	Hacia la construcción de un sistema de Extracción de Información chino como soporte de innovación en los servicios bibliotecarios.
	Zhang Zhixiong, Li Sa, Wu Zhengxin, Lin Ying Biblioteca de la Academia de las Ciencias China Pekín 100080 China Traducido por: Marina Jiménez Piano
Jornadas:	**97 Information Technology with Audiovisual and Multimedia and National Libraries (part 2)**
Traducción simultánea:	No

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL 20-24 August 2006, Seoul, Korea http://www.ifla.org/IV/ifla72/index.htm

Resumen

Siendo conscientes de la importancia de la Extracción de Información (EI) para favorecer la innovación en muchas áreas de los servicios bibliotecarios, los autores construyen un sistema de extracción de información chino para procesar con eficacia los enormes recursos de información chinos. Basados en experiencias y comparaciones de algunos populares sistemas de EI, los autores proponen una solución de EI china, que haga pleno uso del sistema GATE (General Architecture for Text Engineering) de la Universidad de Sheffield, intentando desarrollar un *plug-in* chino para procesar los recursos de información chinos basados en una estructura GATE. Después de más de un año de trabajo, los autores implementaron este sistema.
Este artículo analiza la estructura del sistema GATE, describe la solución de EI (Extracción de información) china basada en ésta, centrándose en tres problemas claves en el proceso de implementación del sistema de extracción de información chino, que son el problema de la segmentación o tokenización[i], los *gazetteers* profesionales y la identificación de las entidades con

[i] Nota del traductor. El *tokenizing* o segmentación en palabras de un texto, no tiene traducción en español, se ha traducido a veces como tokenización

nombres. (1) La segmentación es un problema debido a que la estructura de la lengua china es muy flexible y realizar la segmentación de palabras en chino es muy difícil. Para resolver este problema el software de código libre llamado ICTCLAS (Institute of Computing Technology, Chinese Lexical Análisis systems) de CAS (Academia de las Ciencias china) se integró en el sistema. (2) En el sistema GATE para ayudar a la identificación de las entidades con nombres se proporciona una lista de diccionarios geográficos. Pero la lista proporcionada por GATE está pensada para el reconocimiento de las entidades con nombre ingleses, no chinos. Como base para una buena identificación las entidades chinas con nombres, los autores han recolectado y desarrollado alrededor de 1000 mega bites de *gazetteers* profesionales para ser usados en el sistema GATE. (3) El sistema GATE usa gramáticas JAPE (Java Annotation Patterns Engine) para escribir reglas de reconocimiento de NE (Nombres de entidad). Puesto que la gramática del chino es muy diferente de la del inglés las reglas JAPE que proporciona GATE no son idóneas para los textos chinos. Los autores mantuvieron en uso las gramáticas JAPE y rescribieron unas 100 normas para adaptarlas a la identificación de las entidades con nombres chinas.
También llevaron a cabo un experimento en el cual el sistema de extracción chino recuperó con éxito miles de piezas de novedades de ciencia y tecnología. Los autores creen que este sistema es una prueba significativa y que sienta las bases para futuros trabajos de investigación.

Palabras claves: Extracción de Información china; Biblioteca Digital, Procesamiento del lenguaje natural; GATE; Innovación en los servicios bibliotecarios

1 Introducción

En 2001 la Academia de las Ciencias china (CAS) inició el programa Biblioteca Digital de la Ciencia Nacional china (CSDL)[1] y la biblioteca de la CAS fue una de lasque implementó el CSDL. La misión del CSDL es desarrollar y mantener un entorno digital integrado para los investigadores y los graduados que trabajen en los institutos de investigación CAS del país, proporcionar servicios de información de integrados fiables para ayudar al lector a usar eficientemente recursos de alta calidad.
Después de casi cinco años de funcionamiento, el CSDL llegó a ser uno de los proyectos más notables de biblioteca digital de toda China, con abundantes recursos de información y una amplia gama de servicios de información:

- CSDL proporciona a sus usuarios abundantes recursos de información digital. Incluye recursos como texto completo de revistas STM (scientific, technical, medical), actas de congresos, tesis y disertaciones (ETDs), patentes, obras de referencia y libros electrónicos. En lo que se refiere solo revistas electrónicas, el CSDL cubre en la actualidad más de 6000 revistas STM occidentales, y 10.000 chinas. CSDL establece también un sistema de suministro que permite a sus usuarios obtener artículos de 15000 revistas en el plazo de un día.
- CSDL desarrolla una amplia gama de servicios de información con servicios en red, incluyendo catálogos colectivos, búsquedas federadas en bases de datos, suministro de documentos, referencia digital, personalización de “Mi biblioteca” y autentificación remota.
- Lleva a cabo muchos programas de formación y difusión para ayudar a los profesores y estudiantes a comprender y usar los servicios CSDL

Hoy día el CSDL ha llegado a ser una de las instituciones claves de investigación para los investigadores y graduados de la CAS . Están tan acostumbrados a usar el CSDL que un fallo de sus servicios en red ha llegado a convertirse en el peor desastre para nuestra biblioteca.

Además, con el rápido desarrollo de la ciencia y la tecnología china los requerimientos de los investigadores y graduados han cambiado muy rápidamente. Frente a la enorme cantidad de información académica y de otra información de investigación, los usuarios del CSDL consideran que el uso de los métodos de recuperación de información tradicionales no son suficientes, debido a que el número de documentos recibidos como respuesta a cualquier consulta es enorme. Por tanto necesitan:

- Deshacerse del ruido informativo de modo que puedan identificar correctamente los resultados potencialmente interesantes y localizar certeramente, extraer, recopilar y usar el conocimiento contenido en la bibliografía electrónica disponible.
- Realizar una visual de conjunto efectiva de los desarrollos recientes de sus áreas de interés, incluyendo la realización de resúmenes precisos y personalizados para los investigadores
- Revelar las relaciones significativas de la información, extraer vetas más ricas del material de investigación electrónico y descubrir nuevo conocimiento de la información digital

Desde otro punto de vista, los bibliotecarios del CSDL necesitan también mejorar su servicio normalizado. Además de la recuperación y el suministro de información, los bibliotecarios del CSDL reflexionan acerca de cómo convertir la biblioteca digital en un repositorio de conocimiento, intentando encontrar soluciones adecuadas para hacer buen uso de gran cantidad de bibliografía académica y datos guardados en el CSDL y desarrollar herramientas automáticas para analizar las grandes colecciones textuales.
La Extracción de Información es la tecnología emergente que cubre nuestras necesidades.

2 La EI y su función potencial en la innovación en los servicios en la biblioteca

Desde 2004, el CSDL inició varios proyectos relacionados con el uso de la tecnología de Extracción de Información (E.I.) en el entorno de biblioteca digital, intentando aplicar EI para traer innovación a los servicios bibliotecarios. Desde 2005, tenemos también la colaboración del National Social Sciences Foundation of China (NSSF), centrada en la implementación la extracción de conocimiento de los recursos digitales.

2.1 Extracción de información (EI)

Extracción de Información es un término que ha comenzado a utilizarse para la actividad de extraer automáticamente tipos de información previas especificadas de los textos de lenguaje natural[2]. Sus objetivos son extraer conocimiento estructurado, dependiente del contexto, de la información existente, generalmente texto no estructurado, con el fin de mejorar el uso y la reutilización de esta información. Hamish define la extracción de información como un proceso que toma los textos (y a veces el habla) como entrada y que produce formatos fijos, datos no ambiguos como salida[3]. EI también puede verse como la actividad de poblar una fuente estructurada de información (o base de datos) desde una fuente de información no estructurada o de texto libre. Esta fuente estructurada de información (o base de datos) se usa entonces para otros propósitos: para la búsqueda o el análisis usando consultas convencionales de bases de datos o técnicas de minería de datos; para generar un resumen; para construir índices en los textos fuente.
Iniciativas del gobierno de Estados Unidos como el Message Understanding Conference (MUC)[4], TIPSTER[5] y ACE (Automatic Content Extraction)[6] promueven el desarrollo de tecnología de Extracción de Información y prepara el terreno para la creación de algunos sistemas en curso de Extracción de Información. El programa MUC divide la Extracción de Información en cinco tareas:

- Identificación de entidades con nombre (NE). Encontrar y clasificar nombres, lugares, etc.

- Resolución de co-referencias (CO). Identifica relaciones de identidad entre entidades
- Construcción de la plantilla de elementos (TE). Añade información descriptiva a los resultados NE (usando CO)
- Construcción de la plantilla de relación (TR). Encuentra relaciones entre las entidades TE
- Producción del escenario de plantilla (ST). Encaja los resultados TE y TR en escenarios de eventos específicos.

En términos más simples: NE trata de encontrar entidades; CO sobre las entidades y las referencias (como pronombres) referidas a las mismas cosas; TE sobre los atributos que tienen las entidades; TR sobre qué relaciones existen entre las entidades; ST acerca de los eventos en los que participan las entidades.
La más simple y más fiable tecnología EI es la Identificación de entidades con nombres. El sistema NE identifica todos los nombres de personas, lugares, organizaciones, fechas, cantidades monetarias, etc. En textos científicos y tecnológicos, el reconocimiento de términos de un área única tiene mayor valor. Por supuesto, la identificación de los términos en un texto no es la ayuda esencial: los términos deben relacionarse con el conocimiento existente y/o con otros términos.

2.2 La EI e innovación de los Servicios bibliotecarios

Creemos que la extracción de información jugará un importante papel en el manejo de la enorme colección de información digital y producirá innovaciones en los servicios bibliotecarios. Después de minuciosos estudios, pensamos que EI puede proporcionar ayudas en la anotación automática de los materiales digitales, adquisición automática de metadatos, mejora de la minería de datos en el análisis de información, desarrollo de una base de conocimiento a partir del texto libre y generación de respuestas en los sistemas de referencia digital[7].

Anotación automática y creación de metadatos

La anotación semántica se usa para crear metadatos enlazando el texto con una o más ontologías. Las bibliotecas necesitan anotar la información digital y crear metadatos para permitir una mejor recuperación de la información y potenciar los agentes de recuperación semántica. La mayor parte de la creación actual de metadatos esta basada en la anotación humana, muy a menudo completamente manual. La anotación manual es difícil, exige mucho tiempo y es cara.
Existen varios proyectos de anotación y creación de metadatos automáticos (o semiautomáticos). Por ejemplo MnM[8], S-CREAM (Semi-automatic CREAtion of Metadata)[9] y AERODAML[10] exploran los métodos semiautomáticos para ayudar a los humanos a crear metadatos de los recursos digitales. Mientras que SemTag[11], KIM[12] y hTechsight[13] intentan crear automáticamente metadatos de gran cantidad de volúmenes de texto.
La anotación automática usa técnicas de EI basadasen ontologías. Tomando como ejemplo KIM, es una plataforma que ha sido implementada para la anotación semántica, la indización y los servicios de recuperación. Su objetivo es usar herramientas automáticas masivas de anotación semántica para crear metadatos, algo necesario para que la Web Semántica se haga realidad. Para conseguirlo, KIM reutiliza tecnología de lenguaje humano (HLT), y especialmente tecnología de Extracción de Información (EI). de hecho KIM aplica GATE[14] para construir un sistema de extracción de información semánticamente mejorado para alcanzar su objetivo.

Mejora de la minería de datos en el análisis de información

El análisis de información se está volviendo cada vez más importante para las bibliotecas de investigación. El análisis de datos de larga escala juega un papel cada de creciente importancia en el análisis de información. La detección de muchos tipos de evidencias requiere el reconocimiento y el diseño de inferencias útiles de la información incluida o implícita en una enorme cantidad de datos. Importantes aspectos del análisis de datos incluye la minería de datos (descubrimiento de información relevante en bases de datos). Pero para conseguir datos suficientemente estructurados para el análisis, se debe encontrar un modo de convertir eficientemente texto libre en datos estructurados de formato fijo. La extracción de información, que encuentra patrones estereotipados de información en textos libres o semi-estructurados, puede aportar grandes contribuciones en este tema.

Desarrollo de una base de conocimiento a partir de un texto libre

Una base de conocimiento es útil para el bibliotecario para realizar los servicios bibliotecarios. Para permitir que los científicos lleven a cabo su investigación los bibliotecarios de investigación necesitan mantener algunas bases de datos estadísticas y numéricas, terminológicas y de sucesos o hechos. La Extracción de Información.

Ahora existen varios sistemas que usan Extracción de Información para generar bases de conocimientos. SOBA (SmartWeb Ontology-Based Annotation)[15] es uno de ellos. Se trata de un subsistema del sistema de dialogo multi-modal SmartWeb. SOBA puede llenar automáticamente una base de conocimientos extrayendo información de los informes de partidos de fútbol encontrados en la web. La información extraída es definida con respecto a una ontología subyacente (SWIntO: SmartWeb Integrated Ontology). En SOBA, la extracción de información, las puestas al día de la base de conocimientos y el razonamiento están estrechamente relacionados. También integra información de fuentes heterogéneas (datos semi-estructurados como tablas, texto no estructurado, imágenes y títulos de imágenes) en un nivel semántico en la base de conocimientos.

Generación de respuestas en el sistema de referencia digital

Muchas bibliotecas de investigación establecen un servicio de referencia digital para responder a las preguntas de los lectores en un entorno digital. Casi todos los bibliotecarios referencistas se preocupan de cómo pueden responder las preguntas de las personas que buscan información de forma efectiva y eficiente. Los referencistas necesitan algunos instrumentos útiles de ayuda. ¿Pueden obtener las respuestas directamente de los sistemas de información?

El lenguaje natural QA (Question Answering) es el indicado para estudiar cómo el sistema de información puede generar una respuesta desde una colección de textos potencialmente enorme. Hoy en día muchos investigadores opinan que la extracción de información es muy importante para la generación de respuestas y para la evaluación de éstas[16].

3. Construcción de un Sistema de Extracción de Información Chino

Como podemos ver la extracción de información es muy importante como soporte de la innovación en los servicios bibliotecarios. Así qué ¿cómo construir un sistema de extracción de información que pueda procesar texto chino?. El CSDL intenta conseguir un modo efectivo de crear un sistema de extracción de información adecuado para este fin.

Existen en la actualidad varios sistemas de extracción de información disponibles. Como por ejemplo KEA[17], ANP (Arizona Noun Phraser)[18], TIES (Trainable Information Extration System)[19],

GATE (General Architecture for Text Engineering), etc. Algunos de ellos son software de acceso abierto.
Basándose en experimentos y comparaciones de los sistemas de EI existentes, los autores produjeron un sistema de EI que usase plenamente el sistema GATE (General Architecture for Text Engineering) de la Universidad de Sheffield, intentando desarrollar un *plug-in* de EI chino para procesar los recursos de información chinos basándose en la estructura GATE. Después de más de un año de trabajo, los autores implementaron este sistema.

3. 1 Extracción de Información en GATE

GATE es una arquitectura, un entorno de desarrollo y una estructura para crear sistemas que procesen el lenguaje humano. Como sus desarrolladores declaran GATE es una arquitectura o estructura organizativa para software de procesamiento del lenguaje; un marco o una *class library* que implementa la arquitectura y puede usarse para incrustar las capacidades de procesamiento del lenguaje en diversas aplicaciones; y un entorno de desarrollo construido sobre la estructura realizada con herramientas gráficas adecuadas para desarrollar componentes[20]
En el sistema GATE, todas las cosas están definidas como componentes - la unidad reutilizable y la estructura GATE provee descubrimiento de recursos y facilidades de carga para permitir varios tipos de operaciones de entrada salida. Existen tres tipos de componentes en el sistema GATE:
Recursos lingüísticos (LRs) almacena alguna clases de datos lingüísticos como documentos, corpus, ontologías y proporciona servicios para acceder a ellos.
Recursos de procesamiento (PRs) son recursos cuyo carácter es principalmente programático o algorítmica como un *POS tagger*[ii] o un *parser*[iii].
Recursos visuales (VRs) son componentes gráficos que se visualizan en el interfaz del usuario.

Hay un conjunto de recursos de procesamiento reutilizable proveído con GATE, que forma un sistema de información llamado ANNIE (A Nearly-New IE system). ANNIE consta de los principales de procesamiento de Extracción de Información, como tokenizador, divisor de oraciones o *sentence splitter*, etiquetador de partes del discurso o *POS tagger*, diccionario de nombres propios o *gazetteer*, transductor de estado finito o *finite state transducer* y el *orthomatcher*[iv]

[ii] POS tagger o part-of-speech tagger o etiquetador de parte del discurso
[iii] Procesador o analizador sintáctico
[iv] Tiene como objetivo principal establecer relaciones entre entidades. También tiene un papel secundario en la mejora del reconocimiento de entidades asignando anotaciones a nombres previamente no clasificados basándose en las relaciones con los nombres existentes.

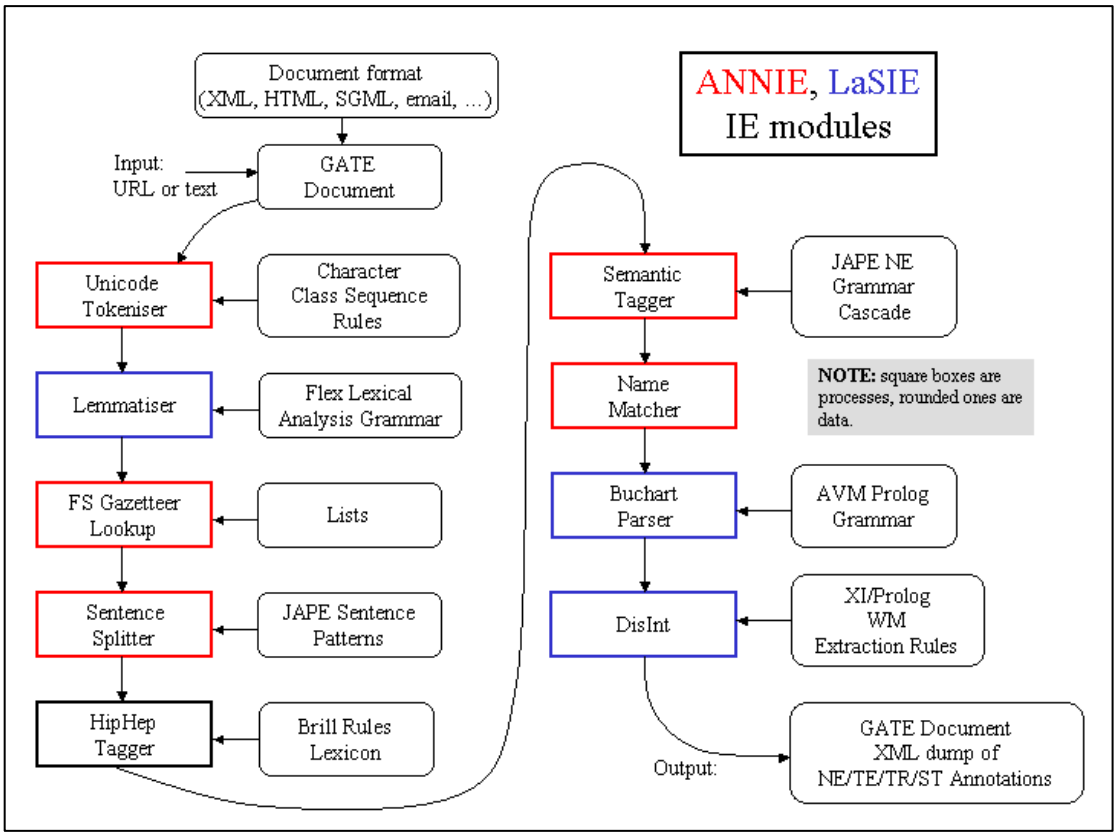

Figura 1:Flujo de ANNIE[21]

También hay muchos recursos de procesos adicionales que no forman parte de ANNIE pero que están incluidos en la instalación por defecto de GATE. Se trata de recolector de *gazetteers*, recursos de procesamiento para el aprendizaje de las máquinas, conjunto de transferencia de anotación, etc.

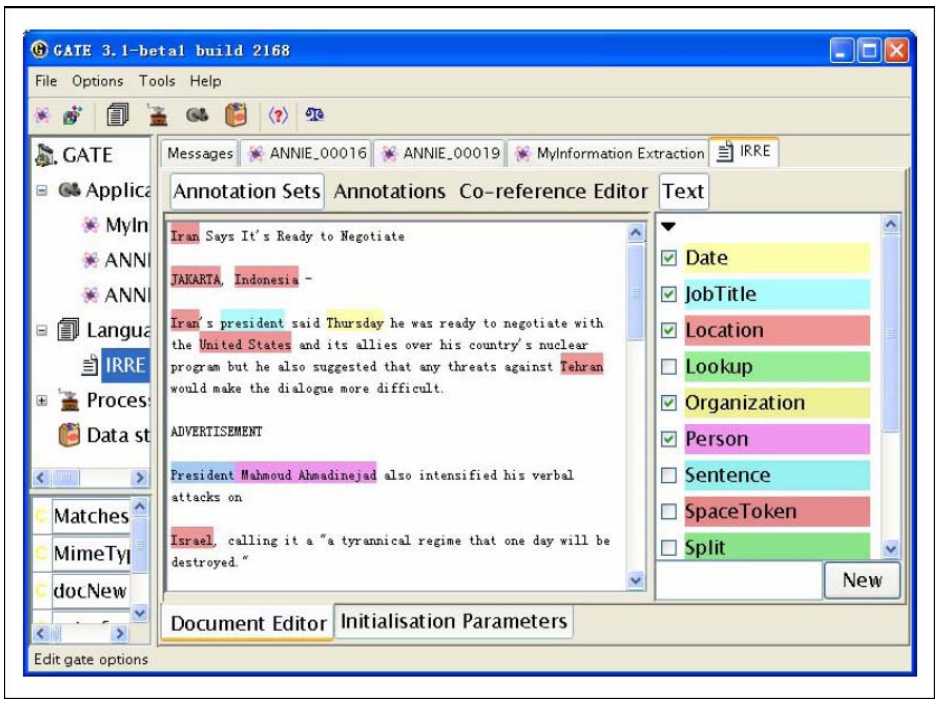

Figura 2. Uso de ANNIE para la Extracción de Información de textos ingleses

ANNIE es adecuado para extraer información de textos ingleses. El tokenizador, el divisor de oraciones o *sentence splitter* y el *orthomatcher* son básicamente independientes del lenguaje del dominio y de la aplicación. En relación con el etiquetador de parte del discurso o *POS tagger*, es dependiente del la lengua pero independiente del dominio y de la aplicación. Generalmente una aplicación nueva puede usar directamente la mayor parte de los componentes principales de ANNIE (ver figura 2) para extraer identificación de entidades con nombre, como fechas, nombre, nombre del trabajo, organización, etc. Pero si se desea realizar una extracción más compleja, por ejemplo extraer términos de un determinado dominio es necesario modificar el *gazetteer* y rescribir las gramáticas JAPE. A veces se requieren también recursos de procesamiento (PRs) adicionales. GATE es una infraestructura basada en UNICODE y, según declara sus desarrolladores, soporta Extracción de Información multilingüe. Por lo tanto, GATE puede procesar texto chino. De hecho el conjunto estándar GATE siempre incluye algunos recursos (como listas de nombres, gramáticas, etiquetador, tokenizador, segmentador) como soporte de la Extracción de Información de textos chinos. pero sus resultados con los textos chinos no son tan buenos como con los ingleses (ver Figura 3)

Figura 3. La extracción de información china de GATE no es muy buena (No disponible)

3.2 Dificultades clave para la extracción de información china basada en GATE

Tras un estudio detallado del sistema GATE y de la naturaleza de la lengua china, los autores llegaron a la conclusión de que hay tres problemas claves que es necesario resolver para mejorar los resultados de la Extracción de Información china en el sistema GATE:

La tokenización china

Con objeto de realizar ésta, la aplicación necesita conocer donde están las palabras en una frase de texto. Para muchas lenguas, esto es una tarea relativamente fácil: las palabras se separan con espacios blancos y puntuación. Por el contrario, el chino se escribe sin ninguna separación entre palabras. Los espacios blancos se usan poco o nada. No se puede encontrar ningún espacio entre cada carácter. por ello, una de las áreas de investigación para el procesamiento de la lengua china es el de la segmentación de palabras en chino, dada una frase sin espacios ni ruptura en las palabras. Debido a que la estructura de la lengua china es muy flexible, realizar su segmentación es muy difícil.
Podemos examinar una frase simple:

我是中国人
(Yo soy chino)

Puede separarse de varias formas con el segmentador

我 是 中国人
我 是 中国 人
我 是 中 国 人

…….

Así pues, podemos ver que no es una tarea fácil romper una oración en la forma correcta. El conjunto estándar GATE no realiza la segmentación de palabras chinas. Entre sus *plug-in*, GATE proporciona un segmentador, pero en nuestra opinión se necesita otro mejor.

***Gazetteers* chinos**

En el sistema GATE, para ayudar a la identificación de entidades con nombre se proporcionan un conjunto de listas de *gazetteers* . Estas listas proporcionadas por GATE son muy abundantes. Las listas del *plug-in* chino toma forma de listas de *gazetter* de ANNIE, pero es más corta y más simple. De hecho, el sistema GATE proporciona algunas listas simples, como fecha, hora, organización, localización, moneda, provincia, etc. Pero para una lengua tan flexible como el chino, la lista es muy limitada.

Para establecer una buena base para la identificación de entidades nombradas necesitamos enriquecer las listas *gazetteers.*

Identificación de entidades con nombre chinas

El sistema GATE usa gramáticas JAPE (Java Annotation Patterns Engine) para escribir normas de reconocimiento NE (entidades nombradas). Un etiquetador semántico de GATE consiste en un conjunto de gramáticas JAPE basadas en normas en curso secuencialmente. Las gramáticas contienen normas de modelos - acciones escritas a mano que reconocen por ejemplo anotaciones del *POS tagger* y del *gazetteer* y las combina para producir nuevas anotaciones NE sobre modelos. JAPE es un lenguaje de comparación con modelos. El LHS (Left Hand Side) de cada norma contiene modelos que deben cumplirse y el RHS (Right Hand Side) contiene detalles de anotaciones (y excepcionalmente características) para ser creadas. por ejemplo, una regla puede reconocer un nombre de pila (del módulo del *gazetteer*) seguido por un nombre propio (del *POS tagger*) y anotar este modelo como una persona. Esta regla se escribirá en JAPE como sigue:

```
Rule:Person1
(
{Lookup.majorType ==firstname}
{Token.category ==NNP}
):lable
-->
:lable.Person ={rule =”Person1 ”}
```

Debido a que la gramática del chino es muy diferente de la del inglés, las reglas JAPE proporcionadas por GATE no son adecuadas para los textos chinos. Necesitamos rescribir las reglas JAPE para implementar la extracción de información china.

3.3 Solución de los problemas.

En el proceso de implementación del sistema de Extracción de Información Chino basado en GATE , necesitamos resolver los tres problemas claves que hemos mencionado más arriba. Después de planificar cuidadosamente se consiguió una solución total de los problemas (figura 4)

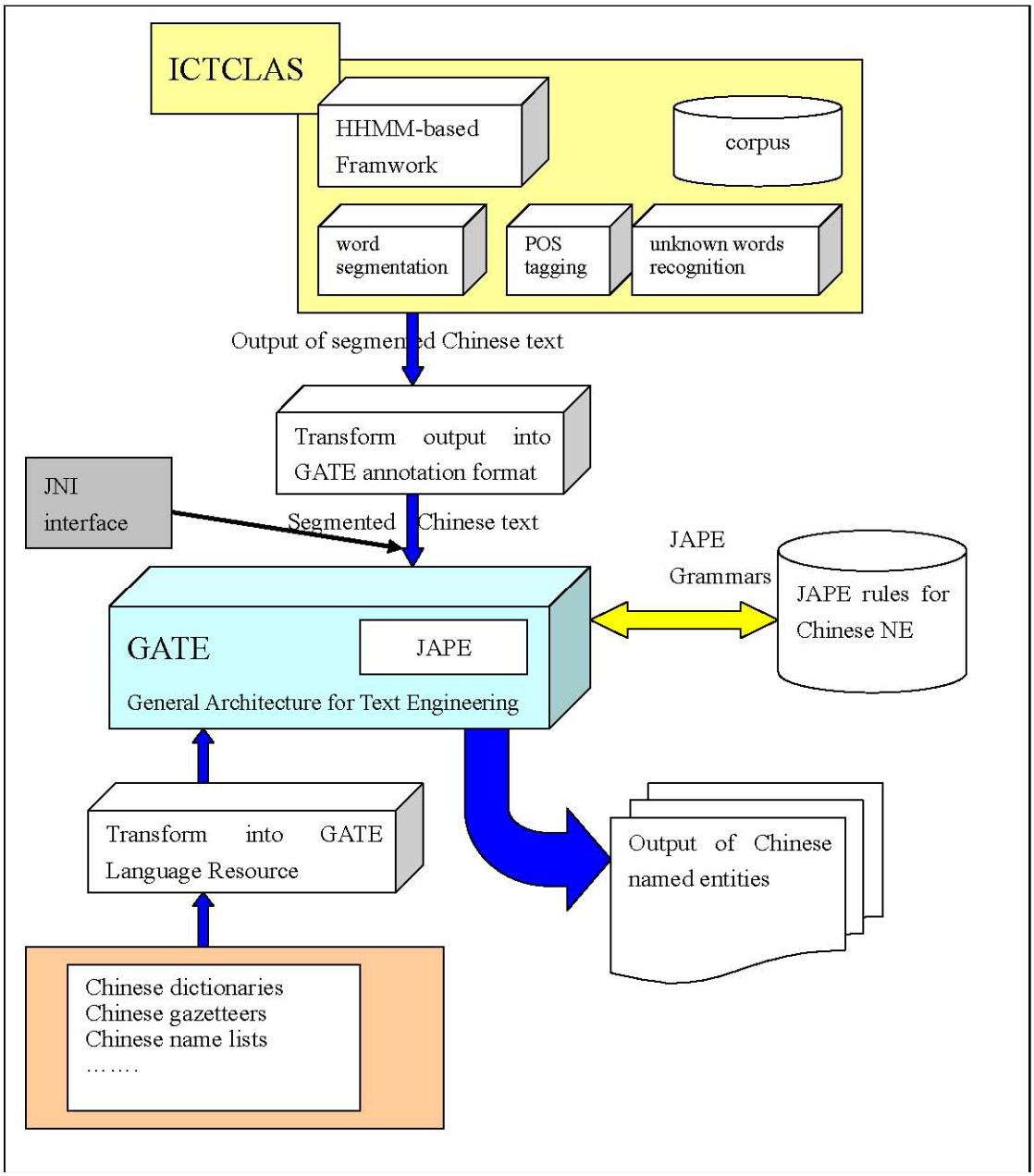

Figura 4.La solución total al sistema de Extracción de Información Chino basado en GATE.

Para implementar esta solución, hemos llevado a cabo tres tareas principales.

Integración de ICTCLAS para realizar la segmentación de palabras
El ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) es un sistema de análisis léxico chino de código abierto desarrollado por el Institute of Computing Technology dela Academia China de las Ciencias. Este usa un esquema basado en HMM multicapa, que incluye segmentación de palabras, etiquetación de parte del discurso y reconocimiento de palabras desconocidas. Su precisión en la segmentación es del 97,58%. El porcentaje de reconocimiento de palabras desconocidas usando el etiquetado defunciones alcanza más del 90%.en concreto el reconocimiento de nombres de persona chino alcanza casi un 98%. La velocidad de segmentación de palabras y de etiquetado de parte del discurso es de 31,5KB/s.
ICTCLAS es una buena elección para mejorar la segmentación de palabras chinas del sistema GATE. Pero para integrarlo en GATE debimos realizar algunos desarrollos. Porque ICTCLAS está escrito en lenguaje C/C++, y GATE en JAVA puro. Para invocar las bibliotecas de enlaces dinámicos de ICTCLAS en el sistema GATE usamos Java Native Interface (JNI) del Java Develpment Kit (JDK) para resolver el problema. En la figura 5, la parte dela izquierda es el texto chino original y la parte de la derecha es la salida ICTCLAS invocada por GATE. Como GATE soporta UNICODE, el texto con palabras separadas por espacios blancos es adecuado para que lo tome GATE como entrada.

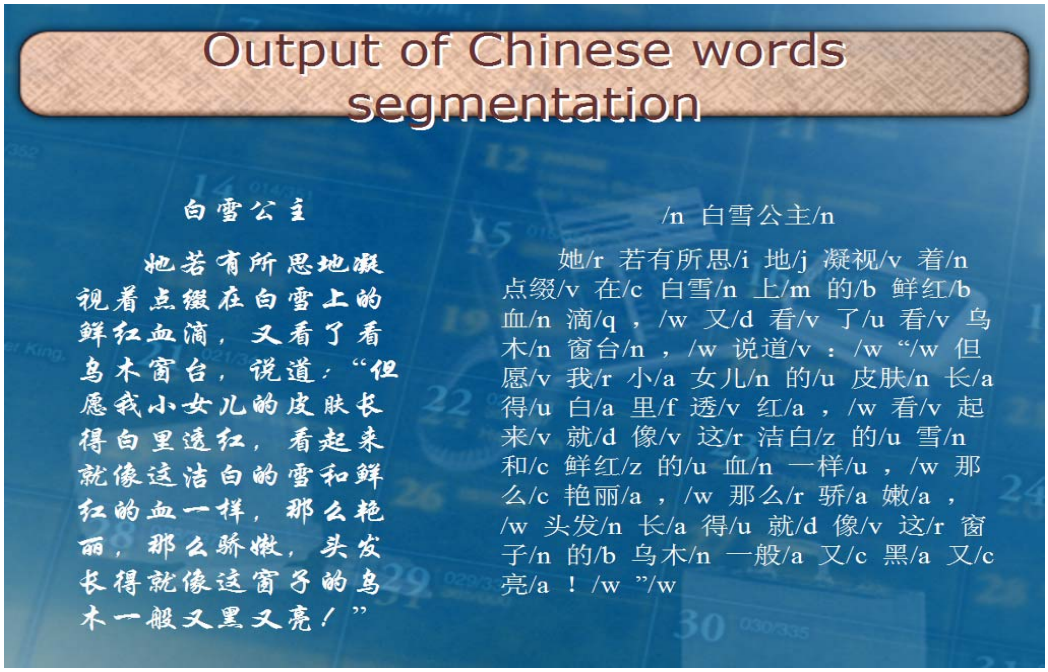

Figura 5. Salida de ICTCLAS invocada por GATE

Desarrollo de *gazetteers* chinos para enriquecer los recursos lingüísticos de GATE
En el sistema GATE las listas de *gazetteers* son ficheros de texto plano con una entrada por línea. Cada lista representa un conjunto de nombres, como nombres de ciudades, de organizaciones, días de la semana, etc. Se usa un fichero de índice (list.def) para acceder a estas listas; para cada lista se especifica un tipo mayor y opcionalmente un tipo menor.

city_china.lst:location:city
city_world.lst:location:city
company.lst:organization:company
company_CHN.lst:organization:company_CHN

Durante el proceso de desarrollo de *gazetteers* chinos hemos usado plenamente de las características de bibliotecas, como diccionarios completos, listas de nombres, *gazetteers*, etc. También recolectamos algunos recursos de Internet. En la actualidad hemos acumulado alrededor de 100MB de *gazetters* chinos indicados para la extracción de Información basada en dominios. La siguiente tabla muestra los *gazetteers* que hemos usado para la identificación de las entidades con nombre chinas.

Tabla 1. *Gazetters* chinos que preparamos para la identificación de entidades con nombre chinos

Gazetteers	**Número de entradas**
Nombre de organización	2100
Nombre de ciudad china	1309
Nombre de ciudad mundial	140
Nombre de compañía exterior	124
Nombre de compañía china	435
Nombre de compañía de media	147
Nombre de país	222
Nombre de región	2189
Universidad china	1003
Nombres de recursos	331
Nombres femeninos	2416
Nombres de institutos	2100
Nombres masculinos	2654
Palabras clave de organizaciones	912

Reescritura de las reglas JAPE para la identificación de las NE chinas

Debido a que la gramática china es muy diferente de la inglesa, necesitamos rescribir las reglas JAPE para adecuar GATE al procesamiento de los textos chinos.

A continuación listamos una comparación sencilla. Por ejemplo la regla JAPE reconoce la hora en inglés, como “10 o’clock”.

```
Rule:TimeOClock
//Recognizing English time like “10 o'clock ”.
(
{Lookup.minorType ==hour}//look into the hour.lst to find 1,2,3 etc.
{Token.string =="o"}
{Token.string =="'"}
{Token.string =="clock"}
)
:time
```

-->

:time.TempTime ={kind ="positive",rule ="TimeOClock"}

En chino, “10 o’clock” es: . Luego la regla JAPE debe rescribirse como vemos a continuación para que GATE pueda extraer la expresión de tiempo china

```
Rule: TimeOClock_cn1
// Recogrizing Chinese time like “10 点钟”or”一点钟” (
(
{Lookup.minorType == hour}//In ”hour.lst” there exist Arabic number likes 1，2，3 and Chinese
number like “一”, “二”,“三”
{Token.string == "点钟"})
)
:time
-->
:time.TempTime = {kind = "positive", rule = "TimeOClock_cn2"}
```

En total hemos reescrito alrededor de 100 reglas JAPE para hacer más precisa la identificación de entidades con nombre chinas.

4. Tests y evaluación

Después de más de un año de trabajo implementamos el sistema. También llevamos a cabo un experimento en el que el sistema chino de EI extrajo con éxito miles de piezas de noticias de ciencia y tecnología.

La figura 6 muestra los resultados de la Extracción de Información china usando el sistema que hemos desarrollado. Comparada con la figura 3 que insertamos anteriormente, puede verse que muchas entidades con nombre que no podían ser reconocidas en el conjunto estándar GATE pueden visualizarse ahora.

Figura 6. Resultados de la extracción de información china usando el sistema que hemos desarrollado

5. Conclusiones

A pesar de que hemos realizado con éxito una experimentación y una buena evaluación de nuestro sistema chino de extracción de información, hay todavía muchas tareas que necesitamos cumplir para un uso más eficiente. Pensamos que es un punto de partida significativo en la aplicación de tecnología de extracción de información en nuestras bibliotecas y que supone una buena base para nuestras futuras investigaciones como soporte de innovación en los servicios bibliotecarios. Ahora vamos a hacer una propuesta para intentar integrar el sistema de extracción de información chino en otros sistemas de servicios bibliotecarios, que sirven a los usuarios, como ayuda en la anotación automática de materiales digitales, adquisición automática de metadatos, etc. También llevaremos a cabo más experimentos de desarrollo y localización de software internacional y uso del código abierto para promover los servicios bibliotecarios.

Bibliografía

[1] CSDL, Chinese national Science Digital Library, http:://www.csdl.ac.cn/ [consultado el 8 de mayo de 2006]

[2] Natural Language Processing Research Group en la Universidad de Sheffield, Extracción de Información, http://nlp.shef.ac.uk/research/areas/ie.html [consultado el 8 de mayo de 2006]

[3] Hamish Cunningham,Information Extraction,Automatic,*Encyclopedia of Language & Linguistics*, 2 edición, 2005, http://gate.ac.uk/sale/ell2/ie/main.pdf [Consultado el 8 de mayo de 2006]

[4] NIST, MUC, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html [Consultado el 8 de mayo de 2006]

[5] NIST, TIPSTER Text Program, http://www-nlpir.nist.gov/related_projects/tipster/ [consultado el 8 de mayo de 2006]

[6] NIST,ACE -Automatic Content Extraction,http://www.nist.gov/speech/tests/ace/ [Consultado el 8 mayo de2006] Zhang Zhixiong, Information Extraction and its Functions in the Digital Library, *New Technology of Library and Information Service*, 2004(6):1-5, 23

[6] The Open University, MnM, http://kmi.open.ac.uk/projects/akt/MnM/ [Consultado el 8 de mayo de 2006]

[6] Siegfried Handschuh, Steffen Staab y Fabio Ciravegna, S-CREAM: Semi-automatic CREAtion of Metadata, En Proc.of the European Conference on Knowledge Acquisition and Management -EKAW-2002. Madrid, Octuber 1-4, 2002. Springer, 2002
http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/ekaw2002scream-sub.pdf
[consultado el 8 de mayo de 2006]

[7] Zhang Zhixiong, Information Extraction and its Functions in the Digital Library, *New Technology of Library and Information Service*, 2004(6):1-5, 23

[8] The Open University, MnM, http://kmi.open.ac.uk/projects/akt/MnM/ [consultado el 8 de mayo de 2006]

[9] Siegfried Handschuh,Steffen Staab and Fabio Ciravegna,S-CREAM:Semi-automatic CREAtion of Metadata,In Proc.of the European Conference on Knowledge Acquisition and

Management -EKAW-2002.Madrid,Spain,October 1-4,2002.Springer,2002

http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/ekaw2002scream-sub.pdf
[consultado el 8 de mayo de 2006]

[10] Paul Kogut, William Holmes: AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages, K-CAP 2001 Workshop Knowledge Markup &Semantic Annotation, Octubre 21,2001, Victoria B.C., Canada
http://semannot2001.aifb.uni-karlsruhe.de/positionpapers/AeroDAML3.pdf
[Consultado el 8 de mayo de 8,2006]

[11] Stephen Dill etc. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation, The Twelfth International World Wide Web Conference 20-24 May 2003, Budapest,HUNGARY. http://www2003.org/cdrom/papers/refereed/p831/p831-dill.html [Consultado el 8 de mayo de 2006]

[12] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, Damyan Ognyanoff: Semantic Annotation, Indexing, and Retrieval, *Elsevier's Journal of Web Sematics*,Vol.2, Issue (1), 2005. http://www.websemanticsjournal.org/ps/pub/2005-10 [consultado el 8 de mayo 2006]

[13] Project hTechsight, http://www.etse.urv.es/~drianyo/hTechSight/projecte.html [Consultado el 8 de mayo de 2006]

[14] GATE, A General Architecture for Text Engineering, http://gate.ac.uk/ [Consultado el 8 de mayo de 2006]

[15] Paul Buitelaar, Philipp Cimiano, Stefania Racioppa, Melanie Siegel, Ontology-based Information Extraction with SOBA, http://www.dfki.de/~paulb/lrec2006.SmartWeb.pdf [Consultado el 8 de mayo de 2006]

[16] Rohini Srihari and Wei Li, Information Extraction Supported Question Answering, http://trec.nist.gov/pubs/trec8/papers/cymfony.pdf [Consultado el 8 de mayo de 2006]

[17] KEA project, http://www.nzdl.org/Kea/ [Consultado el 8 de mayo de 2006]

[18] ANP (Arizona Noun Phraser), http://ai.bpa.arizona.edu/research/multilingual/az.htm [Consultado el 8 de mayo de 2006]

[19] TIES (Trainable Information Extraction System), http://tcc.itc.it/research/textec/tools-resources/ties.html [Consultado el 8 de Mayo de 2006]

[20] Hamish Cunningham etc. Developing Language Processing Components with GATE Version 3 (a User Guide), http://gate.ac.uk/sale/tao/index.html [Consultado el 8 de mayo de 2006]

[21] Diana Maynard,Introduction to ANNIE,March 2004, http://gate.ac.uk/sale/talks/annie-tutorial.ppt [Consultado el 8 de mayo de 2006]