Date : 25/05/2006

Towards Constructing a Chinese Information Extraction System to Support Innovations in Library Services

Zhang Zhixiong, Li Sa, Wu Zhengxin, Lin Ying
The library of Chinese Academy of Sciences
Beijing 100080
China

Meeting:	**97 Information Technology with Audiovisual and Multimedia and National Libraries (part 2)**
Simultaneous Interpretation:	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

Abstract

Being aware of the importance of Information Extraction (IE) in supporting innovation in many areas of library services, the authors begin to construct a Chinese information extraction system to effectively process huge Chinese information resources. Based on experiments and comparisons of some popular IE systems, the authors bring forth a Chinese IE solution which makes full use of GATE (General Architecture for Text Engineering) system from University of Sheffield, trying to develop a Chinese IE plug-in to process Chinese information resource based on GATE framework. After more than one years of working, the authors implemented this system.

The article here analyses the framework of GATE system, describes the Chinese IE solution based on the GATE system, focuses on three key difficulties in the process of implementing Chinese information extraction system, which are Chinese tokenizing problem, professional gazetteers and Chinese named entity recognition. (1) Chinese tokenizing is a problem because language structure of Chinese is very flexible and performing word segmentation of Chinese language is very difficult. To solve this problem, the open source software named ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) from CAS is integrated into this system. (2) In GATE system, to aid named entity recognition, a set of gazetteer lists is provided. But the gazetteer lists *provided by GATE is just for English named entity recognition, not for Chinese. To lay a good foundation for Chinese named entity recognition, the authors have collected and developed over 100 Megabyte of Chinese professional gazetteers that can be used by GATE system. (3) GATE system use JAPE (a Java Annotation Patterns Engine) grammars to write rules to recognize NE (named entity). Because grammar of Chinese is quite different from that of English, the JAPE rules provided by GATE are not suitable for Chinese texts. The authors keep in use JAPE grammars and rewrite over one hundred of JAPE rules to succeed in the Chinese named entity recognition.*

The authors also carry out an experiment in which the Chinese IE system successfully extracts thousands of pieces of science and technology news. The authors believe this system is a significant trial and lays a good foundation for the future research works.

Keywords: *Chinese Information Extraction; Digital Library; Natural Language Processing; GATE; Innovations in Library Services*

1 Introduction

In 2001,Chinese Academy of Sciences (CAS) initiated Chinese National Science Digital Library (CSDL) program[1] and the Library of CAS became the one who implement the CSDL. The mission of CSDL is to develop and maintain a integrated digital information environment for researcher and graduated students working in CAS research institutes across the country, to provide reliable one-stop information services to help reader use high quality resources efficiently.

After nearly 5 years of development, CSDL became one of the most noticeable digital library projects all over China, which have abundant information resources and a wide range of information services:

- CSDL provided abundant digital information resources for their users. Including resources like full text STM journals, conference proceedings, theses and dissertations (ETDs), patents, reference books, and e-books. For e-journals alone, CSDL now covers more than 6000 core western STM journals and 10000 Chinese ones. CSDL also set up a supply chain system, which let user of CSDL could get the document from 15000 journals in one day.
- CSDL developed a wide range of information systems to support networked services including union catalogs, federated database search, document delivery, digital reference, MyLibrary customization, and remote authentication.
- Carried out lots of training and propaganda program to help researchers and students understand and use the services of CSDL.

Now, CSDL become one of the key research facility to researcher and graduated students of CAS. They got so used to using CSDL that failure of network services of CSDL are now become the worst disasters for our library.

With rapid development of Chinese science and technology, the information requirement of researcher and graduated students of CAS also changed rapidly. Facing vast outpouring of academic literature and other research information, the users of CSDL find using traditional information retrieval methods is not sufficient because the number of documents returned in response to a query is huge. They want to:

- Get rid of the information noise so that they can efficiently identify potentially interesting features and accurately locate, extract, gather and make use of knowledge encoded in electronically available literature.
- Effectively get a comprehensive view of recent development of domain related to them, including drawing up precise and tailored summaries personalized to the researchers.
- Disclose significant relationships between information, excavate richer seams of electronic research material and discover new knowledge from digital information.

From another point of view, the librarians of CSDL also want to improve the service standard of CSDL. In addition to information retrieval and information delivery, the librarians of CSDL think about how to turn the digital library into a knowledge repository, try to find suitable solutions to make good use of the vast amounts of academic literature and data held in CSDL and develop automatic tools for analyzing large textual collections.

Information Extraction (IE) is the emerging technology serves to our needs.

2 IE and Its Potential function in Innovations in Library Services

From 2004, CSDL initiates several projects involved in using Information Extraction (IE) technology in digital library environment, trying to apply IE to bring innovation in library services. From 2005, we also got support from National Social Sciences Foundation of China (NSSF), focusing on implementing knowledge extraction from digital resources.

2.1 Information Extraction (IE)

Information extraction (IE) is a term that has come to be applied to the activity of automatically extracting pre-specified sorts of information from natural language texts[2]. Its aim is to extract structured, contextually-dependant knowledge from existing information, typically unstructured text, in order to enhance the use and reuse of that information. Hamish defines information extraction as a process that takes texts (and sometimes speech) as input and produces fixed-format, unambiguous data as output[3]. IE also can be seen as the activity of populating a structured information source (or database) from an unstructured, or free text, information source. This structured information source (or database) is then used for some other purposes: for searching or analysis using conventional database queries or data-mining techniques; for generating a summary; for constructing indices into the source texts.

US Government initiatives such as the Message Understanding Conference (MUC)[4], TIPSTER[5], and ACE (Automatic Content Extraction) [6]promote the development of Information Extraction technology and pave the way for the creation of many current Information Extraction systems. MUC program split Information Extraction into five tasks:

- Named Entity recognition (NE). Finds and classifies names, places, etc.
- Coreference resolution (CO). Identifies identity relations between entities.
- Template Element construction (TE). Adds descriptive information to NE results (using CO).
- Template Relation construction (TR). Finds relations between TE entities.
- Scenario Template production (ST). Fits TE and TR results into specified event scenarios.

In simpler terms: NE is about finding entities; CO about which entities and references (such as pronouns) refer to the same thing; TE about what attributes entities have; TR about what relationships between entities there are; ST about events that the entities participate in.

The simplest and most reliable IE technology is Named Entity recognition. NE systems identify all the names of people, places, organizations, dates, amounts of money, etc. In scientific and technological text, recognition of term in one domain has more value. Of course, recognition of terms in text is not the ultimate aim: terms should be also related to existing knowledge and/or to each other.

2.2 IE and Innovations in Library Services

We believe information extraction will play a very important role in coping with the huge collections of digital information and bring innovations in library services. After carefully study, we find IE can provide helps in automatic annotation of digital materials, automatic acquisition of metadata, improving data mining in information analysis, developing knowledge base from free text, and generating answers in digital reference system[7].

Automatic annotation and metadata creation

Semantic annotation is used to create metadata linking the text to one or more ontologies. Libraries need to annotate digital information and create metadata to both enable better information retrieval and empower semantically aware agents. Most of the current metadata creation is based on human centered annotation, very often completely manual. Manual annotation is difficult, time consuming and expensive.

There are several projects on automatic (or semi-automatic) annotation and metadata creation. For example, MnM[8], S-CREAM (Semi-automatic CREAtion of Metadata)[9], and AERODAML[10] explore semi-automatic methods to help human create metadata from digital resources. While SemTag[11], KIM[12] and hTechsight[13], try to automatically create metadata from large volumes of text.

Automatic annotation generally relies on ontology-based IE techniques. Take KIM for example, KIM is a platform has been implemented for semantic annotation, indexing, and retrieval services. Its aim is to use massive automatic semantic annotation tools to create metadata, which is needed for the Semantic Web to happen. In order to achieve this, KIM reuses existing human language technology (HLT), and especially Information Extraction (IE) technology. In fact KIM apply GATE[14] to build a semantically enhanced information extraction system to reach the goal.

Improving data mining in information analysis

Information analysis is becoming more and more important for research library. Large-scale data analysis plays an increasingly important role in information analysis. Detection of many types of evidence requires recognizing and drawing useful inferences from information embedded or implicit in huge quantities of data. Important aspects of data analysis include data mining (discovering relevant information in databases). But in order to get enough structured data for analysis, one should find a way to effectively turn free texts into structured, fixed-format data. Information extraction, which finds stereotypical patterns of information in free or semi-structured text, can make great contribution to this.

Developing knowledge base from free text

A knowledge base is helpful for librarian to carry out information services. In order to support scientists to carry out their research, some statistical and numeric databases, terminological database, and fact sheets are needed to setup by research librarian. Information Extraction can help librarian build knowledge base from free text.

Now, there exist several systems using Information Extraction to generate knowledge base. SOBA (SmartWeb Ontology-Based Annotation)[15] is one of them. It is a sub-component of the SmartWeb multi-modal dialog system. SOBA can automatically populate a knowledge base by extracting information from soccer match reports found on the web. The extracted information is defined with respect to an underlying ontology (SWIntO: SmartWeb Integrated Ontology). In SOBA, information extraction, knowledge base updates and reasoning are tightly interleaved. It also integrates information from heterogeneous sources (semi-structured data such as tables, unstructured text, images and image captions) on a semantic level in the knowledge base.

Generating answers in digital reference system

Most research libraries establish digital reference service to answer reader's questions in a digital environment. Almost every reference librarian cares about how he (she) can answer information seekers' questions effectively and efficiently. Reference librarians need some useful tools for assisting them. Can they get answers directly from information systems?

Natural language QA (Question Answering) is the right one to study how information system can generate an answer from a potentially huge collection of natural language texts. Many researchers now believe information extraction is very important for generation answer and carry out many tests to prove it[16].

3. Constructing a Chinese Information Extraction System

As we can see information extraction is very important to support innovation in library services. Then how to build an information extraction system that can process Chinese text? CSDL try to find an effective way to build an information extraction system suitable for its use.

There are now several information extraction systems available. Such as KEA[17], ANP (Arizona Noun Phraser)[18], TIES (Trainable Information Extraction System)[19], GATE (General Architecture for Text Engineering) etc. Some of them are open source software.

Based on experiments and comparisons of current IE systems, the authors bring forth a Chinese IE solution which makes full use of GATE (General Architecture for Text Engineering) system from University of Sheffield, trying to develop a Chinese IE plug-in to process Chinese information resource based on GATE framework. After more than one years of working, the authors implemented this system.

3.1 Information Extraction in GATE

GATE is an architecture, development environment and framework for building systems that process human language. As its developer put it, GATE is an architecture, or organizational structure, for language processing software; a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in diverse applications; and a development environment built on top of the framework made up of convenient graphical tools for developing components[20].

In GATE system, every thing is defined as components – the reusable unit and the GATE framework provides resource discovery and loading facilities to supports various kinds of input output operations. There are three kinds of components in GATE system:

- Language Resources (LRs) store some kind of linguistic data such as documents, corpora, ontologies and provide services for accessing it.
- Processing Resources (PRs) are resources whose character is principally programatic or algorithmic such as a POS tagger or a parser.
- Visual Resources (VRs) are graphical components that are displayed by the user interface.

There is a set of reusable processing resources provided with GATE, which forms an information system named ANNIE (A Nearly-New IE system). ANNIE consists of the main processing resources for Information Extraction such as: tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer and orthomatcher.

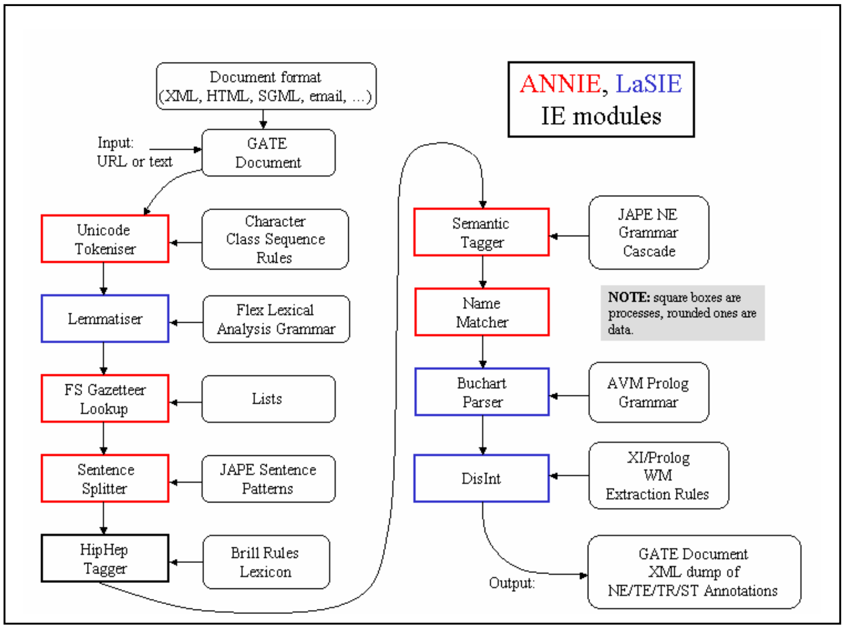
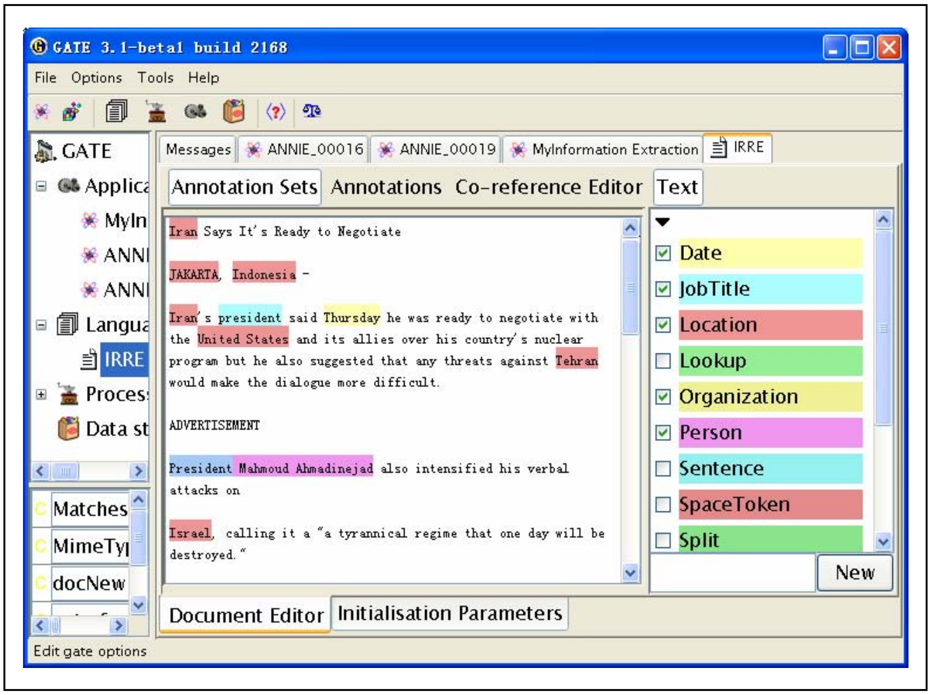


Figure 1. Workflow of ANNIE[21]

There are also lots of additional processing resources which are not part of ANNIE itself but which come with the default installation of GATE. Such as gazetteer collector, processing resources for machine learning, various exporters, annotation set transfer etc.



Iran Says It's Ready to Negotiate

JAKARTA, Indonesia -

Iran's president said Thursday he was ready to negotiate with the United States and its allies over his country's nuclear program but he also suggested that any threats against Tehran would make the dialogue more difficult.

ADVERTISEMENT

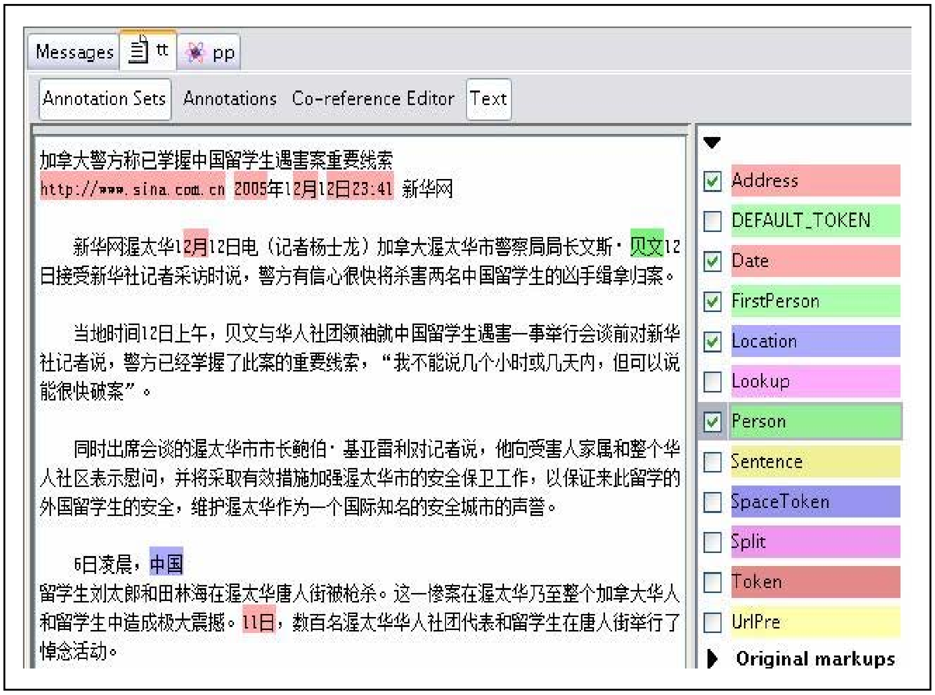
President Mahmoud Ahmadinejad also intensified his verbal attacks on

Israel, calling it a "a tyrannical regime that one day will be destroyed."

Figure 2, Using ANNIE for information extraction from English text

ANNIE is suitable for extracting information from English text. The tokeniser, sentence splitter and orthomatcher of ANNIE are basically language, domain and application-independent. While the POS tagger is language dependent but domain and application-independent. Typically a new application can directly use most of the core components from ANNIE (see figure2) to extract basic named entities such as date, name, job title, organization etc. But if you want to do more complex extraction, for example extracting term from a domain, you need to modify the gazetteer lists and rewrite JAPE grammars. Sometime, you may also require additional PRs.

GATE is a unicode-based infrastructure and, as the developer said, supports multilingual Information Extraction. We find GATE can process Chinese text. In fact, the standard GATE suite already includes some resources (such as gazetteer lists, grammar, tagger, tokeniser, segmenter) to support information extraction from Chinese text. But its performance with Chinese text is not so good as with English one (see figure3).



加拿大警方称已学握中国留学生遇害案重要线索
http://www.sina.com.cn 2005年12月12日23:41 新华网

新华网渥太华12月12日电（记者杨士龙）加拿大渥太华市警察局局长文斯·贝文12日接受新华社记者采访时说，警方有信心很快将杀害两名中国留学生的凶手缉拿归案。

当地时间12日上午，贝文与华人社团领袖就中国留学生遇害一事举行会谈前对新华社记者说，警方已经掌握了此案的重要线索，“我不能说几个小时或几天内，但可以说能很快破案”。

同时出席会谈的渥太华市市长鲍伯·基亚雷利对记者说，他向受害人家属和整个华人社区表示慰问，并将采取有效措施加强渥太华市的安全保卫工作，以保证来此留学的外国留学生的安全，维护渥太华作为一个国际知名的安全城市的声誉。

6日凌晨，中国
留学生刘太郎和田林海在渥太华唐人街被枪杀。这一惨案在渥太华乃至整个加拿大华人和留学生中造成极大震撼。11日，数百名渥太华华人社团代表和留学生在唐人街举行了悼念活动。

☑ Address
□ DEFAULT_TOKEN
☑ Date
☑ FirstPerson
☑ Location
□ Lookup
☑ Person
□ Sentence
□ SpaceToken
□ Split
□ Token
□ UrlPre
Original markups

Figure 3. Chinese information extraction from standard GATE suite is not so good

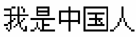
3.2 Key difficulties for Chinese information extraction based on GATE

After carefully study GATE system and the nature of Chinese language, the authors figure out there are three key problems need to be solved to improve the performance of Chinese Information Extraction in GATE system:

Chinese tokenizing

In order to perform tokenizing, the applications need to know where the words are in a sentence of text. For many languages, this is a relatively “easy” task: words are separated by white space and punctuation. Chinese, in comparison, is written without any separation between words. White space serves little or no purpose. You cannot find any spaces between every character at all. So one of the research areas for Chinese language processing is to perform Chinese words segmentation, taking a sentence with no spaces and breaking it into words. Because language structure of Chinese is very flexible , so performing word segmentation of Chinese language is very difficult.

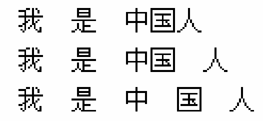
We can look at a simple sentence.



我是中国人

(I am a Chinese)

It can be broken into several forms with segmenter:



我 是 中国人
我 是 中国 人
我 是 中 国 人

……

So we can see it is not an easy task to correctly break a sentence into right form. Standard GATE suite do not perform Chinese word segmentation. In it’s plug-in, GATE provides a segmenter, but we think a better one is needed.

Chinese gazetteers

In GATE system, to aid named entity recognition, a set of gazetteer lists is provided. The gazetteer lists provided by GATE for English named entity recognition is very abundant. The gazetteer lists of Chinese plug-in take the form of the gazetteer lists of ANNIE, but it is much simple and short.

In fact, GATE system provides some simple gazetteers such as date, time, organization, location, money, province etc. But for a flexible language like Chinese, the list is very limited.

To lay a good foundation for Chinese named entity recognition, we need to enrich the GATE gazetteer lists.

Chinese named entity recognition

GATE system uses JAPE (a Java Annotation Patterns Engine) grammars to write rules to recognize NE (named entity). A semantic tagger of GATE consists of a set of rule-based JAPE grammars run sequentially. The grammars contain hand-written pattern-action rules which recognize e.g. annotations from the POS tagger and gazetter, and combine them to produce new NE annotaiton over patterns. JAPE is a pattern-matching language. The LHS (Left Hand Side) of each rule contains patterns to be matched, and the RHS (Right Hand Side) contains details of annotations (and optionally features) to be created.

For example, a rule might recognize a first name (from the gazetter module) followed by a proper noun (from the POS tagger), and annotate this pattern as a person. This rule could be written in JAPE as follows:

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

Because grammar of Chinese is quite different from that of English, the JAPE rules provided by GATE are not suitable for Chinese texts. We need to rewrite JAPE rules to implement Chinese information extraction.

3.3 Solutions to the problems

In the process of implementing Chinese Information Extraction system based on GATE, we need to overcome the three key problems we said just. After carefully planning, a total solution to the problems is brought forth (figure 4).

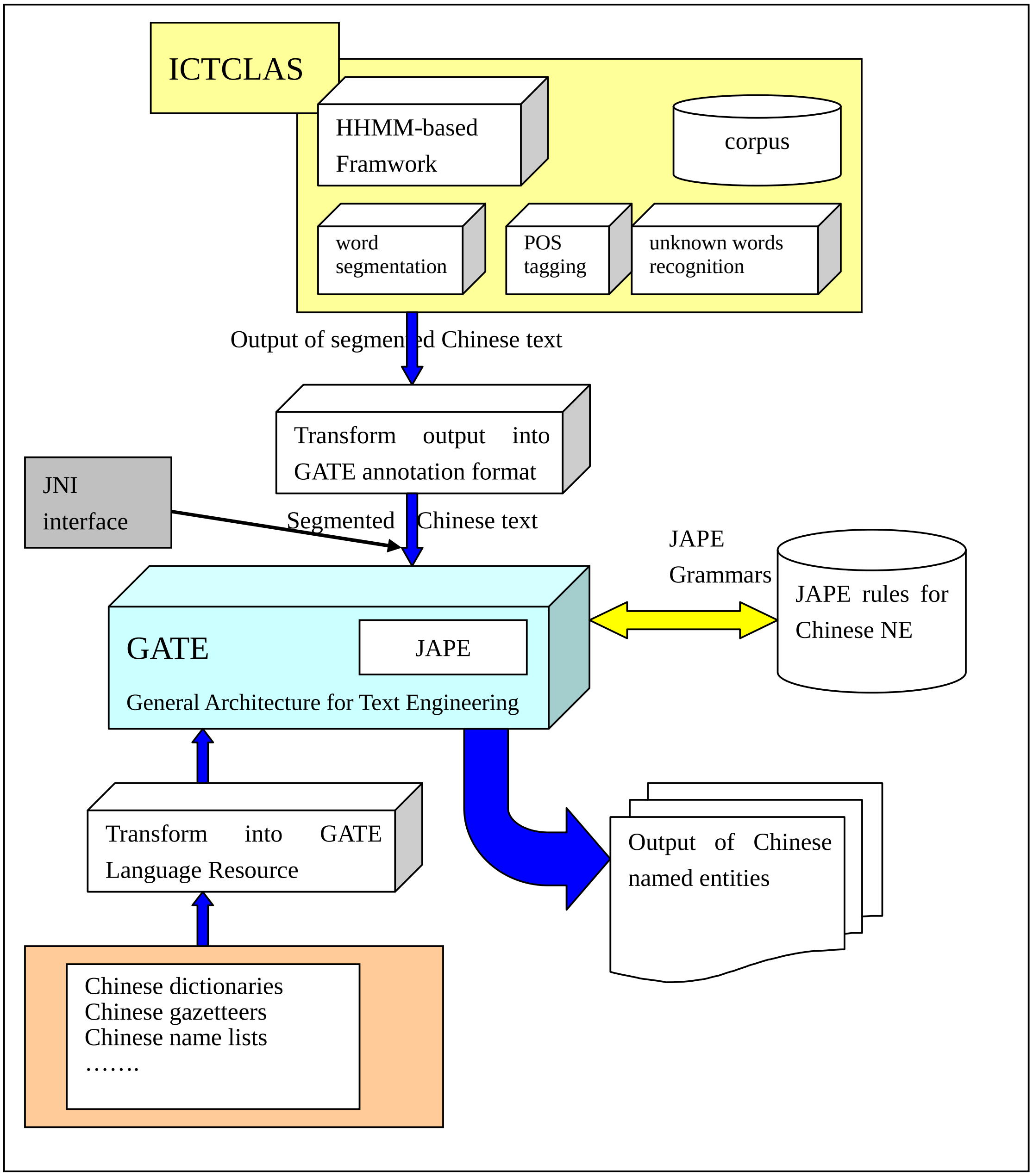

ICTCLAS
HHMM-based Framwork
corpus
word segmentation
POS tagging
unknown words recognition
Output of segmented Chinese text
Transform output into GATE annotation format
JNI interface
Segmented Chinese text
JAPE Grammars
JAPE rules for Chinese NE
GATE
JAPE
General Architecture for Text Engineering
Transform into GATE Language Resource
Output of Chinese named entities
Chinese dictionaries
Chinese gazetteers
Chinese name lists
…….

Figure 4. The total solution to Chinese Information Extraction system based on GATE

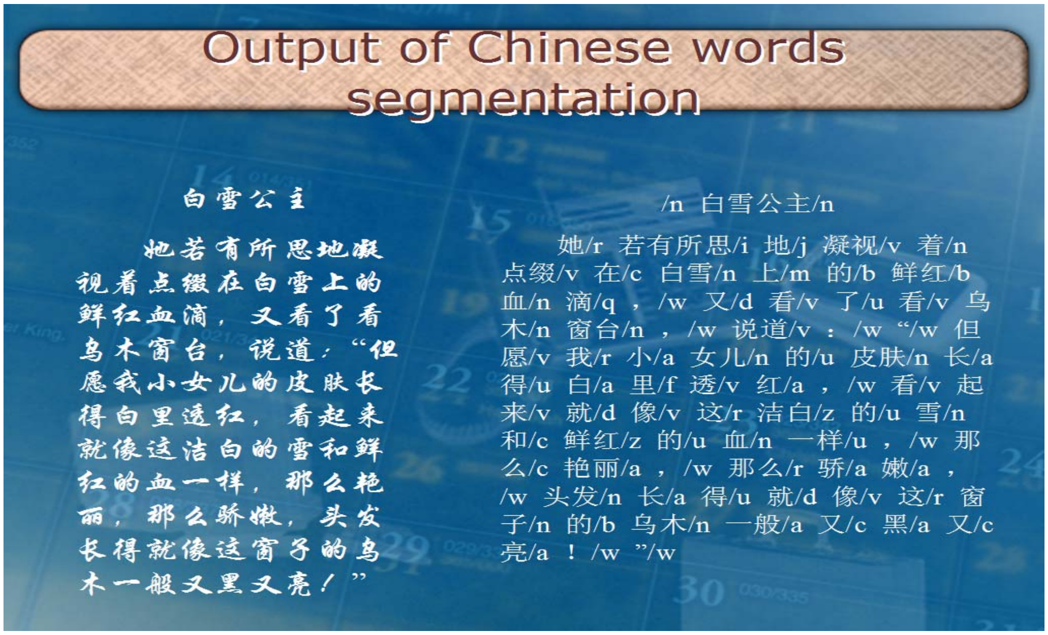
There are three main tasks we have done to implement this solution.

Integrating ICTCLAS to perform words segmentation

ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is an open source Chinese lexical analysis system developed by Institute of Computing Technology of Chinese Academy of Sciences. It uses an approach based on multi-layer HMM, including word segmentation, Part-Of-Speech tagging and unknown words recognition. Its segmentation precision is 97.58%. The recalling rates of unknown words recognized using roles tagging achieve more than

90%. Especially, the recalling of Chinese person names achieve nearly 98%. The speed for word segmentation and POS tagging is 31.5KB/s.

ICTCLAS is a good choice to enhance the Chinese words segmentation for GATE system. But in order to integrate ICTCLAS to GATE system, we have to do some developing. Because ICTCLAS is written by C/C++ language, while GATE is written by pure java language. In order to invoke dynamic link libraries of ICTCLAS in GATE system, we use Java Native Interface (JNI) from Java Development Kit (JDK) to solve the problem. In figure 5, the left part is the original Chinese text, and the right part is output of ICTCLAS invoked by GATE. Since GATE support Unicode, so the text with words separated by white space is suitable for GATE to take as input.



Output of Chinese words segmentation

白雪公主

她若有所思地凝视着点缀在白雪上的鲜红血滴，又看了看乌木窗台，说道：“但愿我小女儿的皮肤长得白里透红，看起来就像这洁白的雪和鲜红的血一样，那么艳丽，那么骄嫩，头发长得就像这窗子的乌木一般又黑又亮！”

```
/n 白雪公主/n
    她/r 若有所思/i 地/j 凝视/v 着/n
点缀/v 在/c 白雪/n 上/m 的/b 鲜红/b
血/n 滴/q ，/w 又/d 看/v 了/u 看/v 乌
木/n 窗台/n ，/w 说道/v ：/w “/w 但
愿/v 我/r 小/a 女儿/n 的/u 皮肤/n 长/a
得/u 白/a 里/f 透/v 红/a ，/w 看/v 起
来/v 就/d 像/v 这/r 洁白/z 的/u 雪/n
和/c 鲜红/z 的/u 血/n 一样/u ，/w 那
么/c 艳丽/a ，/w 那么/r 骄/a 嫩/a ，
/w 头发/n 长/a 得/u 就/d 像/v 这/r 窗
子/n 的/b 乌木/n 一般/a 又/c 黑/a 又/c
亮/a ！/w ”/w
```

Figure 5. Output of ICTCLAS invoked by GATE

Developing Chinese gazetteers to enrich GATE language resources

In GATE system, the gazetteer lists used are plain text files, with one entry per line. Each list represents a set of names, such as names of cities, organizations, days of the week, etc. An index file (lists.def) is used to access these lists; for each list, a major type is specified and, optionally, a minor type.

In the example below, the first column refers to the list name, the second column to the major type, and the third to the minor type. These lists are compiled into finite state machines. Any text tokens that are matched by these machines will be annotated with features specifying the major and minor types.

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

During the process of developing Chinese gazetteer, we have made full use of the features of library, which is full of dictionaries, name lists, gazetteers etc. We also collect many resources from Internet. Now we have accumulated about 100MB of Chinese gazetteers suitable for domain based information extraction. The table below is the gazetteers we used for common Chinese named entities recognition.

Table 1.Chinese gazetteer we prepared
for common Chinese named entities recognition

Gazetteers	Number of entry
Organization name	2100
Chinese city name	1309
World city name	140
Foreign company name	1241
Chinese company name	435
Media company name	147
Country name	222
County name	2189
Chinese university	1003
Resort name	331
Female name	2416
Institutes name	2100
Male name	2654
Keywords of organization	912

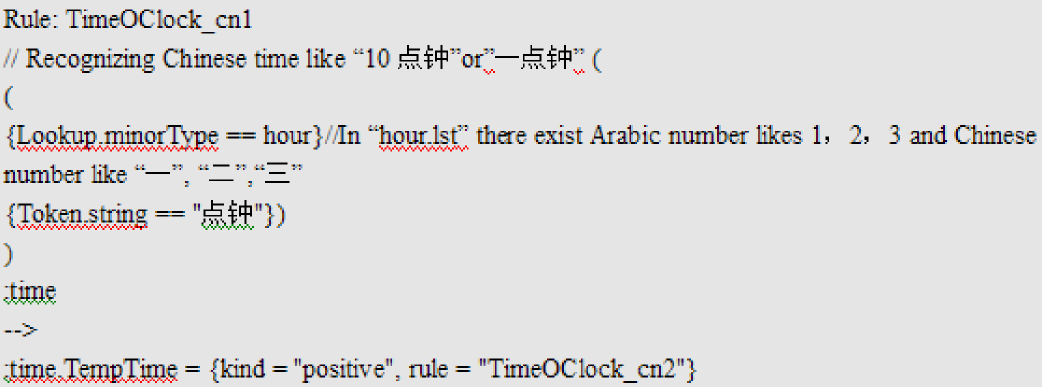
Rewriting JAPE rules to recognize Chinese NE

Because grammar of Chinese is quite different from that of English, we need to rewrite JAPE rules to make GATE suitable for processing Chinese texts.

A simple comparison is listed below. For example the JAPE rule recognizes English time like “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"}
```

In Chinese, “10 o’clock” is “10点钟” or “十点钟”. So the JAPE rule should be rewritten like below to make GATE could extract Chinese way of time expression.



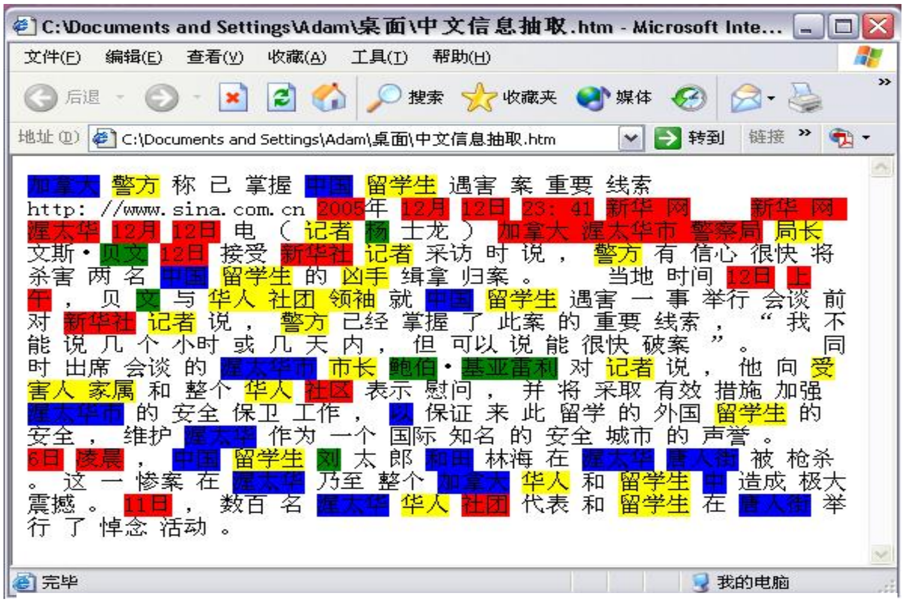
```
Rule: TimeOClock_cn1
// Recognizing Chinese time like “10 点钟”or”一点钟” (
(
{Lookup.minorType == hour}//In “hour.lst” there exist Arabic number likes 1，2，3 and Chinese
number like “一”, “二”,“三”
{Token.string == "点钟"})
)
:time
-->
:time.TempTime = {kind = "positive", rule = "TimeOClock_cn2"}
```

Altogether, we have rewritten about one hundred of JAPE rules to make the Chinese named entity recognition more precise.

4 Tests and Evaluation

After more than one years of working, we implemented the system. We also carry out an experiment in which the Chinese IE system successfully extracts thousands of pieces of science and technology news.

Figure 6 shows the result of Chinese information extraction using the system we developed. Compared with figure 3 we tested before, you can see many named entities could not be recognized in standard GATE suit can be figure out now.


C:\Documents and Settings\Adam\桌面\中文信息抽取.htm - Microsoft Inte...
文件(F)
编辑(E)
查看(V)
收藏(A)
工具(T)
帮助(H)
后退
搜索
收藏夹
媒体
地址(D)
C:\Documents and Settings\Adam\桌面\中文信息抽取.htm
转到
链接
完毕
我的电脑

加拿大 警方 称 已 掌握 中国 留学生 遇害 案 重要 线索
http://www.sina.com.cn 2005年 12月 12日 23:41 新华 网 新华 网
渥太华 12月 12日 电 （ 记者 杨 士龙 ） 加拿大 渥太华市 警察局 局长
文斯·贝文 12日 接受 新华社 记者 采访 时 说 ， 警方 有 信心 很快 将
杀害 两 名 中国 留学生 的 凶手 缉拿 归案 。 当地 时间 12日 上
午 ， 贝 文 与 华人 社团 领袖 就 中国 留学生 遇害 一 事 举行 会谈 前
对 新华社 记者 说 ， 警方 已经 掌握 了 此案 的 重要 线索 ， “ 我 不
能 说 几 个 小时 或 几 天 内 ， 但 可以 说 能 很快 破案 ” 。 同
时 出席 会谈 的 渥太华市 市长 鲍伯·基亚雷利 对 记者 说 ， 他 向 受
害人 家属 和 整个 华人 社区 表示 慰问 ， 并 将 采取 有效 措施 加强
渥太华市 的 安全 保卫 工作 ， 以 保证 来 此 留学 的 外国 留学生 的
安全 ， 维护 渥太华 作为 一个 国际 知名 的 安全 城市 的 声誉 。
6日 凌晨 ， 中国 留学生 刘 太 郎 和田 林海 在 渥太华 唐人街 被 枪杀
。 这 一 惨案 在 渥太华 乃至 整个 加拿大 华人 和 留学生 中 造成 极大
震撼 。 11日 ， 数百 名 渥太华 华人 社团 代表 和 留学生 在 唐人街 举
行 了 悼念 活动 。

🟩 Person 🟦 Address 🟥 Time Organization 🟨 Job title Money

Figure 6. Results of Chinese information extraction using the system we developed

5 Conclusions

Although we carried out a successful experiment and got a good evaluation for our Chinese information extraction system, there are still many works we need to do for more efficient use. We think it is a significant start of applying information extraction technology in our library and it will lay a good foundation for our future research works to support innovations in library services. Now we are going to make a proposal for trying to integrate Chinese information extraction system into other library service systems, which are serving users, to help in automatic annotation of digital materials, automatic acquisition of metadata, etc. Also we got more experiences of Developing & Localizing International Software and using open source software to promote library services.

References

[1] CSDL, Chinese national Science Digital Library, http://www.csdl.ac.cn/ [accessed May 8,2006]
[2] Natural Language Processing Research Group at the University of Sheffield, Information Extraction, http://nlp.shef.ac.uk/research/areas/ie.html [accessed May 8,2006]
[3] Hamish Cunningham, Information Extraction, Automatic, Encyclopedia of Language & Linguistics, 2nd Edition, 2005, http://gate.ac.uk/sale/ell2/ie/main.pdf [accessed May 8,2006]
[4] NIST, MUC, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html [accessed May 8,2006]
[5] NIST, TIPSTER Text Program, http://www-nlpir.nist.gov/related_projects/tipster/ [accessed May 8,2006]
[6] NIST, ACE - Automatic Content Extraction, http://www.nist.gov/speech/tests/ace/ [accessed May 8,2006]
[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/ [accessed May 8,2006]
[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 [accessed May 8,2006]

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

[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 [accessed May 8,2006]

[13] Project hTechsight, http://www.etse.urv.es/~drianyo/hTechSight/projecte.html [accessed May 8,2006]

[14] GATE, A General Architecture for Text Engineering, http://gate.ac.uk/ [accessed May 8,2006]

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

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

[17] KEA project, http://www.nzdl.org/Kea/ [accessed May 8,2006]

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

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

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

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