中华人民共和国国家知识产权局
State Intellectual Property Office
of the People's Republic of China

SIPO's Efforts on Improving Quality of Chinese-English Patent Machine Translation Service

Dan WANG
SIPO
August 30, 2009

www.sipo.gov.cn

Outline

I. Background

II. MT System Overview

III. Test Feedbacks and Quality-improving Measures

IV. Summary

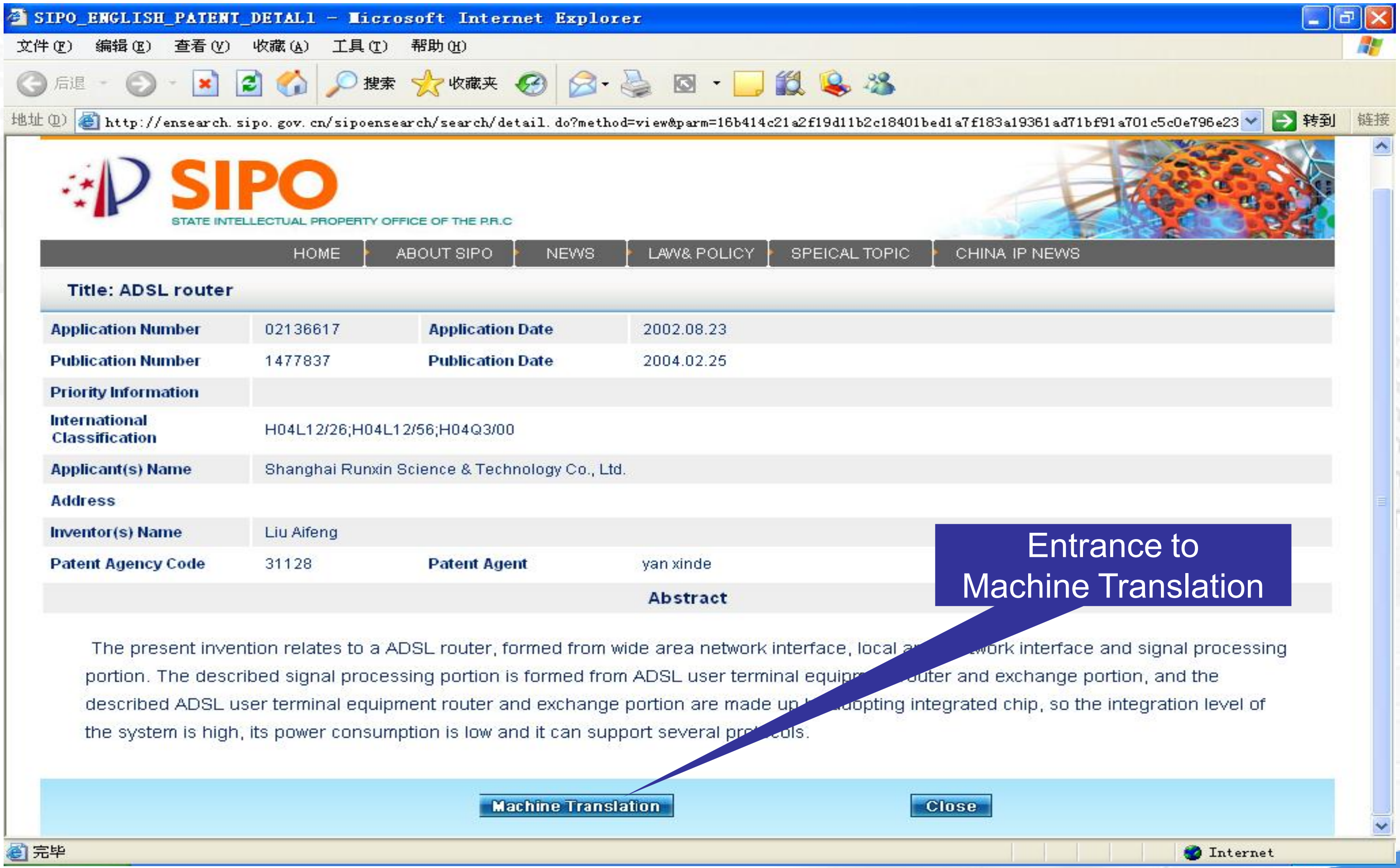
Chinese-English Patent MT Roadmap

2005.6	Chinese-English MT project started in SIPO
2007.4	Single-user version prototype **CPMT** (China Patent Machine Translation)
2008.4	CPMT launched for patent gisting
2008.4 -	Test and improvement for higher quality

Current Status of CPMT

- Integrated use with patent search services of SIPO and CPIC
  - ➲ available on http://www.sipo.gov.cn/sipo_English
  - ➲ available on http://www.cnpat.com.cn
  - ➲ English search of bibliographic data and abstracts, and on-the-fly MT for full text

MT Entrance: Abstracts by Manual Translation


SIPO_ENGLISH_PATENT_DETAIL1 - Microsoft Internet Explorer
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
地址(D)
转到
链接
SIPO
STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.C
HOME
ABOUT SIPO
NEWS
LAW& POLICY
SPEICAL TOPIC
CHINA IP NEWS
完毕
Internet

http://ensearch.sipo.gov.cn/sipoensearch/search/detail.do?method=view&parm=16b414c21a2f19d11b2c18401bed1a7f183a19361ad71bf91a701c5c0e796e23

Title: ADSL router

Application Number	02136617	**Application Date**	2002.08.23
Publication Number	1477837	**Publication Date**	2004.02.25
Priority Information			
International Classification	H04L12/26;H04L12/56;H04Q3/00		
Applicant(s) Name	Shanghai Runxin Science & Technology Co., Ltd.		
Address			
Inventor(s) Name	Liu Aifeng		
Patent Agency Code	31128	**Patent Agent**	yan xinde

Abstract

The present invention relates to a ADSL router, formed from wide area network interface, local a[illegible]work interface and signal processing portion. The described signal processing portion is formed from ADSL user terminal equip[illegible]uter and exchange portion, and the described ADSL user terminal equipment router and exchange portion are made up [illegible]dopting integrated chip, so the integration level of the system is high, its power consumption is low and it can support several pr[illegible]cols.

Machine Translation

Close

Entrance to
Machine Translation

Quality of MT: a Critical Topic

- Problems of uncertainty and ambiguities are common in patent document translation? (e.g. Long NPs and multiple verbs.)
- Linguistic issue in MT:
  - it is very hard to cover all the rules and adjust them to all possible variations
- But in patent document, the language is very specific, thus MT is useful

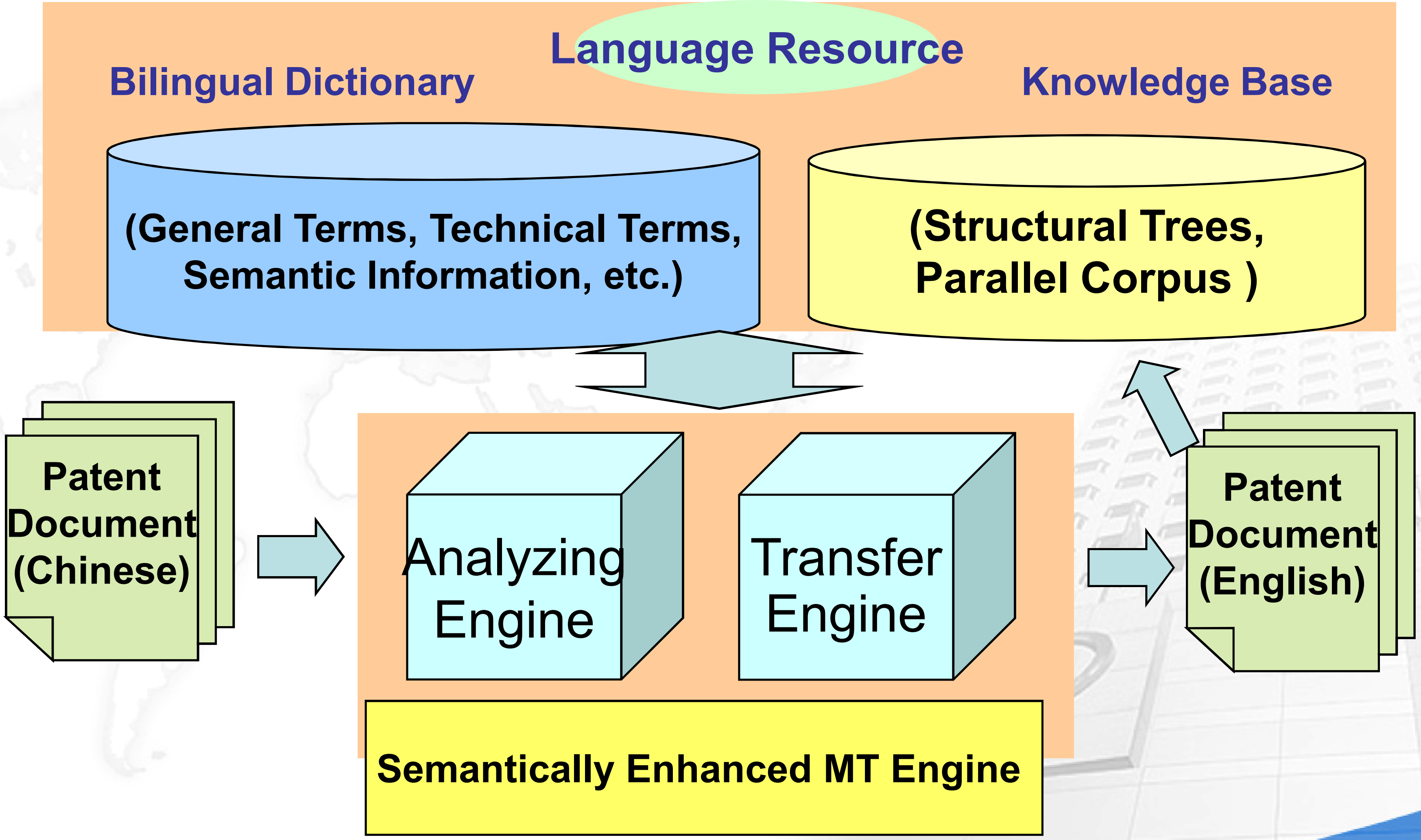
Strategies for Tackling the Problems

- Customization of existed general domain Chinese-English MT engines
- Integration of different machine translation paradigms (rule-based, statistical, and example-based)
- Facilitating and speeding up acquiring and handling of language resources in patent documents

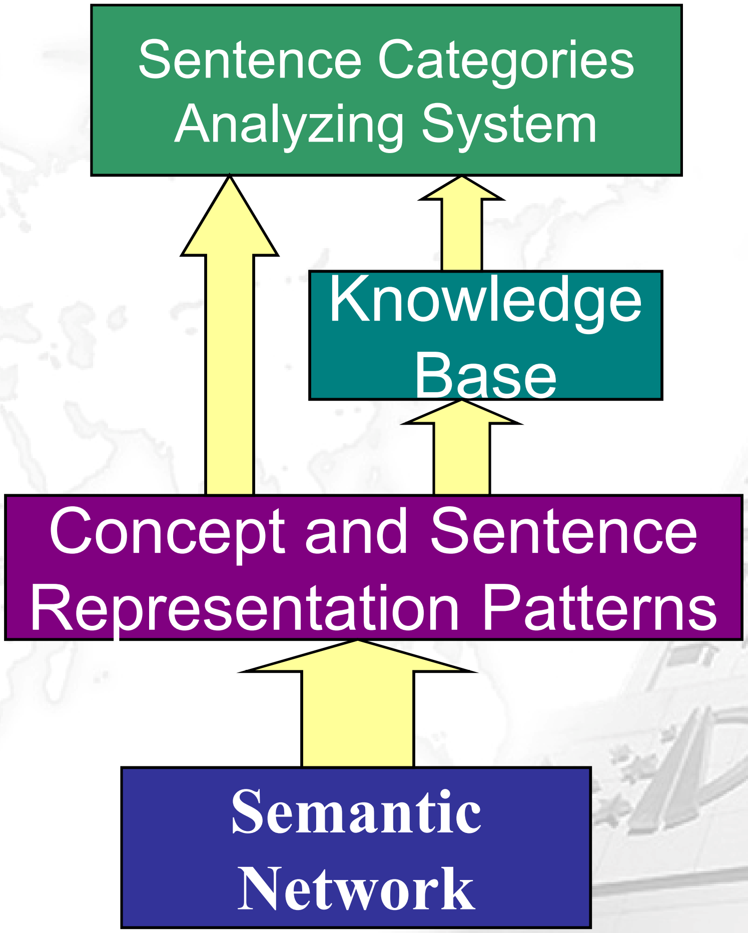
MT System Overview

- Semantic features ad hoc: high-quality semantic-based MT engine
- Semantic analyzing: effective for disambiguation
- Representing the meaning of the source sentence and then generate the translation from the meaning
  - Fix cases of syntactic mismatch

Patent MT System Architecture


Language Resource
Bilingual Dictionary
Knowledge Base
(General Terms, Technical Terms, Semantic Information, etc.)
(Structural Trees, Parallel Corpus)
Patent Document (Chinese)
Analyzing Engine
Transfer Engine
Patent Document (English)
Semantically Enhanced MT Engine

Hierarchical Network of Concepts (HNC)


Sentence Categories Analyzing System
Knowledge Base
Concept and Sentence Representation Patterns
Semantic Network

Sentence Categories in HNC

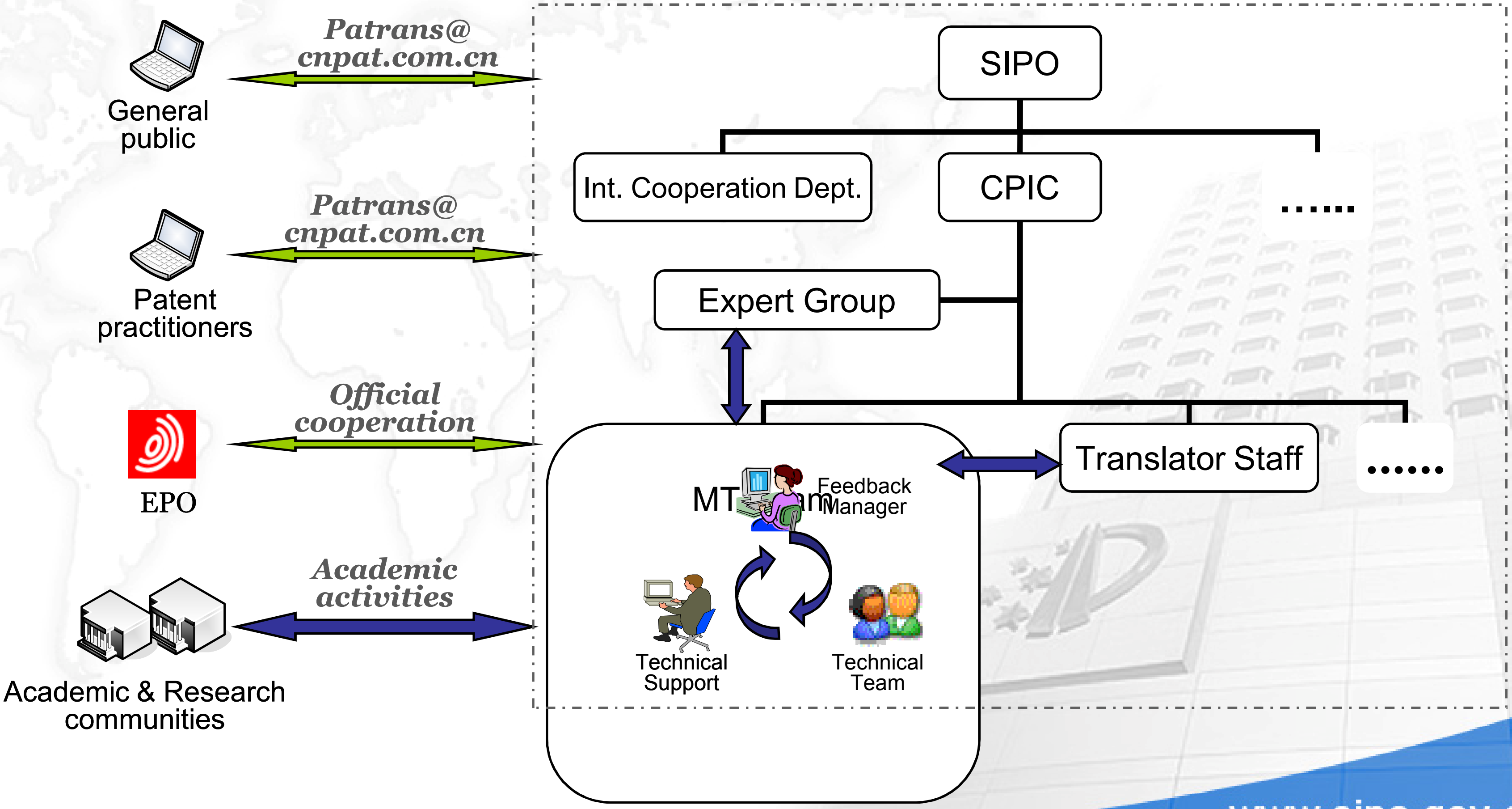
- Sentences are classified into 57 categories, this significantly raises the likelihood that the MT system correctly translates different sentence structures and meanings of words
- The 57 categories cover over 90 percent of the real-text sentences to be translated from Chinese into English
- Grammatical outputs are produced in most cases

HNC Symbols

思考 v80	产生 v311	情感 g713	思维 g80	消除 v31
想法 r80	推动 v361	承担 v901		
力 g008	抑制 v362	保护 v3219	责任 rc010	
力量 gz00	调节 v360	照顾 v653219		圆满 u30a
力度 z00	年 wj10-	维护 v93219	完成 v30a8	
弱 u00c21	月 wj10-0	保卫 vc3219	精力 gz655098	
强 u00c22	日 wj10-00	写作 va31	旺盛 zu5098e71	
萌芽 gv10ac41	体 j20-	幸福 gu50a9ae81		
成长 v10ac42	面 j20-0	治疗 va82	生活 gv50a9	
成熟 vu10ac43	线 j20-00	处方 gwa82		
衰亡 v10ac44	点 j20-000		药物 wa82	

达成 vc249a$(v308|(jlv001/v810))

Improvement via. Analyzing User Feedbacks and Evaluations



General public

Patrans@ cnpat.com.cn

Patent practitioners

Patrans@ cnpat.com.cn

EPO

Official cooperation

Academic & Research communities

Academic activities

SIPO

Int. Cooperation Dept.

CPIC

……

Expert Group

Translator Staff

……

MT Team

Feedback Manager

Technical Support

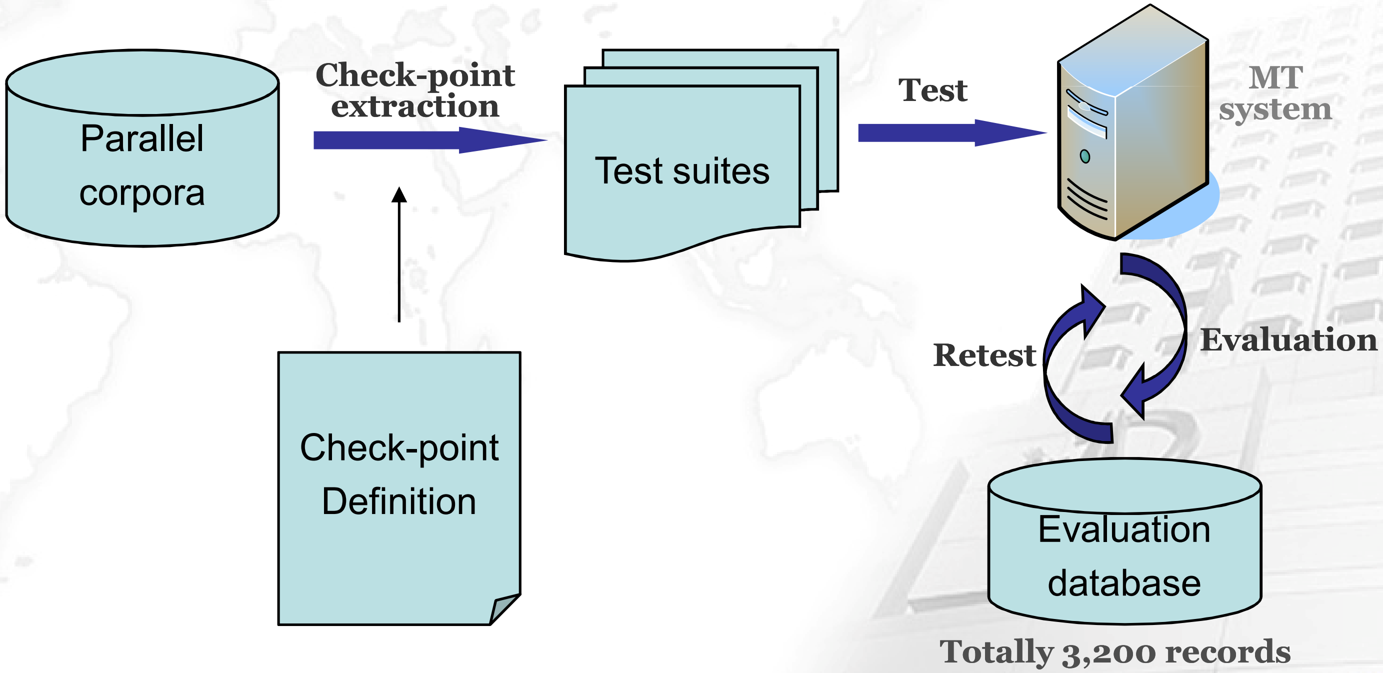
Technical Team

Response to User Feedback

	Issues	Responses
Error Reports	- Terminologies - Inappropriate EN expression - Linguistic errors	Integrated analysis for quality improvement
Thoughts & Recommendations	- MT technique - Evaluation methodology	Strategy research & investigation
Comments & Questions	- Accessibility - Usability	System performance optimization

Standardization of Test & Evaluation Workflow

- Test suites created based on check-point definition
- Full trace of each test case ensured by test->evaluation->retest and evaluation database



Check-point Extraction

- Check-points are now manually extracted

  - Source corpora：

    bilingual versions of PCT documents and priority documents;
    parallel abstracts from daily translation work;
    existing sentence-aligned parallel corpus

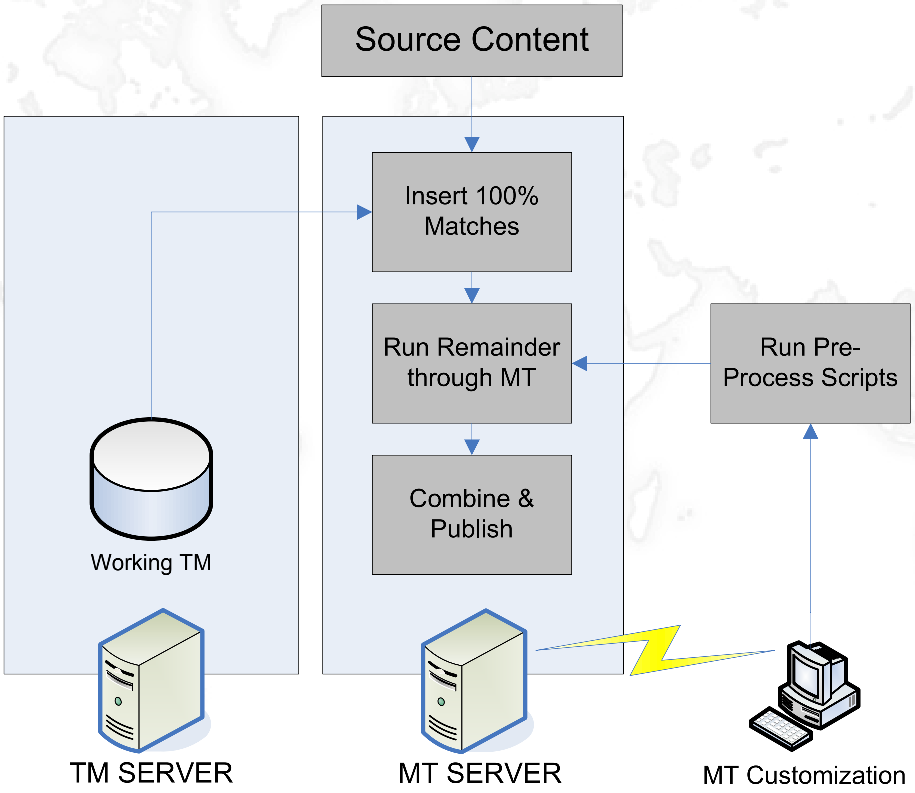
  - Principle of extraction：

    adequate number of test cases per check-point to ensure system stability;
    diversity of test cases for one single check-point

Feedbacks and Check-points Definition

Terminological		Idiomatic		Syntactic		Semantic
Technical term		Patent specific idiom		Indefinite article		Definite article
Patent-specific term		Technical idiom		Ordinal numbers		Word/sentence segmentation
UNK		Fixed collocation		Plurality		POS recognition and conversion
……		Discrete structure		Tense and voice		Physical interrelationship
	➲	……	➲	Adverbial phrase of time	➲	Logical interrelationship
				Adverbial phrase of location		Word/phrase order
				Verb-subject/object collocation		Special-style CN sentence
				Inappropriate EN expression		Long and complex sentence
				……		……

Language Resources in Patent Documents



- Improvement for translation quality is achieved through introduction of large-scale parallel corpus
- MT system output benefited from human translations
- Improvement over time, also through terminology additions

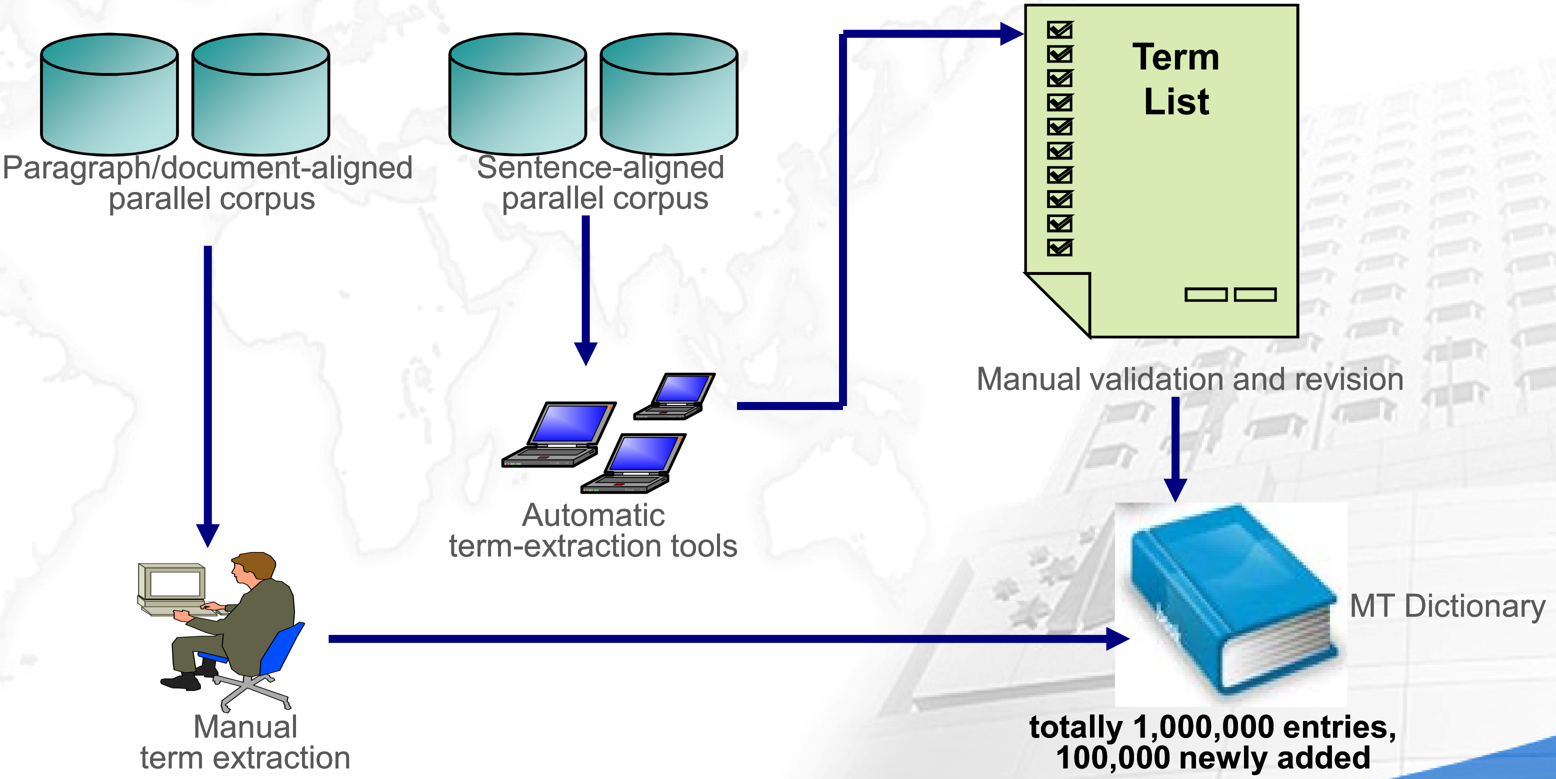
Language Resources in Patent Documents

- Establishing and maintaining bilingual dictionaries for various sub-domains

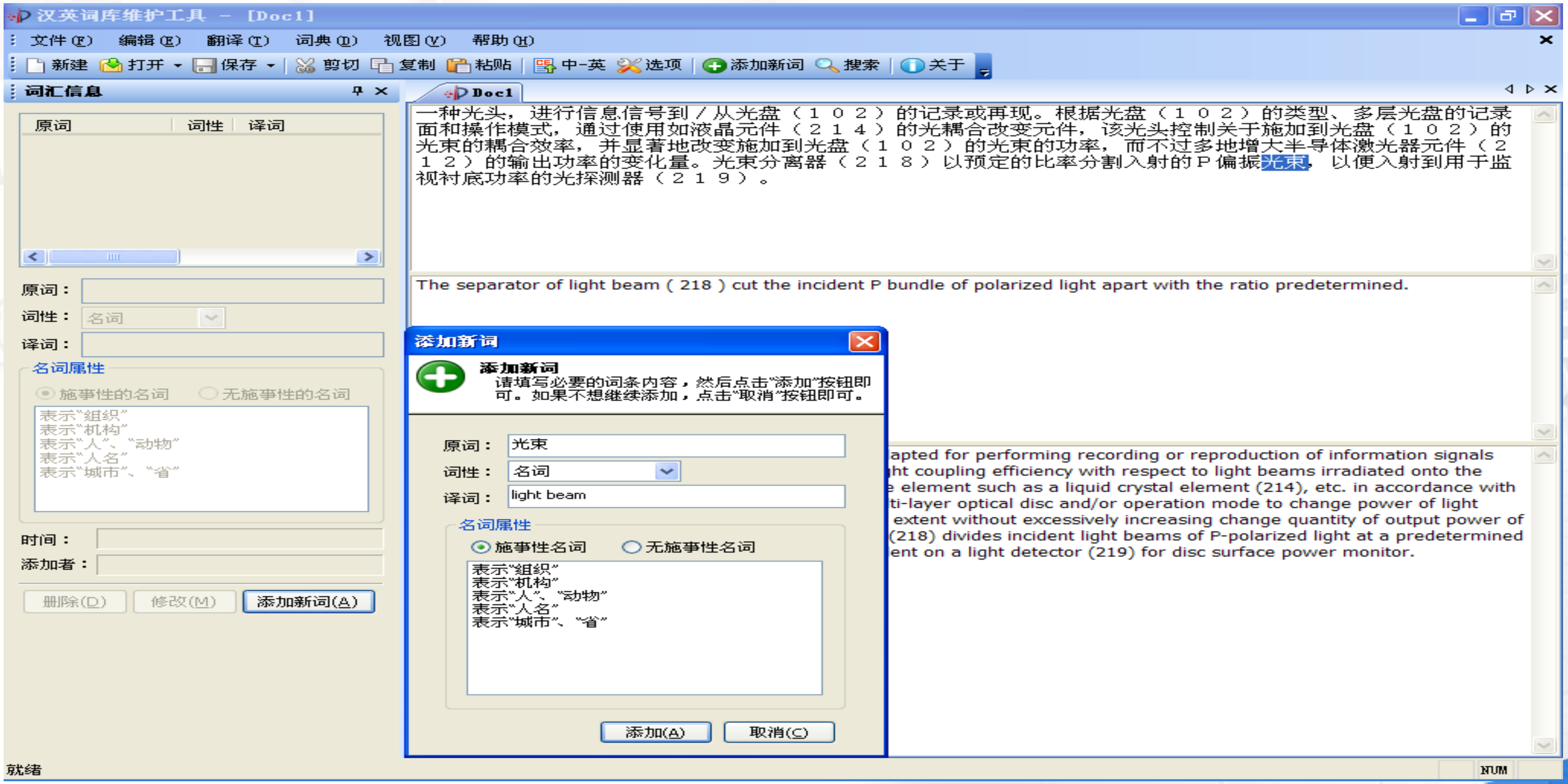
Chinese	English
隔板	board ⇒ baffle
先验信息	*transcendent information* ⇒ prior information
立体	*stereo* ⇒ three-dimensional
中间结果	*middle results* ⇒ intermediary results
:	:

Knowledge Acquisition Techniques

- Statistical-manual approach for term extraction


Paragraph/document-aligned parallel corpus
Sentence-aligned parallel corpus
Term List
Manual validation and revision
Automatic term-extraction tools
MT Dictionary
Manual term extraction
totally 1,000,000 entries, 100,000 newly added

New Words Detection Based on Large-scale Corpus


汉英词库维护工具 － [Doc1]
文件(F) 编辑(E) 翻译(T) 词典(D) 视图(V) 帮助(H)
新建 打开 保存 剪切 复制 粘贴 中-英 选项 添加新词 搜索 关于
词汇信息
原词 词性 译词
原词：
词性： 名词
译词：
名词属性
施事性的名词 无施事性的名词
表示"組织"
表示"机构"
表示"人"、"动物"
表示"人名"
表示"城市"、"省"
时间：
添加者：
删除(D) 修改(M) 添加新词(A)
Doc1
就绪
NUM

一种光头，进行信息信号到／从光盘（１０２）的记录或再现。根据光盘（１０２）的类型、多层光盘的记录面和操作模式，通过使用如液晶元件（２１４）的光耦合改变元件，该光头控制关于施加到光盘（１０２）的光束的耦合效率，并显著地改变施加到光盘（１０２）的光束的功率，而不过多地增大半导体激光器元件（２１２）的输出功率的变化量。光束分离器（２１８）以预定的比率分割入射的Ｐ偏振光束，以便入射到用于监视衬底功率的光探测器（２１９）。

The separator of light beam (218) cut the incident P bundle of polarized light apart with the ratio predetermined.

apted for performing recording or reproduction of information signals
ht coupling efficiency with respect to light beams irradiated onto the
e element such as a liquid crystal element (214), etc. in accordance with
ti-layer optical disc and/or operation mode to change power of light
extent without excessively increasing change quantity of output power of
(218) divides incident light beams of P-polarized light at a predetermined
ent on a light detector (219) for disc surface power monitor.

添加新词

添加新词
请填写必要的词条内容，然后点击"添加"按钮即可。如果不想继续添加，点击"取消"按钮即可。

原词： 光束

词性： 名词

译词： light beam

名词属性

● 施事性名词 ○ 无施事性名词

表示"組织"
表示"机构"
表示"人"、"动物"
表示"人名"
表示"城市"、"省"

Attempts towards a Multi-engine MT System

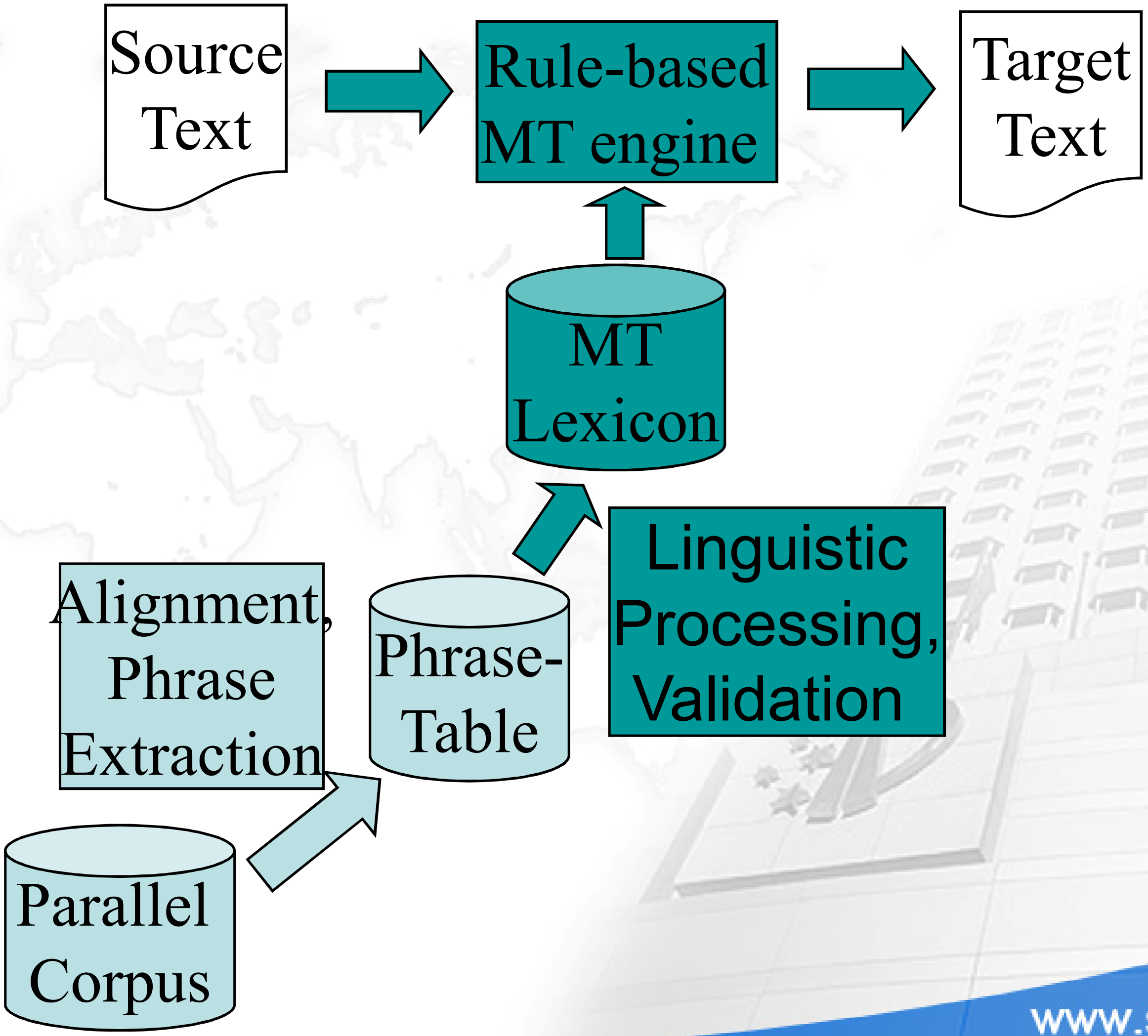
- Selection and comparison of two candidate engines
- Customization of the semantic engine adopting Chinese NLP techniques

Current results of MT engines comparison and customization

RBMT engine A	Semantic engine B
grammatical constituent combination English expression generation	segmentation logical chunk recognition and scheduling CN-EN sentence style transformation parsing of long and complex sentences

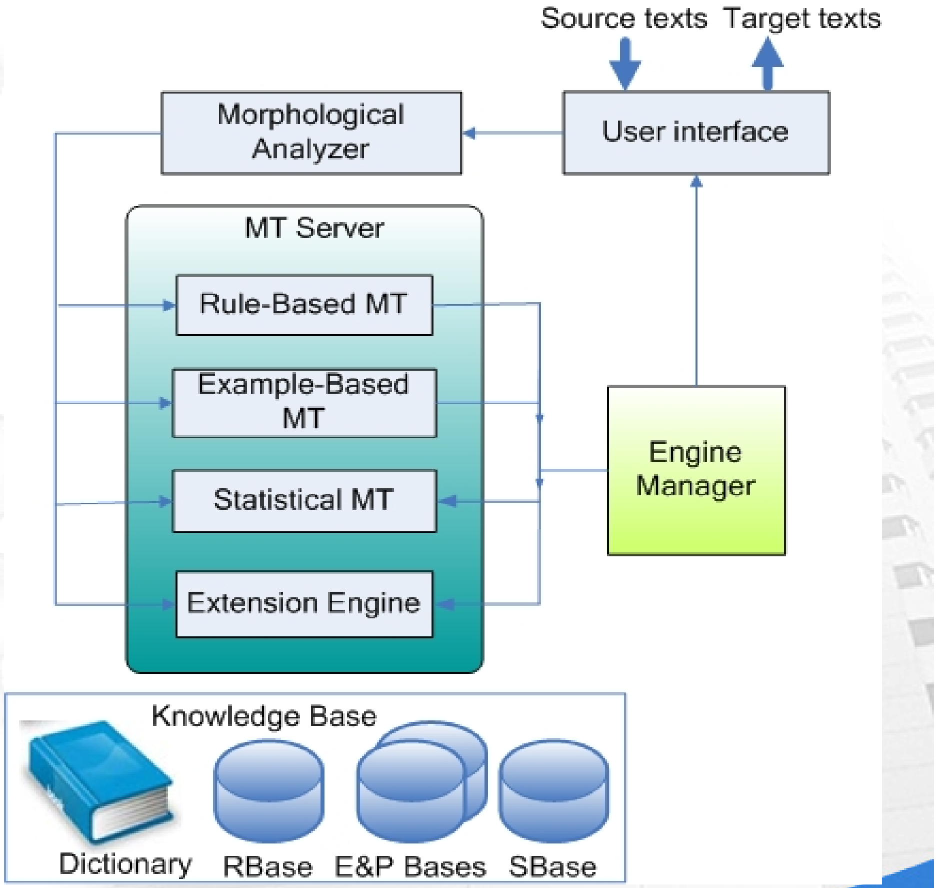
- Future work: to implement multi-engine MT system through exploring system combination strategies.

Statistical Approach Feeds Rule-based MT



MT Engine Combination, How to?

- Rule-based MT and SMT co-exist and show complementary strengths possible?
- System combination, how to combine the advantages?



Future Vision of CPMT

www.sipo.gov.cn

Thank you!