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

CPIC's MT Development: Current Status and Future Directions

WANG Dan
China Patent Information Center, SIPO
Dec. 10, 2010, TOKYO

Outline

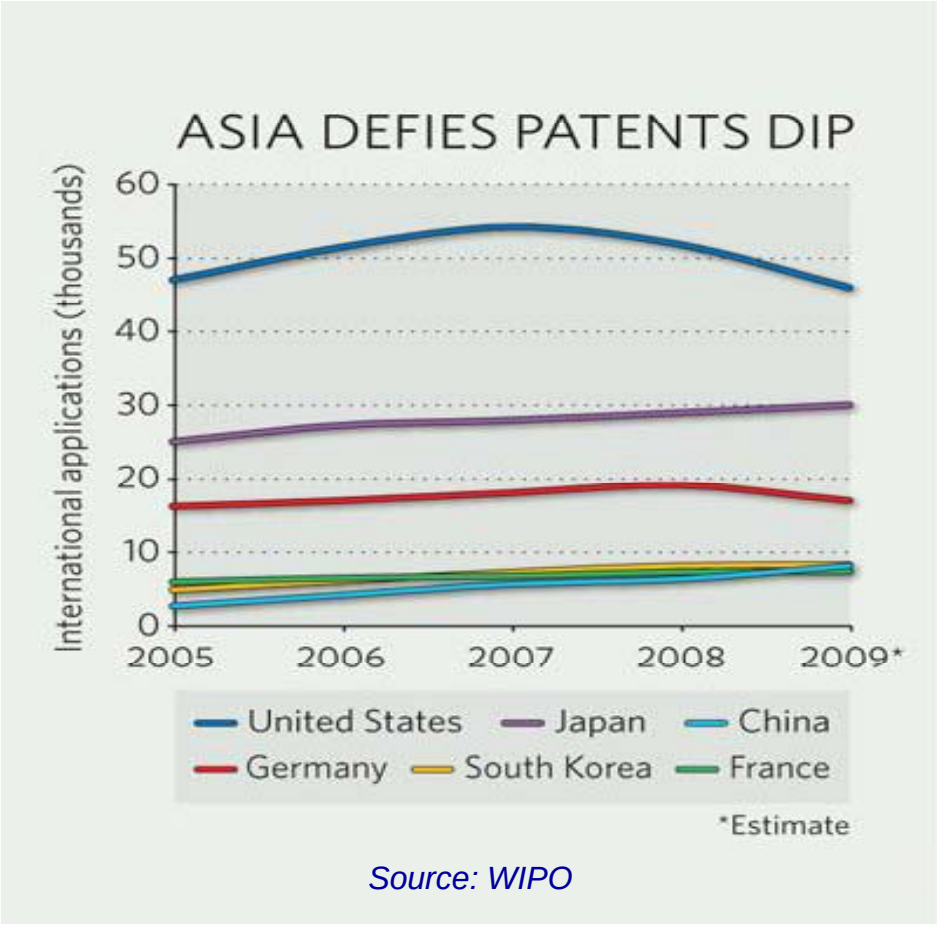
I. Background

II. IP5 Evaluation Analysis

III. Tackling Quality Challenge

IV. Future Directions

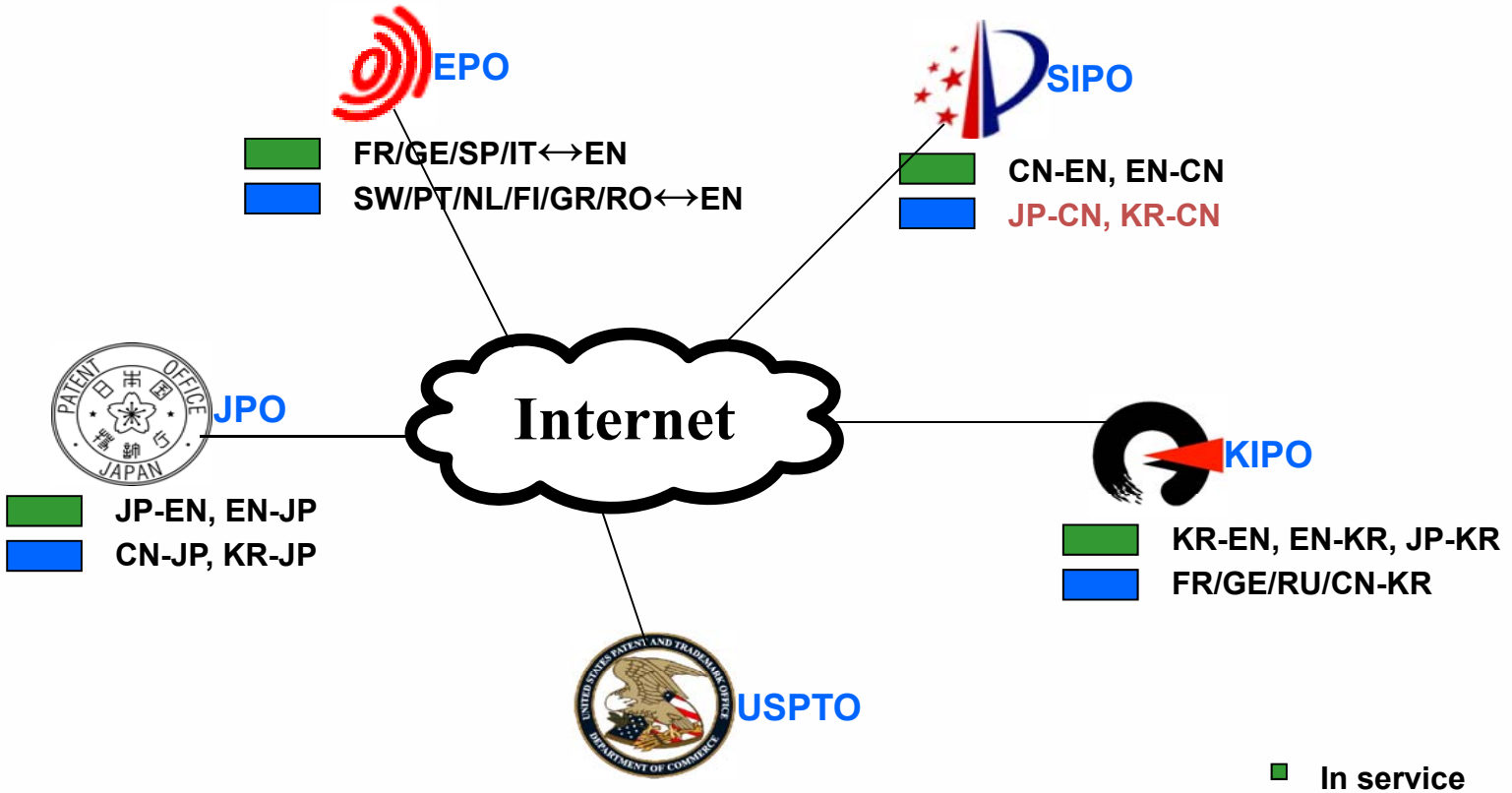
MT : Solution for Barrier-free Access to East Asian Patent Information



Source: WIPO

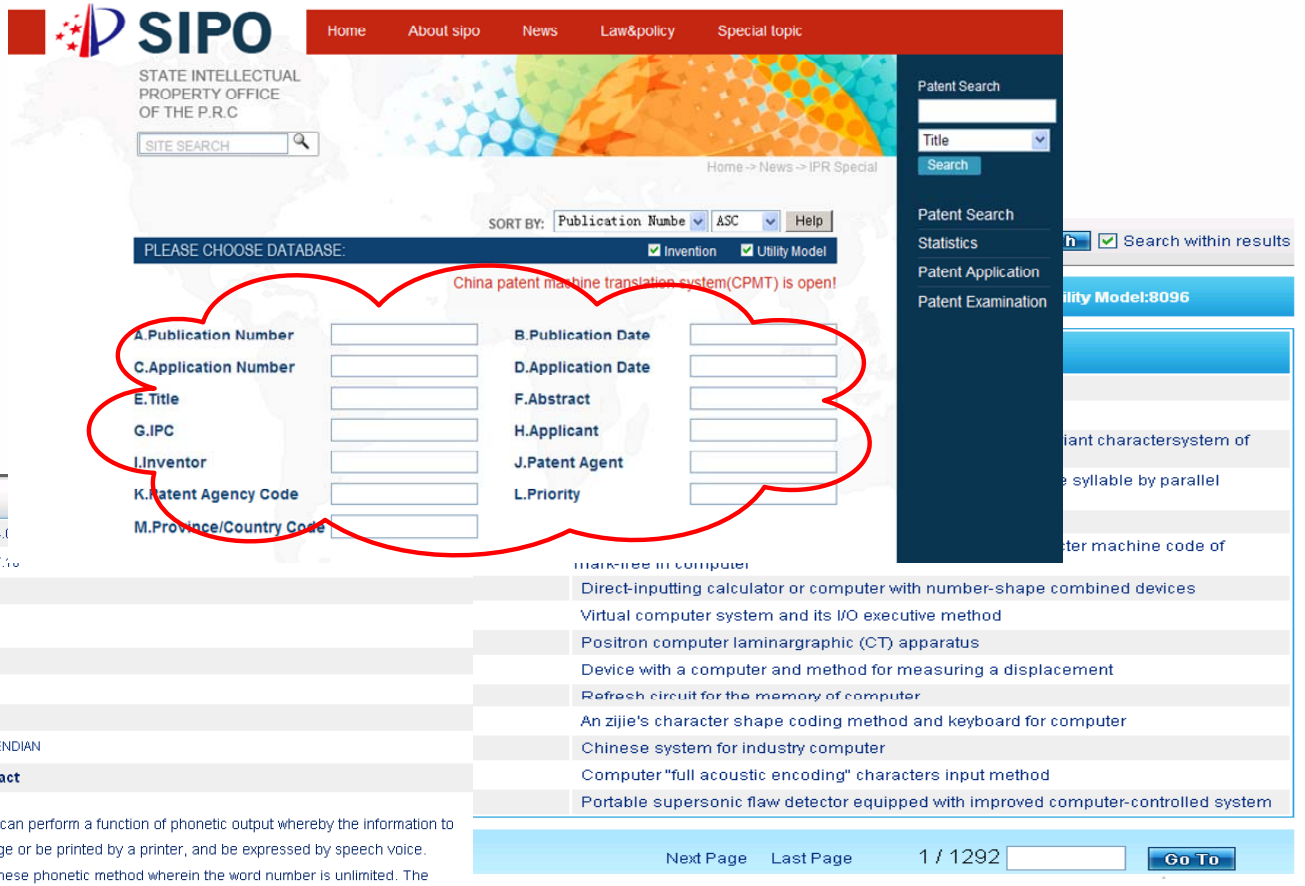
- **Patent application declines across the Western world**
- **PCT applications in 2009: Japan No.2, Korea No.4, China No.5**
- **Guiding principle of EPO's patent information policy: barrier-free access to patent information (including languages)**

MT Language Pairs Underway



SIPO's Chinese to English MT Service (CPMT)

- Covering inventions and utility models
- Updating 3 months after publication



CPMT System Overview

Free Access

Free of charge for the IP community

- Available on http://www.sipo.gov.cn/sipo_English
- And on http://www.cnpat.com.cn

A Production System

On-the-fly CN-EN MT for claims and specifications

- Accepted usability for catching the gist of CN patent information
- Further improvement in progress through testing the system

Translation Engine

Customization of general purpose engines

- Employing semantic analyzing for effective disambiguation
- A rule-based engine in essence

CPIC's MT Roadmap

Outline

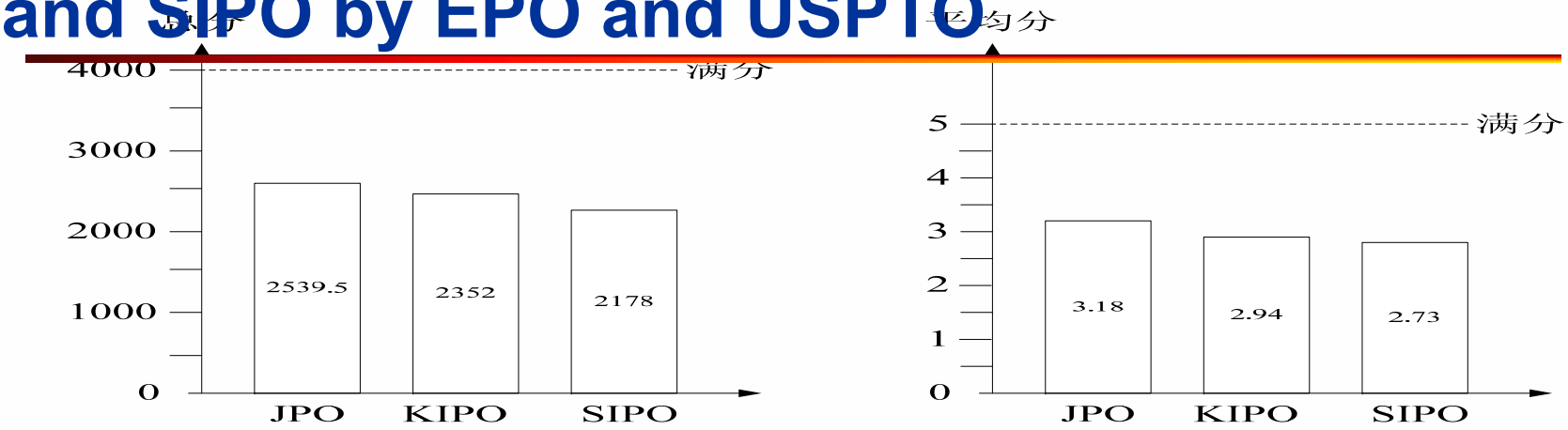
I. Background

II. **IP5 Evaluation Analysis**

III. Tackling Quality Challenge

IV. Future Directions

Scores of Evaluation Results for JPO, KIPO and SIPO by EPO and USPTO



- Score of 3: The translation is not entirely understandable, but is actionable, with some critical information accurately transferred. The text is stylistically or grammatically odd, but the language may still reflect a sound content to a patent professional.

Analysis of SIPO's Score

Different Development Stages

- Machine translation @ JPO and KIPO: more mature services as result of earlier launching
- SIPO' CN-EN MT powered by CPIC's engine: still undergoing improvement

Influence of Evaluation Organization

- No source language sentences are considered during evaluation
- No pre-edited texts are used in the process of Chinese to English machine translation
- CPIC's earlier development efforts were focused on grammatical output of the engine

Difficulty of Chinese to English Translation

- Japanese & Korean: hypotactic languages, Chinese: a paratactic one
- Japanese & Korean contain much more morphological information than Chinese

Outline

I. Background

II. IP5 Evaluation Analysis

III. **Tackling Quality Challenge**

IV. Future Directions

Plan for Quality Improvement

Evaluation Methodology	To explore an appropriate quality evaluation methodology; and to formulate standards for each language pair(C2E, E2C)
Engine Customization	To apply the methodology in the "test-evaluate-improve" cycle of MT engine customization
Domain Terms Adding	Add terms through user's feedbacks, active testing and human translation

Exploration for Better Evaluation Methodology

Expected Role

A tool useful for:

- Monitoring results of system development and improvement process
- Selecting MT engine in introducing new language pairs

Expected Features

Appropriate

- Evaluation for “Improvement” instead for “Evaluation” itself

Sufficient

- Adequate number of check-points, and diversity of test cases under one checkpoint for system capability

Diagnostic

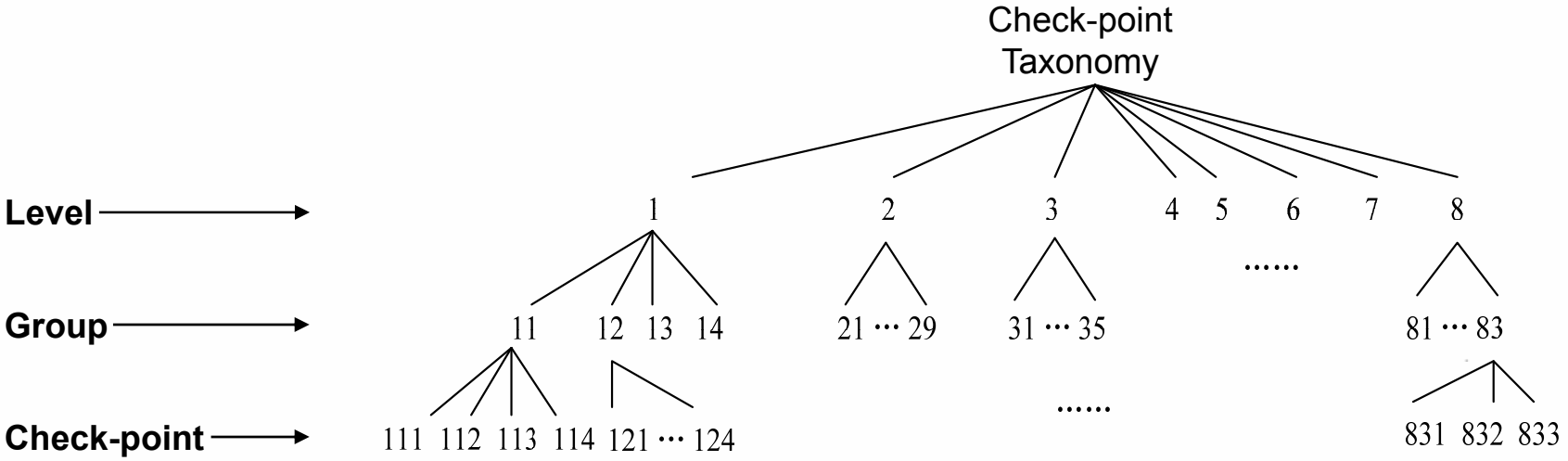
- Linguistically comprehensive for testing capabilities

Effective

- Efficient for evaluation workflow during development

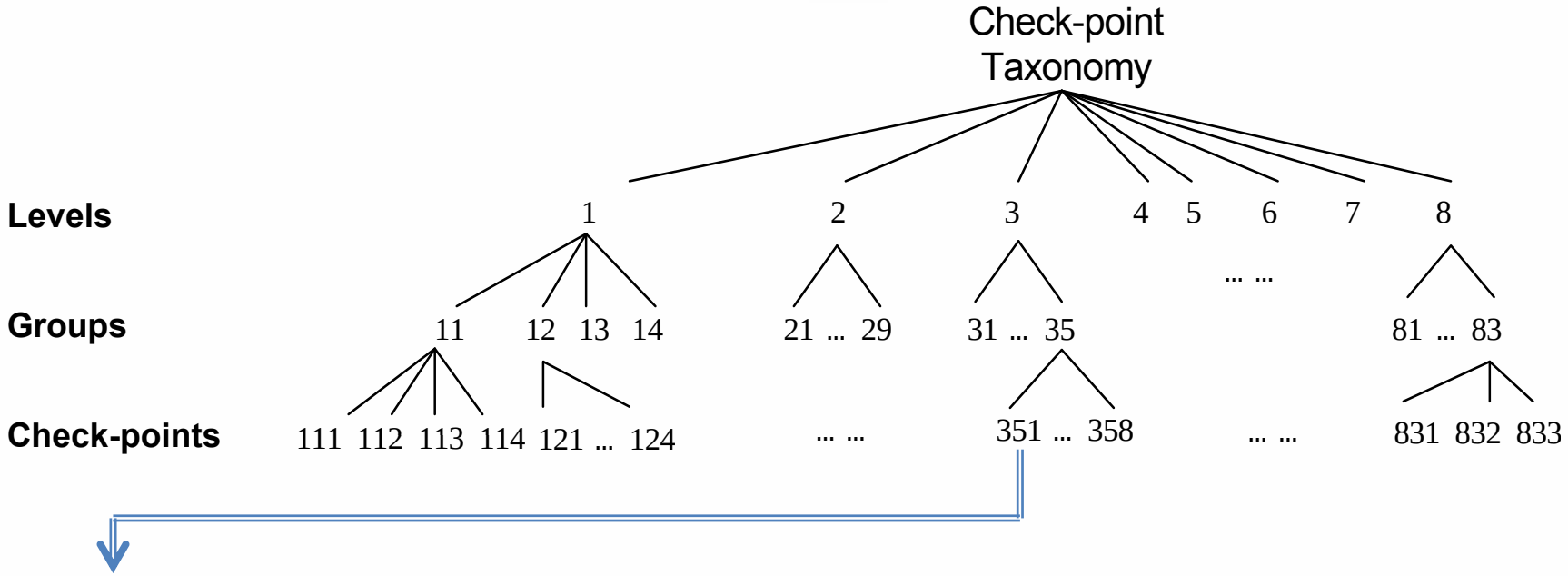
Current Work on Check-point Taxonomy

- **C2E, E2C check-point taxonomies finished**
  - Each taxonomy organized as a tree structure
  - Each leaf node is a check-point representing a linguistic unit
- **Improvement of all taxonomies will be continued**
- **Design of J2C check-point taxonomy in process**



Schematic view of numbering of check-points

Check Point Example 1: Solved



Source	一条指定水平线的像素数据的扫描级被有次序地存储在一个地址存储器中。
Reference	A scanning level of pixel data for a given horizontal line is regularly stored in an address memory.
Last version of MT	Article one, appoint the scanning stage of the pixel data of horizontal line regularly to be saved in an address memory.
Current version of MT	The scanning stage of the pixel data of an appointed horizontal line is regularly saved in an address memory.

Check point 351 - 数量短语（Quantifier phrase）

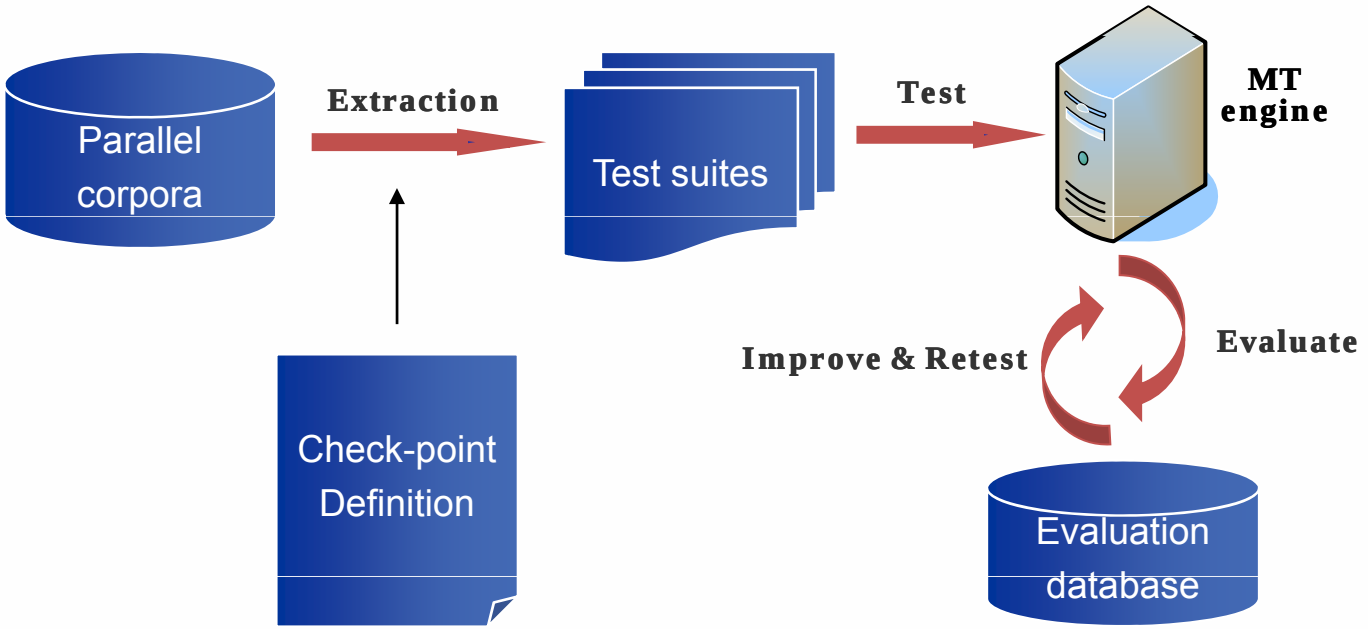
Check Point Example 2: Unsolved

Check point 723 – 叠词（Repetitive word）

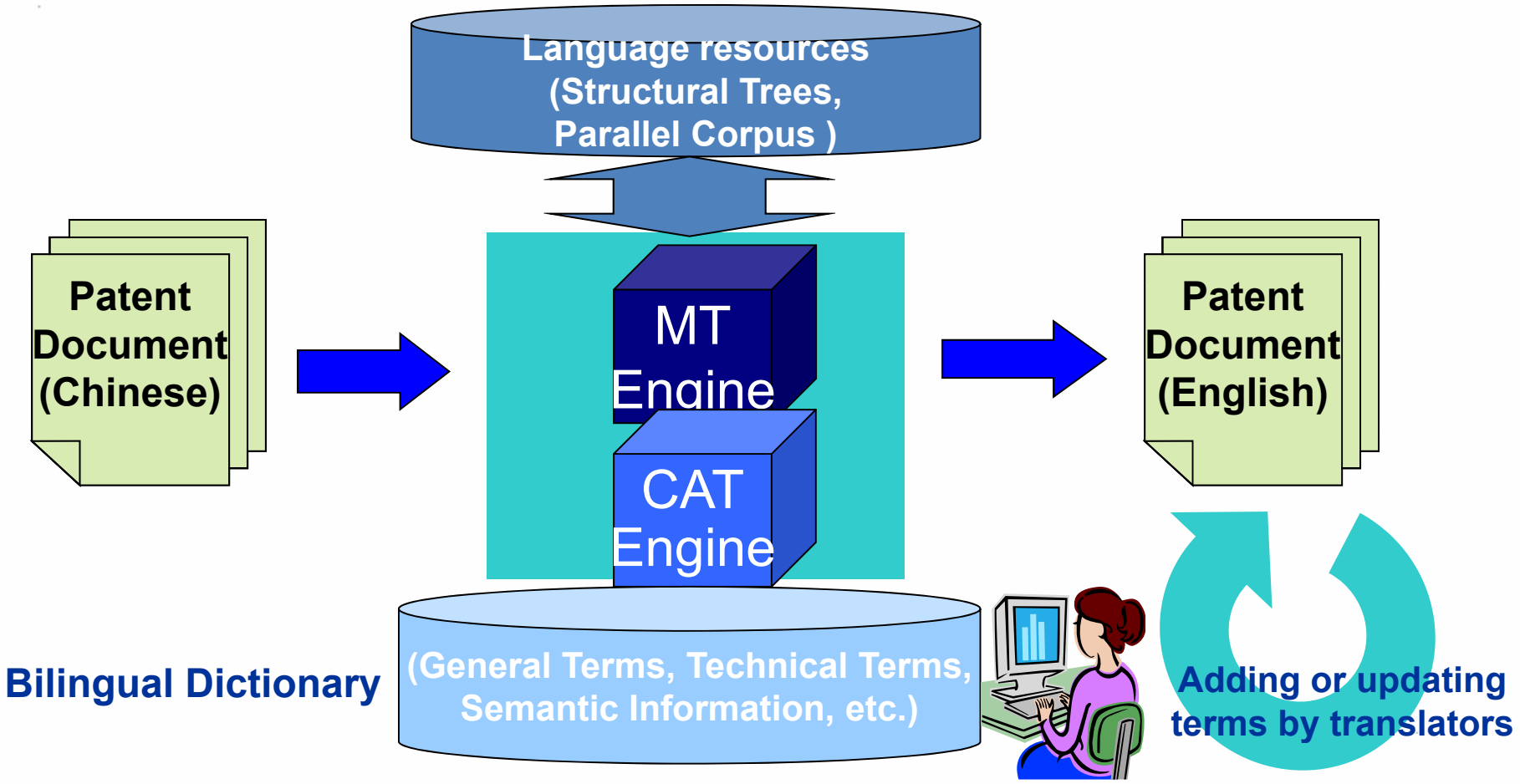
Source	基于设置为在每个时间调度帧中一帧一帧地提供多种类别的内容的提供时间表，将与请求时间相匹配的在时间调度帧中提供的多种内容顺序提供给客户ＰＣ４。
Reference	Many kinds of contents to be provided in a time schedule frame matching the time demanded are sequentially provided to a client PC (4) based on a providing schedule set up to provide plural kinds of contents in each time schedule frame **one by one**.
Last version of MT	Based on setting up as the timetable that provides that provides to **a frame** content of multiple classification in every time dispatch frame, will offer customer PC4 in proper order with the multiple content that provides of request time phase-match in time dispatch frame.
Current version of MT	Based on setting up as the timetable that provides that **one by one** provides the content of multiple classification in every dispatch frame time, will with the dispatching the multiple content that provides in the frame in the time and offer customer PC4 in proper order of request time phase-match.

Current Work on Engine Customization

- Engine customization of all 3 language pairs in process
- Standardized workflow of "Test→Evaluate→Improve" cycles to be continued by applying the quality standards
- Error feedbacks to be collected and addressed within the workflow



Adding Terminology Though CAT Route



Integrating MT with CAT

Integration into translation production environment

- Making better usage of linguistic assets maintained by human translators accumulated over the years
- An enhanced mechanism for terminology control and addition
- Less and less unknown terms for the machine translation engine

Outline

I. Background

II. IP5 Evaluation Analysis

III. Tackling Quality Challenge

IV. Future Directions

Future Directions

Works ahead

Further quality improvement

- Continue the "Test – Evaluate – Improve" cycle
- Enhancing results by adding statistical and example-based approaches

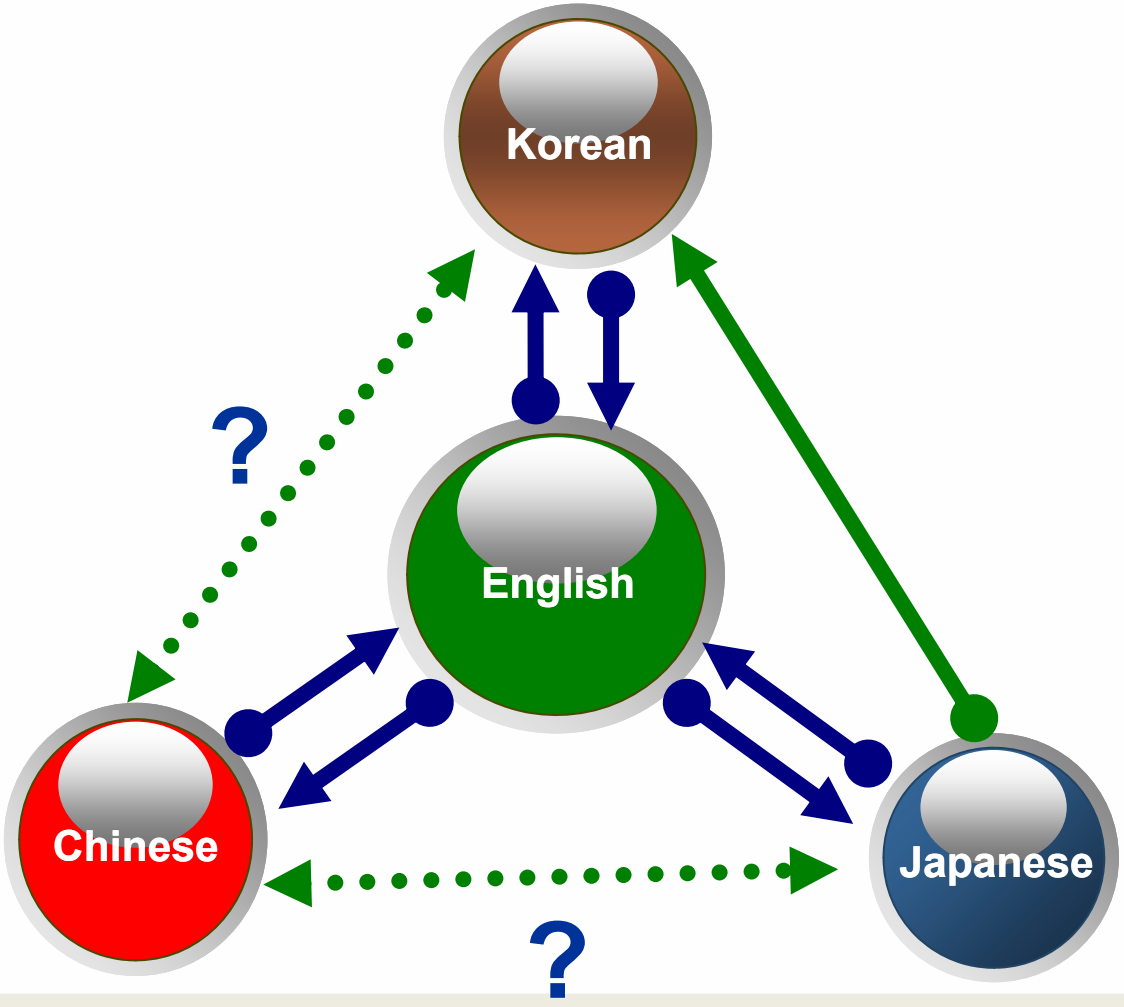
Towards other language pairs

- First priority: Japanese and Korean to Chinese
- Other language pair in the future

Supporting SIPO's Collaboration

- IP5-MMT Project
- SIPO's cooperation with other IP offices
- R&D Organizations

MT from and into Chinese: English as a Pivot?



English as a Pivot: Example 1

Route	Translation results
Source (Japanese Abstract)	光吸収性反射防止構造体、それを備えた光学ユニット及びレンズ鏡筒ユニット、並びにそれらを備えた光学装置 （修正有） 【課題】反射が十分に抑制された光吸収性反射防止構造体を提供する。 【解決手段】レンズ鏡筒の内周面10は入射光の波長よりも大きな表面粗さの粗面に形成されており、内周面10には、入射光の波長以下の周期で規則的に配列された複数の微小凹凸部11が形成されている。微小凹凸部11は、例えば、略円錐状（頂部が面取り又はR面取りされていてもよい）、角錐台状、線状（断面形状が、三角形状、台形状、矩形状等）の凹部又は凸部でもよい。
Via Japanese-English-Chinese	吸收光防反射结构、光学装置和装备有同样透镜镜筒单元和光器件，其装备有它们 摘要： 被解决的问题：提供一次吸收光防反射结构其中反射令人满意地被压制。 解决方案：一个透镜镜筒的内部的圆周表面10作为具有一个大于入射光的波长表面粗度的一个粗糙表面被形成和在内部的圆周表面10上，定期地由一个比入射光的波长更小的周期排列成的多个粉末崎岖的部件11。每一崎岖的第11部分的粉末可以包括，例如，一凹部部分或一突起部分是基本上锥形(顶点部分可以是倒角或R刻槽的)，锥体或线性的(横断面形状可以是三角的，不规则四边形，矩形的等)
Direct Japanese-Chinese	光吸收性反射防止构造体，具备其光学单元及透镜透镜电子管单元，以及具备其光学装置（修正有） 【课题】反射提供被充分抑制的光吸收性反射防止构造体。 【解决办法】透镜镜筒内周面10被形成比入射光的波长更大的表面粗糙度的粗面，在内周面10中，形成用入射光波长以下的周期被有规则的排列的复数的微小凹凸部11。微小凹凸部11，例如，近似圆锥状，(顶部也可以倒角或被R倒角)，角锥台状，线状(断面形状，三角形状，梯形状，进行矩形状等)的凹部或凸部。

English as a Pivot: Example 2

Route	Translation results
Source (Japanese Abstract)	測位装置、測位装置の制御方法、その制御プログラム及び記録媒体 【課題】信頼性を有し、かつ、精度の高い位置を出力することができる測位装置等を提供すること。 【解決手段】測位衛星からの信号である衛星信号に基づいて、測位を行う測位装置20であって、参照位置Pを保持する位置保持手段と、参照位置Pが静止条件Bを満たすか否かを判断する静止条件判断手段と、静止条件Bを満たす参照位置Pと、測位によって算出した現在の測位位置Pgを平均化して平均位置Pavを算出する平均位置算出手段と、平均位置Pavを出力する位置出力手段と、平均位置Pavを参照位置Pとして位置保持手段に格納する位置格納手段と、を有する。
Via Japanese-English-Chinese	标题：定位器、定位器的调节方法、其控制程序，和记录媒介 被解决的问题：提供一个定位器等，其可以输出一个位置高精度，与可靠性一起。 解决方案：定位器20执行一次定位，基于一个卫星信号从一个定位卫星。装置包括用于保持一个参考位置P的一个位置支持手段;静态条件决定方式，其决定是否参考位置P满足停止条件B;一个平位置计算装置，其计算平位置Pav，与满足通过定位计算的停止条件B和现在位置位置Pg的参考位置P的平均一起;输出平位置Pav的一个位置输出装置;一个位置存储器装置，其用于在位置支持手段中存储作为参考位置P的平位置Pav。
Direct Japanese-Chinese	测定位置装置，测定位置装置的控制方法，该控制程序及进行记录媒体 【课题】提供一种测定位置装置等，具有可靠性，且，能输出精度高的位置测定位置装置等。【解决办法】根据从定位卫星来的信号的卫星信号，进行测定位置测定位置装置20，具有保持引用位置P的位置保持手段、及判断引用位置P是否满足静止条件B的静止条件判断手段、及满足静止条件B的引用位置P、及由测定位置平均化算出现在的测定位置位置Pg，计算出平均位置Pav的平均位置计算方法、及输出平均位置Pav的位置输出手段、及作为引用位置P在位置保持手段存储平均位置Pav的位置存储手段。

Possible Benefits of Using English as a Pivot

Expected facilitation for MT system development and deployment

- MT engines to and from English already available
- English translations in JPO, KIPO and SIPO are pooled
- Technical terms are first written in English in many cases
- The approach already proved to be effective by the academic field

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

Thank you !