiclase.
  - Es fácilmente entendible e interpretable.
  - La clasificación provee una explicación.
- Desventajas:
  - Optimización NP-completa. Sólo algoritmos greedy.
  - Expresividad limitada.
  - Riesgo de baja generalización (*overfitting*).
  - La clasficación no provee un valor probabilístico.

Clasificadores: Árboles de decisión

Ejemplo

... Teniente 1ro D ERNESTO HANSE LARRAMENDI ...
O O O ???

```
if ('prev(label)=O' == 1)
    if ('prev(word)=D' == 1 or 'prev(word)=Da' == 1)
        return 'B-PER'  // (1)
    elif ('isname' == 1 and 'isupper' == 1)
        return 'B-PER'
    else
        return 'O'
else if ('prev(label)=B-PER' == 1 or 'prev(label)=I-PER' == 1)
    if ('isupper' == 1 and 'next(isuppper)' == 1)
        return 'I-PER'  // (2)
    else
        return 'O'
```

Clasificadores: Árboles de decisión

Ejemplo

... Teniente 1ro D ERNESTO HANSE LARRAMENDI ...
O O O B-PER ???

```
if ('prev(label)=O' == 1)
    if ('prev(word)=D' == 1 or 'prev(word)=Da' == 1)
        return 'B-PER'  // (1)
    elif ('isname' == 1 and 'isupper' == 1)
        return 'B-PER'
    else
        return 'O'
else if ('prev(label)=B-PER' == 1 or 'prev(label)=I-PER' == 1)
    if ('isupper' == 1 and 'next(isuppper)' == 1)
        return 'I-PER'  // (2)
    else
        return 'O'
```

Clasificadores: Árboles de decisión

Ejemplo

... Teniente 1ro D ERNESTO HANSE LARRAMENDI ...
O O O B-PER I-PER ???

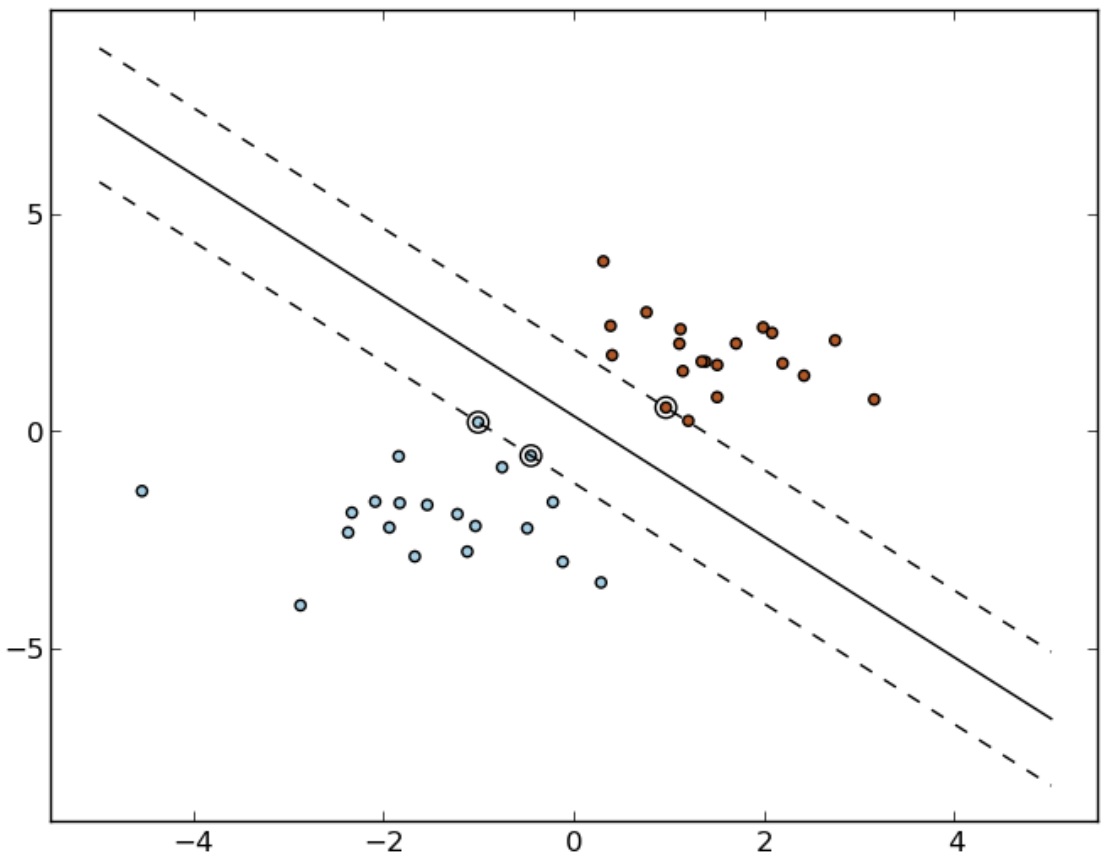
```
if ('prev(label)=O' == 1)
    if ('prev(word)=D' == 1 or 'prev(word)=Da' == 1)
        return 'B-PER'  // (1)
    elif ('isname' == 1 and 'isupper' == 1)
        return 'B-PER'
    else
        return 'O'
else if ('prev(label)=B-PER' == 1 or 'prev(label)=I-PER' == 1)
    if ('isupper' == 1 and 'next(isuppper)' == 1)
        return 'I-PER'  // (2)
    else
        return 'O'
```

Clasificadores: Support Vector Machines (SVMs)

- Una división en dos partes del espacio de vectores de features utilizando un hiperplano (clasificador binario).
- Clasificación multiclase simulada: 1-vs-1 ó 1-vs-el-resto.
- Ventajas:
  - Alta expresividad: división no-lineal del espacio a través de kernels.
  - Alta generalización: para versión lineal y kernels no muy raros.
  - Optimización tratable: problema convexo.
  - Muy buenos resultados experimentales.
- Desventajas:
  - Difícil de interpretar.
  - La clasficación no provee un valor probabilístico.

Clasificadores: Support Vector Machines (SVMs)

- Algoritmo de aprendizaje: maximizar el tamaño de la franja entre los dos grupos de puntos (vectores) de entrenamiento.
- Variante soft-margin: tolera ruido y errores.



Extracción de Relaciones

- Identificación de menciones de relaciones entre entidades en textos de lenguaje natural (presencia de persona en un lugar, vínculos entre personas, etc.).
- Variantes:
  - Cantidad de entidades relacionadas: binarias, n-arias.
  - Tipos de relaciones: fijos o libres.
  - Dominio: abierto o cerrado.
- Dos acercamientos al problema:
  - Sistemas basados en Reglas.
  - Sistemas basados en Aprendizaje por Computadora.

Ejemplo

(Boletín 5037)

Reglas: Combinación de Expresiones Regulares sobre Entidades

Ejemplo

- Extracción de **pases**:
  1. Una entidad **fecha**.
  2. Un nominal de pase: "pase a continuar sus servicios", "pase a prestar servicios" y un par más.
  3. Uno o más de los siguientes:
     1. 'AL' ó 'A LA'.
     2. Una entidad **lugar**.
     3. Uno o más entidades **persona**.
- Se admiten palabras "sueltas" (sin entidad) intercaladas.

El Software

- Carga de los boletines:
  - Segmentación de boletines (detección de portadas) basada en *machine learning*.
  - Extracción de metadatos (número de boletín, fecha, números de página, etc.).
  - Corrección automática de metadatos.
- Reconocimiento de entidades basado en reglas.
- Reconocimiento de personas basado en aprendizaje por computadora.
- Reconocimiento de pases basado en reglas.
- Interfaz de edición de entidades y ocurrencias: Agregar, eliminar, corregir, etc.
- Motor de búsqueda.

El Software

Algunos números:

- Sistema basado en reglas:
  - Personas: 66939 menciones.
  - Lugares: 22335 menciones.
  - Fechas: 12639 menciones.
  - Pases: 10394.
- Sistema basado en Aprendizaje por Computadora:
  - Personas: 141929 menciones (más del doble!).
  - Pases: 21653 (más del doble!).

Conclusiones

- El uso de herramientas de software basadas en PLN permite facilitar el trabajo de sistematización, mantenimiento y consulta de información contenida en grandes volúmenes de documentos.
- El uso de Aprendizaje por Computadora permite dar un salto cualitativo en los sistemas obtenidos y el costo asociado.
- Todo sistema automático de PLN es imperfecto por la naturaleza del lenguaje humano. La interacción con el usuario permite obtener nueva información para mejorarlos.

Trabajo Futuro

- Procesamiento de relaciones entre cursos y personas (realizar, aprobar).
- Desambiguación automática de entidades.
- Procesamiento de pases usando machine learning.
- Reconocimiento de la estructura de los boletines.
- Comunicación con “Presentes”.
- Re-entrenamiento y re-etiquetado a partir de correcciones de los usuarios.

¡Gracias! ¿Preguntas?