Incorporando Inteligencia a la Extracción de Información

Cristian A. Cardellino • Laura Alonso i Alemany

Jornada de Doctorandos 2013
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba

2 de Diciembre de 2013

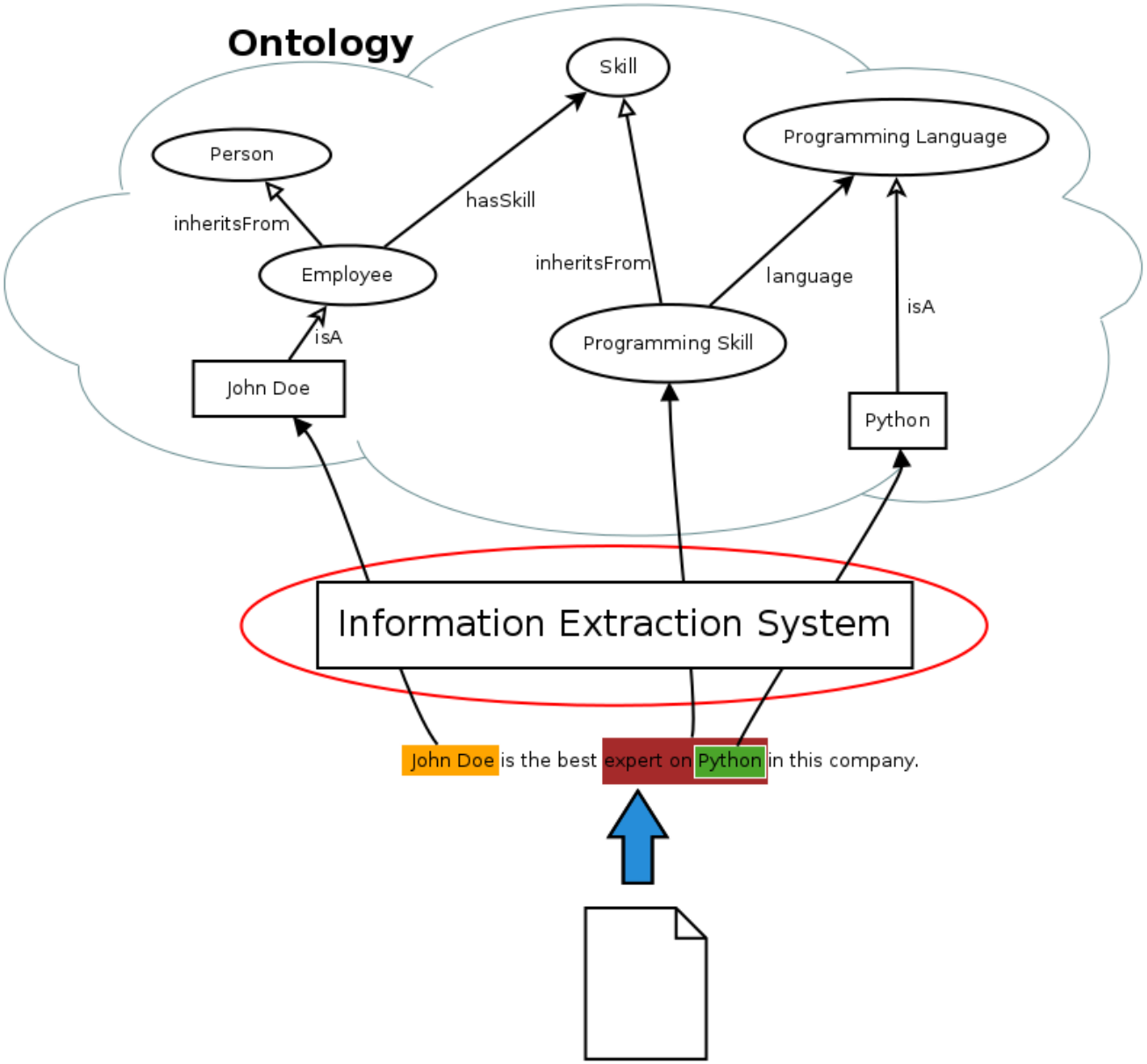
Section 1

Introducción

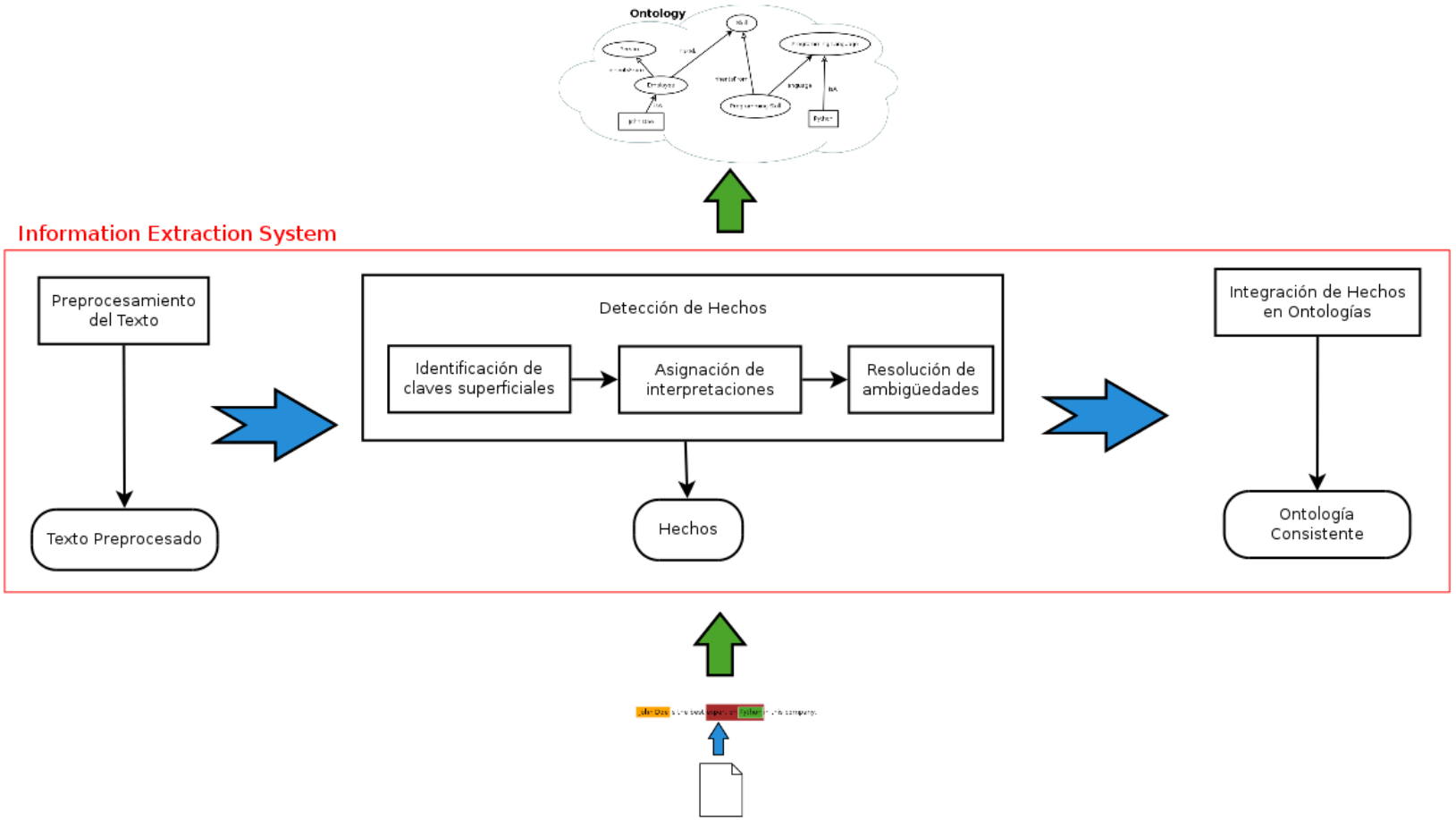
Objetivos

- Incorporar información de lexicones verbales a procesos de Extracción de Información (EI).
- Enriquecer información de lexicón a partir de corpus.
- Integrar técnicas semi-supervisadas para mejorar la desambiguación automática de sentidos verbales.
- Automatizar la creación de robots de EI a partir de recursos lingüísticos.

Objetivo de Extracción de Información



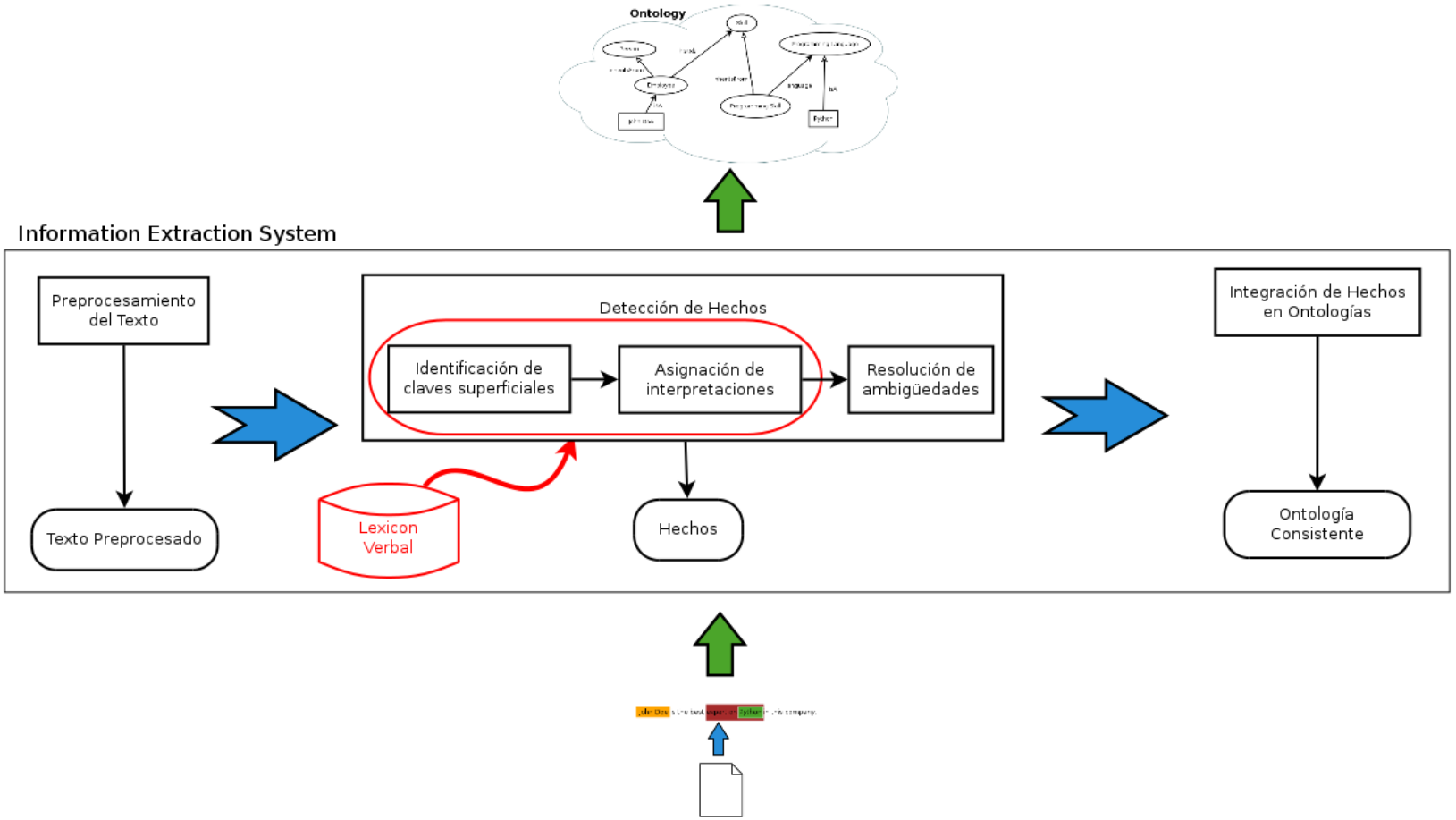
Arquitectura de Extracción de Información



Section 2

Incorporando lexicones verbales

Incorporando lexicones verbales



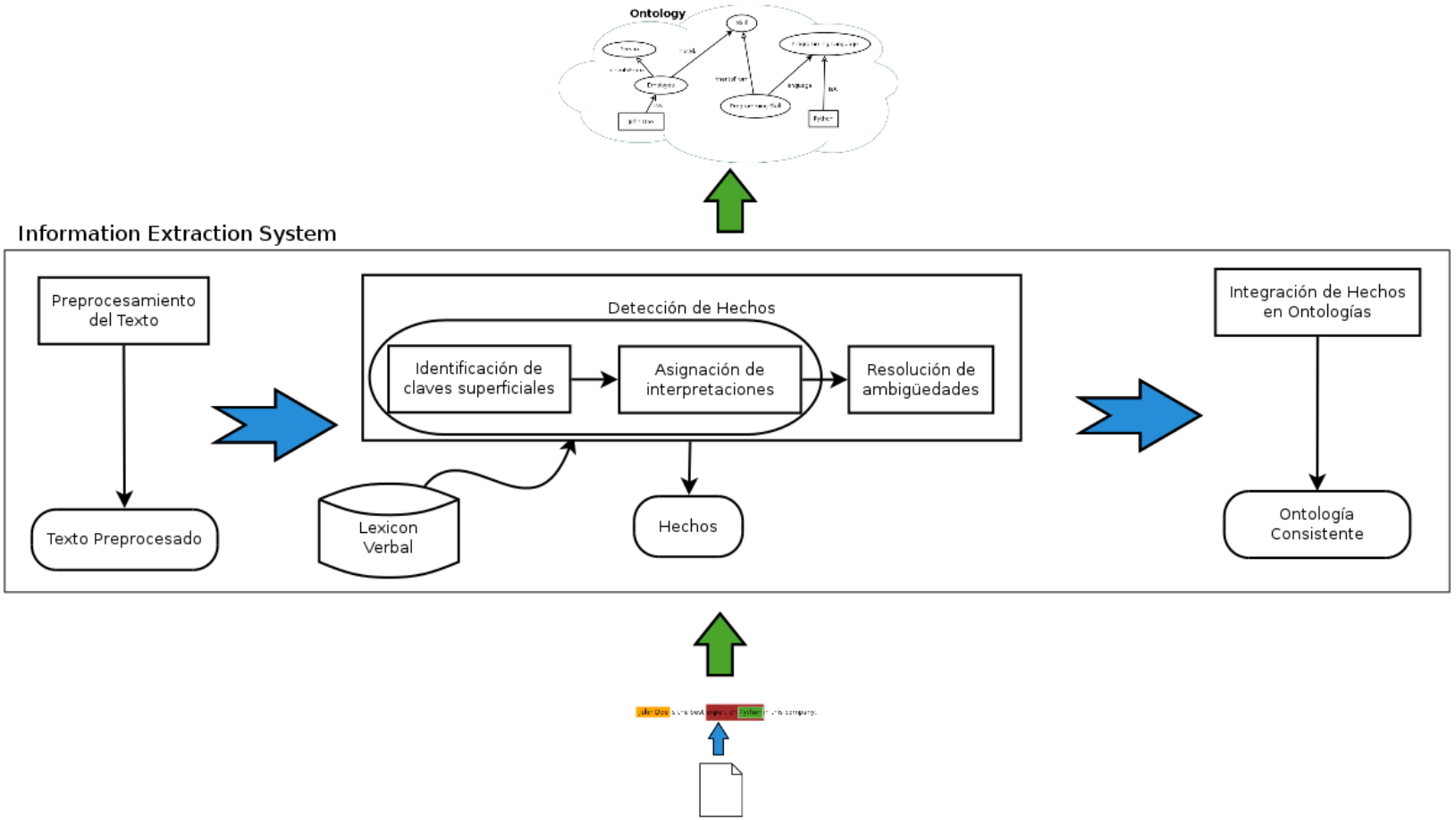
Incorporando lexicones verbales

- Un lexicón verbal es una base de datos donde los predicados verbales se describen como escenas con sus participantes y las características de sus participantes.
- Hay lexicones desarrollados para inglés, castellano y otras lenguas.
- Integran información lingüística profunda y rica.
- En general, información teórica, introspectiva, estática, poco conectada con la realidad del uso del lenguaje.

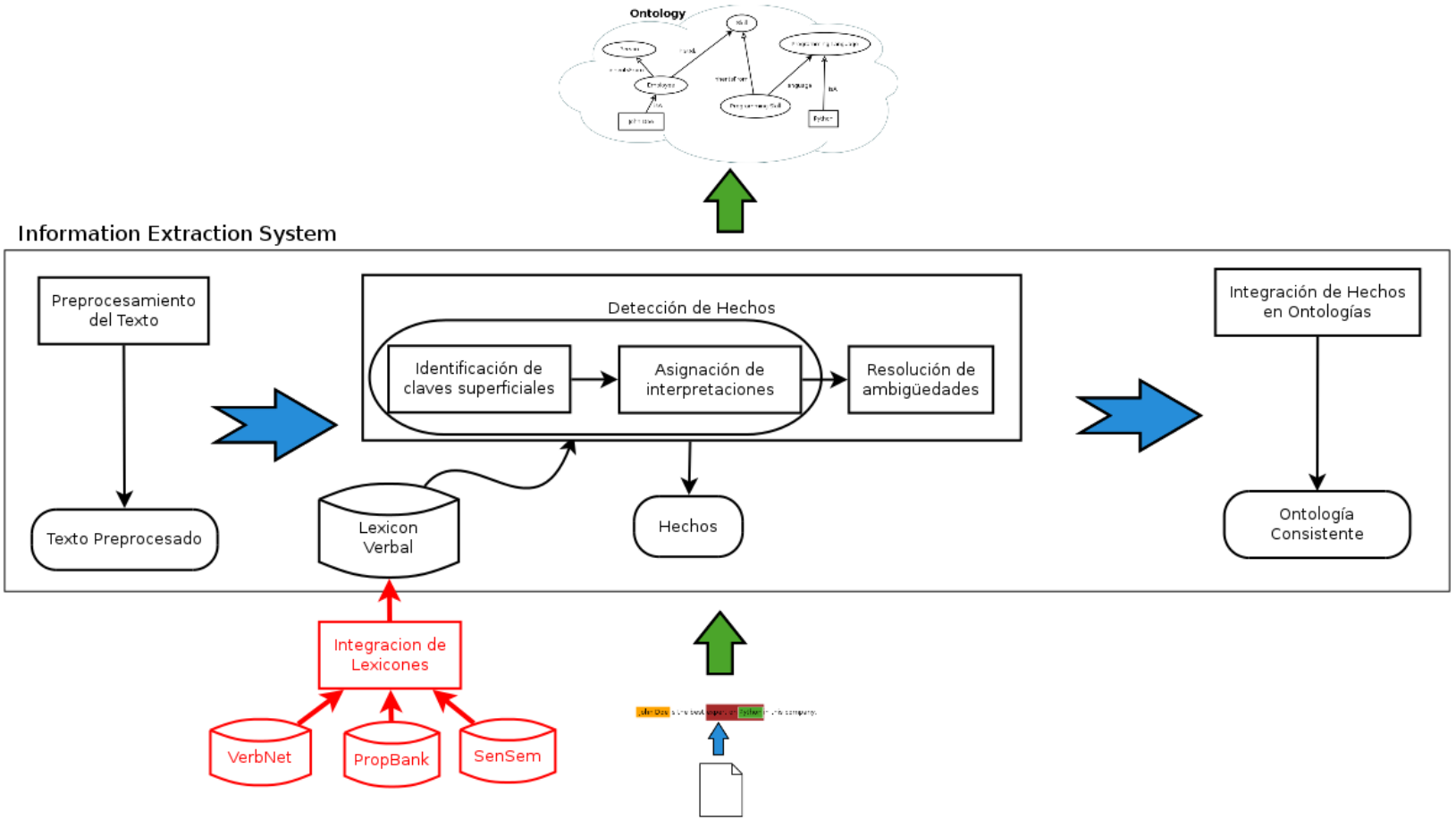
Integrando lexicones verbales

- Existen muchos y diversos lexicones verbales, cada uno con propiedades distintas.
- Creamos una representación común y mappings de los lexicones existentes a una representación común.
- Integramos la información de algunos lexicones seleccionados.
- Integramos información / procedimientos asociados a los lexicones: herramientas de análisis, mappings a otros recursos, etc.

Integración de lexicones



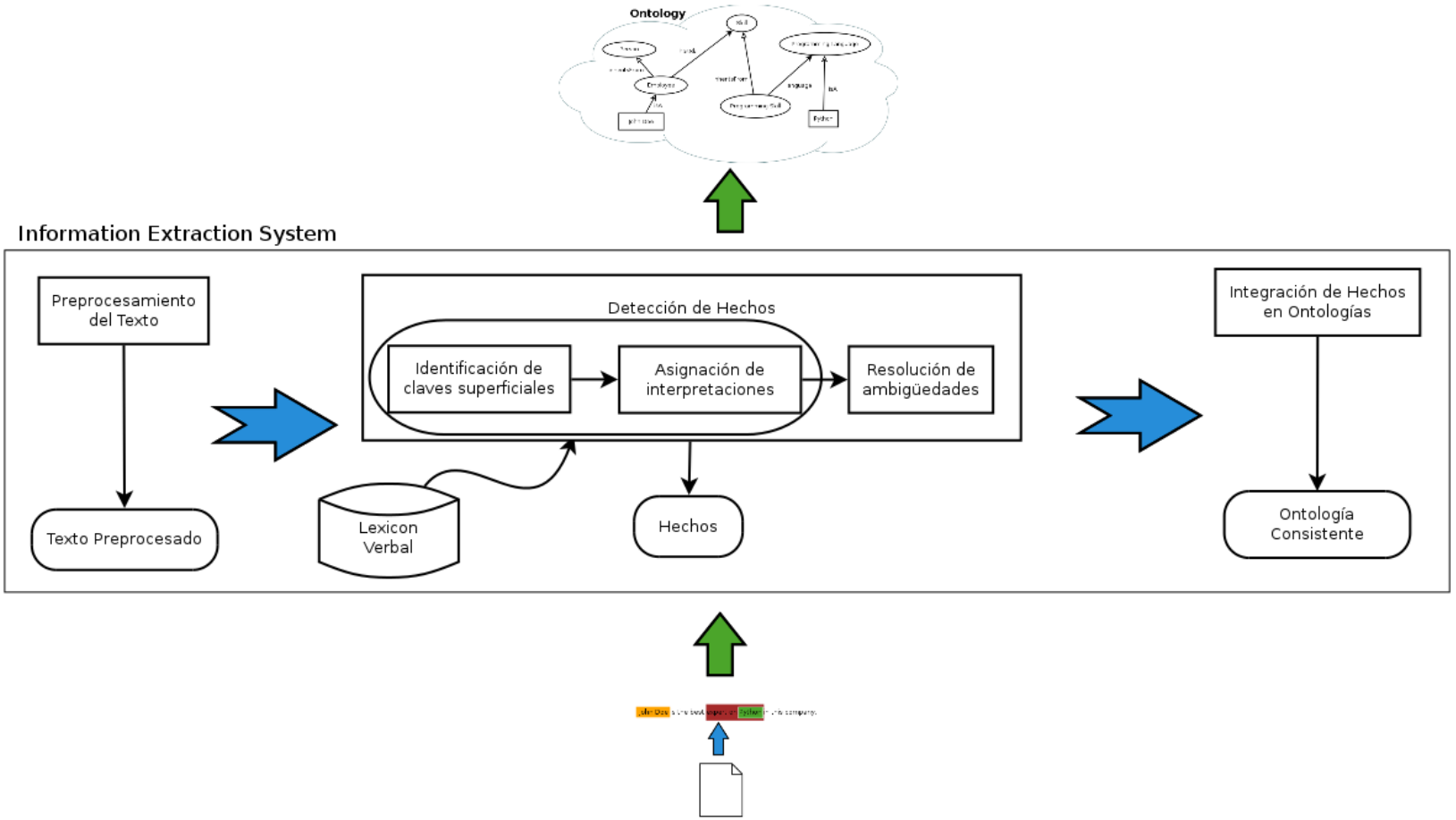
Integración de lexicones



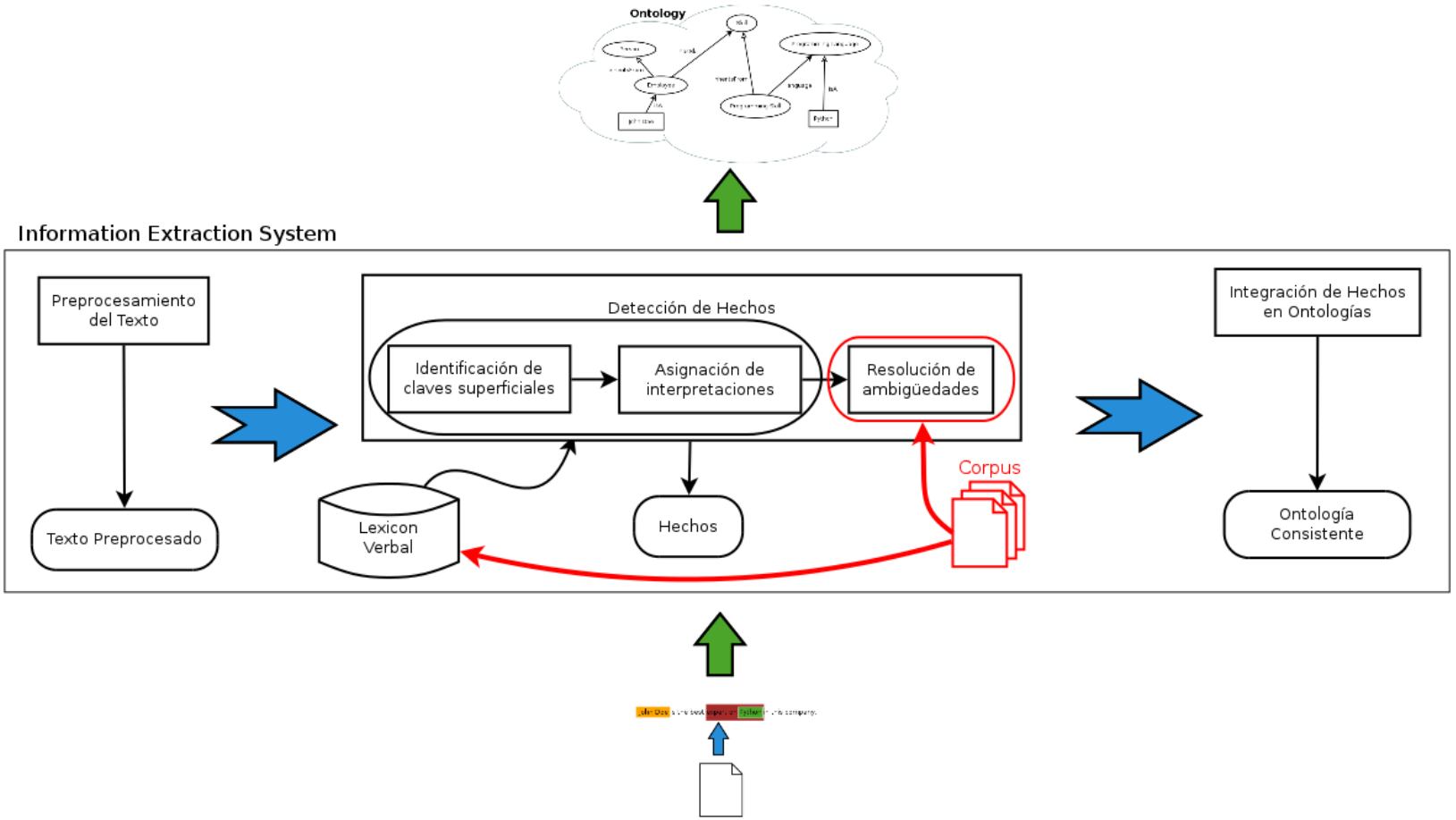
Section 3

Enriqueciendo el lexicón con corpus

Enriqueciendo el lexicón con corpus



Enriqueciendo el lexicón con corpus



Enriqueciendo el lexicón con corpus

- Buscamos facilitar la resolución de ambigüedades.
- Vemos el impacto de incorporar probabilidades obtenidas de frecuencias de ocurrencias en corpus.

Ejemplo: Verbo "dar"

Para el verbo "dar", la estructura transitiva sujeto-verbo-obj. dir. (e.g. "el profesor dio una charla") es de mayor frecuencia de aparición en corpus que la estructura ditransitiva sujeto-verbo-obj. dir.-obj. ind. (e.g "el estudiante dio una charla a los alumnos").

- Sumamos el uso de preferencias léxicas.

Ejemplo: Verbo "ladrar"

El verbo "ladrar", en la mayoría de las ocasiones es precedido por el léxico "perro".

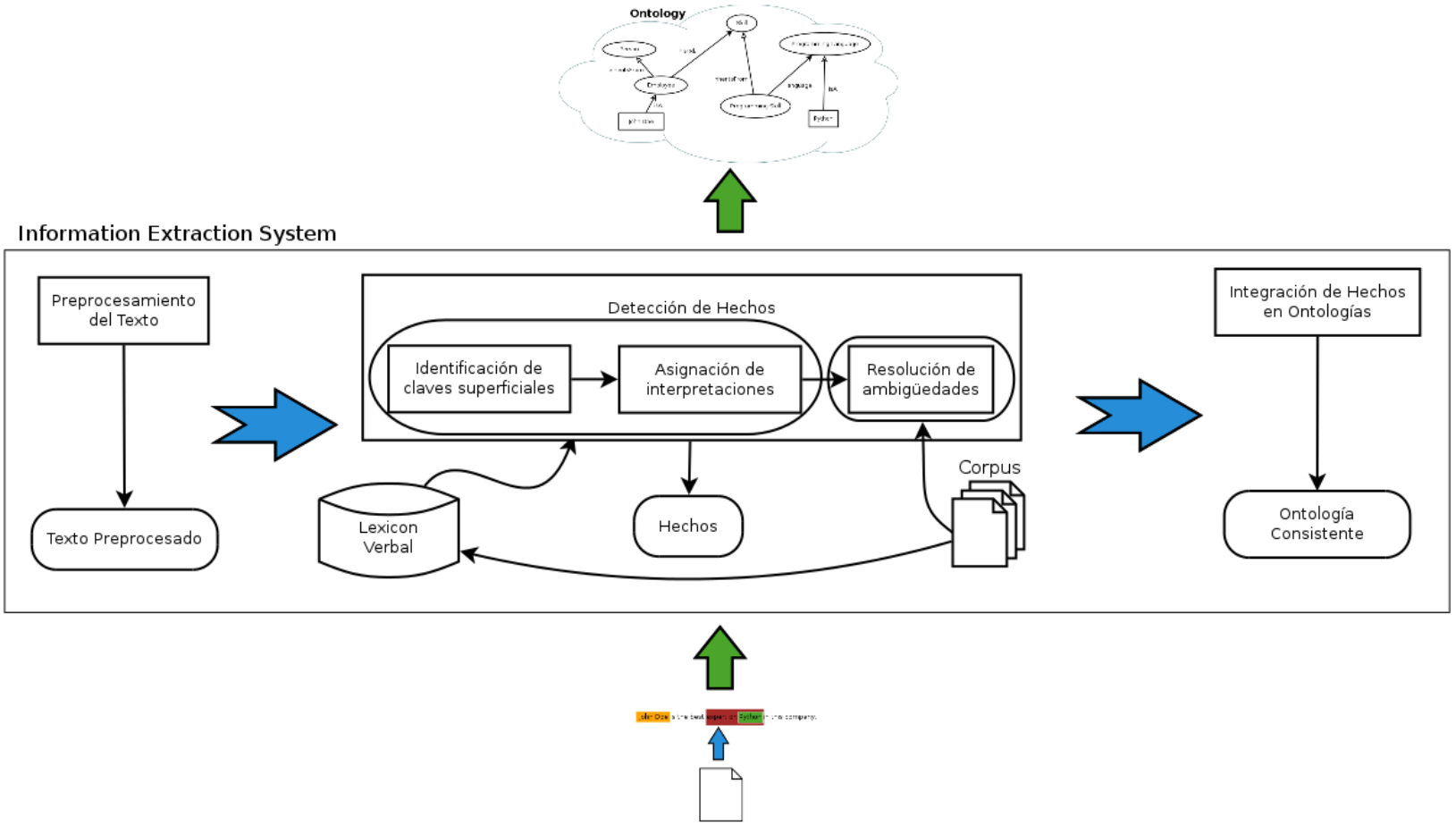
Enriqueciendo el lexicón con corpus

- Los corpus de mayor tamaño no están anotados.
- Para mejorar la cobertura se necesita un corpus de gran tamaño.
- El corpus es analizado automáticamente, lo que introduce error.
- Se busca extender las propiedades de un corpus anotado (que son de mayor confianza) a un corpus no anotado.

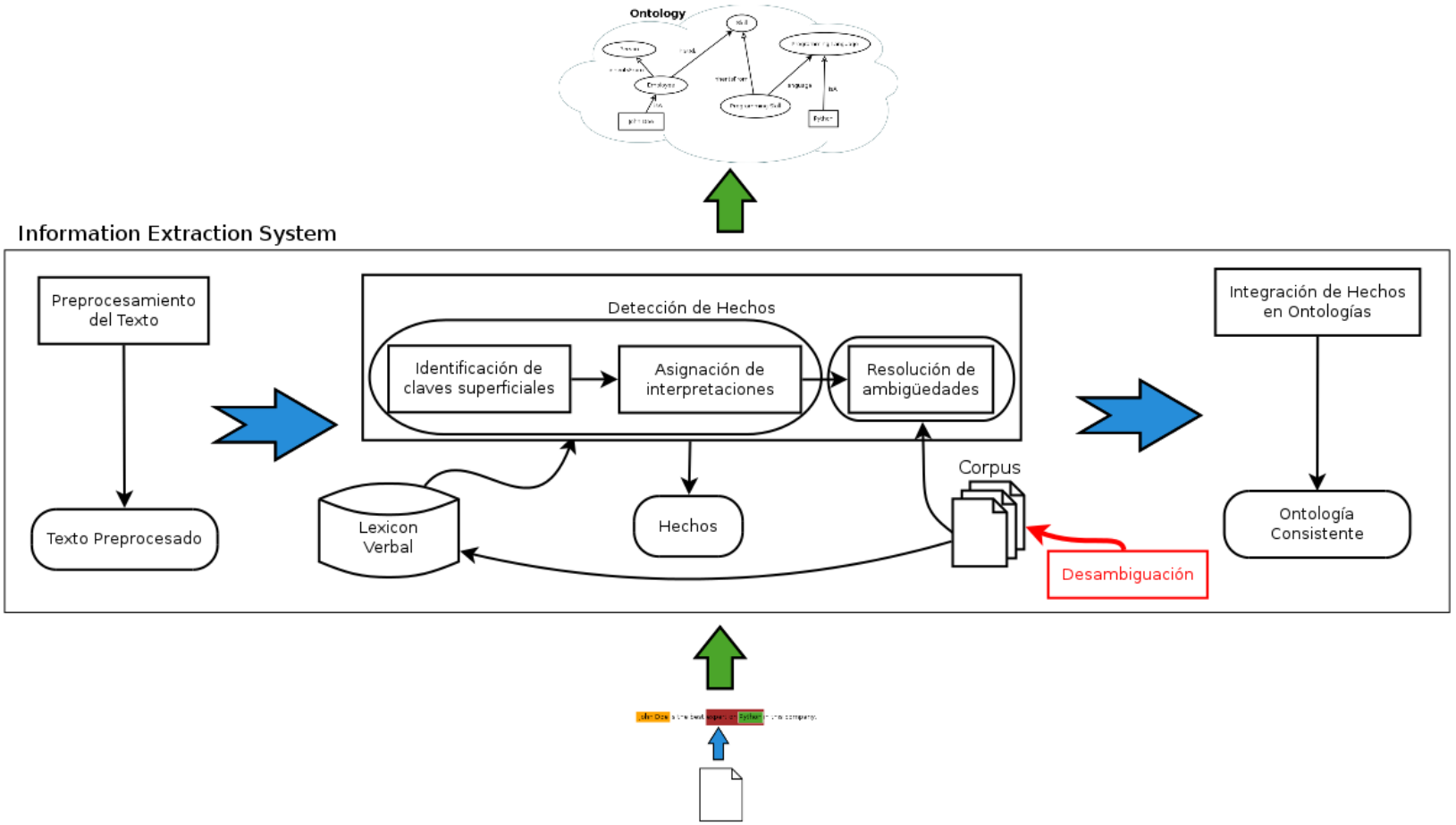
Section 4

Desambiguación de sentidos verbales

Desambiguación de sentidos verbales



Desambiguación de sentidos verbales



Desambiguación de sentidos verbales

- Es necesario saber el sentido de cada verbo en cada frase analizada, para utilizar mejor la información del corpus.
- La desambiguación de sentidos de palabras es difícil pero ha mejorado con el tiempo... ¡Pero sólo para nombres!
- Muy poco trabajo en desambiguación de sentidos verbales.
- Involucra mucha información: sobre los argumentos, sobre estructuras sintácticas, etc.
- Esta información la provee el lexicón.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

1. Tomamos ejemplos analizados manualmente.
2. Identificamos las palabras significativamente asociadas con un sentido (preferencias léxicas).
3. Algunos problemas:
   - Los ejemplos anotados son pocos.
   - Ley de Zipf: en Lenguaje Natural hay una gran dispersión de los datos (pocas palabras ocurren muchas veces y la mayoría ocurre una o ninguna en todo el corpus).
4. Construimos un clasificador con estas palabras.
5. Con el clasificador asignamos sentidos a ejemplos no vistos previamente.
   - El clasificador no posee la información para tratar la mayor parte de ejemplos no vistos.
6. Los ejemplos etiquetados por el clasificador se incorporan al conjunto de entrenamiento inicial y se repite el procedimiento.

Desambiguación de sentidos verbales por *bootstrapping*

- ¿Cuándo detener las iteraciones?
  - Cuando no se incorporen nuevos ejemplos o bien cuando el clasificador produzca error.
- ¿Cómo saber si el clasificador está divergiendo?
  - Se reservan parte de los datos etiquetados manualmente como monitores del modelo.

Desambiguación de sentidos verbales por *bootstrapping*

- ¿Cuándo detener las iteraciones?
  - Cuando no se incorporen nuevos ejemplos o bien cuando el clasificador produzca error.
- ¿Cómo saber si el clasificador está divergiendo?
  - Se reservan parte de los datos etiquetados manualmente como monitores del modelo.

Desambiguación de sentidos verbales por *bootstrapping*

- ¿Cuándo detener las iteraciones?
  - Cuando no se incorporen nuevos ejemplos o bien cuando el clasificador produzca error.
- ¿Cómo saber si el clasificador está divergiendo?
  - Se reservan parte de los datos etiquetados manualmente como monitores del modelo.

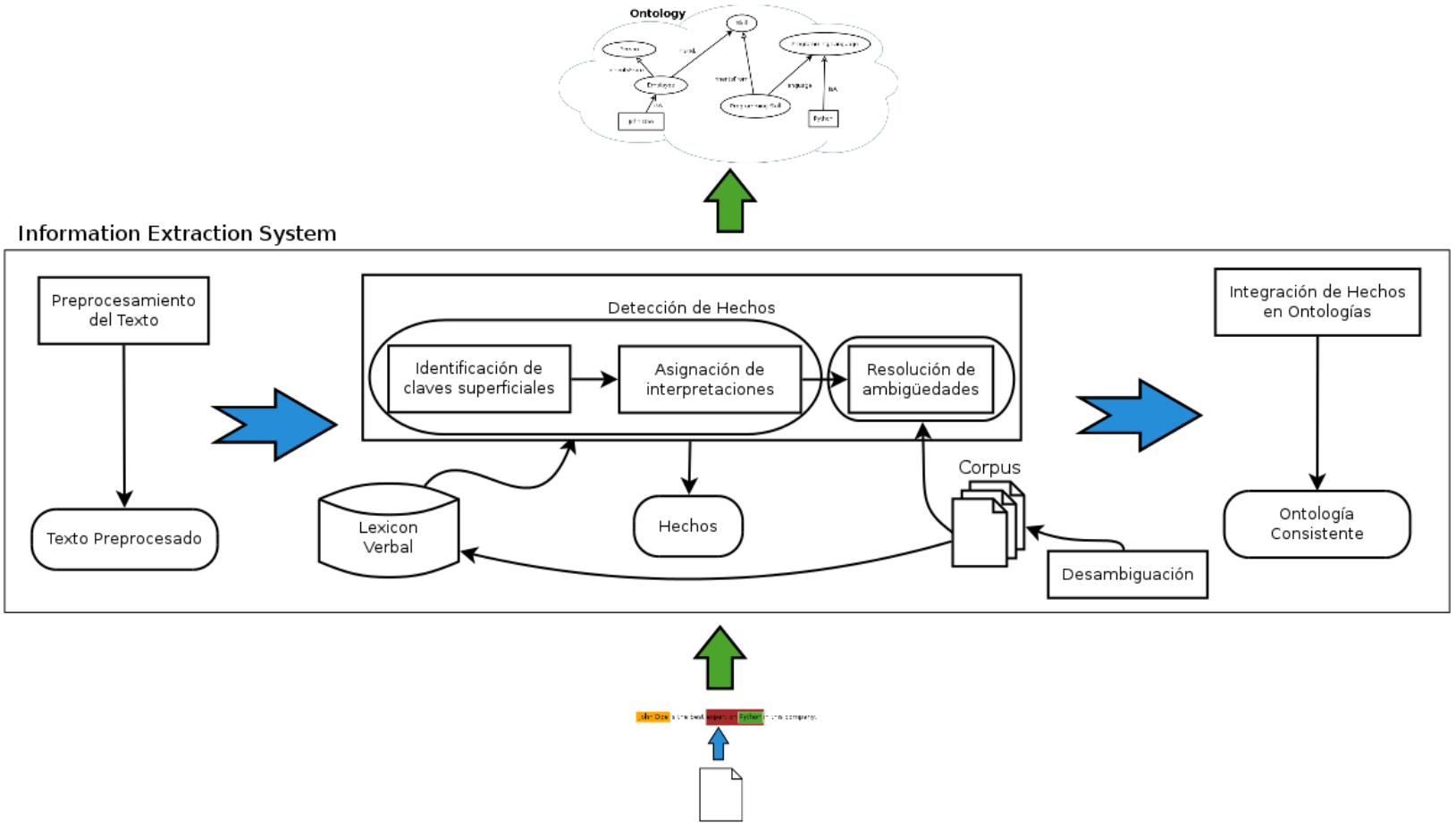
Desambiguación de sentidos verbales por *bootstrapping*

- ¿Cuándo detener las iteraciones?
  - Cuando no se incorporen nuevos ejemplos o bien cuando el clasificador produzca error.
- ¿Cómo saber si el clasificador está divergiendo?
  - Se reservan parte de los datos etiquetados manualmente como monitores del modelo.

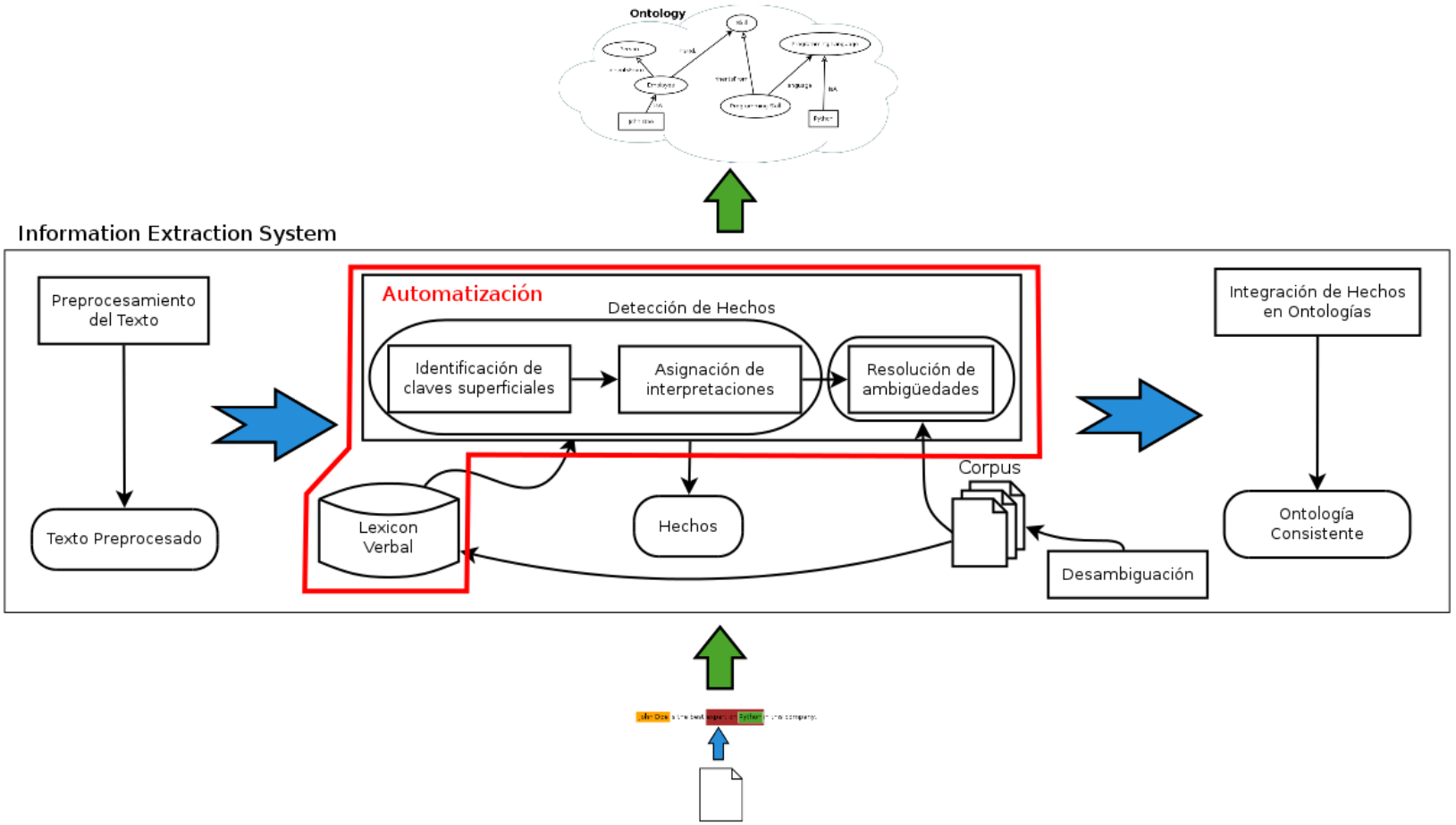
Section 5

Creando robots

Creando robots



Creando robots



Creando robots

- El lexicón provee la mayor parte de la información necesaria para detección y desambiguación de hechos.
- Podemos generar automáticamente los procedimientos para detectar y desambiguar a partir de la información del lexicón $\rightarrow$ robots de extracción de información.
- Robots están integrados dentro del marco del proyecto Kyoto: *kybots* (knowledge yielding robots).
- Son creados en base al esquema de un verbo (sus argumentos y la posición de los mismos en la estructura).
- Se generaron alrededor de 10 mil para el español y 20 mil para el inglés, todavía pendientes de evaluación.
- Arquitectura del Proyecto Kyoto (Link)

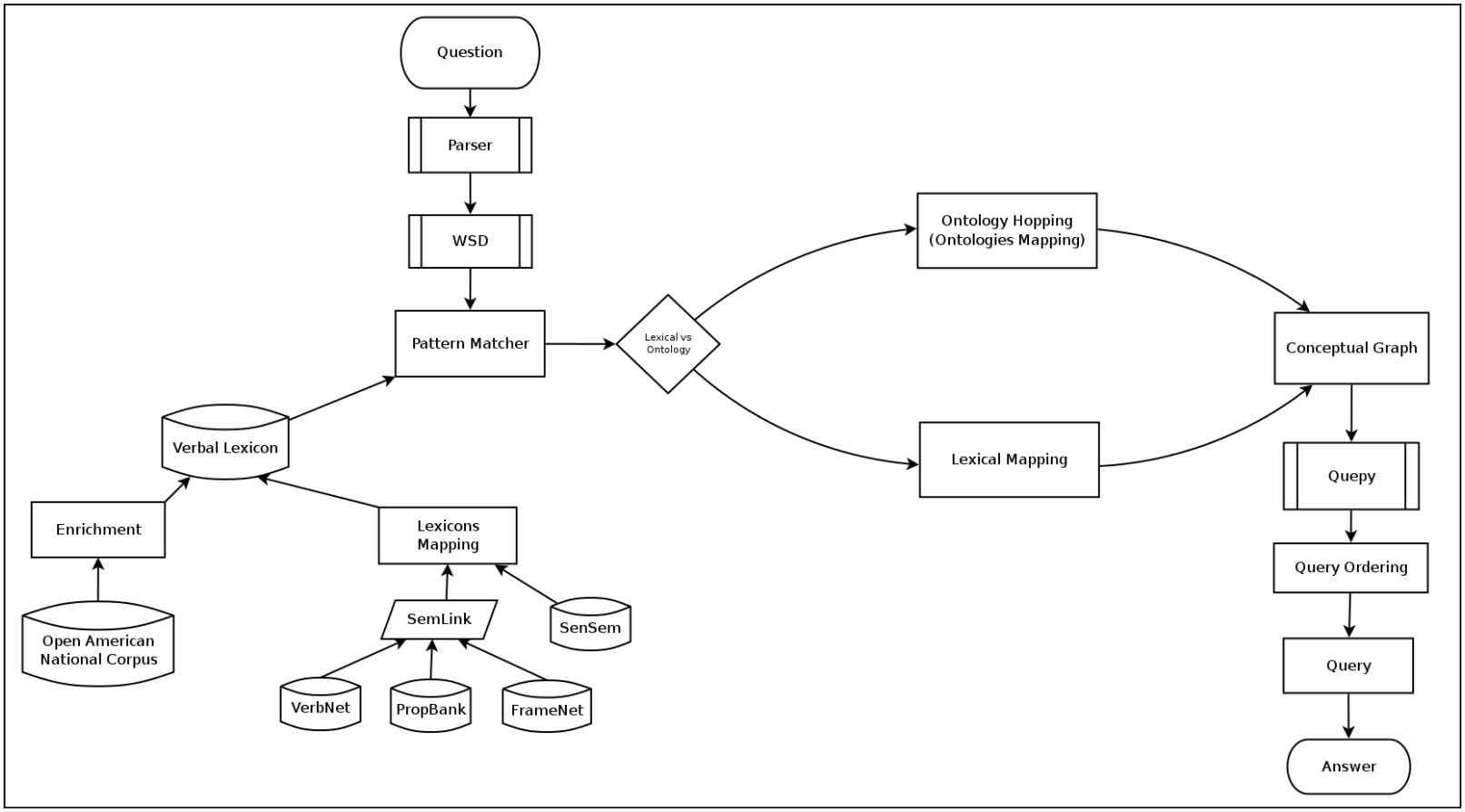
Section 6

Aplicaciones Prácticas

Aplicaciones Prácticas

- Buscamos su integración en Question Answering over Linked Data (QALD).
- Para esto se diseñó un sistema que toma como base a la librería para QALD de Python: quepy.
- La idea es que dada una pregunta en lenguaje natural, se use la Semantic Web para brindar una respuesta.
- Nuestro aporte es en el análisis de las preguntas y desambiguación de las mismas para poder brindar respuestas más acertadas.
- Con base en este diseño, los alumnos de la materia de PLN realizaron proyectos relacionados que tienen como objetivo integrarse a Quepy.

Arquitectura del sistema para integrar a quepy



Section 7

Trabajo Futuro

Trabajo Futuro

- Integración de los kybots en español con validación humana.
- Aplicar técnicas de bootstrapping para Verb Sense Disambiguation.
- Mejorar la integración de información de corpus en el lexicón.
- Evaluación del impacto de los lexicones en la tarea de QALD.