Generación de Lenguaje Natural y Aplicaciones

Carlos Areces y Luciana Benotti

{carlos.areces,luciana.benotti}@gmail.com

INRIA Nancy Grand Est, Nancy, France
Universidad Nacional de Córdoba, Córdoba, Argentina

ELiC 2010 - Buenos Aires - Argentina

De que se Trata este Curso?

- Vamos a hablar de Generación Automática de Lenguaje Natural (GLN)
- Es decir, el diseño e implementación de sistemas que
  - producen texto comprensible en lenguaje natural (e.g., Castellano, Inglés, etc.)
  - a partir de una representación no lingüística de información
  - usando conocimiento acerca del lenguaje y del dominio de aplicación.

Objetivos del Curso

- Dar un panorama amplio del área y de lo que es posible hacer hoy en día.
- Introducir en detalle algunas de las técnicas (algunas básicas y otras más avanzadas) del área.
- Discutir algunos temas que son importantes para la aplicación de técnicas de GLN en proyectos concretos.

Estructura del Curso

- Primera Parte: Carlos Areces
  - Lunes: El Problema de Generación de Lenguaje Natural. Algunos sistemas de GNL. GNL Pipeline. Representación de Información e Inferencia para GLN.

Estructura del Curso

- Primera Parte: Carlos Areces
  - Lunes: El Problema de Generación de Lenguaje Natural. Algunos sistemas de GNL. GNL Pipeline. Representación de Información e Inferencia para GLN.
  - Martes: Tree Adjoining Grammars (TAG). Interface Sintáctica-Semántica. Realización. Realización via Charts.

Estructura del Curso

- Primera Parte: Carlos Areces
  - Lunes: El Problema de Generación de Lenguaje Natural. Algunos sistemas de GNL. GNL Pipeline. Representación de Información e Inferencia para GLN.
  - Martes: Tree Adjoining Grammars (TAG). Interface Sintáctica-Semántica. Realización. Realización via Charts.
  - Miércoles: Algoritmos de Generación de Expresiones Referenciales. Información Proposicional vs. Información Relacional. Optimización de Algoritmos. Evaluación.

Estructura del Curso

- Primera Parte: Carlos Areces
  - Lunes: El Problema de Generación de Lenguaje Natural. Algunos sistemas de GNL. GNL Pipeline. Representación de Información e Inferencia para GLN.
  - Martes: Tree Adjoining Grammars (TAG). Interface Sintáctica-Semántica. Realización. Realización via Charts.
  - Miércoles: Algoritmos de Generación de Expresiones Referenciales. Información Proposicional vs. Información Relacional. Optimización de Algoritmos. Evaluación.
- Segunda Parte: Luciana Benotti
  - Jueves: Entornos Virtuales (e.g., Second Life) y Aplicaciones (e.g., Tutoring) para Sistemas de GNL. Inferencia Orientada a Metas. Algoritmos de Planning y su uso en Entornos Virtuales.

Estructura del Curso

- Primera Parte: Carlos Areces
  - Lunes: El Problema de Generación de Lenguaje Natural. Algunos sistemas de GNL. GNL Pipeline. Representación de Información e Inferencia para GLN.
  - Martes: Tree Adjoining Grammars (TAG). Interface Sintáctica-Semántica. Realización. Realización via Charts.
  - Miércoles: Algoritmos de Generación de Expresiones Referenciales. Información Proposicional vs. Información Relacional. Optimización de Algoritmos. Evaluación.
- Segunda Parte: Luciana Benotti
  - Jueves: Entornos Virtuales (e.g., Second Life) y Aplicaciones (e.g., Tutoring) para Sistemas de GNL. Inferencia Orientada a Metas. Algoritmos de Planning y su uso en Entornos Virtuales.
  - Viernes: Generación de Referencias en un Entorno Virtual. Estrategias de Referencia. Supervisión de la Interpretación. Evaluación.

Evaluación

Viene en dos sabores

- Examen Takehome: Con preguntas teóricas y ejercicios prácticos sobre los contenidos del curso. Lo publicaremos a más tardar el Lunes, en la página del curso. Se resuelve en forma individual. Se envia a las 15 días.
- Projectos de Desarrollo: Definimos tres proyectos de desarrollo de sistemas de GNL extendiendo un baseline dado. El framework esta en Java. Se trabaja en grupos de dos personas. Se entregará código + documento explicando las ideas y testing. Se envía a las 3 semanas.
  (Si hay más interesados podemos definir algunos proyectos más de este tipo.)
  (Quizás más trabajo, pero viene con bonus.)

Workshop Satelite de Iberamia

Si los proyectos son interesantes (i.e., si funcionan!) los invitamos a una presentación en la sesión de estudiantes del:

Workshop on Natural Language Processing and Web-based Technologies

Collocated with: iberamia 2010 Bahía Blanca - Argentina

1 November
Bahia Blanca
Argentina

http://cs.uns.edu.ar/iberamia2010/

Si quieren charlar sobre los proyectos nos vienen a ver en cualquier recreo.

Lo que Veremos Hoy

- Introducción a GLN
- Un Caso de Estudio
- Las Tareas básicas de GLN
- GLN en Ambientes Multimedia y Multimodales

Lo que Veremos Hoy

- Introducción a GLN
  - Que es GLN?
  - Ejemplos
  - Aplicaciones típicas de GLN
  - Cuando es apropiado usar GLN?
  - La Arquitectura de un sistema de GLN
- Un Caso de Estudio
- Las Tareas básicas de GLN
- GLN en Ambientes Multimedia y Multimodales

Qué es GLN?

- Natural language generation is the process of deliberately constructing a natural language text in order to meet specified communicative goals.

  [McDonald 1992]

Qué es GLN?

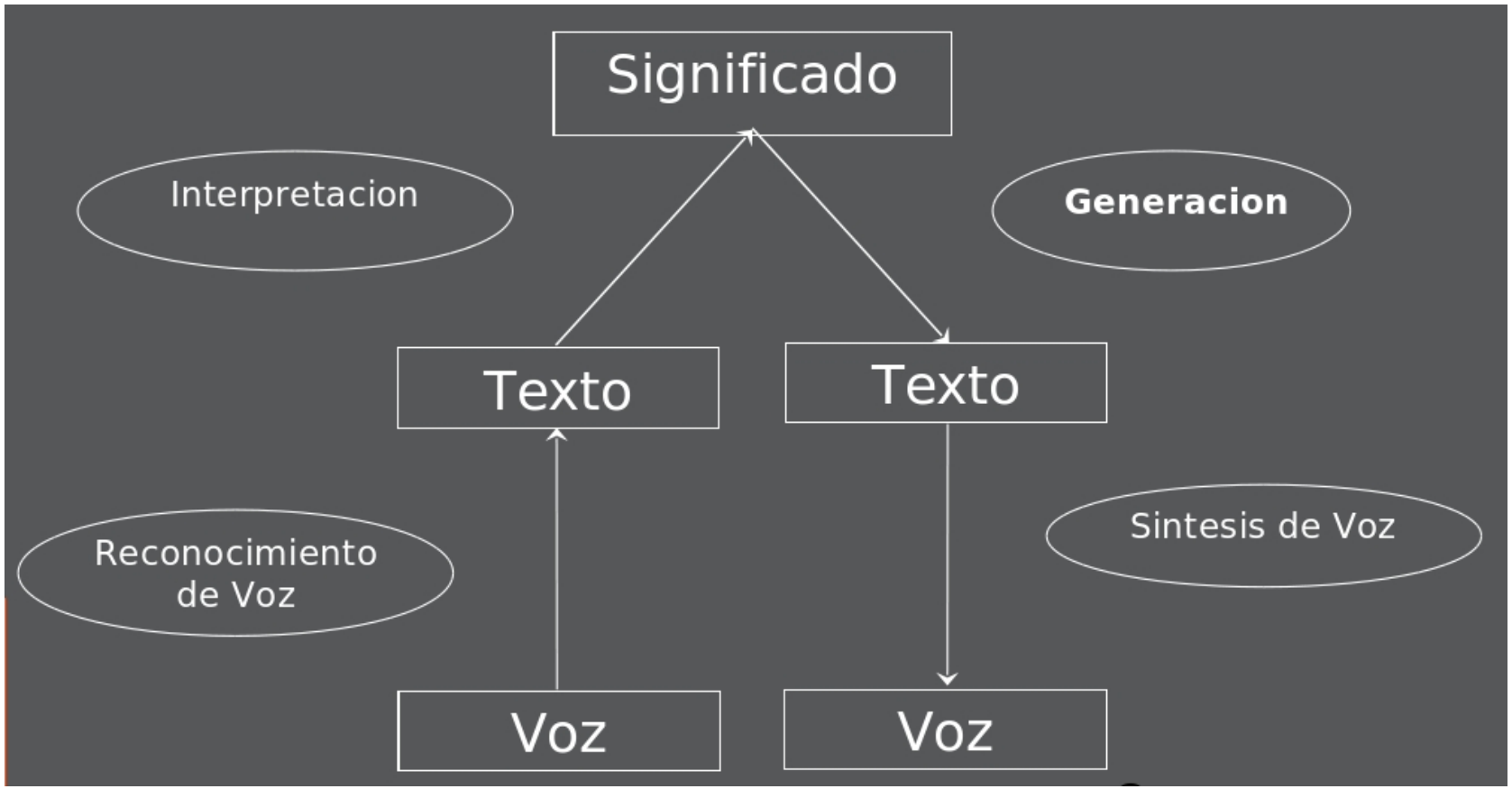
- Natural language generation is the process of deliberately constructing a natural language text in order to meet specified communicative goals.

[McDonald 1992]

Qué es GLN?

- Objetivo:
  - software que produce texto entendible y adecuado en lenguaje natural (e.g., Inglés).
- Input:
  - Información no lingüística (e.g., una base de datos)
- Output:
  - documentos, reportes, explicaciones, mensajes de ayuda, etc.
- Información requerida:
  - Conocimiento del lenguaje y del dominio de aplicación

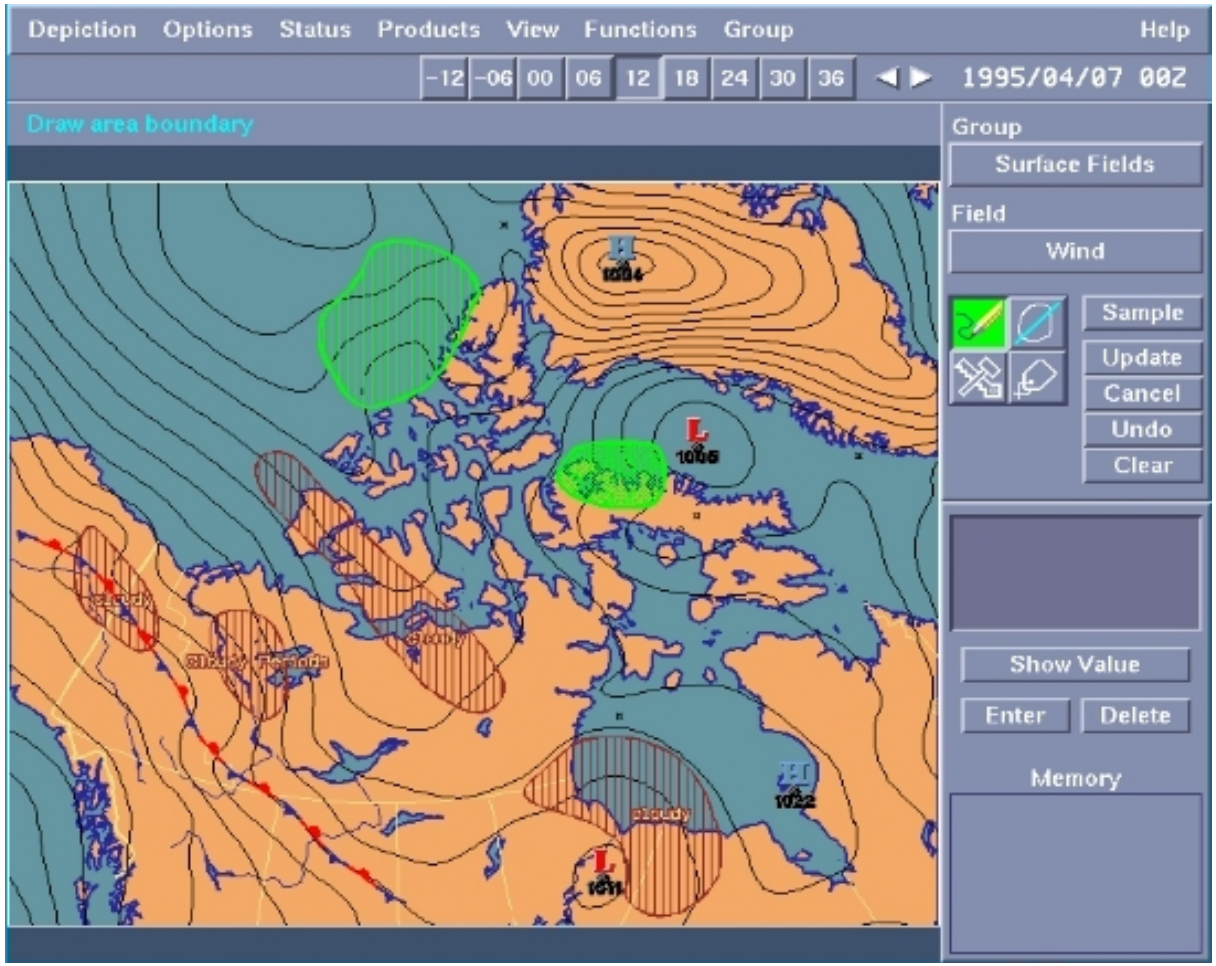
Generación vs. Interpretación



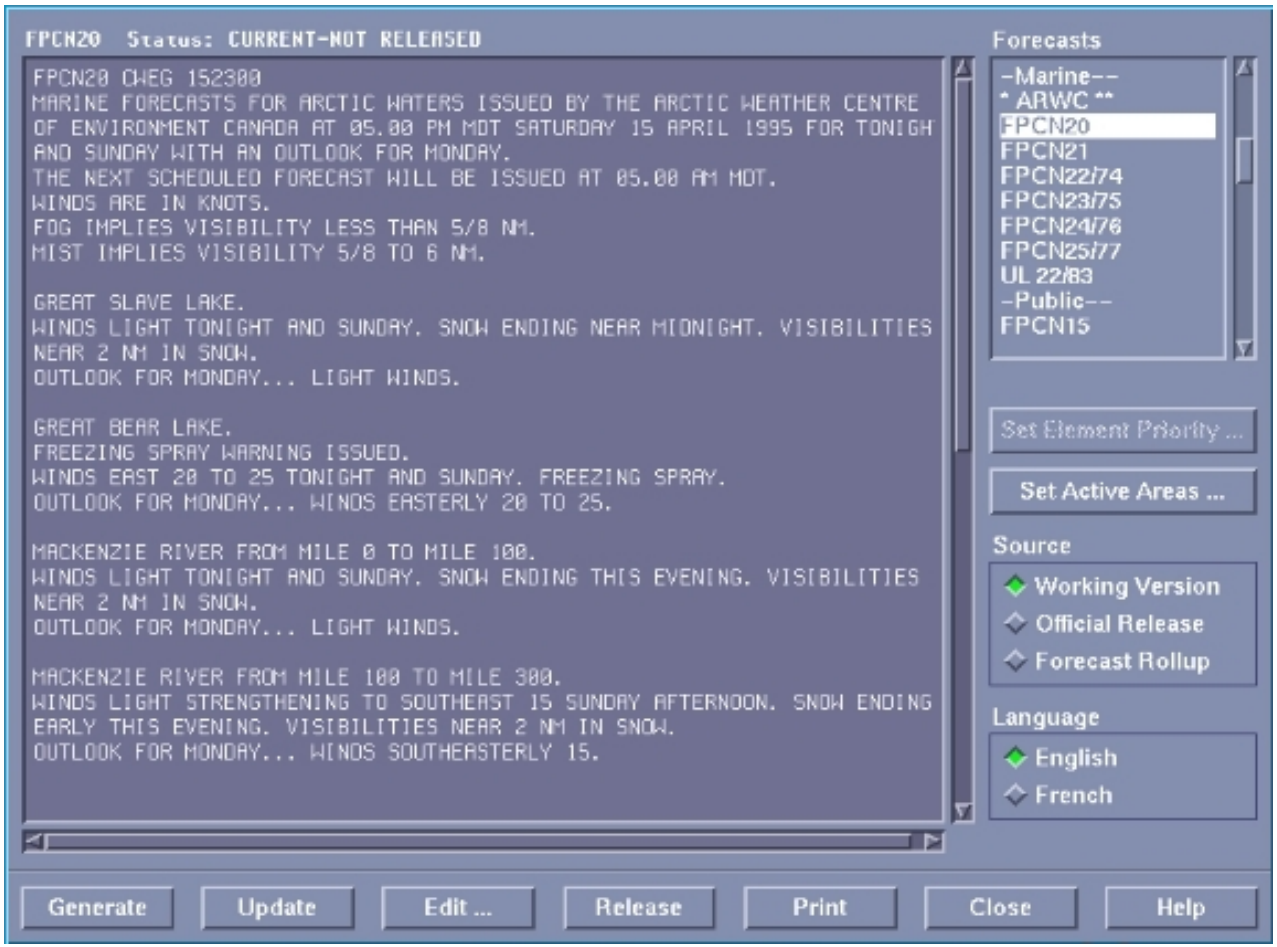
Sistema Ejemplo #1: FoG

- Función:
  - Producir reportes climáticos en formato texto en Inglés y en Francés.
- Input:
  - Imágen gráfica climática con información numérica
- Usuario:
  - Environment Canada (Servicio Climático Canadiense)
- Status:
  - Funcionando desde 1992

FoG: Input


Depiction
Options
Status
Products
View
Functions
Group
Help
-12
-06
00
06
12
18
24
30
36
1995/04/07 00Z
Draw area boundary
Group
Surface Fields
Field
Wind
Sample
Update
Cancel
Undo
Clear
Show Value
Enter
Delete
Memory

FoG: Output


FPCN20 Status: CURRENT-NOT RELEASED
FPCN20 CWEG 152300
MARINE FORECASTS FOR ARCTIC WATERS ISSUED BY THE ARCTIC WEATHER CENTRE
OF ENVIRONMENT CANADA AT 05.00 PM MDT SATURDAY 15 APRIL 1995 FOR TONIGH
AND SUNDAY WITH AN OUTLOOK FOR MONDAY.
THE NEXT SCHEDULED FORECAST WILL BE ISSUED AT 05.00 AM MDT.
WINDS ARE IN KNOTS.
FOG IMPLIES VISIBILITY LESS THAN 5/8 NM.
MIST IMPLIES VISIBILITY 5/8 TO 6 NM.
GREAT SLAVE LAKE.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING NEAR MIDNIGHT. VISIBILITIES
NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.
GREAT BEAR LAKE.
FREEZING SPRAY WARNING ISSUED.
WINDS EAST 20 TO 25 TONIGHT AND SUNDAY. FREEZING SPRAY.
OUTLOOK FOR MONDAY... WINDS EASTERLY 20 TO 25.
MACKENZIE RIVER FROM MILE 0 TO MILE 100.
WINDS LIGHT TONIGHT AND SUNDAY. SNOW ENDING THIS EVENING. VISIBILITIES
NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... LIGHT WINDS.
MACKENZIE RIVER FROM MILE 100 TO MILE 300.
WINDS LIGHT STRENGTHENING TO SOUTHEAST 15 SUNDAY AFTERNOON. SNOW ENDING
EARLY THIS EVENING. VISIBILITIES NEAR 2 NM IN SNOW.
OUTLOOK FOR MONDAY... WINDS SOUTHEASTERLY 15.
Forecasts
--Marine--
* ARWC **
FPCN20
FPCN21
FPCN22/74
FPCN23/75
FPCN24/76
FPCN25/77
UL 22/83
--Public--
FPCN15
Set Element Priority ...
Set Active Areas ...
Source
Working Version
Official Release
Forecast Rollup
Language
English
French
Generate
Update
Edit ...
Release
Print
Close
Help

Sistema Ejemplo #2: PlanDoc

- Función:
  - Producir un reporte describiendo las opciones de simulación que un ingeniero ya ha explorado
- Input:
  - Un archivo de log de simulaciones
- Usuario:
  - Southwestern Bell
- Status:
  - Funcionando desde 1996

PlanDoc: Input

```
RUNID fiberall FIBER 6/19/93 act yes
FA 1301 2 1995
FA 1201 2 1995
FA 1401 2 1995
FA 1501 2 1995
ANF co 1103 2 1995 48
ANF 1201 1301 2 1995 24
ANF 1401 1501 2 1995 24
END. 856.0 670.2
```

PlanDoc: Output

This saved fiber refinement includes all DLC changes in Run-ID ALLDLC. RUN-ID FIBERALL demanded that PLAN activate fiber for CSAs 1201, 1301, 1401 and 1501 in 1995 Q2. It requested the placement of a 48-fiber cable from the CO to section 1103 and the placement of 24-fiber cables from section 1201 to section 1301 and from section 1401 to section 1501 in the second quarter of 1995. For this refinement, the resulting 20 year route PWE was $856.00K, a $64.11K savings over the BASE plan and the resulting 5 year IFC was $670.20K, a $60.55K savings over the BASE plan.

Sistema Ejemplo #3: STOP

- Function:
  - Producir un folleto personalizado para ayudar a dejar de fumar
- Input:
  - Questionario sobre historia, creencias, actitudes, etc. sobre el cigarrillo
- Usuario:
  - NHS (British Health Service)
- Status:
  - Utilizado por varios años

STOP: Input

SMOKING QUESTIONNAIRE

Please answer by marking the most appropriate box for each question like this: ¤

Q1 Have you smoked a cigarette in the last week, even a puff?

YES ¤ — Please complete the following questions

NO ☐ — Please return the questionnaire unanswered in the envelope provided. Thank you.

Please read the questions carefully. If you are not sure how to answer, just give the best answer you can.

Q2 **Home situation**:

Live alone ☐ | Live with husband/wife/partner ¤ | Live with other adults ☐ | Live with children ¤

Q3 **Number of children** under 16 living at home ..... boys .....1..... girls

Q4 **Does anyone else in your household smoke?** *(If so, please mark all boxes which apply)*

husband/wife/partner ¤ | other family member ¤ | others ☐

Q5 **How long have you smoked for?** ...10... years

Tick here if you have smoked for less than a year ☐

STOP: Output

Dear Ms Cameron

Thank you for taking the trouble to return the smoking questionnaire that we sent you. It appears from your answers that although you're not planning to stop smoking in the near future, you would like to stop if it was easy. You think it would be difficult to stop because *smoking helps you cope with stress, it is something to do when you are bored, and smoking stops you putting on weight*. However, you have reasons to be confident of success if you did try to stop, and there are ways of coping with the difficulties.

Sistema Ejemplo #4: TEMSIS

- Función:
  - Sumarización de información sobre contaminación
- Input:
  - Datos ambientales + una pregunta específica
- Usuario:
  - Agencias ambientales en Francia y Alemania
- Status:
  - Prototipos fueron instalados en la region Saar/Alsacia (borde entre Alemania y Francia).

TEMSIS: Input Query

```
((LANGUAGE FRENCH)
  (GRENZWERTLAND GERMANY)
  (BESTAETIGE-MS T)
  (BESTAETIGE-SS T)
  (MESSSTATION \"Voelklingen City\"
  (DB-ID \"#2083\"
  (SCHADSTOFF \"#19\"
  (ART MAXIMUM)
  (ZEIT ((JAHR 1998)
         (MONAT 7)
         (TAG 21))))
```

TEMSIS: Output Summary

- Francés:
  Le 21/7/1998 à la station de mesure de Völklingen-City, la valeur moyenne maximale d'une demi-heure (Halbstundenmittelwert) pour l'ozone atteignait 104.0 μg/m^3. Par conséquent, selon le decret MIK (MIK-Verordnung), la valeur limite autorisée de 120 μg/m^3 n'a pas été dépassé.
- Alemán:
  Der höchste Halbstundenmittelwert für Ozon an der Meßstation Völklingen-City erreichte am 21.7.1998 104.0 μg/m^3, womit der gesetzlich zulässige Grenzwert nach MIK-Verordnung von 120 μg/m^3 nicht überschritten wurde.

Tipos de Aplicaciones de GLN

- Producción automática de documentos
  - reportes climáticos, reporte de simulaciones, cartas, ...
- Presentación de información al público en forma entendible
  - imformes médicos, sistemas expertos de inferencia, ...
- Enseñanza
  - educación a distancia
- Entretenimiento/Arte
  - bromas (?), historias (??), poesía (???)

El Rol de la Computadora

Dos posibilidades

- El sistema produce un documento automáticamente (sin ayuda humana)
  reportes climáticos, reportes de simulaciones, cartas a pacientes, resúmenes de datos estadísticos, explicaciones en sistemas expertos.
- El sistema ayuda a un redactor humano a crear un documento:
  reportes climáticos, reportes de simulaciones, cartas a pacientes, pedidos de patentes, documentos técnicos (manuales), pedidos de empleo

En qué Casos son las Técnicas de GLN Adecuadas?

Opciones a Considerar:

- Texto vs. Gráficos
  - Qué medio es mejor?
- Generación Automática vs. Autoría Humana
  - Son los datos necesarios accesibles?
  - Vale la pena (e.g., económicamente)?
- GLN vs. Concatenación de strings
  - Cuánta variación hay en el texto?
  - Que impacto tiene la calidad gramatical del texto?

Calidad Gramatical

- La generación de texto lingüísticamente bien formado requiere la verificación de constraints
  - ortográficos, morfológicos, sintácticos
  - referencia, elección de palabras, pragmáticas
- Estos constraints se verifican automáticamente por un sistema de GLN
  - en forma automática, el 100% de los casos
- Los desarrolladores de sistemas basados en concatenación de strings tienen que verificar el cumplimiento de estos strings manualmente y vía testing
  - Muy trabajoso
  - Difícil de garantizar exactitud del 100%

Ejemplo: Syntaxis, agregación

- Output de sistemas de IA Medical existentes:

  The primary measure you have chosen, CXR shadowing, should be justified in comparison to TLC and walking distance as my data reveals they are better overall. Here are the specific comparisons:

  TLC has a lower patient cost TLC is more tightly distributed TLC is more objective walking distance has a lower patient cost

Ejemplo: Pragmática

- Output de un sistema que da versiones en inglés de consultas a una base de datos:

  The number of households such that there is at least 1 order with dollar amount greater than or equal to $100.

Ejemplo: Pragmática

- Output de un sistema que da versiones en inglés de consultas a una base de datos:

  The number of households such that there is at least 1 order with dollar amount greater than or equal to $100.

  - Se interpreta como "number of households which have placed an order $>=$ \$100"

Ejemplo: Pragmática

- Output de un sistema que da versiones en inglés de consultas a una base de datos:

  The number of households such that there is at least 1 order with dollar amount greater than or equal to $100.

  - Se interpreta como "number of households which have placed an order $>=$ \$100"
  - La consulta inicial era el número total de casas en la base de datos, si habia alguna orden en la base de datos (de cualquier casa) por mas de \$100

La Arquitectura de un Sistema de GLN

- Las tareas básicas en un sistema de GLN
- Arquitectura de Pipeline
- Alternative Architectures

Las Tareas Básicas en un sistema de GLN

1. Determinación de Contenido (Content determination)
2. Estructuración del Documento (Document structuring)
3. Agregación (Aggregation)
4. Lexicalización (Lexicalisation)
5. Generación de Expresiones Referenciales (Referring expression generation)
6. Realización Lingüística (Linguistic realisation)
7. Realización de Estructura (Structure realisation

Las Tareas Básicas en un sistema de GLN

Content Determination Document Structuring	Document planning
Aggregation Lexicalisation Referring Expression Generation	Micro-planning
Linguistic Realisation Structure Realisation	Surface realization

Una Arquitectura Pipeline

Otras Arquitecturas

- Variaciones de la arquitectura "standard":
  - Alterar el orden de las tareas
  - Permitir feedback entre las distintas etapas
- Un sistema integrado donde todas las tareas se combinan:
  - representar todas las tareas en forma uniforme: e.g. como constraints, axiomas, operadores de planning, ...
  - alimentar estas especificaciones a un constraint-solver, demostrador de teoremas, ...

Temas de Investigación

- Cuándo es texto la mejor forma de cominicarse con el usuario?
- Cuándo es GLN mejor que concatenación de strings?
- Existe alguna arquitectura que combine la elegancia teórica y la flexibilidad de un sistema integrado, con la simplicidad de una arquitectura pipeline?
- Como debemos representar Document Plans y Text Specifications?

Lo que Veremos Hoy

- Introducción a GLN
- Un Caso de Estudio
- Las Tareas básicas de GLN
- GLN en Ambientes Multimedia y Multimodales

Lo que Veremos Hoy

- Introducción a GLN
- Un Caso de Estudio
  - Generando Resúmenes Climáticos
  - Recopilación y Uso de Corpus
- Las Tareas básicas de GLN
- GLN en Ambientes Multimedia y Multimodales

Un Caso de Estudio en GLN Aplicada

- Cada mes, un periódico institucional publica un resumen climático del mes
- El resumen se basa en datos meteorológicos recolectados automáticamente
- La persona que hasta el momento hacía este trabajo deja la institución

Un Resumen Climático

MARSFIELD (Macquarie University No 1)
On Campus, Square F9

TEMPERATURES (C)

Mean Max for Mth:	18.1 Warmer than average
Mean Max for June (20 yrs):	17.2
Highest Max (Warmest Day):	23.9 on 01
Lowest Max (Coldest Day):	13. On 12
Mean Min for Mth:	08.2 Much warmer than ave
Mean Min for June (20 yrs):	06.4
Lowest Min (Coldest Night):	02.6 on 09
Highest Min (Warmest Night):	13.5 on 24

RAINFALL (mm) (24 hrs to 09:00)

Total Rain for Mth:	90.4 on 12 days. Slightly below average.
Wettest Day (24h to 09:00):	26.4 on 11
Average for June (25 yrs):	109.0 on 10
Total for 06 mths so far:	542.0 on 72 days. Very depleted.
Average for 06 mths (25 yrs):	762.0 on 71 days
Annual Average Rainfall (25 yrs):	1142.8 on 131 days

WIND RUN (at 2m height) (km) (24 hrs to 09:00)

Total Wind Run for Mth:	1660
Windiest Day (24 hrs to 09:00):	189 on 24, 185 on 26, 172 on 27
Calmest Day (24 hrs to 09:00):	09 on 16

SUNRISE & SUNSET

Date	Sunrise	Sunset	Difference
01 Jun	06:52	16:54	10:02
11 Jun	06:57	16:53	09:56
21 Jun	07:00	16:54	09:54
30 Jun	07:01	16:57	09:56

(Sunset times began to get later after about June 11)
(Sunrise times continue to get later until early July)
(Soon we can take advantage of the later sunsets)

SUMMARY

The month was warmer than average with average rainfall, but the total rain so far for the year is still very depleted. The month began with mild to warm maximums, and became cooler as the month progressed, with some very cold nights such as June 09 with 02.6. Some other years have had much colder June nights than this, and minimums below zero in June are not very unusual. The month was mostly calm, but strong winds blew on 23, 24 and 26, 27. Fog occurred on 17, 18 after some rain on 17, heavy rain fell on 11 June.

Output: Un Resumen Climático

The month was warmer than average with average rainfall, but the total rain so far for the year is still very depleted. The month began with mild to warm maximums, and became cooler as the month progressed, with some very cold nights such as June 09 with 02.6. Some other years have had much colder June nights than this, and minimums below zero in June are not very unusual. The month was mostly calm, but strong winds blew on 23, 24 and 26, 27. Fog occurred on 17, 18 after some rain on 17, heavy rain fell on 11 June.

Los Datos de Input

- Un conjunto de 16 datos recolectados automáticamente cada 15 minutos: presión del aire, temperatura, velocidad del viento, lluvia caída, ...
- Preprocesados para obtener un instancia de DailyWeatherRecords:

```
((type dailyweatherrecord)
   (date ((day ...)
          (month ...)
          (year ...)))
 (temperature ((minimum ((unit degrees-centigrade)
                         (number ...)))
               (maximum ((unit degrees-centrigrade)
                         (number ...)))))
 (rainfall ((unit millimetres)
            (number ...))))
```

Otros Datos Disponibles

- Datos Históricos: E.g. Temperaturas máximas y mínimas registradas para los distintos meses.
  Nos permite generar cosas como "La temperatura excedió la máxima histórica para Mayo"
- Datos Promedio: E.g. Valores promedio de temperatura y lluvia en lo que viene del año.
  Nos permite generar cosas como "El mes fue más cálido que el anterior"

Análisis de Requirimientos basado en Corpus

Un corpus

- consiste de ejemplos de textos generados anteriormente con sus correspondientes datos de entrada
- especifica 'mediante ejemplos' la funcionalidad esperada del sistema de GLN
- servirá de patrón para los mensajes que queremos generar

Análisis de Requirimientos basado en Corpus

Cuatro Actividades:

- recolectar un corpus inicial de textos generados a mano con sus correspondientes datos de input
- analizar el contenido del corpus en termino de los datos de input
- desarrollar un corpus target
- especificar formalmente el mapeo de datos a texto

Paso 1: Crear el Corpus Inicial

Recolectar corpus de texto (generados anteriormente a mano) y los correspondientes datos de input

- Una fuente pueden ser ejemplos archivados
- Si no existen ejemplos anteriores se deberá recurrir a expertos del dominio para que produzcan ejemplos
- El corpus debe proveer ejemplos de la totalidad de casos que se esperan manejar con el sistema de GLN

Texto Inicial

- SUMMARY
- The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Paso 2: Analizar el Contenido del Corpus

- Objetivo:
  - determinar de donde viene la información contenida en el texto, y en qué medida el sistema de GLN tendrá que manipular esta información
- Resultado:
  - un entendimiento detallado de la correspondencia entre los datos de entrada existentes y el texto generado en cada caso en el corpus

En particular queremos clasificar el texto a generar en 4 clases: Texto fijo, texto generado directamente a partir de los datos, texto generado a partir de datos computables, texto no soportado por los datos.

Ejemplo

SUMMARY
The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Texto Fijo

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Obtenido Directamente de los Datos

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Obtenido de Datos Computables

SUMMARY
The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Sin Datos de Soporte

SUMMARY
The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Resolviendo el Problema de la Falta de Datos

- Quizas podemos dar datos adicionales al sistema
  - agregar censores en Mars Creek?
- Si el sistema en realidad esta ayudando en la redacción a un humano, esta información podrá ser agregada más tarde
  - el sistema produce las primeras dos sentencias, el redactor humano agrega luego las últimas dos
- El corpus target es revisado para eliminar las frases vinculadas con este tipo de información.
  - produciremos solamente las primeras dos sentencias

Paso 3: Construyendo el Corpus Target

- Cambios Obligatorios:
  - eliminar texto generado a partir de datos inaccessible
  - especificar las porciones que serán generadas por el redactor humano
- Cambios Opcionales:
  - simplificar el texto para que sea más fácil de generar
  - mejorar la coordinaciń entre el texto generado automáticamente y el texto generado por el redactor humano

Texto Target

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again.

Paso 4: Especificación Funcional

- Basada en el corpus target obtenido
- Define en forma explícita el role del redactor humano (si corresponde)
- Define en forma explícita la estructura y el rango del input que será utilizado

Texto Inicial vs. Texto Target

Texto Initial: The month was our driest and warmest August in our 24 year record, and our first 'rainless' month. The 26th August was our warmest August day in our record with 30.1, and our first 'hot' August day (30). The month forms part of our longest dry spell 47 days from 18 July to 02 September 1995. Rainfall so far is the same as at the end of July but now is very deficient.

Texto Target: The month was the driest and warmest August in our 24 year record, and the first rainless month of the year. 26th August was the warmest August day in our record with 30.1, and the first hot day of the month. Rainfall for the year is now very deficient.

Vale la Pena usar GLN?

- Para un resumen por mes probablemente no. Sobre todo teniendo en cuanta las simplificaciones que debimos introducir en los textos para hacerlso fácil de generar.
- Pero nuestro cliente esta interesado en un caso piloto porque:
  - en el futuro los reportes se haran de forma semanal.
  - hay varios sitios de recolección automática de datos, y el sistema podría utilizarse en todos ellos.

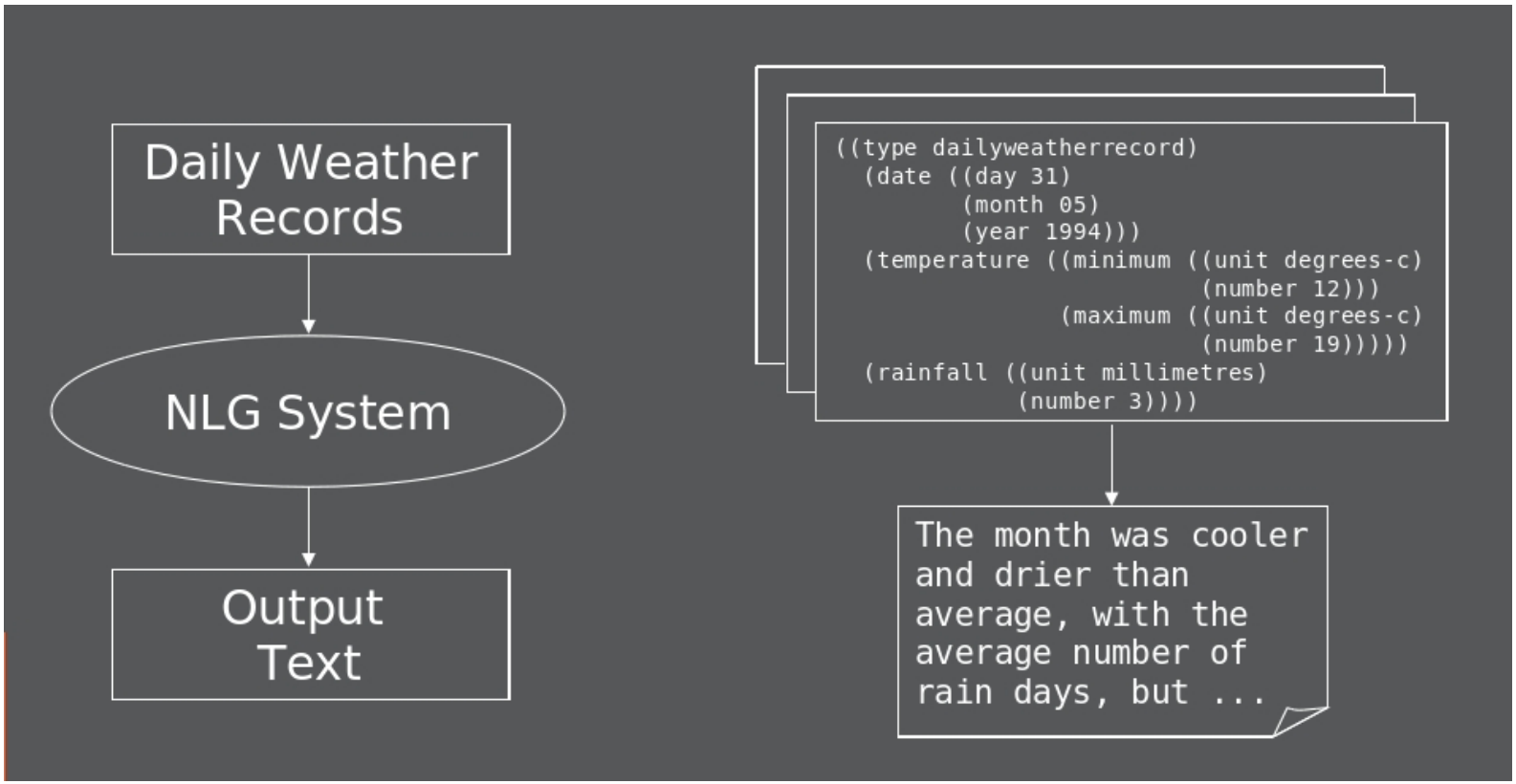
Lo que Veremos Hoy

- Introducción a GLN
- Un Caso de Estudio
- Las Tareas Básicas de GLN
- GLN en Ambientes Multimedia y Multimodales

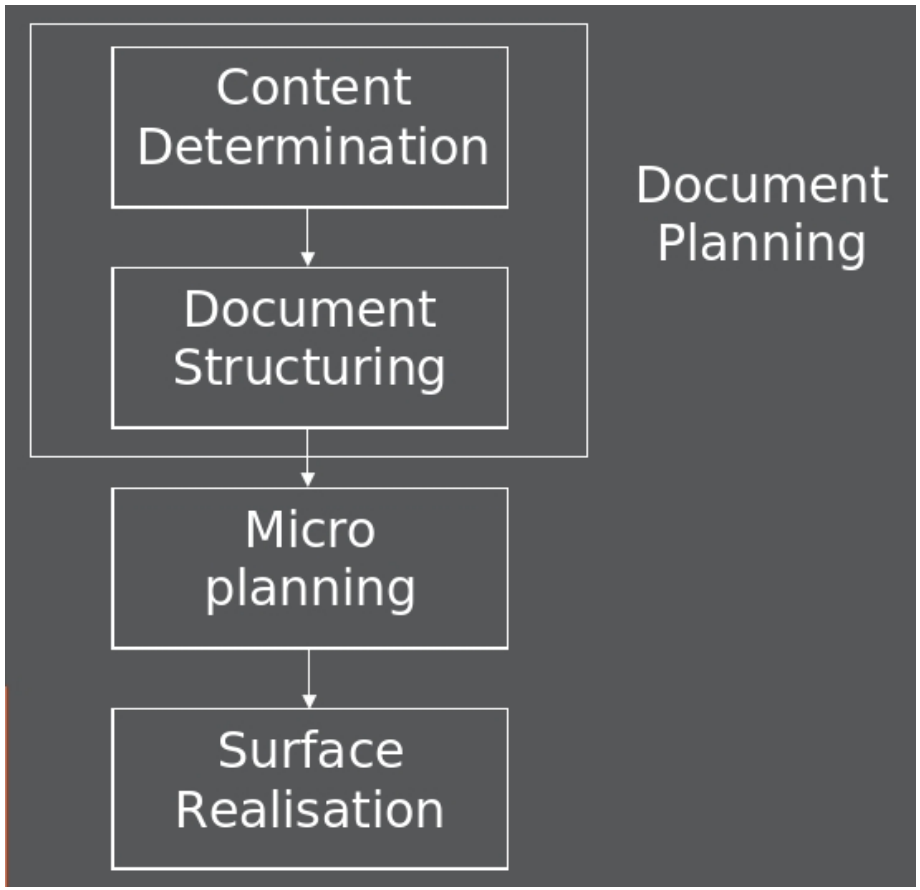
Lo que Veremos Hoy

- Introducción a GLN
- Un Caso de Estudio
- Las Tareas Básicas de GLN
  - Document Planning
  - Microplanning
  - Surface Realization
- GLN en Ambientes Multimedia y Multimodales

Inputs y Outputs



La Arquitectura



Document Planning

- Objetivos:
  - determinar que información debe comunicarse
  - determinar como estructurar esta información para obtener un texto coherente
- Existen dos enfoques usuales:
  - métodos basados en observaciones directas de cómo se estructura el texto en ejemplos
  - métodos basados en razonamiento sobre coherencia del discurso y el objetivo comunicativo del texto

Content Determination

- Usualmente basado en mensajes: estructuras de información predefinidas que
  - se corresponden con bloques de información en el texto
  - agrupan elementos de información de forma de facilitar su expresion textual
- Idea Fundamental:
  - A partir del análisis del corpus, identificar los agrupamientos de elementos de información lo más grande posibles, que no limiten nuestra flexibilidad al querer generarlos.

Content Determination en WeatherReporter

- Mensajes Rutinarios
  - MonthlyRainFallMsg,
  - MonthlyTemperatureMsg,
  - RainSoFarMsg,
  - MonthlyRainyDaysMsg
- Se incluyen en todos los resumenes a generar

Content Determination en WeatherReporter

MonthlyRainfallMsg:

```
((message-id msg091)
 (message-type monthlyrainfall)
 (period ((month 04)
          (year 1996)))
 (absolute-or-relative relative-to-average)
 (relative-difference ((magnitude ((unit millimeters)
                                   (number 4)))
                       (direction +))))
```

Content Determination en WeatherReporter

- Mensajes de Eventos Significativos
  - RainEventMsg,
  - RainSpellMsg,
  - TemperatureEventMsg,
  - TemperatureSpellMsg
- Sólo se general cuando los datos lo indiquen: e.g., si se registran lluvias en un número de días consecutivos mayor a una cantidad especificada.

Content Determination en WeatherReporter

A RainSpellMsg:

```
((message-id msg096)
 (message-type rainspellmsg)
 (period ((begin ((day 04)
                  (month 02)
                  (year 1995)))
          (end ((day 11)
                (month 02)
                (year 1995)))
          (duration ((unit day)
                     (number 8)))))
 (amount ((unit millimetres)
          (number 120))))
```

Document Structuring mediante Esquemas

La idea básica

- Los textos de un determinado tipo siguen (usualmente) patrones convencionalizados
- estos patrones pueden ser expresados mediante 'gramáticas de texto' que indican el contenido a generar y aseguran una estructura coherente.
- estos patrones especifican como se construirá el plan de un documento particular usando esquemas más chicos o mensajes atómicos.

Document Structuring mediante Esquemas

Implementando esquemas:

- los esquemas más simples se especifican mediante gramáticas
- esquemas más flexibles se especifican como macros, o clases de librerías sobre lenguajes de programación convencionales, donde cada esquema es un procedimiento.
- este es, hoy en día, el método de document planning mas usual en sistemas de GLN

Derivando Esquemas a Partir del un Corpus

Usando el corpus target:

- tomar un cierto número (pequeño) de textos similares
- identificar los mensajes, y determinar como cada mensaje puede ser computado a partir de los datos de input
- proponer reglas o estructuras que expliquen por que el mensaje x es en el texto A pero no en el B. (Esta tarea puede ser más fácil si los mensajes se organizan en una taxonomía)
- discutir este análisis con expertos del dominio, e iterar
- repetir los pasos anteriores con conjuntos cada vez mas grandes de texto del corpus

Document Structuring en WeatherReporter

Un esquema simple:

```
WeatherSummary →
        MonthlyTempMsg
        MonthlyRainfallMsg
        RainyDaysMsg
        RainSoFarMsg
```

Document Structuring en WeatherReporter

Un conjunto de esquemas más interesante

```
WeatherSummary →
    TemperatureInformation RainfallInformation

TemperatureInformation →
    MonthlyTempMsg [ExtremeTempInfo] [TempSpellsInfo]

RainfallInformation →
    MonthlyRainfallMsg [RainyDaysInfo] [RainSpellsInfo]

RainyDaysInfo →
    RainyDaysMsg [RainSoFarMsg]
...
```

Esquemas: Pros and Cons

- Ventajas:
  - Computacionalmente eficientes
  - Relativamente simples de obtener a partir de un corpus
  - Permiten especificar naturalmente particularidades de un determinado dominio (i.e., customizables)
  - Pueden ser arbitrariamente complejos
- Desventajas
  - Flexibilidad Limitada: requieren la especificación a priori de todas las estructuras posibles
  - Portabilidad Limitada: en general, son particulares al dominio

Document Structuring mediante Razonamiento Explícito

- Observación:
  - La coherencia de un texto se obtiene a partir de ciertas relaciones que existen entre las distintas partes. Relaciones como secuencia narrativa, elaboración, justificación
- Idea:
  - organizar el conocimiento de que es lo que hace un texto coherente en forma de reglas
  - usar estas reglas para construir textos dinámicamente a partir de fragmentos elementales mediante razonamiento del rol de cada elemento en el texto a construir

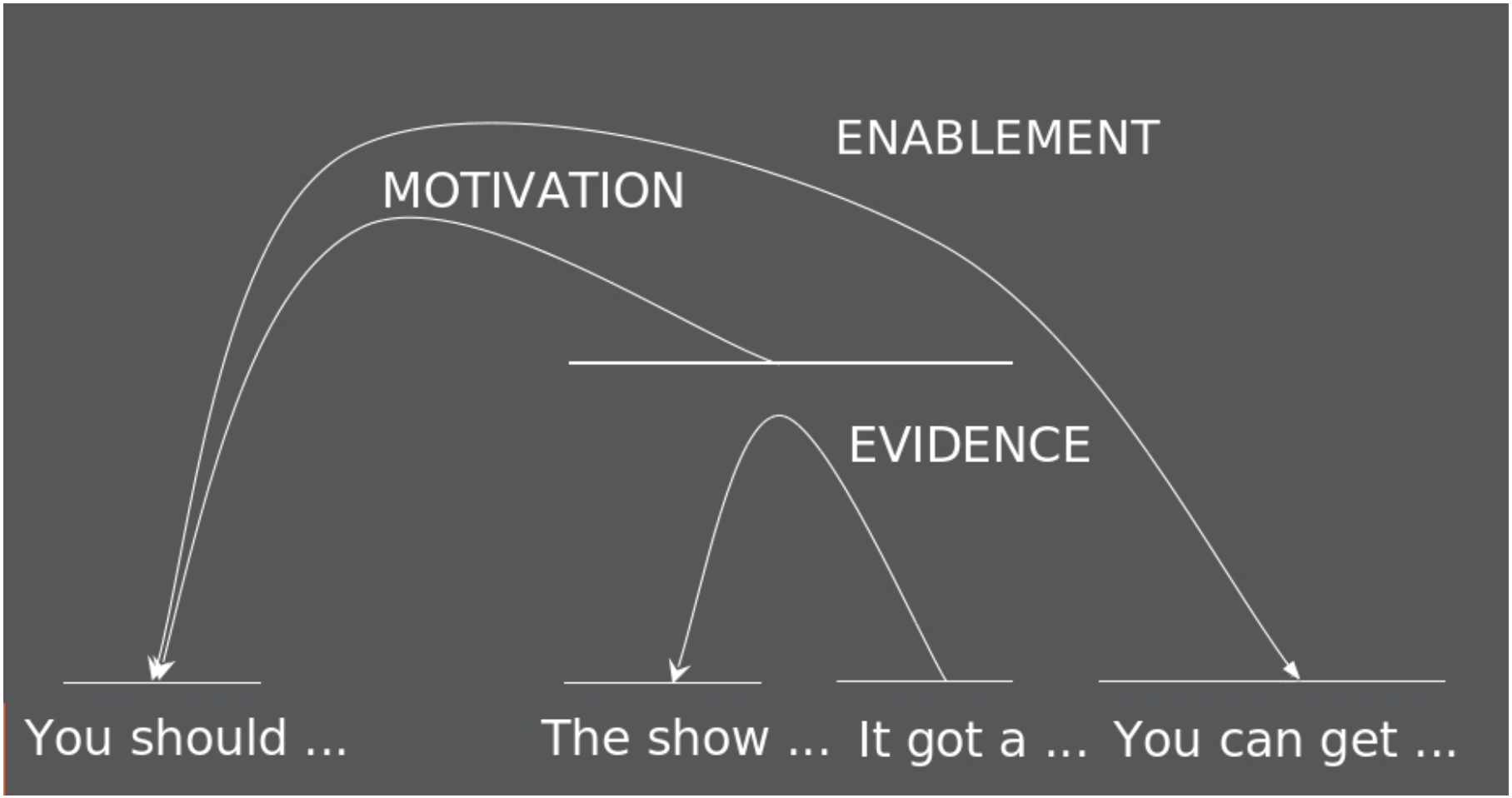
Document Structuring mediante Razonamiento Explícito

- Típicamente usan tecnicas de AI planning
  - Goal = el efecto comunicativo deseado
  - Elementos del Plan = mensajes o estructuras que combinan mensajes (subplans)
- Puede requerir razonamiento explícito sobre el conocimiento del usuario.
- Usualmente basados en ideas de Rethorical Structure Theory

Rethorical Structure Theory

- D1: You should come to the Northern Beaches Ballet performance on Saturday.
- D2: The show is really good.
- D3: It got a rave review in the Times.
- D4: You can get the tickets from the shop next door.

Rethorical Structure Theory



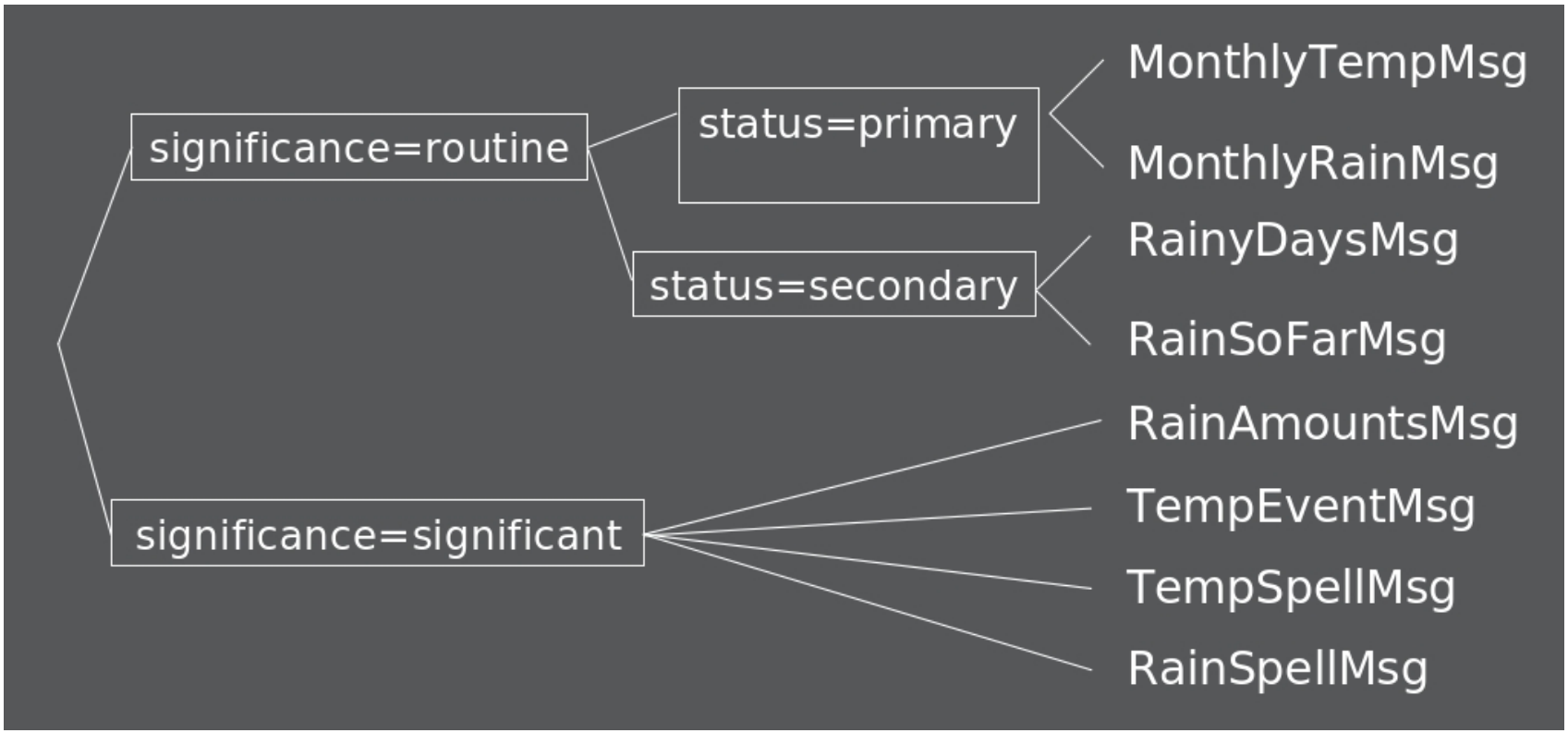
Definición de una Relacion en RST

```
Relation name:  Motivation

Constraints on N:
Presents an action (unrealised) in which the hearer
is the actor

Constraints on S:
Comprehending S increases the hearers desire to
perform the action presented in N

The effect:
The hearers desire to perform the action presented
in N is increased
```

Document Structuring en WeatherReporter

- Tres relaciones RST basicas
  - SEQUENCE
  - ELABORATION
  - CONTRAST
- Las reglas de aplicación de cada una de estas relaciones se definen a partir de atributos de los mensajes

Atributos de Mensajes



Document Structuring en WeatherReporter

- SEQUENCE
  Dos mensajes pueden conectarse mediante una relación de SEQUENCE si ambos tienen el atributo message-status = primary
- ELABORATION
  Dos mensajes pueden conectarse mediante la relación ELABORATION si: ambos tienen el mismo message-topic, el nucleo tiene menssage-status = primary
- ...

Document Structuring en WeatherReporter

- Selecionar un mensaje para comenzar con atributo message-significance = routine
- Aplicar relaciones retóricas a dos mensajes en esta estructura hasta que todos los mensajes hallan sido consumidos o hasta que no puedan aplicarse más relaciones

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

The Message Set:

- MonthlyTempMsg ("cooler than average")
- MonthlyRainfallMsg ("drier than average")
- RainyDaysMsg ("average number of rain days")
- RainSoFarMsg ("well below average")
- RainSpellMsg ("8 days from 11th to 18th")
- RainAmountsMsg ("amounts mostly small")

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

MonthlyTempMsg – SEQUENCE $\rightarrow$ MonthlyRainfallMsg

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

MonthlyTempMsg – SEQUENCE $\rightarrow$ MonthlyRainfallMsg
MonthlyRainfallMsg – ELABORATION $\rightarrow$ RainyDaysMsg

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

MonthlyTempMsg – SEQUENCE $\rightarrow$ MonthlyRainfallMsg
MonthlyRainfallMsg – ELABORATION $\rightarrow$ RainyDaysMsg
RainyDaysMsg – CONTRAST $\rightarrow$ RainSoFarMsg

Ejemplo

The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

MonthlyTempMsg – SEQUENCE $\rightarrow$ MonthlyRainfallMsg
MonthlyRainfallMsg – ELABORATION $\rightarrow$ RainyDaysMsg
RainyDaysMsg – CONTRAST $\rightarrow$ RainSoFarMsg
...

Document Planning

- El resultado de este paso es un Plan del Documento: una estructura en forma de árbol que tiene mensajes en sus nodos terminales.
- El siguiente paso es realizar estos mensajes como texto

Temas de Investigación

- Por el momento, la mayor parte del trabajo durante document structuring se hace ad-hoc
- Cómo podemos extraer esquemas a partir de un corpus?
- Mejor entendimiento de las relaciones retóricas
- Cómo podemos integrar esquemas y relaciones retóricas?
- Knowledge acquisition