TJ Watson Research Center

Procesamiento de Lenguaje Humano y Posibles Repercusiones a Niveles Laborales y Gubernamentales

Pablo Ariel Duboue

Research Staff Member

Objetivos de esta charla

- **Presentar dos aplicaciones del procesamiento de lenguaje humano**
  - Descubrimiento de información seleccionada en texto
    - Parte de mi tesis doctoral
  - Respuesta automática a preguntas
    - Parte de mi trabajo actual
- **Comentar brevemente el impacto de las nuevas tecnologías**
  - A nivel laboral
    - Nuevas oportunidades
    - Nuevos desafíos
  - A nivel gubernamental

Procesamiento de Lenguaje Humano

- ***"Natural Language Processing"***
- **Hace hincapié en la idea de**
  - Procesar, trabajar de manera ordenada, computar, realizar operaciones automáticas
  - Lenguaje humano o "natural", en vez de lenguajes artificiales
  - Resolución de problemas concretos más allá de mejorar la comprensión del proceso cerebral de comunicación (o imitarlo)
- **A nivel de investigación, es un proceso experimental**
  - La computadora contiene un modelado del proceso natural
  - Se prefieren modelos que mejor describan los datos (palabras, textos, etc.)

Mi Educación

- **Escuela N° 270 El Bolsón (Río Negro)**
- **Escuela Florencio Sánchez (Córdoba)**
- **Colegio Nacional de Monserrat**
  - Olimpíadas de Matemática
- **Facultad de Matemática, Astronomía y Física, UNC**
  - Licenciado en Computación,
    tesis con Javier Blanco
- **Columbia University in the city of New York**
  - Doctor en Computación,
    tesis con Kathy McKeown (AAAI fellow)

Hacer un doctorado en Nueva York

- **Muy interesante**
  - Nueva York es un universo en una ciudad
- **Bien preparado por la Fa.M.A.F.**
  - Sobre todo a nivel emocional de entender lo que es hacer un doctorado
- **Dualidad jefe / director de tesis**
  - Muy marcada en EE. UU.
- **Proyectos bien “aterrizados”**
  - Algunos generaban también problemas de índole ético
    - Genética (minería de textos)
    - Medicina (generación de reportes médicos)
    - Inteligencia Militar (generación de biografías)

Mi Tesis Doctoral

- **Defendida el 16 de Enero del 2005**
- **Aprendizaje automático aplicado a la generación de texto**
  - Estimación de parámetros de una formula compleja
    - Similar a resolver una ecuación con varias incógnitas
- **Generación de textos**
  - La computadora tiene datos
  - A partir de ellos genera (produce, “escribe”) textos
- **Dos cuestiones:**
  - ¿Qué decir?
  - ¿En qué orden?

¿En qué orden?

- **Tarea realmente muy compleja**
  - Muchos órdenes posibles
- **Enfocado a textos cuya estructura es en algún sentido muy repetitiva**
  - Buenos resultados en problemas más sencillos
    - Reportes médicos
  - No muy buenos en biografías,
    - La estructura también es fija pero es más “recursiva”

¿Qué decir?

- **Aprendizaje de unidades de información que son “interesantes” para ser incluidas en una biografía**
- **Ejemplo (para actores)**
  - obtener oscares ⇨ *interesante*
  - obtener premios de la asociación de críticos de Montreal ⇨ *no*
- **Aprendizaje**
  - Totalmente automático
  - Sin necesidad de ayuda por parte del creador del sistema
    - Poco versado en cinematografía ☺

Selección Automática de Información

- **Datos sobre 1,100 “celebridades”**
  - De las páginas de E! Online entertainment television
  - Datos estructurados
  - A su vez los estructuré aún más,
    - Forman grafos (mapas de conocimiento) con cientos de nodos por persona
- **Por otro lado, biografías de las celebridades**
  - (Madonna! Britney Spears!)
- **El sistema desarrollado puede darse cuenta si una determinada pieza de información aparecía en la biografía o no.**
  - Sin “entender” las biografías (hoy por hoy eso no es posible) sino analizando las distribuciones de palabras

E! online

July 24, 2003

HOME | NEWS | FEATURES | GOSSIP | REVIEWS+ | CELEBS | FUN&GAMES | MULTIMEDIA | E!TV

TODAY'S NEWS

FIRST LOOK: The News in Brief

Report: Schlesinger Off Life Support

Yo! Joey Scores "Friends" Spinoff

Missy Elliott Works MTV Vid Magic

BE ON TV

Have you got the goods for one of our cool shows in production? Find out!

FRESH FEATURES

Love Chain: Follow the links (and kinks) in J.Lo's history

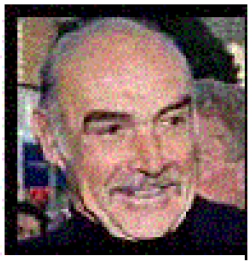
Watch with Kristin: Love for Everwood, a CSI shocker and lots more TV dish

Music Reviews: We rate new stuff from Mya, 311, Jane's Addiction, more

The Awful Truth: It's two guys

THE FACTS

Sean Connery



the facts | credits | stories | multimedia | fanclubs

get the goods

search for Sean Connery products:

- movies
- collectibles

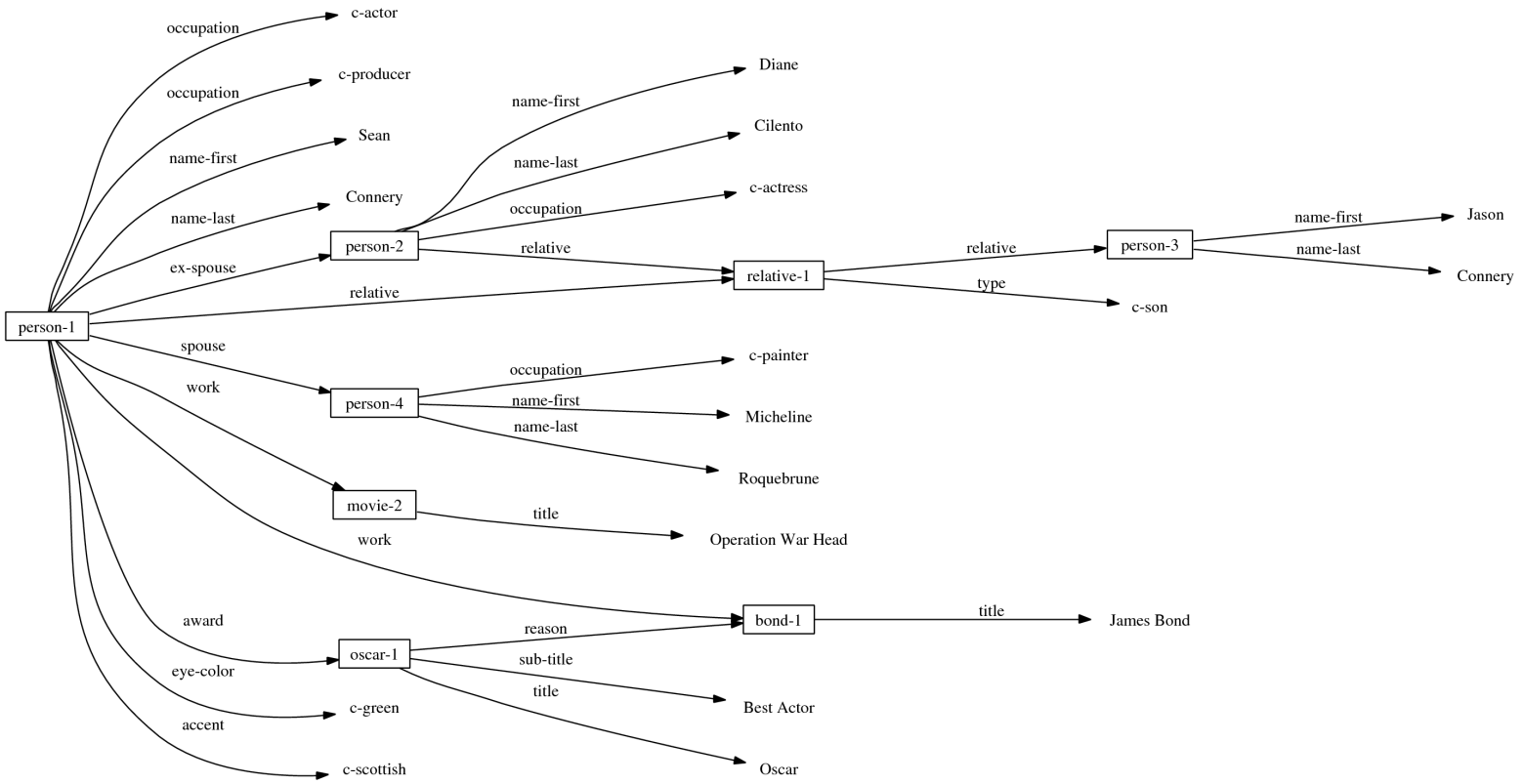
Birth Name: Thomas Sean Connery
Birthdate: August 25, 1930
Birthplace: Edinburgh, Scotland
Occupations: Actor, Director, Model, Producer
Quote: "I would drink Sean Connery's bath water." --Whoopi Goldberg, Cable Magazine, 1989

"He's...one of the best actors there is, simple as that... With Sean, in addition to brilliant talent, there is a persona that every great star has. When Sean's...on the screen, it's hard to look at anything else. To be a great star, you have to be a first- rate

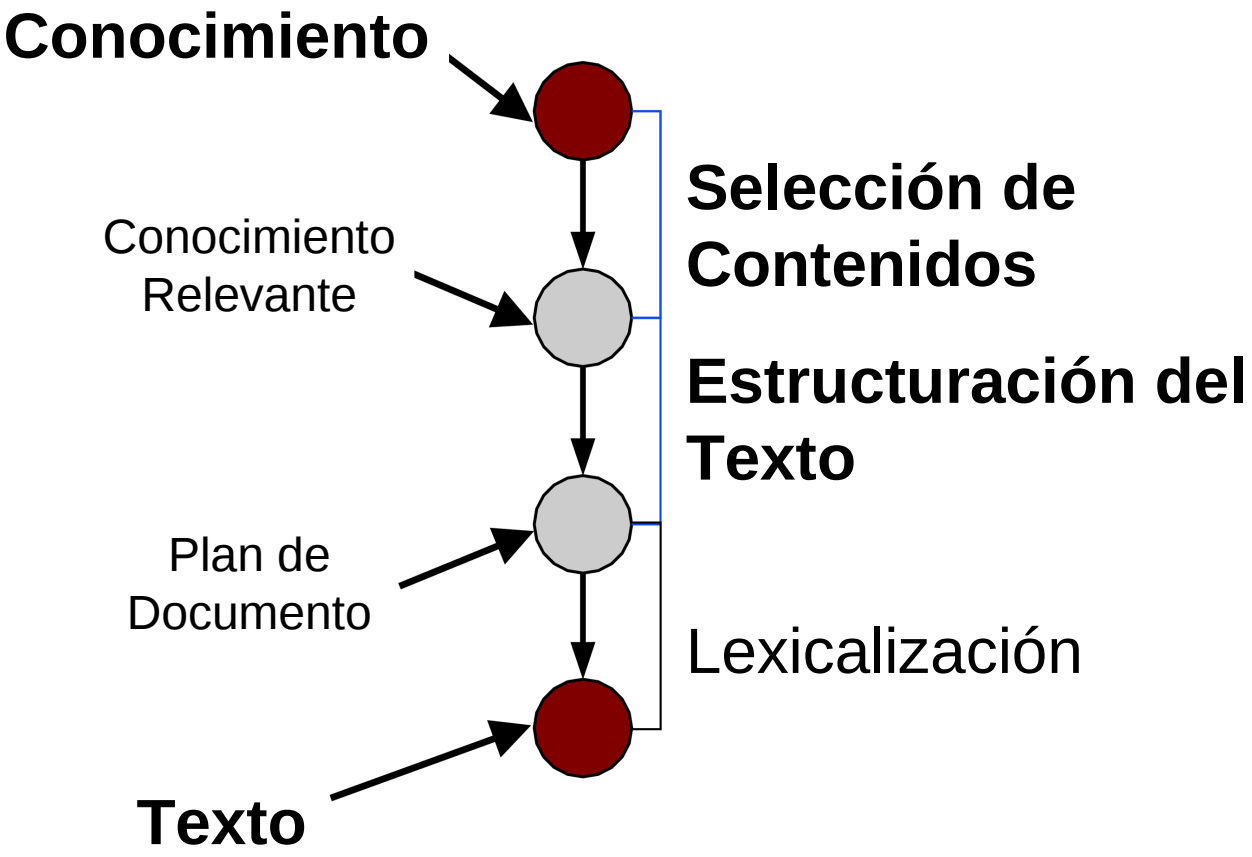
tonight on E!

Shannen Doherty
THS: Discover the fires that fueled her notorious tueds; 8 p.m.

Grafo de Conocimiento



Estructura de mi tesis



Marcado Automático de Información

- **Dado un texto y un grafo de conocimiento**
  - Marcar que información del grafo aparece en el texto
  - Fácil para las personas
  - Difícil para las computadoras
- **Ejemplo:**
  - Si el nodo del grafo dice “15” y el texto menciona “quince”
  - Buscar por el mismo texto exacto no sirve
- **Solución: agregar código que transforme “15” en “quince”**
- **Pero…**
  - Lo mismo con “premios de la academia” y “oscares”
  - Y un largo etcétera.

Un Paréntesis

- **Elementos comunes con otros problemas del procesamiento del lenguaje humano**
  - En el problema:
    - Sencillo para los humanos
    - Difícil para la computadora
    - Hay que salvar grandes dificultades para llegar a una aplicación útil.
  - En la solución:
    - Las computadoras realmente todavía no "entienden" los textos
    - Se trata de aproximar mediante técnicas tales como
      - Conteos gruesos de palabras sobre textos
      - Soluciones que funcionan parcialmente

La Técnica

- **Supongamos que tenemos un cierto dato sobre la vida de una persona**
  - Ese dato es irrelevante para un estilo particular de biografías
  - Por ejemplo: el peso
- **Si ese dato es realmente irrelevante:**
  - Si tomamos todas las biografías de personas similares con respecto a ese dato (con un peso similar),
  - Dichas biografías no deberían ser “parecidas” al conjunto total de biografías
    - “Parecidas” si las vemos como distribuciones de probabilidades sobre palabras
    - Agh!

Ejemplo

- **Supongamos que tenemos ocho personas:**
  - Pocho, Toto, Cholo, Tom, Moncho, Rodolfo, Otto, Pololo.

- **Sobre ellos sabemos los siguientes datos:**
- **Pocho es odontólogo, pesa 80 kilos y nació en Jujuy.**
- **Toto es comerciante, pesa 103 kilos y nació en San Luis.**
- **Cholo es comerciante, pesa 80 kilos y nació en San Luis.**
- **Tom es médico, pesa 76 kilos y nació en Buenos Aires.**
- **Moncho es comerciante, pesa 70 kilos y nació en San Luis.**
- **Rodolfo es cantante, pesa 72 kilos y nació en Buenos Aires.**
- **Otto es comerciante, pesa 83 kilos y nació en Jujuy.**
- **Pololo es comerciante, pesa 120 kilos y nació en Buenos Aires.**

- **Si tenemos las siguiente biografías:**
- **Pocho: Odontólogo jujeño.**
- **Toto: Comerciante puntano.**
- **Cholo: Comerciante nacido en San Luis.**
- **Tom: Médico porteño.**
- **Moncho: Comerciante puntano.**
- **Rodolfo: Cantante nacido en Buenos Aires.**
- **Otto: Comerciante jujeño.**
- **Pololo: Comerciante porteño.**

Conteos

- **Todas las biografías tienen estos conteos de palabras:**
  - Odontólogo: 1
  - Comerciante: 5
  - Médico: 1
  - Cantante: 1
  - Jujeño: 2
  - Puntano: 2
  - Porteño: 2
  - San Luis: 1
  - Buenos Aires: 1

Conteos parciales

- **Si nos fijamos en las biografías de las personas que pesan alrededor de 70 kilos**
  - Moncho y Otto
- **Los conteos son:**
  - Comerciante: 1
  - Cantante: 1
  - Puntano: 1
  - Buenos Aires: 1
- **Similares al conjunto completo**
  - “pesa-70-kilos” no aparece mencionado

- **Si nos fijamos en nacido-en-jujuy**
  - Pocho y Otto
- **Los conteos son:**
  - Odontólogo: 1
  - Comerciante: 1
  - **Jujeño: 2**
- **Diferentes del conjunto completo**
  - “nacido-en-jujuy” aparece mencionado

Comentarios sobre la Técnica

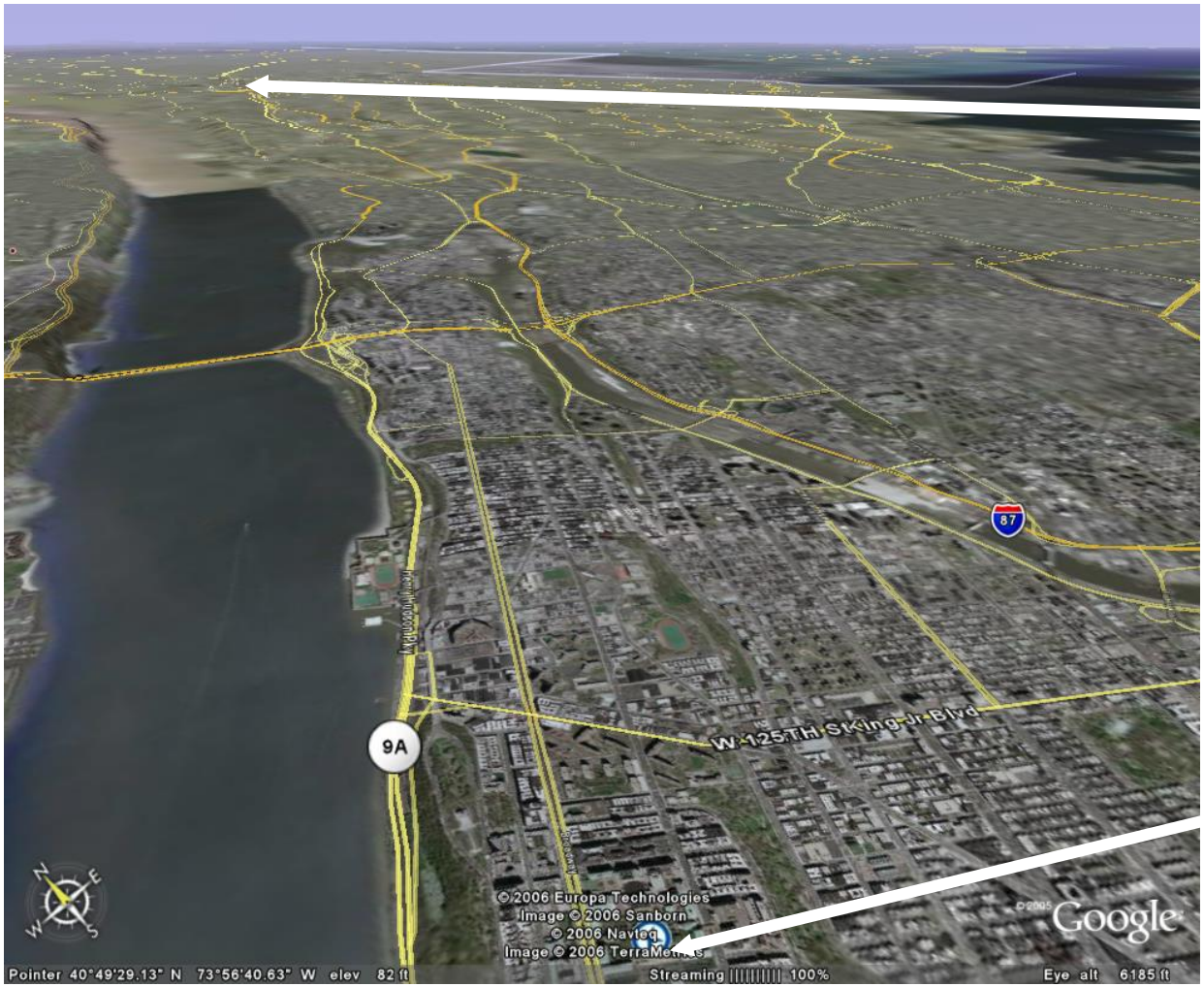
- **Mediante esta técnica fue posible construir alineación de texto y conocimiento**
  - Mejora sobre la búsqueda simple
- **El algoritmo presentado es una simplificación**
  - Se utiliza un muestreo sobre los datos
  - Se comparan las distribuciones entre el conjunto y el resto
- **Permite el aprendizaje de reglas de selección de información**
- **Ejemplo:**
  - No es necesario decir que la persona nació en EE. UU.
  - (Para biografías de celebridades en E! Online se sobreentiende.)

Problemas Técnica

- **Las palabras asociadas con la divergencia de las distribuciones no son necesariamente lexicalizaciones del concepto**
  - En el ejemplo anterior, el sistema puede inferir que ser nacido en San Luis puede expresarse con la palabra "comerciante"
  - Un producto secundario de los datos, no una señal real en sí.
- **También es posible que una cierta información no sea mencionada explícitamente en el texto pero que en algún modo cambie la distribución de las palabras en el texto**
  - Por ejemplo, personas obesas son más predispuestas a enfermedades como la diabetes, con lo cual el sistema podría inferir que el peso es mencionado si el texto menciona la causa de la muerte en biografías históricas.

Respuestas automáticas

- **La segunda parte técnica de la charla**
- **Comentarios sobre algunas cosas en las que estoy trabajando en el centro TJ Watson de IBM.**
  - El mismo donde se originaron los fractales en la década de los 50
  - Y donde se desarrolló la computadora Deep Blue que le ganó a Kasparov.
  - Un centro con unos 2,000 investigadores y una cafetería llena de elementos de origen dudoso.



Trabajar como investigador corporativo

- **Lo bueno**
  - No hay que preocuparse (tanto) del financiamiento
  - Se trabaja en cosas útiles a la sociedad
  - Equipos medianos y una estructura plana
  - Discusiones constantes entre pares, trabajo en conjunto
  - Mucha estabilidad laboral
- **Lo malo**
  - Muchísima confidencialidad
  - Para tener estabilidad, que hay ajustarse a los lineamientos generales de la empresa

El Problema

- **En mi grupo tratamos de mejorar la búsqueda de información**
  - Sobre todo en presencia de búsquedas complejas
- **Estas mejoras son gracias a tecnologías del lenguaje**
  - En comparación a otros grupos que tratan de mejorar la búsqueda de información usando otras técnicas
- **La directora del grupo es nacida en Canadá, educada en Taiwán, hizo doctorado en EE. UU. y después de casarse se convirtió al judaísmo**
  - En el grupo somos cinco, incluyendo a otro argentino, un inglés y un norteamericano.

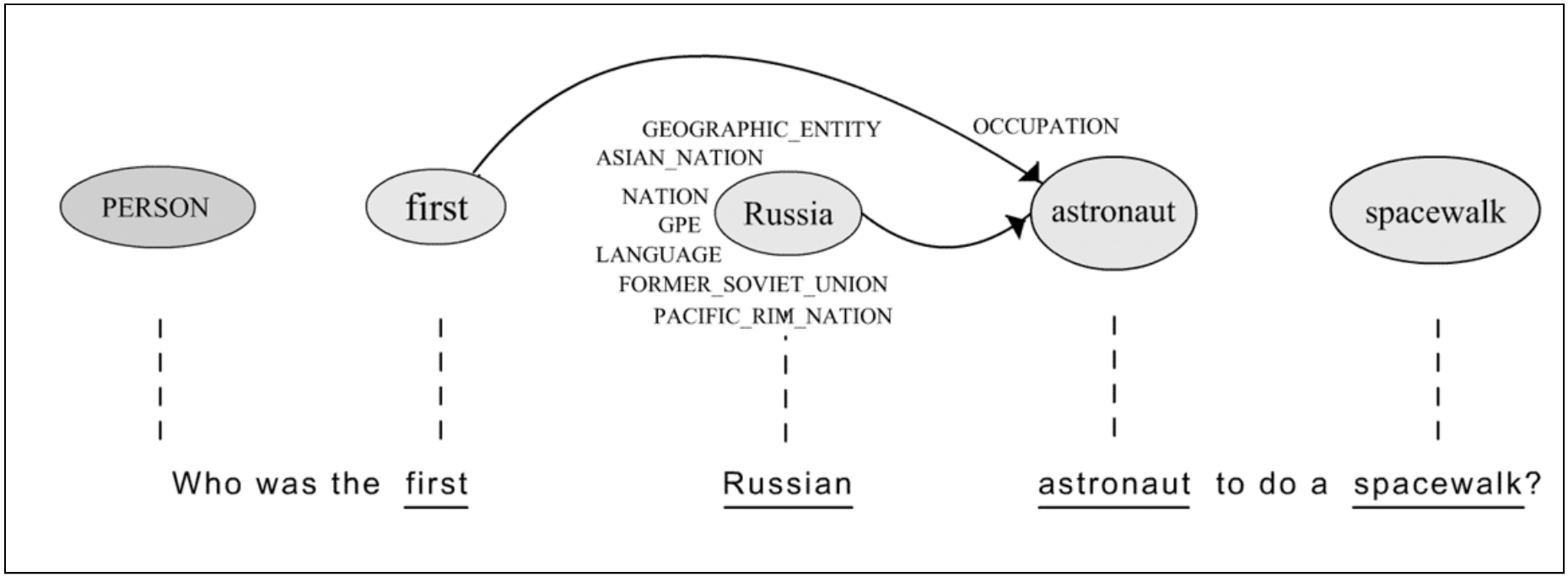
El Sistema

- **En particular, nos dedicamos al tema de preguntas y respuestas automáticas.**
  - El sistema desarrollado en el grupo se llama "Picante“ (Piquant)
  - El año pasado terminó tercero en una competencia internacional de sistemas automáticos de preguntas y respuestas (20+ participantes)
  - El sistema funciona así (grosso modo):
    1. Analiza la pregunta usando un analizador sintáctico automático (un programa que asigna “sujeto,” “verbo,” etc. como en la escuela, pero por computadora), y utiliza un conjunto de reglas para predecir qué tipo de dato está buscando la persona (sobre un total de 90 tipos de datos que conoce el sistema)
    2. Transforma la pregunta en una conjunto de búsquedas como las búsquedas que uno hace normalmente en Google, pero más complejas y con operadores que involucran también tipos de información
    3. Se procesan los documentos devueltos por la búsqueda y se extraen dos o tres oraciones donde el sistema considera que puede estar la respuesta a la pregunta
    4. Se extraen respuestas de las oraciones seleccionadas

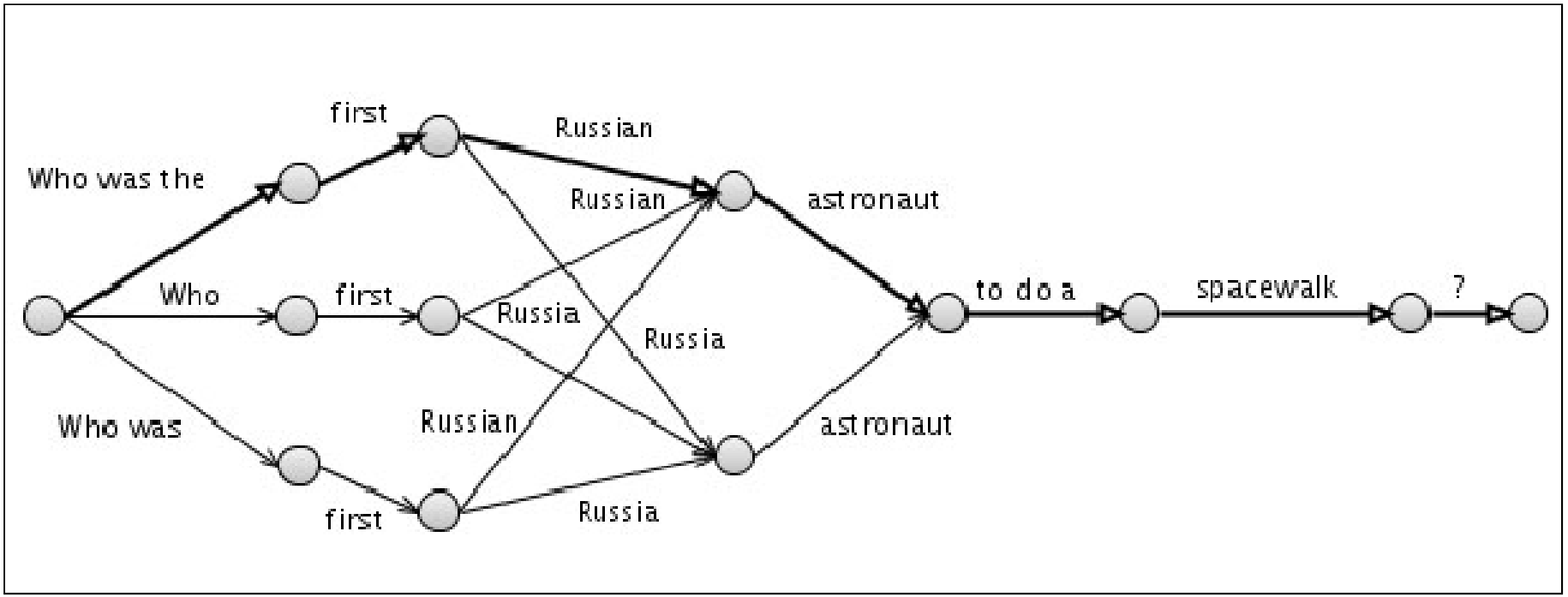
Verificación de Preguntas por Inversión

- **El sistema tiene un alto nivel de error**
  - De alrededor del 60%
- **Pero los errores se pueden asumir independientes**
- **Si se tiene una lista de respuestas candidatas**
  - Es posible verificarlas creando preguntas “invertidas”
- **Ejemplo**
  - Pregunta original “¿Cuál es la capital de Francia?”
  - Respuestas candidatas: Berlín, París, …
  - Preguntas de verificación (pregunta “invertida”)
    - ¿De qué país es Berlín la capital?
    - ¿De qué país es París la capital?
  - La respuesta correcta debe tener “Francia” como resultado
- **Trabajo conjunto con John Prager, presentado en el congreso de la asociación de lingüística computacional este año**

Generación de Preguntas



Reticulado Probabilístico de Lexicalizaciones



Repercusiones a Nivel Laboral

- **Estas nuevas tecnologías están haciendo posibles nuevo negocios**
  - Motores de búsqueda para nichos específicos
  - Acopio y análisis de noticias a gran escala
  - Nuevas posibilidades para los lingüistas de carrera
- **A su vez, hacen obsoletos o reemplazan a gran cantidad de personas en otras áreas**
  - Particularmente en áreas intensivas de trabajo humano
  - Por ejemplo, los centros de atención al público vía operadores telefónicos.
    - Los sistemas actuales pueden contestar una pregunta correctamente de cada tres (sin contar errores en la comprensión de voz)
    - Eso significa una posible reducción de personal en un tercio del mismo
- **Esto es similar a otros avances tecnológicos**
  - Por ejemplo, el diseño gráfico por computadora revolucionó las imprentas
- **Problema**
  - Las características del proceso de comunicación hacen que los trabajadores que van a ser reemplazadas ni se imaginen que ello sea factible
  - Hay que tener las tecnologías del lenguaje presentes

Repercusiones a Nivel Gubernamental

- **Las tecnologías del lenguaje permiten a los gobiernos analizar de manera automática cantidades ingentes de texto y audio.**
  - Un proceso que ya hacían a mano,
    - Pero sobre un conjunto restringido por una contención económica
  - Hoy en día es cada vez más fácil de hacer ese tipo de análisis con medios más bien modestos (desde el punto de vista de los gastos gubernamentales)
- **Esto empieza a permitir un flujo más fácil de la información desde todos hacia los gobernantes**
  - Podría remover la excusa de “no haber entendido al pueblo”
  - Tiende hacia una democracia más directa
- **Pero puede ser utilizado con fines de persecución ideológica.**
- **Debates para establecer en la sociedad**
  - ¿Cuál es la tecnología de vigilancia electrónica que el gobierno usa?
  - ¿Es aceptable a nivel social?
    - Ejemplo: EE. UU. y Holanda

Conclusión

- **Procesamiento del lenguaje humano**
  - Fascinante
  - Complejo
  - Todavía lejos de ser resuelto
- **A nivel individuo**
  - Nuevos servicios
  - Nuevas posibilidades de ser reemplazados por máquinas
- **A nivel social**
  - Nuevas posibilidades para hacerse escuchar
  - Nuevos desafíos respecto del rol del gobierno