

# Overview of the Patent Translation Task at the NTCIR-8 Workshop

Atsushi Fujii (Tokyo Tech, Japan)

Masao Utiyama (NICT, Japan)

Mikio Yamamoto (Univ. of Tsukuba, Japan)

Takehito Utsuro (Univ. of Tsukuba, Japan)

Terumasa Ehara (Yamanashi Eiwa College)

Hiroshi Echizen-ya (Hokkai-Gakuen Univ.)

Sayori Shimohata (Oki Co., Ltd.)

# Motivation

- Systematic evaluation in NLP / IR is crucial
  - Large Test Collections are needed
- We have produced test collections for patent retrieval at NTCIR since 2001
  - What is the next task for patent information?

# History of Patent IR at NTCIR

- NTCIR-3 (2001-2002)

- Technology survey
  - Applied conventional IR problems to patent data

2 years of JPO patent applications \*

\* JPO = Japan Patent Office

- NTCIR-4 (2003-2004)

- Invalidity search
  - Addressed patent-specific IR problems

5 years of JPO patent

Both document sets were published in 1993-2002

- NTCIR-5 (2004-2005)

- Enlarged invalidity search

10 years of JPO patent applications

- NTCIR-6 (2006-2007)

- Added English patents

10 years of USPTO patents granted \*\*

\*\* USPTO = US Patent & Trademark Office <sup>3</sup>

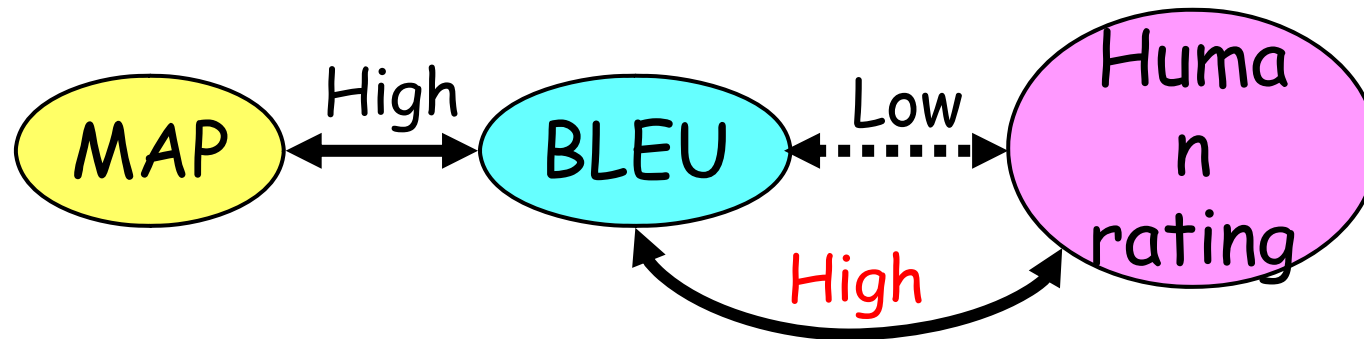
# PATMT at NTCIR-7 (2007-2008)

- Patent machine translation (MT) is realistic
  - **Parallel corpus** can potentially be produced from JPO / USPTO patent document sets
  - **Decoders** for Statistical MT are available and easily trained by our parallel corpus
- Participants = research groups
  - Any types of MT can be used:
    - Statistical MT (**SMT**)
    - Rule-based MT (**RBMT**)
    - Example-based MT (**EBMT**)
- Utility of patent MT
  - Cross-lingual patent retrieval
  - Filing patent applications in foreign countries

Important from science, engineering, & industry points of view

# Findings at NTCIR-7





- Which MT method was effective?
  - **BLEU**: Phrase-based SMT
  - **Human rating**: Rule-based MT
- Correlation b/w evaluation measures



with multiple reference translations (RTs)

- MT for regular sentences was effective for translating patent claims

# PATMT at NTCIR-8

- Larger document sets 
  - 15 years of JPO/USPTO patent documents
- We miss human rating & multi RTs for BLEU
- Subtasks 
  - Machine translation
  - Cross-lingual IR **Cancelled (no participation)** 
    - MT results by other participants can be used
  - Evaluation 
    - Developing automatic evaluation methods highly correlated with human rating

Overview for Evaluation subtask  
will be given by Prof. Ehara later

# Contents

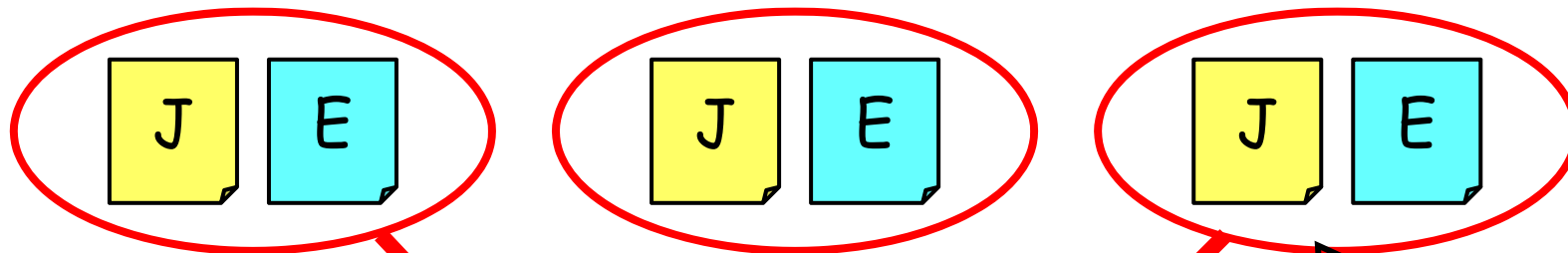
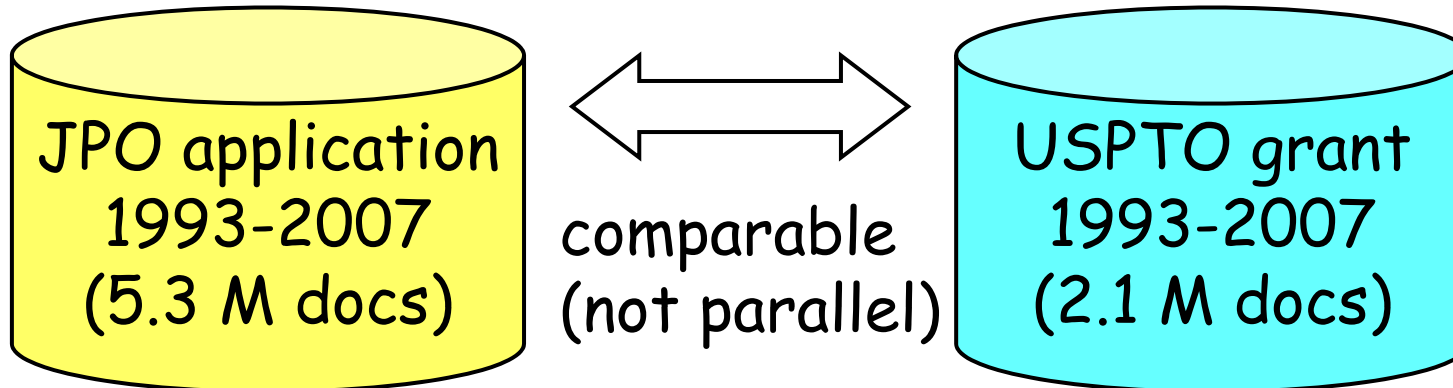
1. Method to produce parallel corpus
2. Method to evaluate participating MT systems
3. Results of Formal run

# Contents

1. Method to produce parallel corpus
2. Method to evaluate participating MT systems
3. Results of Formal run



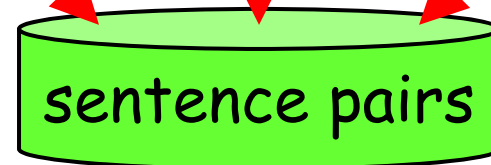
# Producing parallel corpus



Sentence alignment method  
[Utiyama and Isahara, 2007]

Targeting  
"Background" and  
"Description"

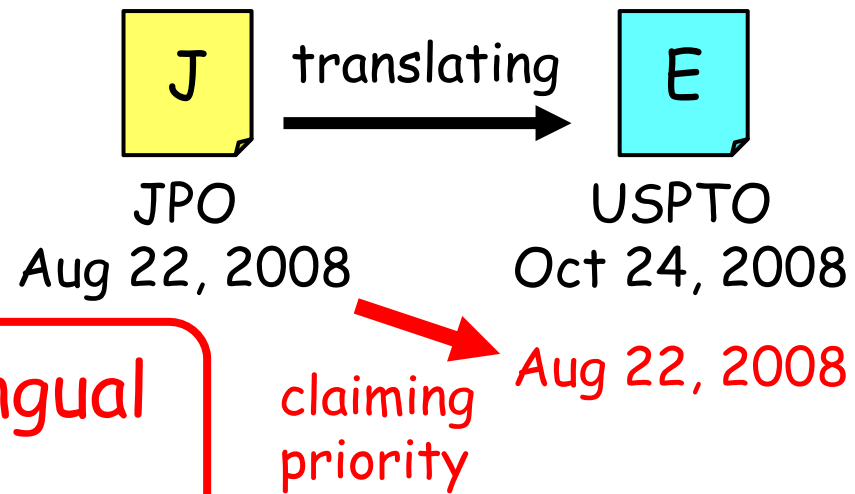
patent family  
patent set for  
same invention



parallel (alignment accuracy = 90%)<sup>9</sup>

# Patent families

- Member patents often claim "priority" under the Paris Convention
  - Related patents can easily be identified by priority numbers
- Merit of priority-based patent families
  - The application date is retroactive to the original date
  - First-to-file system in many countries



Free (or inexpensive) bilingual corpora are growing!!!

# Example of patent family

Invention related to Microactuators



Patent family can be identified by priority number

(51) Int. Cl.  
F 1 6 1 3 1 / 0 0  
G 0 5 1 7 / 0 3  
H 0 1 1 2 1 / 3 0 6

(21) 出願番号 特願平7-233220  
(22) 出願日 平成7年(1995)年11月10日  
(31) 優先権主張番号 2 9 5, 1 2 7  
(32) 優先日 1994年8月24日  
(33) 優先権主張国 米国 (U.S.)

(71) 出願人 59000400  
ヒューレット・パッカード・カンパニー  
アメリカ合衆国カリフォルニア州パロアルト  
ハノーバー・ストリート 3000  
(72) 発明者  
クリストファー・シー・ベティ  
アメリカ合衆国ペンシルバニア州ランデン  
バーグ、ビルズ・ウェイ 23  
(72) 発明者  
ジェイムズ・ダブリュー・ペイカー  
アメリカ合衆国メリーランド州エルクト  
ン、カーター・ロード 110  
(74) 代理人 弁理士 上野 英夫

(54) 【発明の名称】 マイクロアクチュエータ  
(57) 【要約】  
【課題】 加熱構造を備えるマイクロアクチュエータにおいて、熱エネルギーを効率的に伝達する超小型バルブの形態をなすマイクロアクチュエータであり、サーマルアクチュエータによって選択的に駆動される熱膨張部材を有し、これが駆動されることによって熱エネルギーを生成する第1基板と、対向する第1、第2主要面を有する第2基板よりなる。第2基板が第1主要面で第1基板に取り付けられる。第2の主要面は第2基板が支持体に取り付けられると絶縁セルを固定し、これによってマイクロアクチュエータの熱容量を減少させ、第1基板を支持体から熱断絶する。

(11) Patent Number: 5,529,279  
(45) Date of Patent: Jun. 25, 1996

(54) THERMAL ISOLATION STRUCTURES FOR MICROACTUATORS  
(75) Inventors: Christopher C. Beatty, Landsberg, Pa.; James W. Baker, Elton, Md.  
(73) Assignee: Hewlett-Packard Company, Palo Alto, Calif.  
(21) Appl. No.: 295,127  
(22) Filed: Aug. 24, 1994  
(51) Int. Cl. F16K 31/02; F03G 7/06  
(52) U.S. Cl. 251/11; 251/129.01; 251/368; 60/528; 60/529  
(58) Field of Search 251/11; 251/129.01; 251/368; 60/528; 60/529  
(56) References Cited  
5,581,624 O'Connor, et al.  
5,050,838 Beatty, et al.

5,058,856 10/1991 Gordon, et al.  
5,069,419 12/1991 Jernan  
5,161,774 11/1992 Engelsdorf et al. 251/11  
5,333,831 8/1994 Barth, et al.  
5,344,117 9/1994 Truh et al. 251/11

Primary Examiner—Kevin Lee  
Attorney Agent, or Firm—Mark Z. Dudley

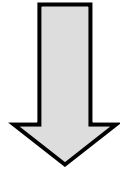
(57) ABSTRACT  
A microactuator preferably in the form of a microminature valve for controlling the flow of a fluid carried by a flow channel includes a first substrate having a thermally-actuated member selectively operated by a thermal actuator such that the first substrate thereby develops thermal energy, and a second substrate having opposed first and second major surfaces. The second substrate is attached to the first substrate at the first major surface. The second major surface defines an isolation cell for enclosing a volume when the second substrate is attached to the support to thereby reduce the thermal mass of the microactuator and to thermally isolate the first substrate from the support.

18 Claims, 16 Drawing Sheets

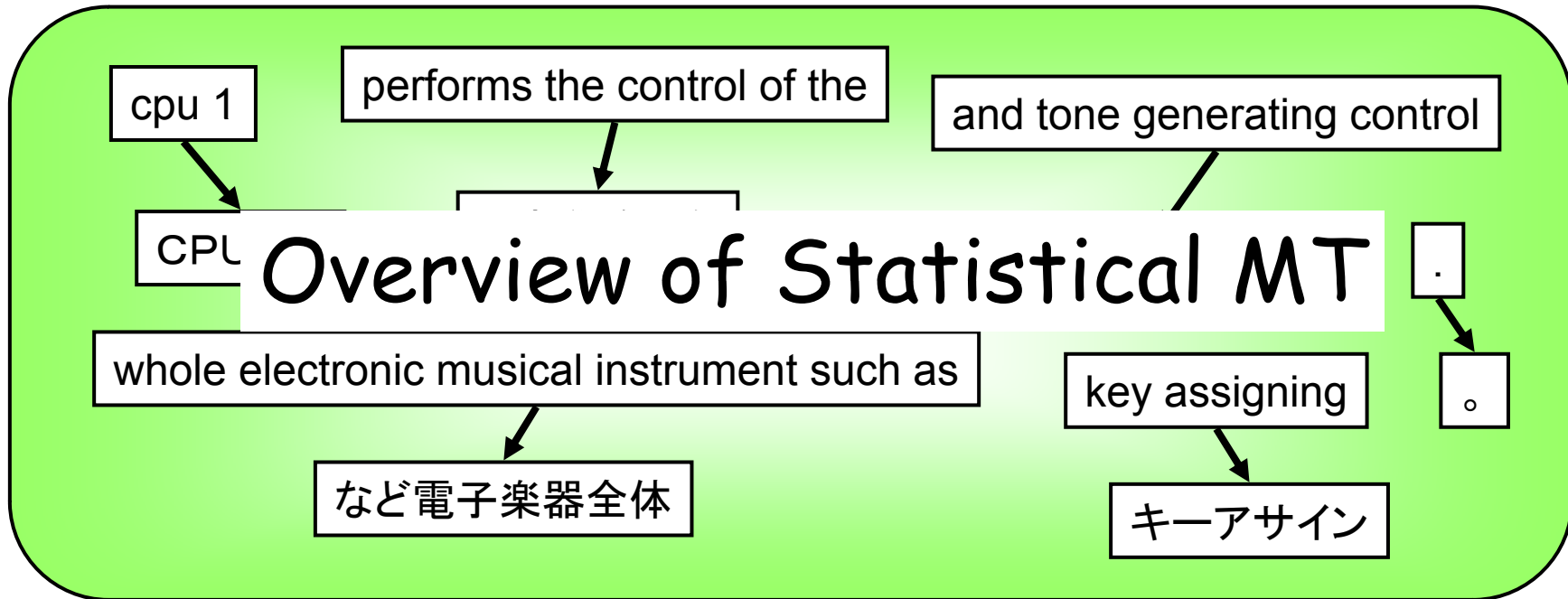
J-E sentence pairs can be extracted from corresponding fields

**English:** cpu 1 performs the control of the whole electronic musical instrument such as key assigning and tone generating control .

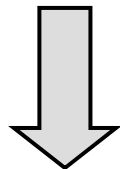
Step 1



- Translating Eng to Jpn on a word/phrase basis
- Eng-Jpn alignments are trained by parallel corpus



Step 2



- Reordering words to produce a fluent sentence
- Jpn word N-gram is trained by Jpn corpus

**Japanese:** CPU1はキーアサイン、発音制御など電子楽器全体の制御を行う。

# Contents

1. Method to produce parallel corpus
2. Method to evaluate participating MT systems
3. Results of Formal run

# Evaluation methods

- **Intrinsic evaluation**

- BLEU

- Reference translation (= aligned counterparts)

Same as existing MT WS  
e.g., NIST, ACL, & IWSLT

- **Extrinsic evaluation**

- Contribution to Cross-Lingual Patent Retrieval (CLPR)

- Invalidity search

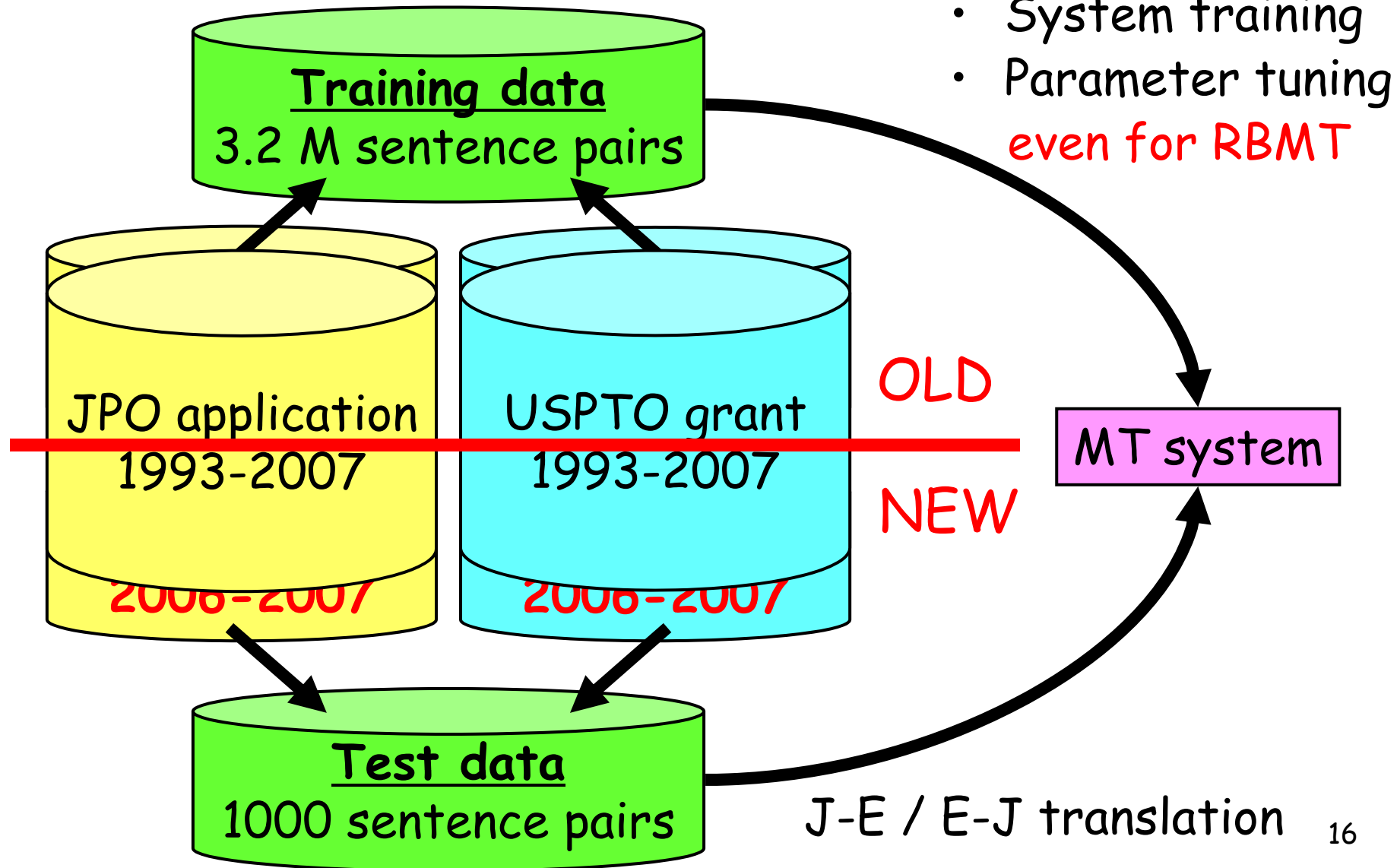
- BLEU

Distinctive feature  
of NTCIR

# BLEU: BiLingual Evaluation Understudy

- BLEU = Automatic evaluation measure for MT
  - Comparing MT and reference translation (RT)
- However, exact match b/w MT and RT is rare
  - Comparing MT and RT on an Ngram-by-Ngram basis
    - Ngram = N consecutive words in sentence
  - BLEU score = Geometric mean of Ngram-based scores
- More than one acceptable translation exist
  - For each test sentence, multiple RTs are required
  - However, in NTCIR-8 only single RT was used

# Intrinsic evaluation

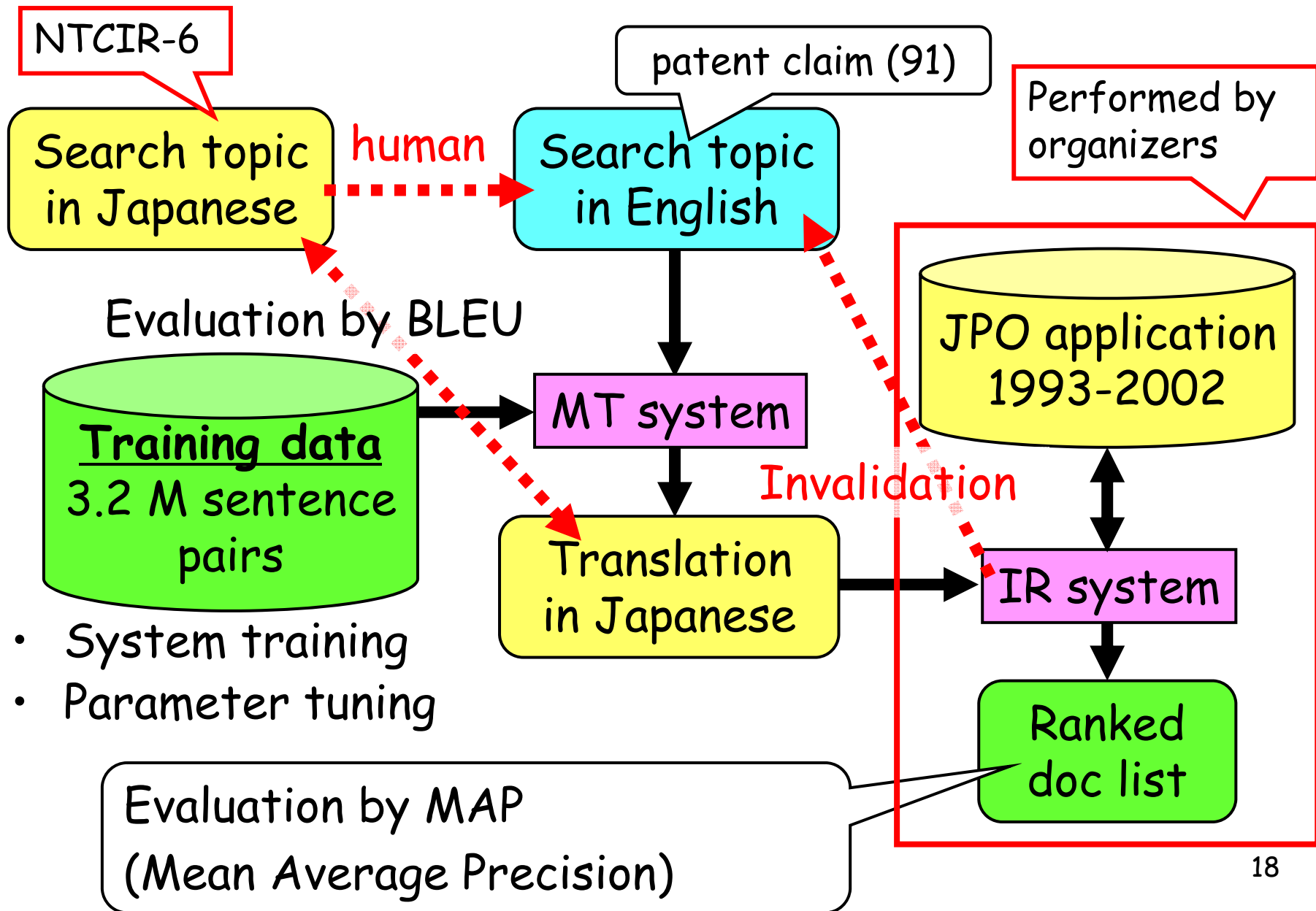




# Example J/E test sentences

- さらに、図4に示すように、システム全体を制御するホストコンピュータ46も通信ネットワーク47上に接続することによって、ホストコンピュータ46とプロセッサ44とを接続する専用線をなくすることができる。
- Further, by connecting the host computer 46, which controls the whole system, also onto the communication network 47 as shown in FIG. 4, the exclusive line for connecting the host computer 46 and the processor 44 to each other can be eliminated.
- また、圧縮機ユニット33にも、I/O変換部(図示せず)が搭載されている。
- An I/O conversion section (not shown) is mounted also on the compressor unit 33.

# Extrinsic evaluation



# Example search topic

- claim = long and complex noun phrase

An ultrasonic probe including an ultrasonic oscillator and an oscillator retaining member for retaining the ultrasonic oscillator provided within a flexible sheath, an ultrasonic transmitting medium filled around the ultrasonic oscillator to obtain an ultrasonic image by ultrasonic scanning, the flexible sheath having, at the tip of its cavity, an ultrasonic transmitting medium sealing member having a curved surface protruded to the oscillator holding member side located closer to this side in the sheath axial direction from this tip.

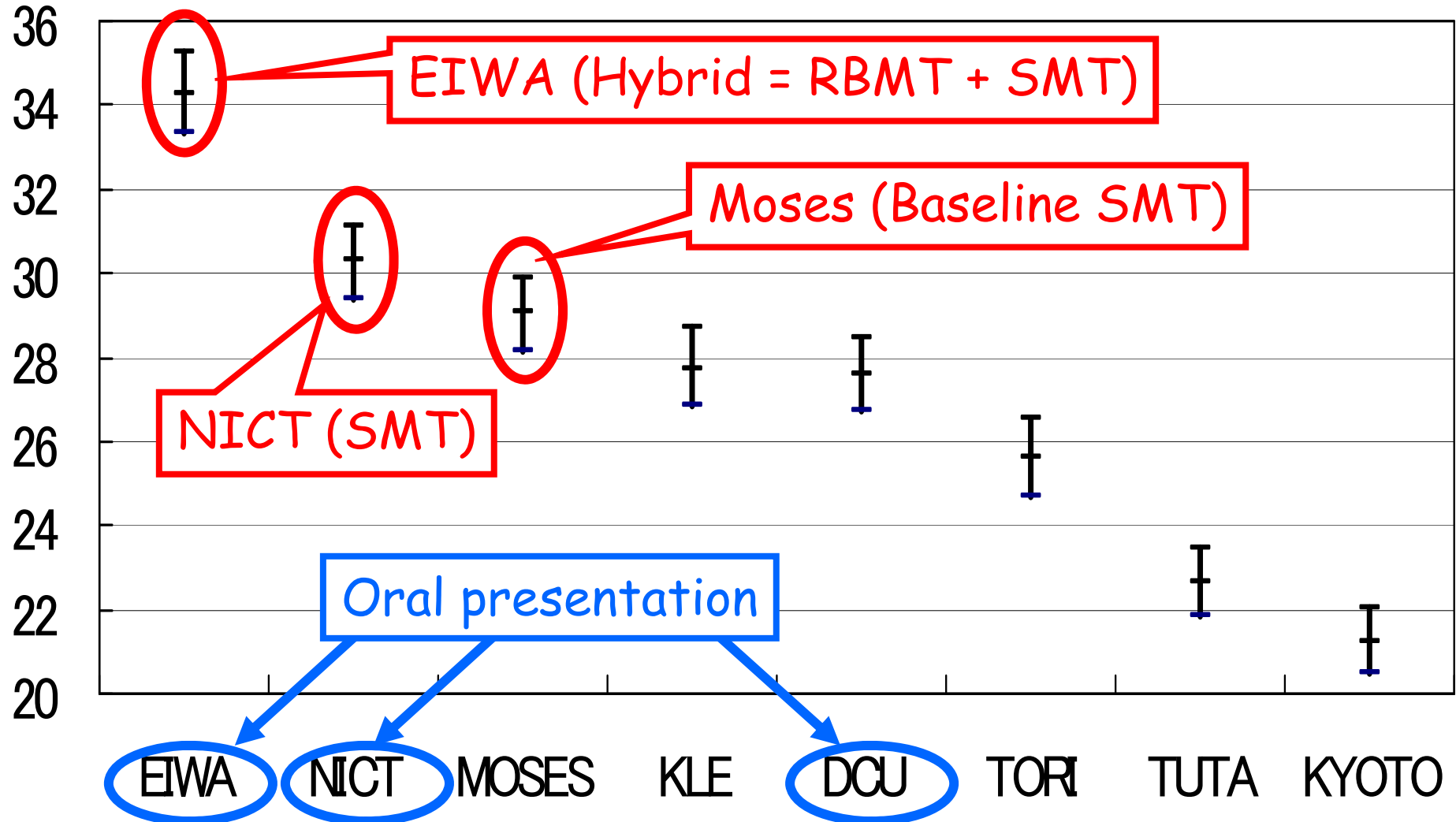


79 words

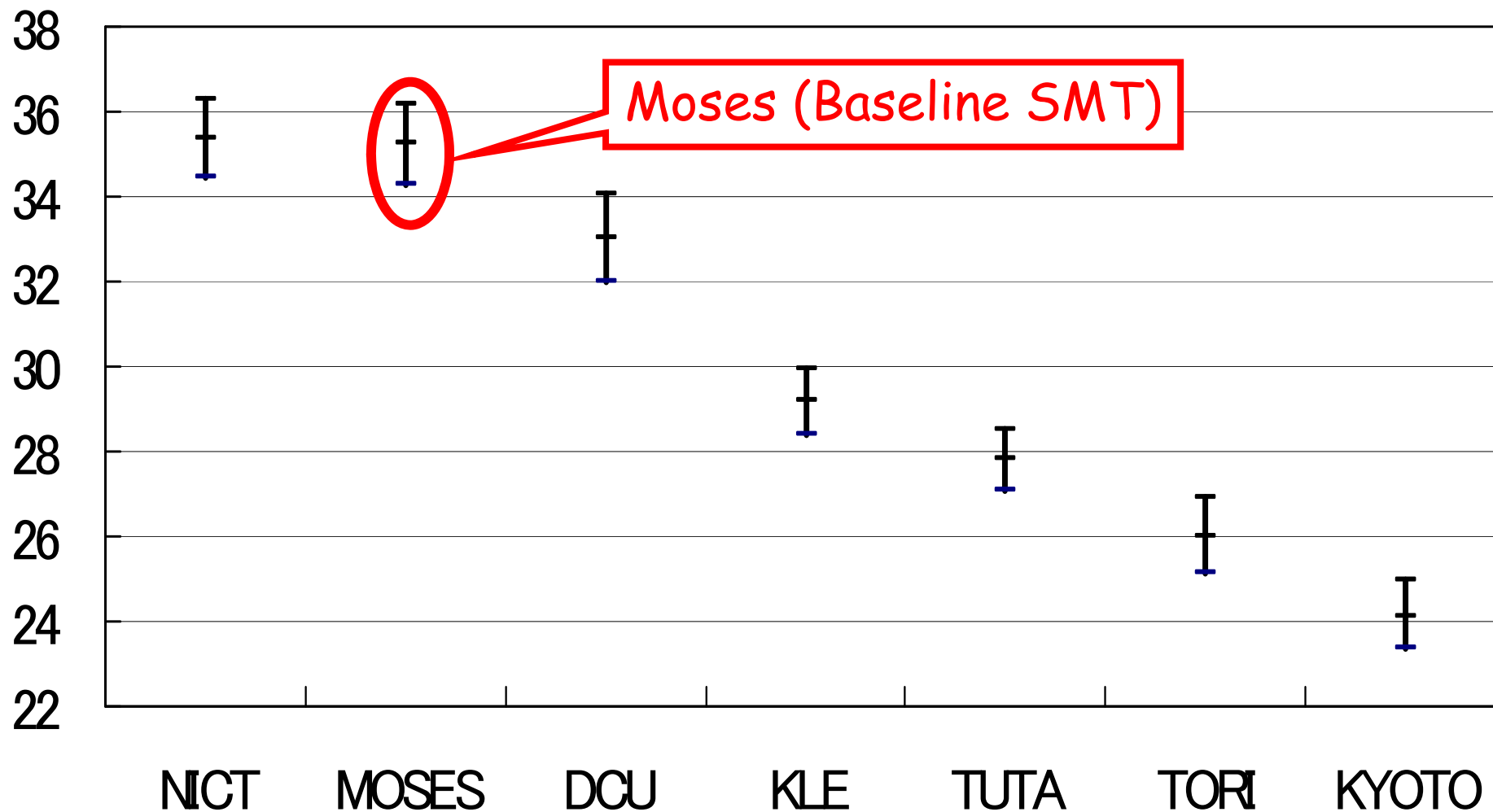
# Contents

1. Method to produce parallel corpus
2. Method to evaluate participating MT systems
3. Results of Formal run

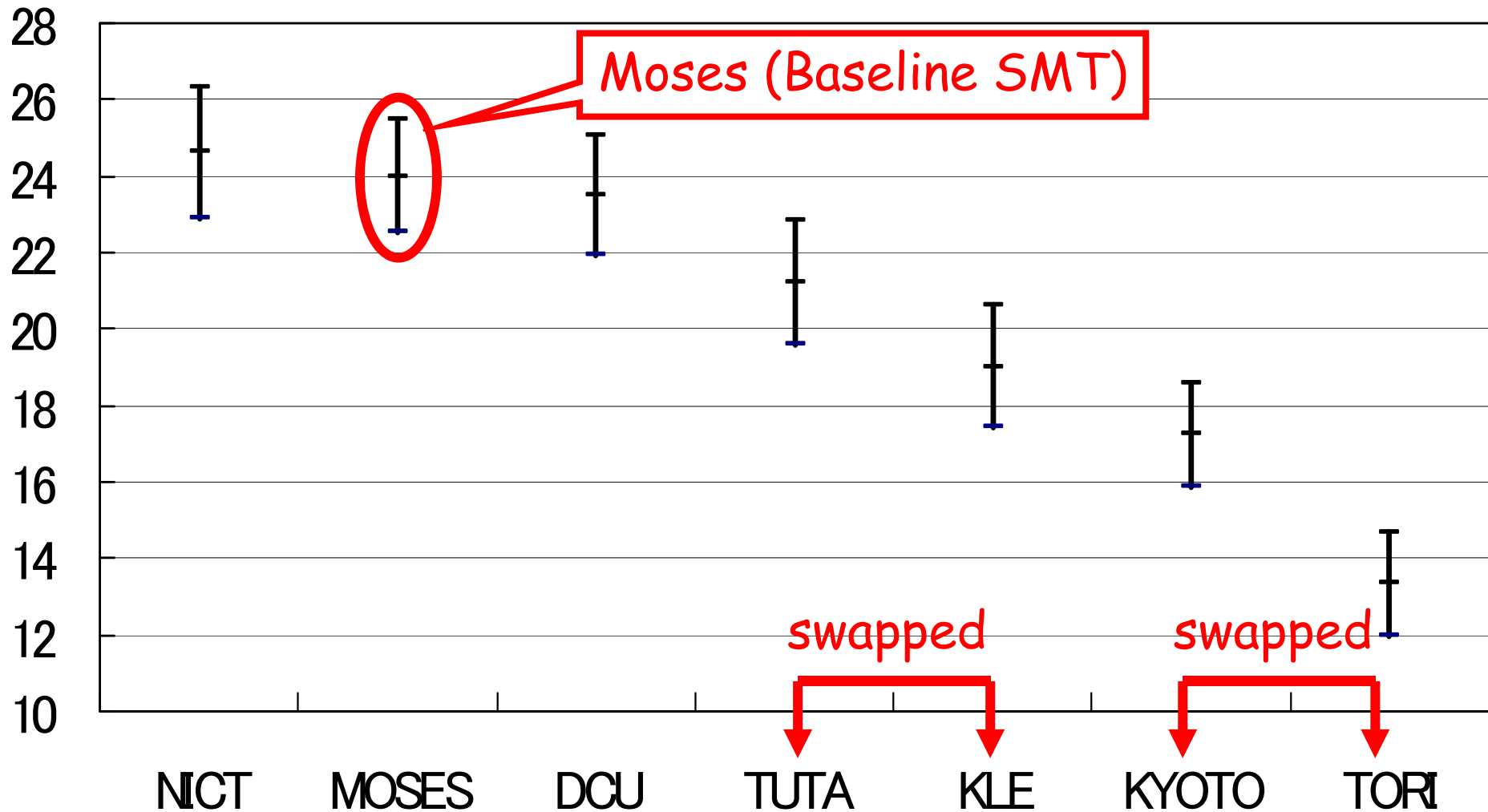
# Intrinsic J-E: BLEU with 95% confidence interval



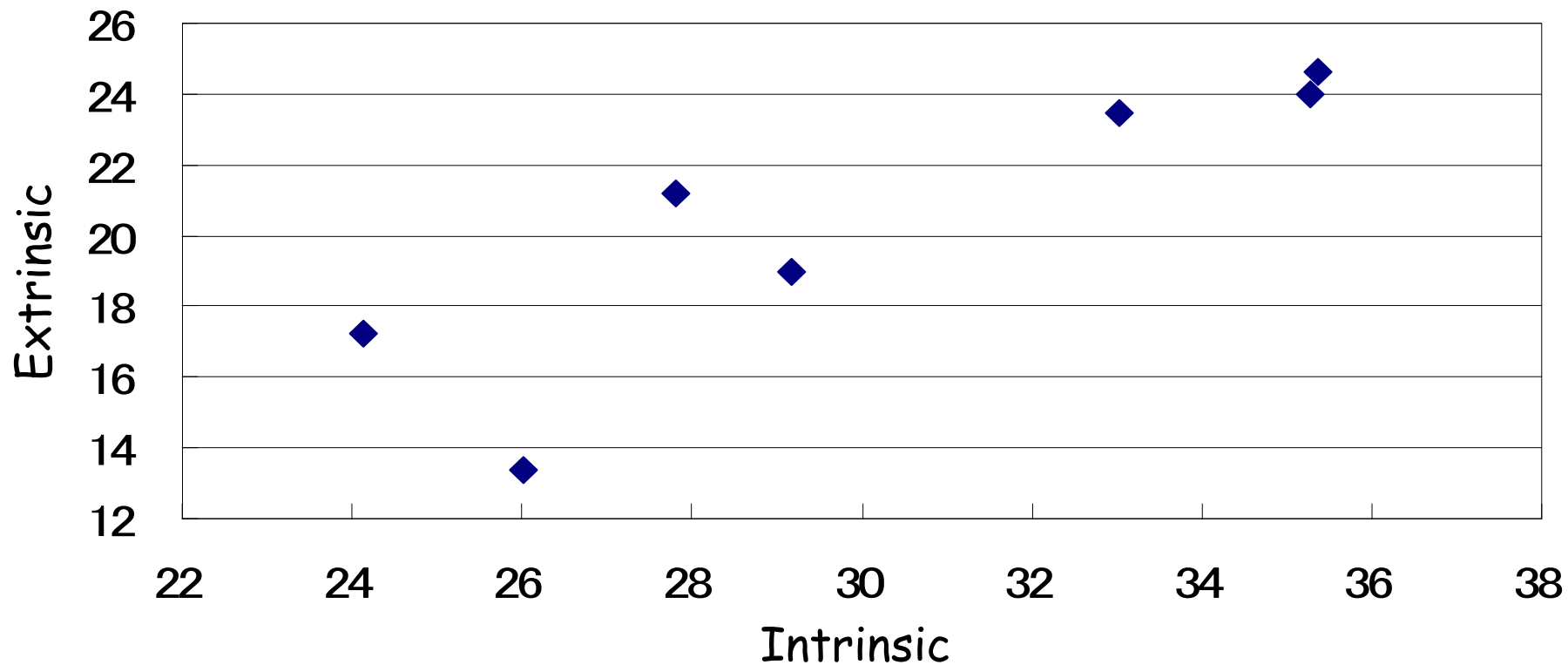
# Intrinsic E-J: BLEU with 95% confidence interval



# Extrinsic E-J: BLEU with 95% confidence interval



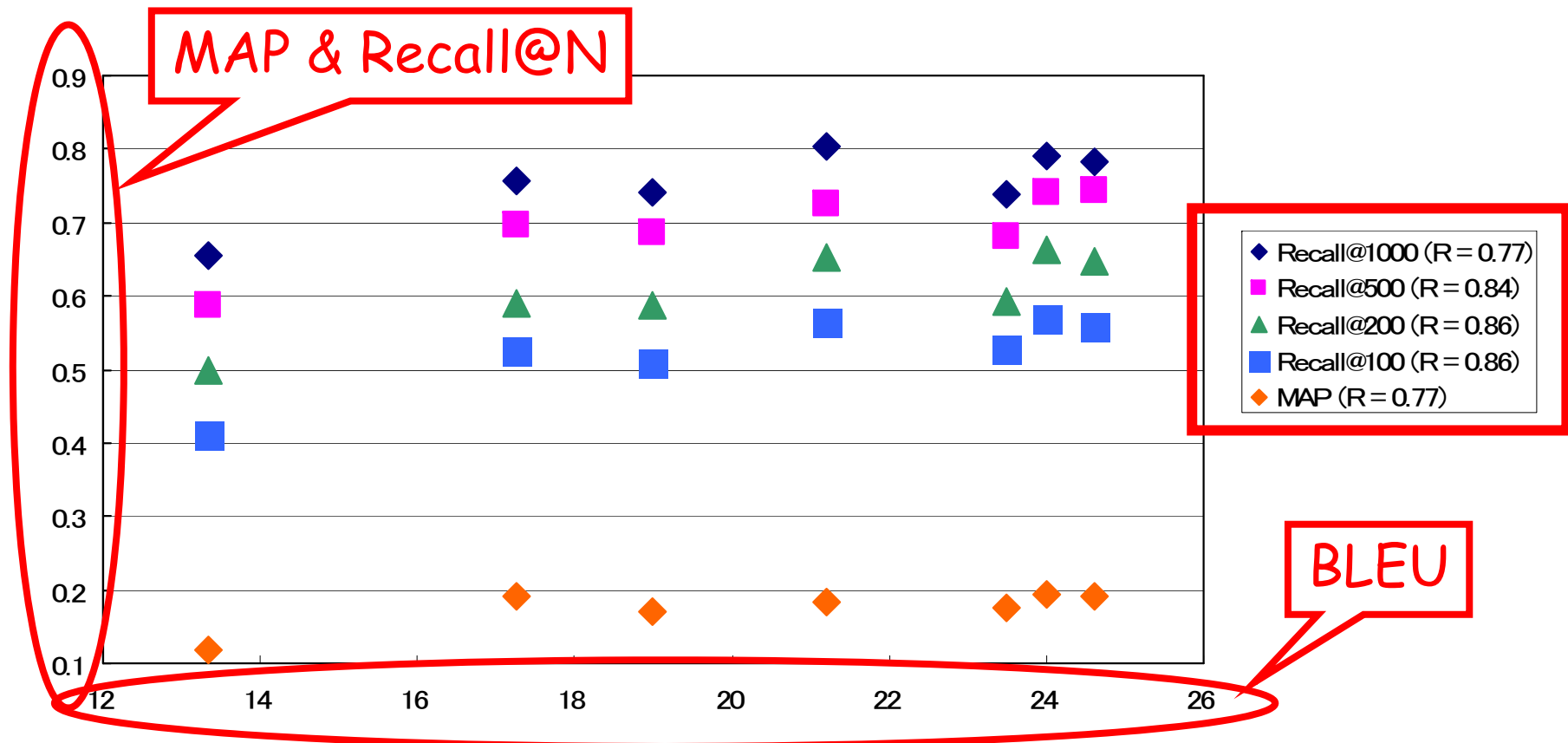
# Comparing BLEU: Int E-J& Ext E-J



- Correlation is high b/w Int and Ext ( $R = 0.87$ )
- MT for regular sentences was also effective for translating claims



# Extrinsic E-J: BLEU & IR measures



Recall@100,200 are highly correlated with BLEU (R = 0.86)

# Summary of Translation subtask

- Produced a large test collection for J/E patent MT
  - 15 years of JPO/USPTO documents
  - 3.2 M J/E aligned sentences for training
  - 1000 sentences for evaluation
- BLEU and IR measures were highly correlated, as in NTCIR-7
- A hybrid method (EIWA) substantially outperformed Moses for J-E MT

# Overview of the Patent Translation Task Evaluation Subtask (AE subtask)

Terumasa EHARA  
Yamanashi Eiwa College

# Aim of the subtask

- To improve automatic evaluation methods for machine translation accuracy
- To overcome the difference of automatic evaluation and human evaluation [Fujii, 2008]

[Fujii, 2008] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro: Overview of the Patent Translation Task at the NTCIR-7 Workshop, Proceedings of NTCIR-7 Workshop Meeting, Dec. 2008, Tokyo, Japan.

# Data 1

- The results of the NTCIR-7 patent translation task (only JE direction)
- Training data: NTCIR-7 patent translation task, dry run data
  - Source and reference data: 100 sentences
  - MT output data: 100 (sent.) × 11 (systems)
  - Human evaluation results (adequacy and fluency): 100 (sent.) × 11 (systems) × 3 (human raters)

# Data 2

- Test data: NTCIR-7 patent translation task, formal run data
  - Source and reference data: 100 sentences
  - MT output data: 100 (sent.) × 12 (systems)
  - Human evaluation result (adequacy and fluency):  
100 (sent.) × 12 (systems) × 3 (raters)

# Data sample

Source

- プリント機構は、感光体ドラム11を備えている。
- 定着ローラ39によって定着処理が施された転写紙は、図示しない排出機構を介して排出トレイ43上に送られる。

Reference

- the printing mechanism has a photoconductor drum 11 .
- the transfer paper and fixed by a fixing roller 39 thus processed is discharged through a discharge tray 43 is sent to a mechanism ( not shown ) .

Test (MT)

- the printing mechanism comprises the photosensitive drum 11 .
- the copy paper which has been subjected to fixing processing by the fixing roller 39 is fed onto a discharge tray 43 through a discharge mechanism ( not shown ) .

# Meta-evaluation measures

- Pearson's correlation coefficients between adequacy/fluency and automatic evaluation result
- Spearman's rank correlation coefficients between adequacy/fluency and automatic evaluation result



# Result

- AE subtask has only one participant

participant	Correlation coefficients to the adequacy data				Correlation coefficients to the fluency data			
	Pearson		Spearman		Pearson		Spearman	
	Avg	All	Avg	All	Avg	All	Avg	All
HCU-1	0.2992	0.2463	0.2712	0.2234	0.2608	0.2285	0.2486	0.2126

Avg : average value for the correlation coefficients for the 12 test systems

All : correlation coefficient for all data of the 12 test systems

# Remarks

- We plan to provide all human evaluation data composed in the NTCIR-7 and NTCIR-8 patent translation task.
- They will be provided from NII's Informatics Research Data Repository.