Capítulo 5. Experimentando con el Parser del CAT y una gramática en Español de México para altas académicas

5.1 Introducción

Con el fin de presentar la capacidad de operación de este módulo, en un idioma diferente, y en un dominio diferente de aplicación, se efectuó la recolección de datos necesarios para diseñar una gramática en español de México y que esta fuera integrada Parser del CAT.

Una vez logrado lo anterior también con el apoyo de otra pequeña herramienta, El CMU Statistical Language Modeling implementada en la Universidad de Carnegie Mellon, la cual es una herramienta que ocupa modelos estadísticos para crear archivos de texto que pueden estructurar y presentar el modelo de lenguaje que se presenta a partir de archivos de texto. Estos archivos contienen por ejemplo el vocabulario, índices de probabilidad de los modelos bigramas, trigramas, etc. Con estos archivos se procedió a formular nuevos conjuntos de ejemplos para poder probar y evaluar el desempeño de este módulo, en este capítulo se presentara un ejemplo de los datos recolectados y la metodología empleada para la recolección de los mismos, los resultados que arrojo el CMU Statistical Language Modeling y de cómo fueron empleados para la prueba y evaluación de este módulo con la nueva gramática.

5.2 Elección del domino de aplicación para la gramática

Con el fin de mostrar el buen funcionamiento de esta herramienta, y de su buena capacidad de adaptación a otro idioma, se propone un dominio de aplicación en el cual se puede efectuar la inscripción de materias académicas a un determinado semestre. Es para el caso de estudio del presente trabajo, enfocarse a mostrar la capacidad de esta herramienta, para soportar la o las gramáticas empleadas por los usuarios, que efectúan intervenciones hechas de manera espontánea sin la restricción del uso de vocabulario.

Como un módulo de Entendimiento de Lenguaje Natural debe manejar oraciones incompletas, mal formadas o con errores de reconocimiento, y que obtiene de ellas la información substancial, la cual sirve para efectuar la acción correspondiente a lo deseado por el usuario, o para crear una representación lo suficientemente significativa para otro módulo del sistema (el manejador de diálogo).

5.2.1 Obtención de los datos necesarios para el diseño de la gramática

La forma de cómo se obtuvieron los ejemplos necesarios para poder diseñar la gramática empleada en esta aplicación, fue hecha a través de la grabación de diálogos con estudiantes de esta misma institución de la facultad de ingeniería en su mayoría, esto se efectuó utilizando una metodología tipo mago de OZ [Bersen 1998]. Un ejemplo de la trascripción de un diálogo obtenido se muestra en la figura 5.1. En total se lograron grabar 40 muestras de diálogo, con los cuales se generó la gramática que contemplase los diversos fenómenos de lenguaje que se presentan en dichas muestras (Ver Apéndice B).

Los medios de grabación y almacenamiento de estos diálogos para obtener una mejor calidad de audio, fueron grabadoras y cintas de formato DAT (Digital Audio Tape Recorder), micrófono y audífonos.

Obtenidas las muestras, éstas fueron capturadas en formato *.wav* para ser manipuladas en computadora, posteriormente los diálogos fueron transcritos a archivos de texto para que estos fueran procesadas por el CMU Statistical Language Modeling Toolkit para obtener su modelo de lenguaje y vocabulario (ver Apéndice B).

De las muestras obtenidas se diseño la gramática que se agregó al Parser del CAT(ver Apéndice A), la que posteriormente fue probada con nuevos modelos hechos de los resultados arrojados por el CMU SLM Toolkit.

Diálogo 110702

S1. Podrías darme tu nombre y numero de estudiante por favor	**U1**. Sí David Castell Guevara i ese once cero siete cero dos
S2. ¿Quieres tomar tus materias de acuerdo al nombre de los profesores, los horarios, o tienes idea de que horario quieres llevar?	**U2**. Este... quiero meter... creo que llevo calculo, quiero saber... ¿calculo uno en que horario esta?
S3. Calculo uno, hay nueve secciones, ¿quieres que las liste por profesor o por horario?	**U3**. Este... por profesor, quiero tomar con Palou
S4. El profesor Palou solamente tiene una sección de once a doce los martes y jueves	**U4**. Bueno a ver inscríbeme a esa sección
S5. ¿Quieres tomar otra materia?	**U5**. ¿Que otras materias tengo disponibles?
S6. Podrías tomar álgebra lineal	**U6**. A ver álgebra lineal, ¿En que horarios esta?
S7. Álgebra lineal hay dos secciones una de nueve a nueve cincuenta y la otra de cuatro a cinco y cuarto de la tarde	**U7**. ¿La de cuatro a cinco y cuarto quien la da?
S8. El profesor Soucek	**U8**. A ver con ese profesor
S9. ¿Quieres tomar otra materia?	**U9**. Este si este que es ele e doscientos cincuenta y uno es francés dos, ¿en que horarios esta esa materia?
S10. Tengo cinco secciones quieres las de la mañana o las de la tarde	**U10**. ¿En la mañana que horarios hay?
S11. La sección uno es de nueve a nueve cincuenta es la única sección en la mañana	**U11**. A ver esa sección
S12. ¿Quieres tomar otra materia?	**U12**. ¿Que otra materia puedo tomar?
S13. Podrías tomar estructura de datos	**U13**. Ahh.. ¿Estructura de datos con que profesores esta?

S14. Podrías tomarla con la profesora Ingrid Kirschning o con el profesor Cerón	**U14**. ¿Con Ingrid a que hora esta esa sección?
S15. La doctora Ingrid tiene dos secciones, una de ocho a ocho cincuenta y otra de cuatro a cuatro cincuenta de la tarde	**U15**. La de ocho a ocho cincuenta
S16.¿Quieres tomar otra materia?	**U16**. Este... Sí también tópicos selectos
S17. Tópicos selectos	**U17**. ¿Que horario tiene esa materia?
S18. Es los viernes de una a una cincuenta es la única sección	**U18**. A bueno sí esta bien
S19. ¿Quieres tomar ora materia?	**U19**. ¿Que otra materia me falta?
S20. Hasta el momento llevas treinta y dos unidades, ¿Estas seguro de que quieres tomar otra materia?	**U20**. No, así esta bien entonces
S21. Es todo quedaste inscrito a álgebra lineal calculo uno francés dos tópicos selectos y estructuras de datos. Gracias	**U21**. Sí esta bien gracias

Figura 5.1 Ejemplo del Trancript de uno de los diálogos obtenidos para la construcción de la gramática en español de México para el Parser del CAT

5.2.2 Resultados del comportamiento de los datos recabados

Se analizaron las muestras de diálogos obtenidas y además fueron depuradas por el CMU SML Toolkit para hallar el vocabulario y el modelo de lenguaje en bigramas y trigramas los cuales son muy útiles para la construcción de la gramática (ver Apéndice D)

Total diálogos	40
Total intervenciones	1613
Intervenciones sistema	841
Intervenciones usuario	772
Prom. de int. por diálogo	40.3

Tabla 5.1 estadísticas de los diálogos recabados

Los datos obtenidos por el CMU SML toolkit nos sirvieron para poder contemplar el vocabulario empleado por los usuarios en este tipo de interacción, y los resultados de los modelos bigrama y trigrama para modelar frases nuevas que presenten indices altos y bajo para estas frases aplicarlas al Parser y observar como se comporta.

Por último la gramática implementada finalmente responde a preguntas que el usuario desea realizar al sistema de acuerdo a ciertos criterios de búsqueda, esta capacidad puede ser aumentada con los ajustes necesarios, debido a que contemplar otro tipo de intervenciones (respuestas mayores, saludos, etc.) implicaría un consumo de tiempo del cual no se dispone sin embargo esta gramática puede servir como caso base, del cual pueden desprenderse trabajos que complementen o refinen la misma.

5.2.3 Ejecutando el Parser con la nueva gramática

Como se habia mencionado en el capítulo anterior el CAT contiene una interfaz que permite probar el desempeño aislado de el Parser, a través de un interfaz. El resultado que se obtiene de ejecutar un análisis de una entrada hecha de manera escrita, se muestra de la siguiente forma:

```
[tonartiuh@Xavier Grammar]$./run_parse

READY

ehhh quisiera ver los horarios disponibles de Ingrid en la
tarde por favor

Pregunta.[Criterio]:HORARIO
Pregunta.[Criterio]:HORARIO.[PARTDIA]:TARDE
Pregunta.[Criterio].[NOMBPROFESOR]
        .[Nombre.Profesor]:INGRID
```

Figura 5.2 Ejemplo de la salida de la interfaz run_parse

5.2.4 Los resultados del proceso de análisis del Parser

La gramática permite ejecutar preguntas y determinarlas como tal, al efectuar una intervención de este tipo el Parser detecta, que el usuario desea hacer una consulta debido a las palabras ”quisiera ver, disponibles, etc. ”, y establece dos criterios de búsqueda de acuerdo con las palabras claves para las redes HORARIO que contempla la palabra “horarios”, PARTDIA para “tarde” y NOMBPROFESOR a “Ingrid”, esta a su vez hace llamado a la subred [Nombre.profesor] que puede contener el valor “Ingrid”.

Es necesario mostrar que en la última parte hablamos de un concepto que implícitamente hace referencia a una subred por que como lo argumenta [Ward 1991] el tamaño del complejo de redes puede distribuirse a través de varias de ellas para tener mayor control de la información, asimismo agregar mayor capacidad de captación de los posibles patrones de frases y de sus tamaños.

Lo cual para este caso si el manejador de diálogo requiere del nombre completo del profesor se puede construir una relación del valor solo de “Ingrid” con “Ingrid Kirshning Albers” lo cual implicaría el agregar mas slots a la subred para contemplar [Apellido.Paterno], [Apellido.Materno] y puedan ser relacionados con cualquier valor dado. Con tiempo disponible todos estos detalles pueden ser cubiertos.

En este caso podemos ver que el Parser esta determinando una pregunta que va implicar una consulta para una Base de Datos, además también se observa que toma como palabras importantes los criterios de búsqueda los cuales aparecen en letras mayúsculas.

La forma en que se presentan estos resultados (HORARIOS, TARDE, INGRID) están estructurados de forma tal que sirvan para que el manejador de diálogo vaya integrando esta información en frames de contexto o bien para determinar una transacción directa a la BD. Algunas de las palabras no las determina por que no son utilizadas para llenar campos en el frame (por ejemplo: ehhh, por favor, etc.), sin embargo otros campos pueden ser implícitos, por ejemplo para el valor TARDE, aún cuando no establece una hora en específico, si

establece un lapso (14:00 a 18:00 por ejemplo) que puede ser manejado por el manejador de diálogo como parámetros de valor, o bien el mismo manejador a través de un proceso de clarificación determina en específico la hora.

En este capítulo se expuso la capacidad que el Parser del CAT tiene para poder soportar una gramática en otro idioma, esto involucro diseñar una gramática que contuviera fenómenos que se presentan en la interacción de un usuario que desea hacer altas de materias académicas.

Con la ayuda de otra herramienta, el CMU Statistical Language Modeling Toolkit pudimos crear una gramática base para formular preguntas al sistema de acuerdo a ciertos criterios, por razones de tiempo esta gramática no es muy extensa, pero se trato de abarcar el mayor numero de casos, en que los usuarios formulan estas consultas. Además gracias al CMU SLM toolkit pudieron obtenerse el vocabulario de los diálogos recabados, su modelo de lenguaje para poder determinar índices de probabilidad de la construcción de ciertas frases y en un momento dado poder aplicarlas al Parser con la nueva gramática.

Por último se da un breve análisis de los resultados que obtiene el Parser de las entradas hechas por el usuario, y como es que estos pueden ser empleados por el manejador de diálogo para crear conceptos o efectuar actividades