Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop

Isao Goto (NICT)
Ka Po Chow (Hong Kong Institute of Education)
Bin Lu (City Univ. of Hong Kong / Hong Kong Institute of Education)
Eiichiro Sumita (NICT)
Benjamin K. Tsou (Hong Kong Institute of Education / City Univ. of Hong Kong)

Table of Contents

- Motivation and Goals
- Previous tasks and comparison
- Notable Findings at NTCIR-10
- PatentMT at NTCIR-10
- Intrinsic Evaluation
- Patent Examination Evaluation
- Summary

Motivation

- There is a significant **practical need** for patent translation.
  - to understand patent information written in foreign languages
  - to apply for patents in foreign countries
- Patents constitute one of the **challenging domains** for MT.
  - Patent sentences can be quite **long** and contain **complex structures**

Goals of PatentMT

- To develop **challenging** and **significant practical** research into patent machine translation.
- To **investigate** the **performance** of state-of-the-art machine translation systems in terms of patent translations involving Chinese, Japanese, and English.
- To **compare** the effects of **different methods** of patent translation by applying them to the same test data.
- To **explore practical MT performance** in appropriate fields for patent machine translation.
- To **create** publicly-available **parallel corpora of patent documents** and human evaluations of MT results for patent information processing research.
- To **drive machine translation research**, which is an important technology for cross-lingual access of information written in unknown languages.
- The ultimate goal is **fostering scientific cooperation**.

Findings of Previous Patent Translation Tasks

NTCIR-7	**Human evaluation**	**RBMT** was better than **SMT** for **JE** and **EJ**.
	CLIR evaluation	SMT was better than RBMT for EJ word selection.
NTCIR-8	Automatic evaluation	A hybrid system (RBMT with statistical post edit) achieved the best score for JE.
NTCIR-9	**Human evaluation**	**SMT** caught up with **RBMT** for **EJ** **RBMT** was better than **SMT** for **JE** **SMT** was better than **RBMT** for **CE**

Comparison of NTCIR-7, 8, 9, and 10

	NTCIR-7	NTCIR-8	NTCIR-9	**NTCIR-10**
Language	**Japanese to English (JE)** **English to Japanese (EJ)**		**Chinese to English (CE)** **Japanese to English (JE)** **English to Japanese (EJ)**	
Intrinsic evaluation by human	**Adequacy** **Fluency**	No human evaluation	**Adequacy** **Acceptability**	
Other evaluations	**CLIR**	**CLIR**	No other evaluation	•**Patent Examination Evaluation** New •Chronological Evaluation •Multilingual Evaluation
Number of participants	15	8	21	21

Notable Findings at NTCIR-10

- The best MT systems for JE and CE were **useful** for **patent examination**

- The top **SMT** **outperformed** the top-level RBMT for **EJ** patent translation.

- **RBMT** is still **better** than SMT for **JE**, but the translation quality of the top **SMT** for JE has greatly improved.

PatentMT at NTCIR-10

Four Types of Evaluations

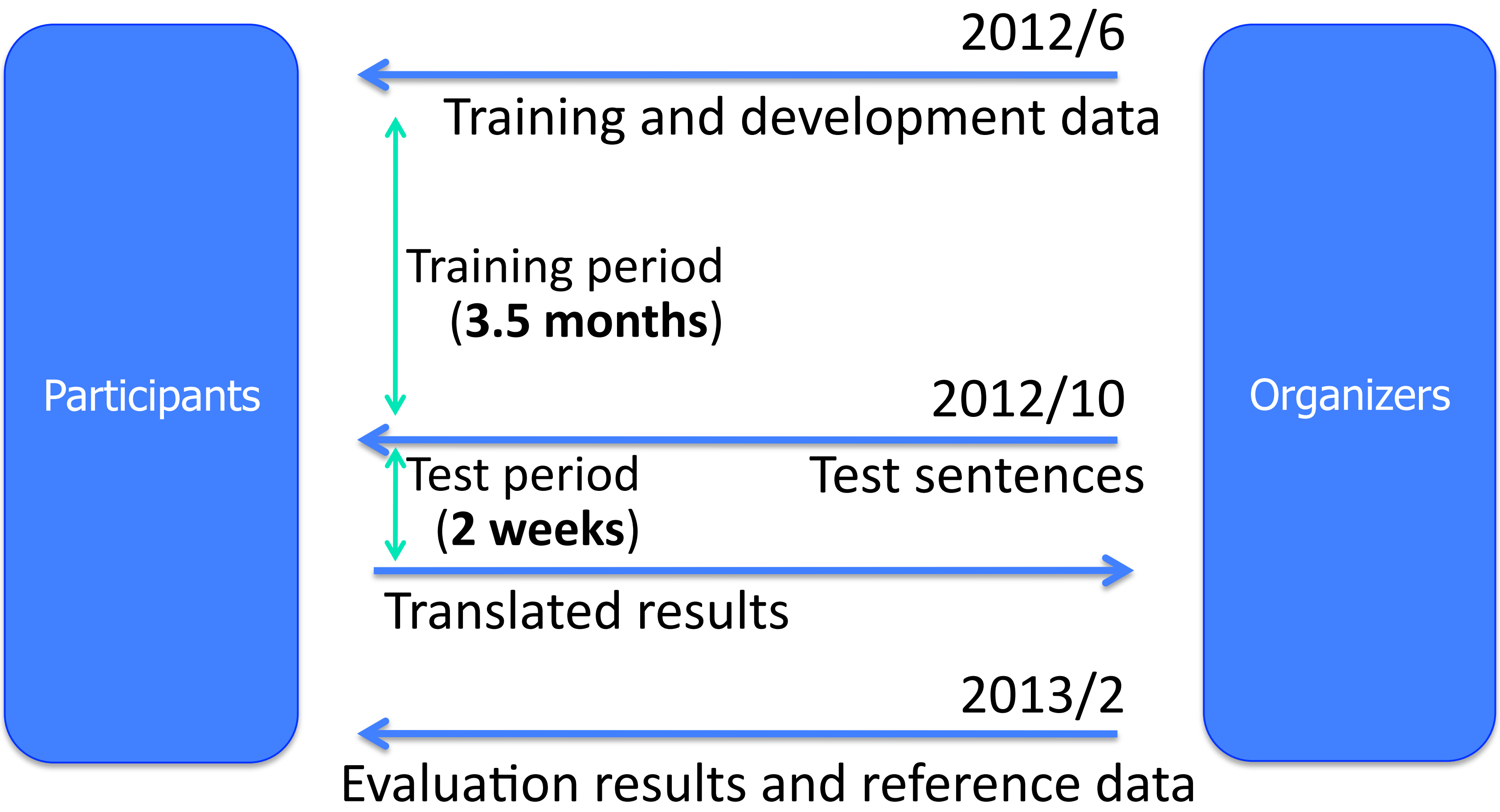
Evaluation Type	Description	Subtask
Intrinsic Evaluation (IE)	The quality of translated sentences were evaluated. Human evaluation: Adequacy and Acceptability	All
Patent Examination Evaluation (PEE)	New: The usefulness of machine translation for patent examination was evaluated.	CE/JE
Chronological Evaluation (ChE)	New: A comparison between NTCIR-10 and 9 to measure progress over time, using the NTCIR-9 test sets	All
Multilingual Evaluation (ME)	New: A comparison of CE and JE translations using the same English references to see the source language dependency.	CE/JE

Provided Data

Training	CE	**1 million** patent **parallel** sentence pairs
		Over 300 million patent monolingual sentences in English
	JE	Approximately **3.2 million** patent **parallel** sentence pairs
		Over 300 million patent monolingual sentences in English
	EJ	Approximately **3.2 million** patent **parallel** sentence pairs
		Over 400 million patent monolingual sentences in Japanese
Development	All	2,000 patent description parallel sentence pairs
Test (IE)	All	**2,300 patent description sentences** (New)
Test (PEE)	CE/JE	**29 patent documents** (New)
Test (ChE)	All	2,000 patent description sentences
Test (ME)	CE/JE	**2,000 patent description sentences** (New)

The periods for the training and test data (IE, ChE, ME) are different
(Training data: 2005 or before, Test data: 2006 or later)

Flow and Schedule



Participants

Group ID	Organization	Nationality	CE	JE	EJ
JAPIO	Japan Patent Information Organization (Japio)	Japan		✓	✓
KYOTO	Kyoto University	Japan		✓	✓
NTITI	NTT Corporation / National Institute of Informatics	Japan		✓	✓
OKAPU	Okayama Prefectural University	Japan		✓	
TORI	Tottori University	Japan		✓	
TSUKU	University of Tsukuba	Japan			✓
EIWA	Yamanashi Eiwa College	Japan	✓	✓	✓
FUN-NRC	Future University Hakodate / National Research Council Canada	Japan/Canada		✓	✓
BUAA	BeiHang University, School of Computer Science & Engineering	P.R. China	✓		
BJTUX	Beijing Jiaotong University	P.R. China	✓	✓	✓
ISTIC	Institute of Scientific and Technical Information of China	P.R. China	✓	✓	✓
SJTU	Shanghai Jiao Tong University	P.R. China	✓		
TRGTK	Torangetek Inc.	P.R. China	✓	✓	✓
MIG	Department of Computer Science, National Chengchi University	Taiwan	✓		
HDU	Institute for Computational Linguistics, Heidelberg University	Germany	✓	✓	
RWTH	RWTH Aachen University	Germany	✓	✓	
RWSYS	RWTH Aachen University / Systran	Germany/ France	✓		
DCUMT	Dublin City University	Ireland			✓
UQAM	UQAM	Canada		✓	✓
BBN	Raytheon BBN Technologies	USA	✓		
SRI	SRI International	USA	✓		

Baseline Systems

SYSTEM-ID	System	Type	CE	JE	EJ
BASELINE1	Moses hierarchical phrase-based SMT system	SMT	✓	✓	✓
BASELINE2	Moses phrase-based SMT system		✓	✓	✓
RBMTx	The Honyaku 2009 premium patent edition	RBMT		✓	✓
RBMTx	ATLAS V14			✓	✓
RBMTx	PAT-Transer 2009			✓	✓
ONLINE1	Google online translation system	SMT	✓	✓	✓

- These commercial RBMT systems are well known for their language pairs.
  - The SYSTEM-IDs of the commercial RBMT systems are anonymized.
- The translation procedures for BASELINE1 and 2 were published on the PatentMT web page.

Intrinsic Evaluation (IE)

Human Evaluation for IE

- Evaluation methods
  - Human evaluations were carried out by **paid evaluation experts**.
  - **300 sentences** were evaluated per system.
    - Number of evaluators: three.
    - Each evaluator evaluated 100 sentences per system.
- Evaluation criteria
  - Adequacy
    - The main purpose is **comparison between the systems**.
    - At least **all of the first priority submissions** before the deadline were evaluated.
  - Acceptability
    - The main purpose is to clarify the **percentage** of translated sentences whose **source sentence meanings can be understood**.
    - Due to budget limitations, only selected systems were evaluated.

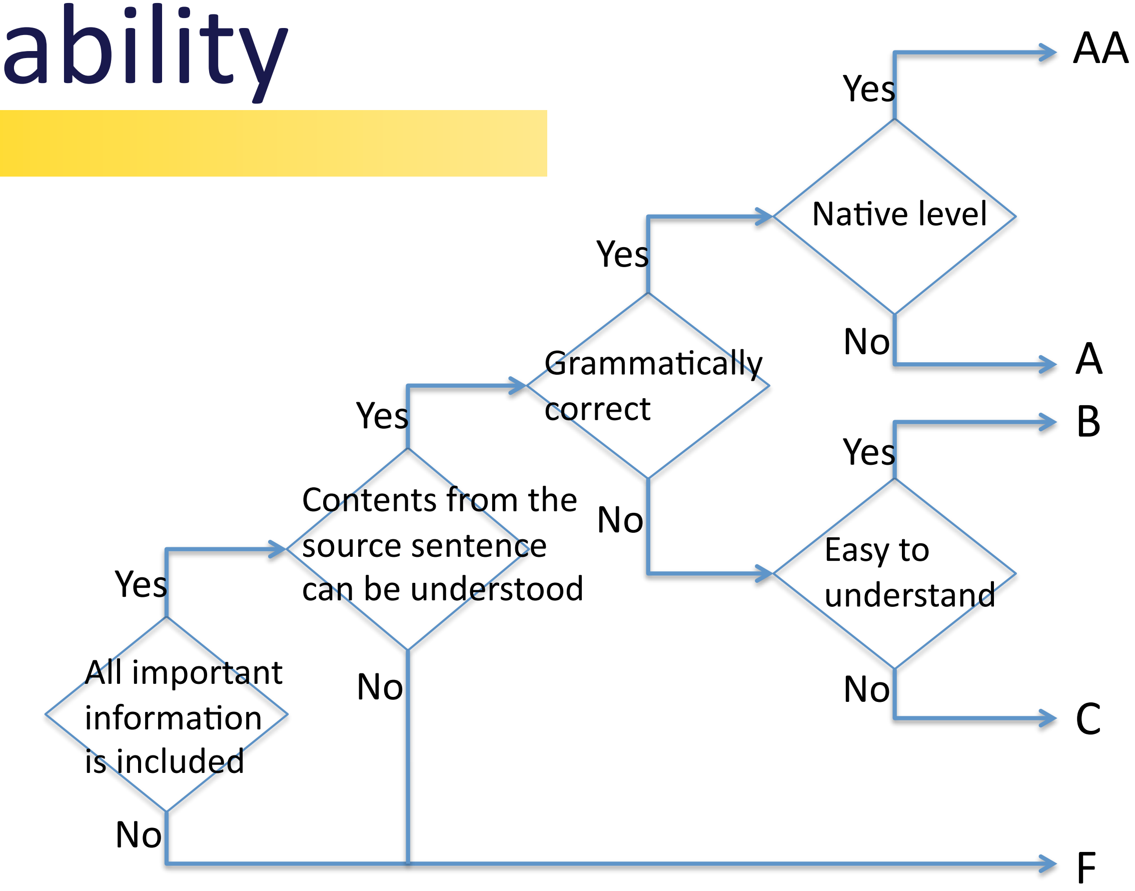
Adequacy

- The criterion of adequacy used for this evaluation
  - A 5-scale (1 to 5) evaluation.
  - **Clause**-level meanings were considered.

- Characteristics
  - This evaluation is effective for system comparison.
  - It is **unknown** what **percentage** of the translated sentences express the **correct meaning of the source sentence**.
    - This is because the scoring criterion for scores of between 2 to 4 is unclear.

Acceptability

- Criterion


AA
Yes
Native level
Yes
No
A
Grammatically correct
Yes
B
Yes
Contents from the source sentence can be understood
No
Easy to understand
Yes
All important information is included
No
No
C
No
F

- Characteristics
  - This evaluation aims more at **practical** evaluation than adequacy.
  - What **percentage** of the translated sentences express the **correct meaning of the source sentence** is known.

    (The rate of C-rank and above)

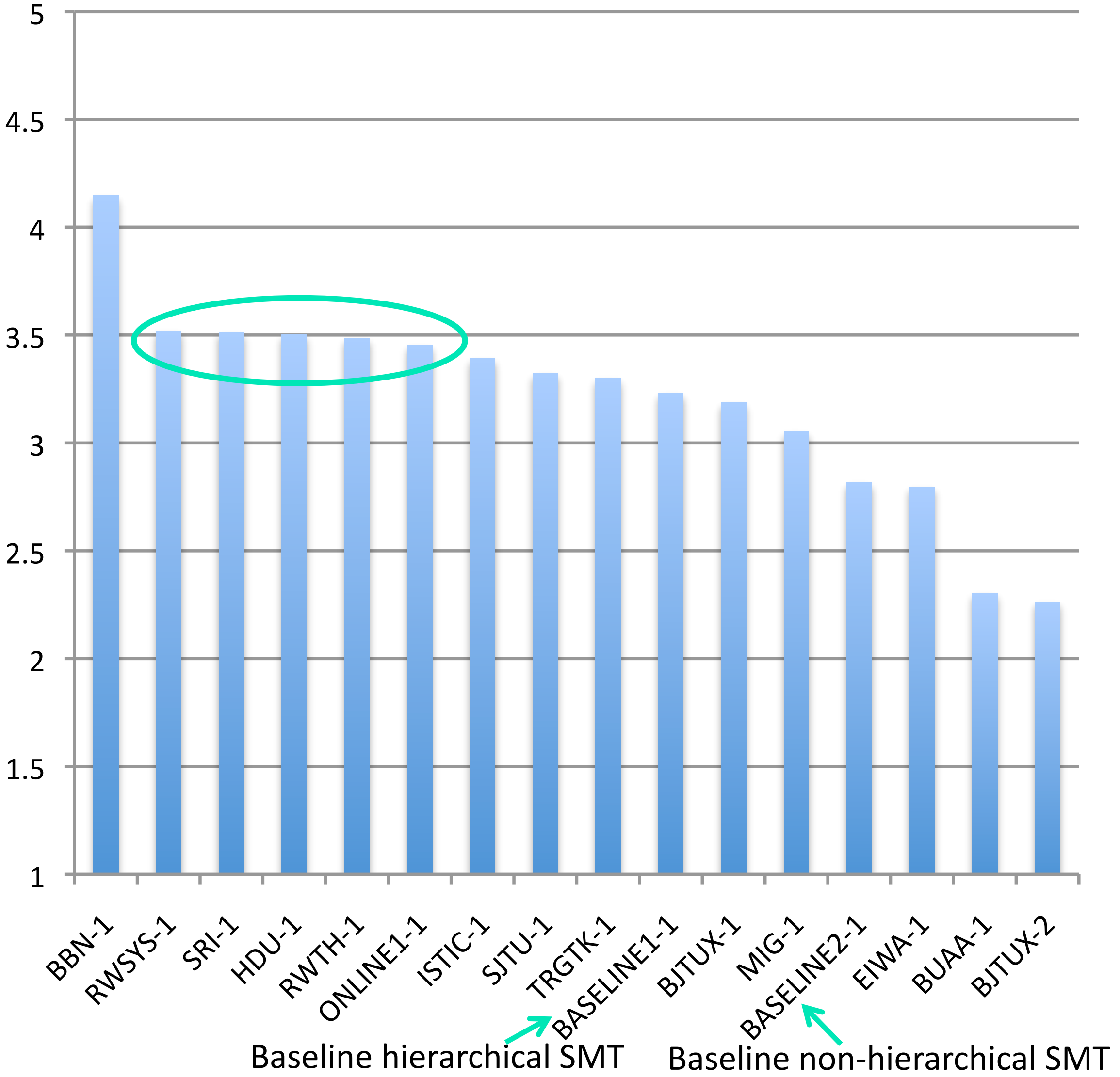
Explored Ideas for CE Subtask (1/2)

Type	Ideas
Adaptation	Sentence-level LM adaptation (BBN) LM adaptation (SRI, SJTU)
Language model	Recurrent neural network LM (BBN)
Feature	Sparse features (SRI)
Tuning	Tuning as reranking with SVM (SRI) Development data selection (SJTU)
Reordering	Soft syntactic constraints (HDU)
Translation model	Context dependent translation probability (BBN)
Decoding	String-to-dependency translation (BBN, SRI) Inverse direction decoding (RWTH) Example-based translation (BJTU)
Hybrid decoder	Statistical post-editing (RWSYS, EIWA)

Explored Ideas for CE Subtask (2/2)

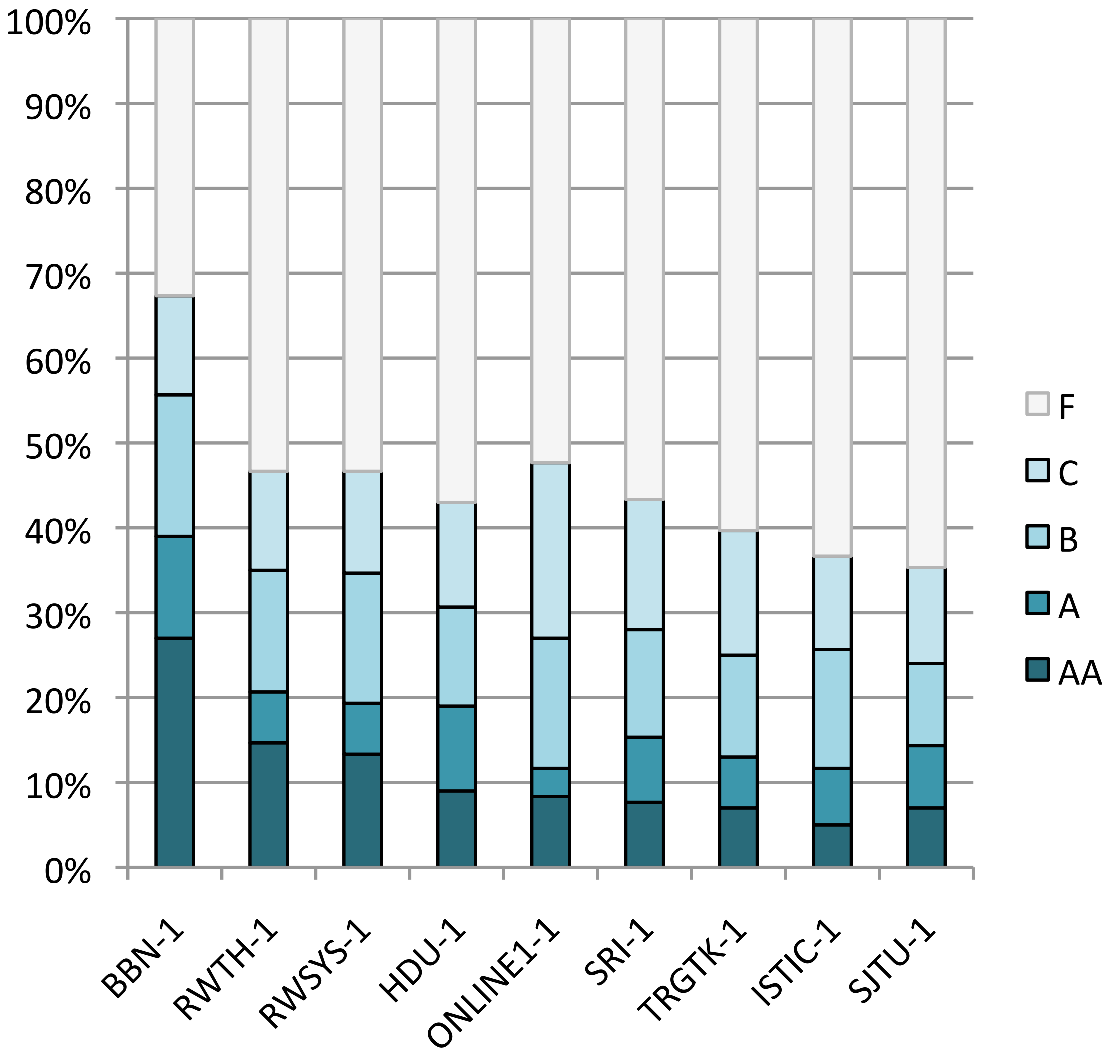
Type	Ideas
Preprocessing	Categorization of numbers (RWTH)
Tokenization	Segmentation using bilingual resources (MIG) Optimized word segmentation (SJTU) Word Segmentation on GPU (TRGTK)
True caser	Translation-based true caser (BBN)
Utilizing Context	Document-level decoding (TRGTK)
System combination	System combination using word graph (RWSYS, RWTH, ISTIC) System combination using reverse translation (EIWA)
Dictionary	Bilingual chemical dictionary (BJTU)

CE Adequacy Results



- The top system (BBN-1) achieved a **significantly better** score than those of the other systems.
- The second group were not statistically significant.

CE Acceptability Results



- **67%** sentences could be understood (C-rank and above) in the **best system** (BBN-1).
- This evaluation demonstrated the effectiveness of the BBN system.

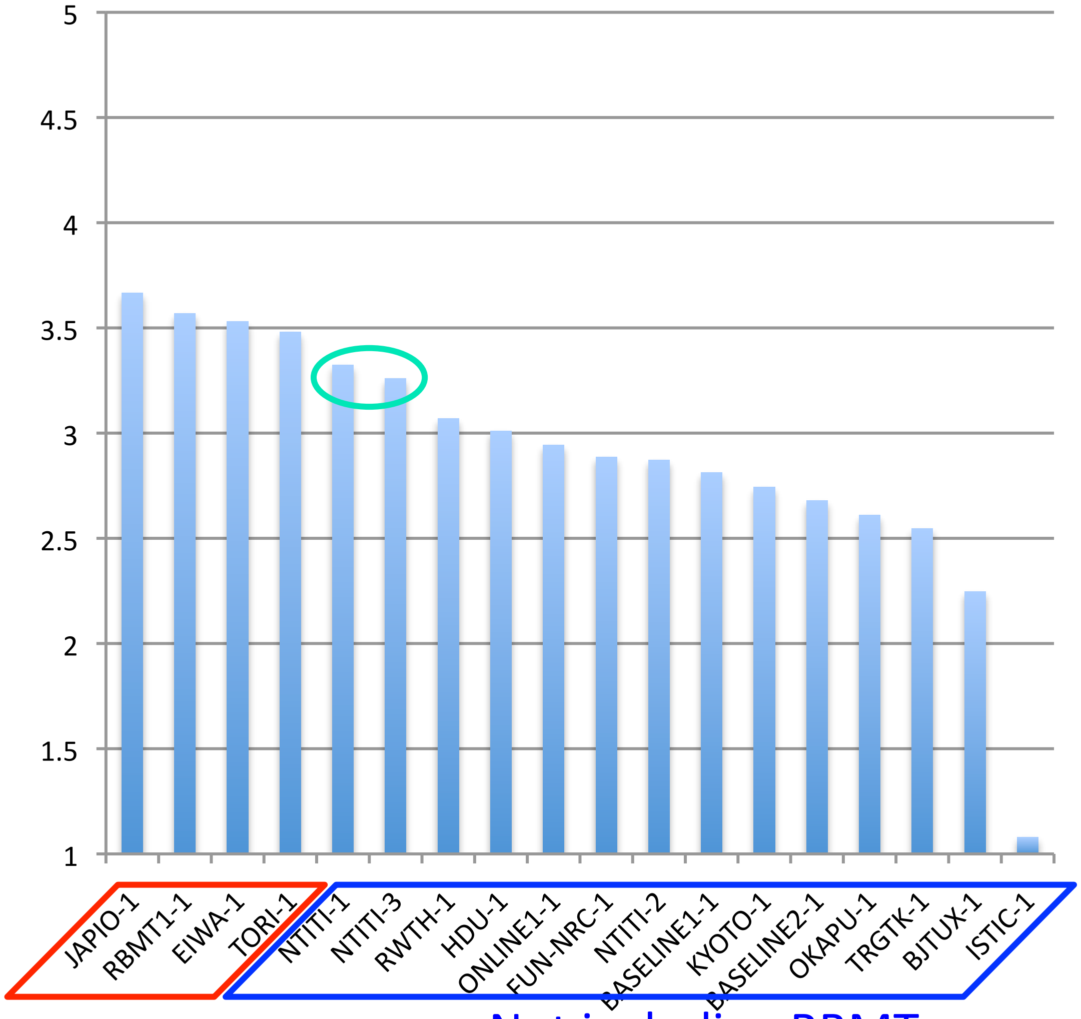
Explored Ideas for JE Subtask (1/2)

Type	Ideas
Post-ordering	Post-ordering by a syntax-based SMT (NTITI)
Reordering model	Hierarchical lexicalized reordering model (RWTH, FUN-NRC)
Pre-ordering	Pre-ordering based on case structures (NTITI) Pre-ordering without syntactic parsing (OKAPU)
Paraphrase	Paraphrase-augmented phrase-table (FUN-NRC)
Feature	Sparse features and feature selection via regularization (HDU)
Tuning	Discriminative training (HDU)
Preprocessing	Categorization of numbers (RWTH)
Alignment	Bayesian treelet alignment model (KYOTO)
Transliteration	Back-transliteration (NTITI)

Explored Ideas for JE Subtask (2/2)

Type	Ideas
Language model	Word class language model (RWTH) 7-gram language model (NTITI) Two language models (ISTIC)
Decoding	Pattern-based translation (TORI) Example-based translation (KYOTO) Inverse direction decoding (RWTH)
Hybrid decoder	Statistical post-editing (EIWA)
Utilizing Context	Document-level Decoding (TRGTK)
System combination	Generalized minimum Bayes risk system combination (NTITI) System combination using reverse translation (EIWA)
Dictionary	Adding technical field dictionaries to RBMT (JAPIO)

JE Adequacy Results

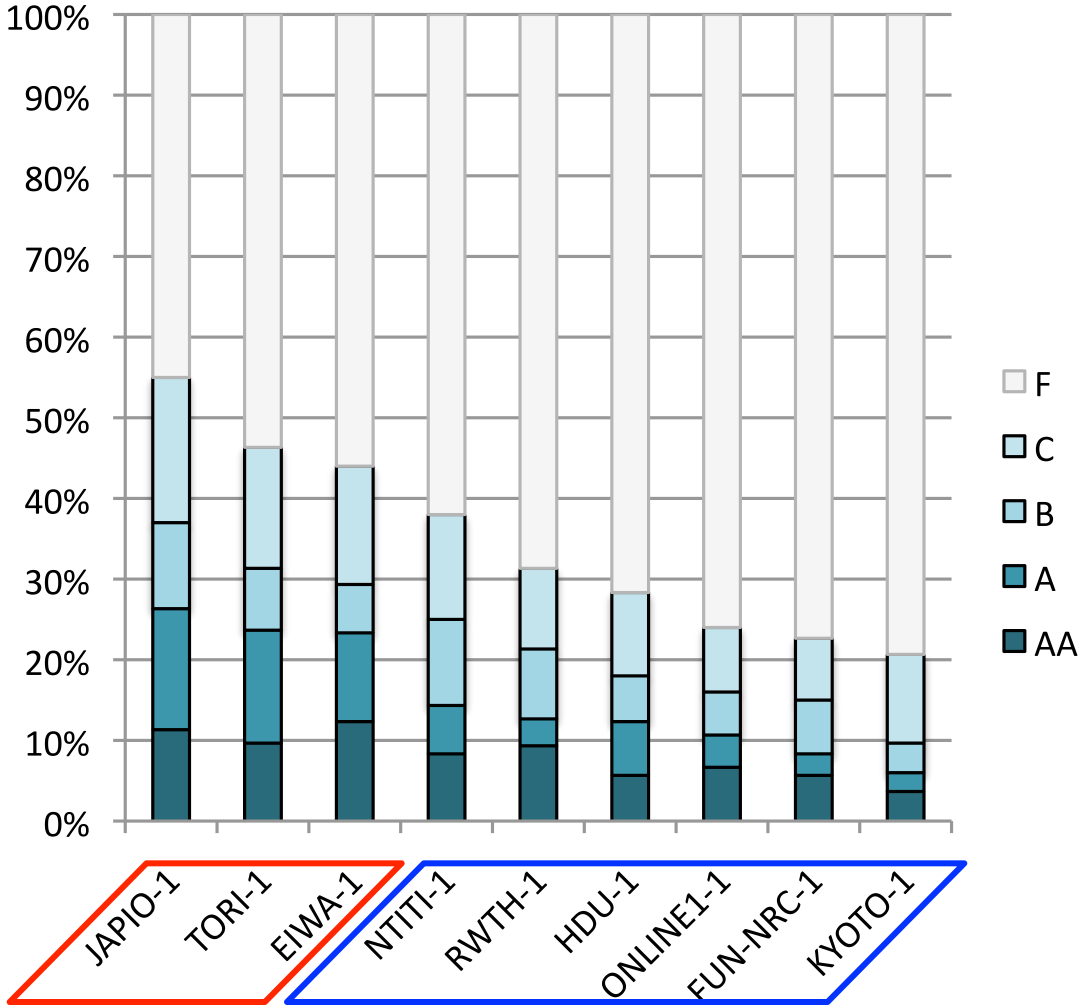


RBMT or including RBMT

Not including RBMT

- The **RBMT** systems were still **better** than the state-of-the-art SMT systems.
- The top SMT systems (NTITI-1 and 3) used **post-ordering**.

JE Acceptability Results



RBMT or including RBMT

Not including RBMT

- **55%** sentences could be understood (C-rank and above) in the **best system** (JAPIO-1) using **RBMT**.
- **38%** sentences could be understood for the **best SMT** (NTITI-1).
- S(･) = The rate of C-rank and above

$$\frac{\text{S(the top SMT)}}{\text{S(the top RBMT)}}$$

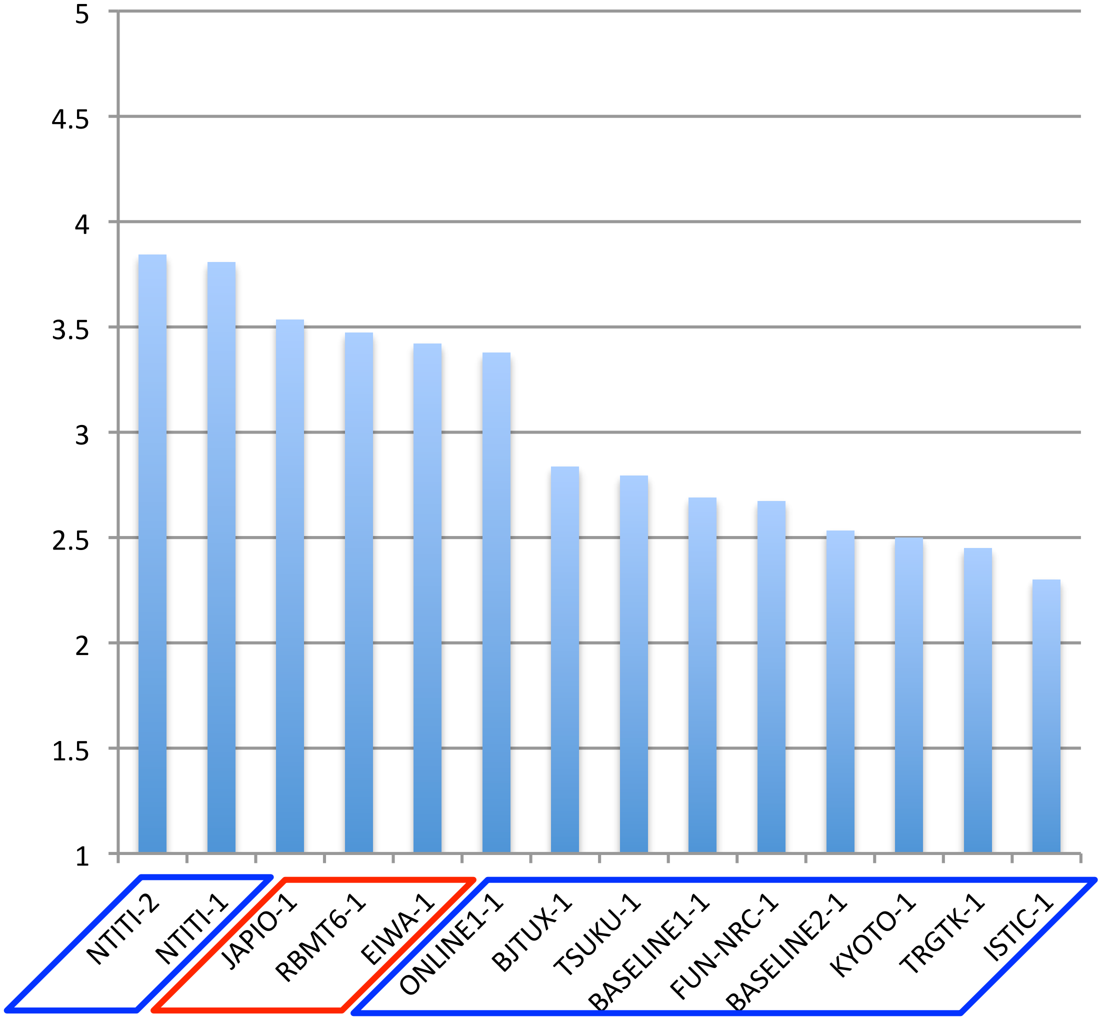
- NTCIR-10: 69% (=38/55)
- NTCIR-9: 39% (=25/63.3)

There is a large improvement in the top SMT performances.

Explored Ideas for EJ Subtask

Type	Ideas
Pre-ordering	Head finalization using dependency structure (NTITI)
Parsing	Dependency parser based on semi-supervised learning (NTITI) Combining a constituency tree and a dependency tree (TSUKU)
Corpus	English patent dependency corpus (NTITI)
Paraphrase	Paraphrase-augmented phrase-table (FUN-NRC)
Reordering	Hierarchical lexicalized reordering model (FUN-NRC)
Alignment	Bayesian treelet alignment model (KYOTO)
Language model	6-gram language model (NTITI) Two language models (ISTIC)
Decoding	Tree-to-string translation model (TSUKU) Example-based translation (KYOTO)
Hybrid decoder	Statistical post-editing (EIWA)
Utilizing Context	Document-level decoding (TRGTK)
System combination	Generalized minimum Bayes risk system combination (NTITI) System combination using reverse translation (EIWA)
Dictionary	Adding technical field dictionaries to RBMT (JAPIO)

EJ Adequacy Results



Not including RBMT

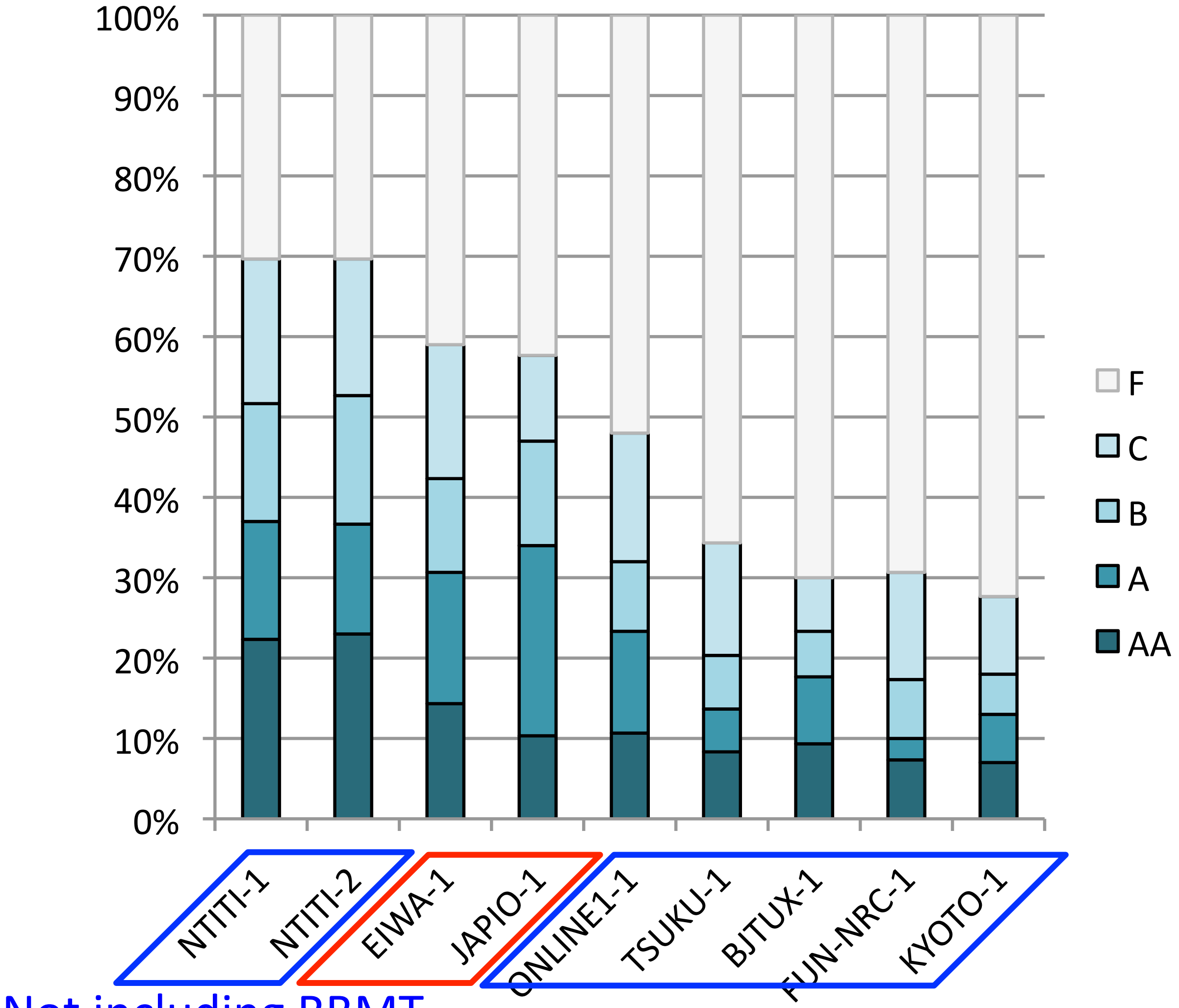
RBMT or including RBMT

- The top **SMT** systems (NTITI-2 and 1) were **better than** the top-level commercial **RBMT** systems.

- At NTCIR-9, the top SMT caught up with RBMT.

  At NTCIR-10, the top SMT **outperformed** RBMT.

EJ Acceptability Results



Not including RBMT

RBMT or including RBMT

- **70%** sentences could be understood (C-rank and above) for the top systems (NTITI-1 and 2).
- The translation quality of the top **SMT** systems **surpassed** those of the top-level RBMT systems **in retaining the sentence-level meanings**.

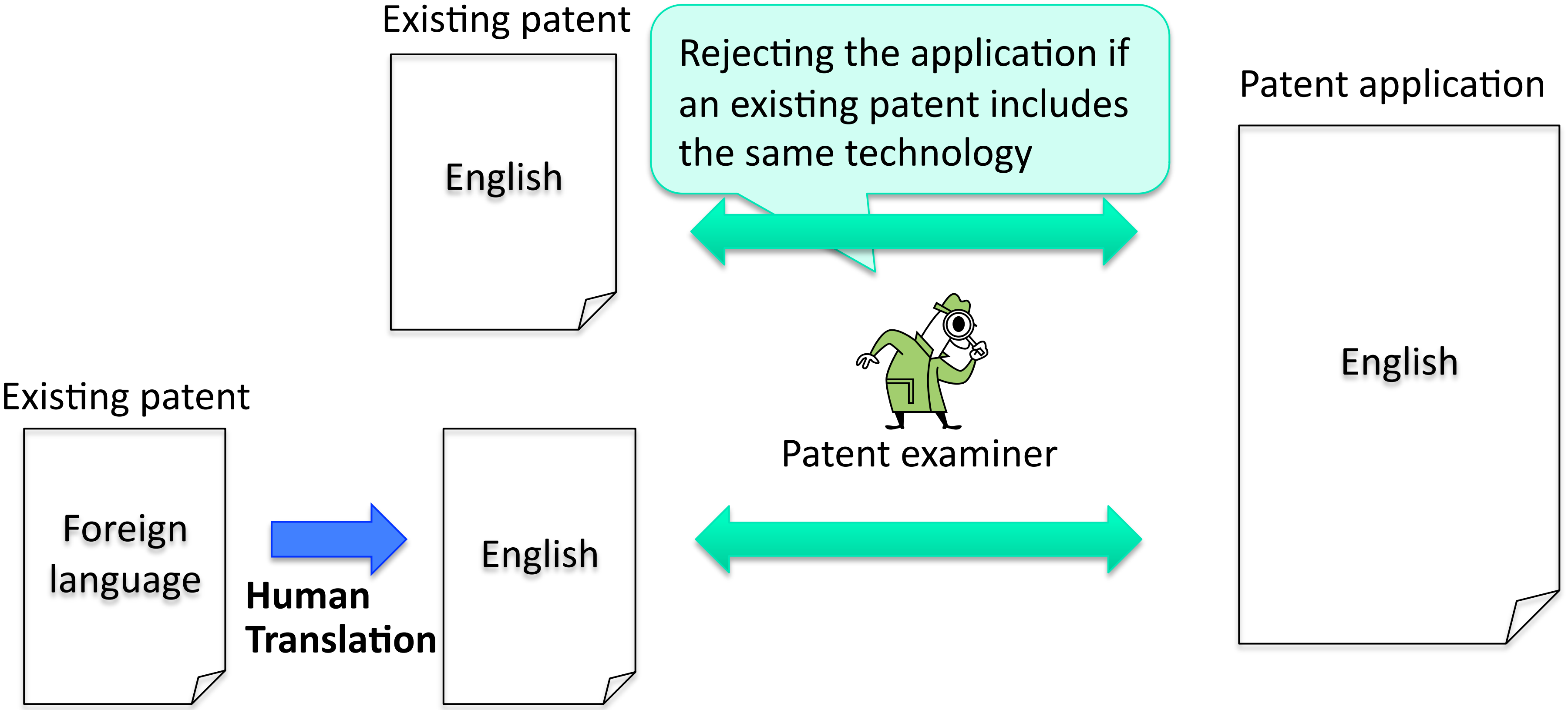
Patent Examination Evaluation (PEE)

Motivation of PEE

- At NTCIR-9, the top systems **achieved high performance** for **sentence-level evaluations**.
- Therefore, we would like to see **how useful** the top systems are for **practical situations**.

- Patent examination is a practical situation.
- Patent Examination Evaluation measures the **usefulness** of MT systems for **Patent Examinations**.

Patent Examination Flow



Existing patent

English

Rejecting the application if an existing patent includes the same technology

Patent application

English

Patent examiner

Existing patent

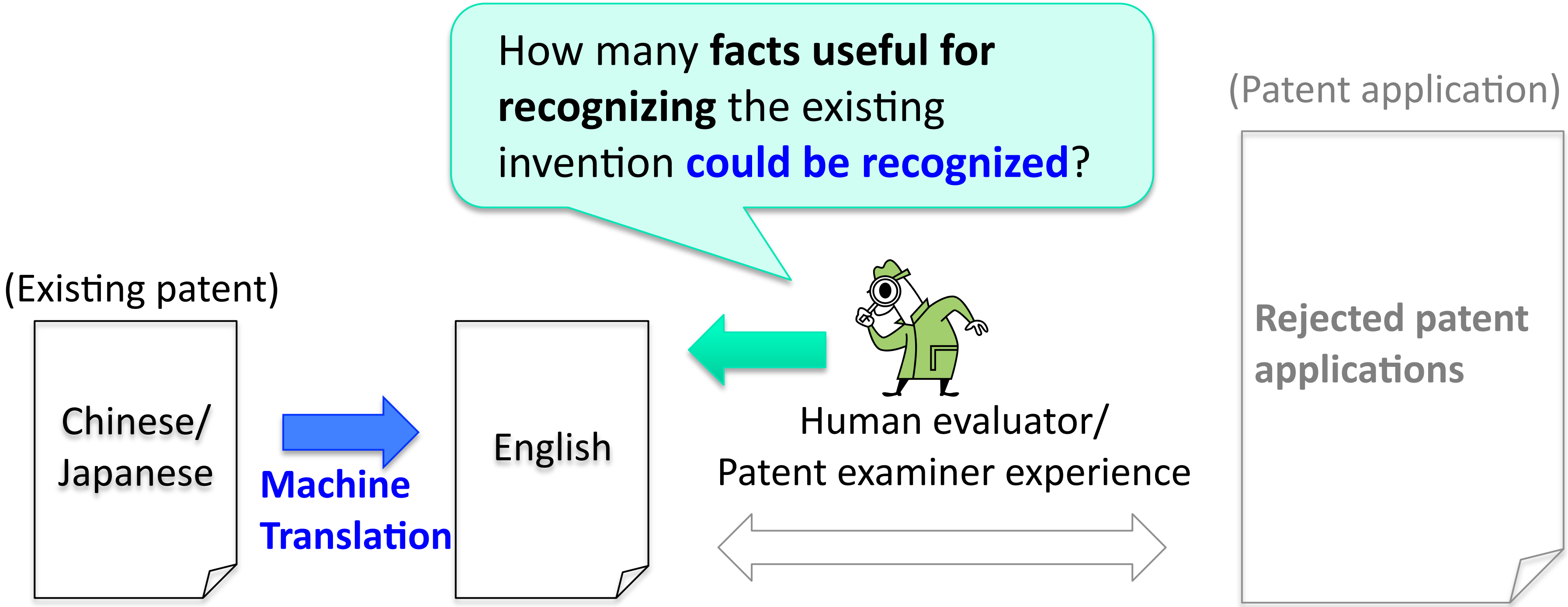
Foreign language

Human Translation

English

How useful is the top level MT for the translation?

Outline of Real Framework



How many **facts useful for recognizing** the existing invention **could be recognized**?

(Existing patent)

Chinese/ Japanese

Machine Translation

English

Human evaluator/ Patent examiner experience

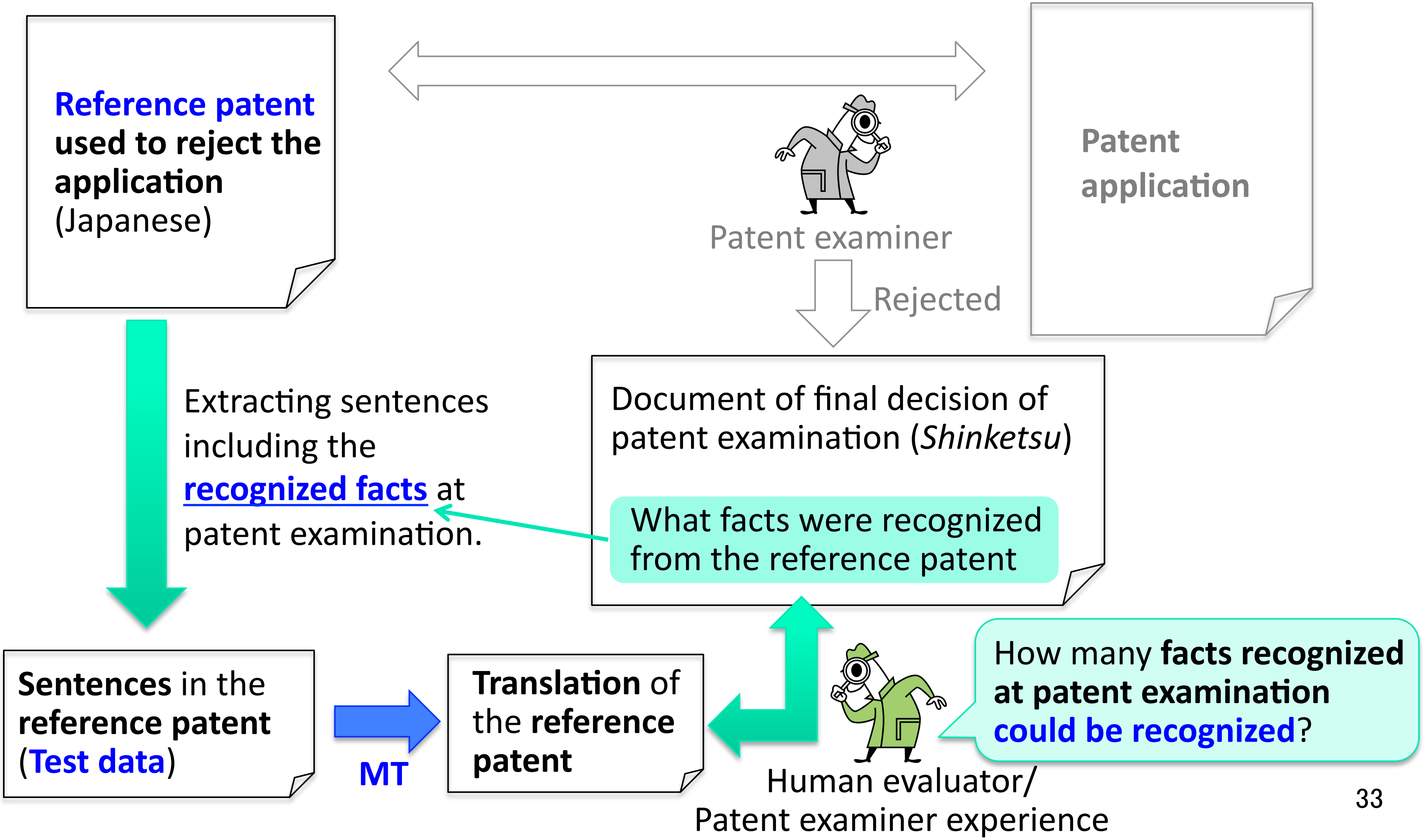
(Patent application)

Rejected patent applications

Reference patents used to reject the applications

- There were two evaluators.
- Test data were 29 reference patents used to reject patent applications.
- Each evaluator evaluated 20 patents.

Real Framework



Reference patent **used to reject the application** (Japanese)

Patent application

Patent examiner

Rejected

Document of final decision of patent examination (*Shinketsu*)

What facts were recognized from the reference patent

Extracting sentences including the **recognized facts** at patent examination.

Sentences in the **reference patent** (**Test data**)

MT

Translation of the **reference patent**

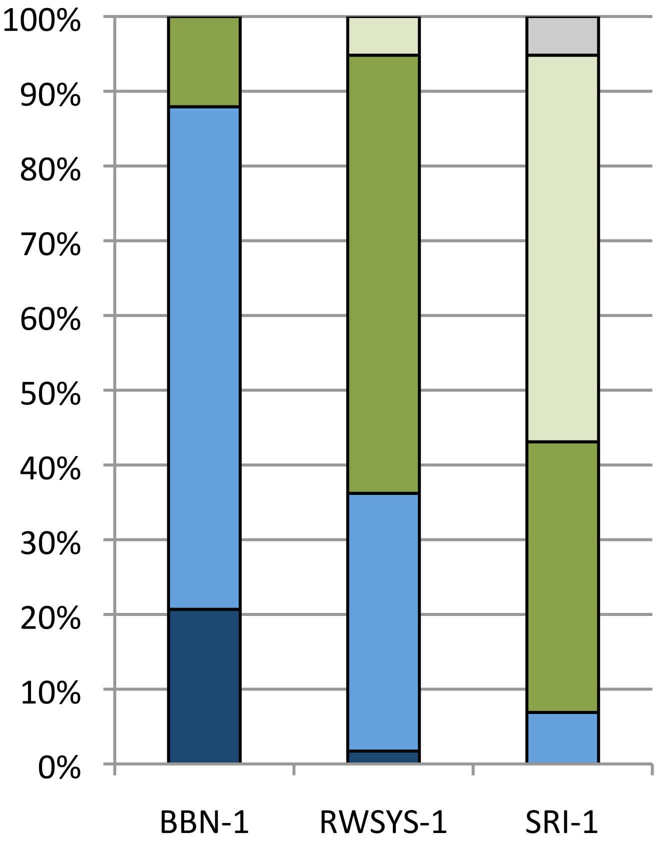
Human evaluator/ Patent examiner experience

How many **facts recognized at patent examination** **could be recognized**?

Example Data

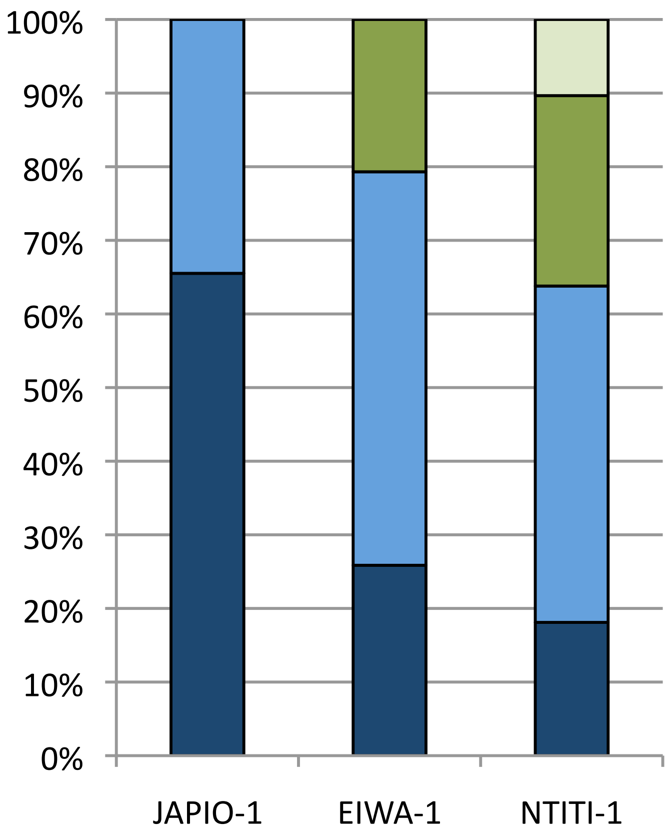
The description of the **facts that a patent examiner recognized** from the reference patent	The description was divided into each component	The **sentences** including each component **in the reference patent** (Japanese **test data**)
これらの記載事項によると、引用例には、 「内部において、先端側に良熱伝導金属部43が入り込んでいる中心電極4と、 中心電極4の先端部に溶接されている貴金属チップ45と、 中心電極4を電極先端部41が碍子先端部31から突出するように挿嵌保持する絶縁碍子3と、 絶縁碍子3を挿嵌保持する取付金具2と、 中心電極4の電極先端部41との間に火花放電ギャップGを形成する接地電極11とを備えたスパークプラグにおいて、 中心電極4の直径は、1．2～2．2mmとしたスパークプラグ。」 の発明が記載されていると認められる。	内部において、先端側に良熱伝導金属部43が入り込んでいる中心電極4	また、図3に示すごとく、中心電極4の内部においては、上記露出開始部431よりも先端側にも良熱伝導金属部43が入り込んでいる。
	中心電極4の先端部に溶接されている貴金属チップ45	また、中心電極4の先端部には、貴金属チップ45が溶接されている。
	中心電極4を電極先端部41が碍子先端部31から突出するように挿嵌保持する絶縁碍子3	上記中心電極4は、電極先端部41が碍子先端部31から突出するように絶縁碍子3に挿嵌保持されている。
	絶縁碍子3を挿嵌保持する取付金具2	上記絶縁碍子3は、碍子先端部31が突出するように取付金具2に挿嵌保持される。
	中心電極4の電極先端部41との間に火花放電ギャップGを形成する接地電極11	上記接地電極11は、図2に示すごとく、電極先端部41との間に火花放電ギャップGを形成する。
	中心電極4の直径は、1．2～2．2mm	また、上記碍子固定部22の軸方向位置における中心電極4の直径は、例えば、1．2～2．2mmとすることができる。

PEE CE Results



I	None of the facts were recognized and the translation results were **not useful** for examination.
II	Parts of the facts were recognized but the translation results could **not** be seen as **useful** for examination.
III	Falls short of reaching IV, but parts of the facts were recognized and it was **proved** that the cited invention **could not be disregarded** at the examination.
IV	**One or more facts** useful for recognizing the cited invention **were recognized** and the translation results were **useful** for examination.
V	**At least half of the facts** useful for recognizing the cited invention **were recognized** and the translation results were **useful** for examination.
VI	**All facts** useful for recognizing the cited invention **were recognized** and examination could be done **using only the translation results**.

PEE JE Results



I	None of the facts were recognized and the translation results were **not useful** for examination.
II	Parts of the facts were recognized but the translation results could **not** be seen as **useful** for examination.
III	Falls short of reaching IV, but parts of the facts were recognized and it was **proved** that the cited invention **could not be disregarded** at the examination.
IV	**One or more facts** useful for recognizing the cited invention **were recognized** and the translation results were **useful** for examination.
V	**At least half of the facts** useful for recognizing the cited invention **were recognized** and the translation results were **useful** for examination.
VI	**All facts** useful for recognizing the cited invention **were recognized** and examination could be done **using only the translation results**.

Comprehensive Comments (Evaluator 1)

CE	BBN-1	**Second-most consistent** after JAPIO-1 in its translation quality. The system seemed to try to **translate complicated input sentences depending on context** and I would like to **applaud this**.
	RWSYS-1	There were fragmental translations. To understand the translations, sentences before or after, or common knowledge of technology were needed for many parts.
	SRI-1	Hard to read. It would not be practical for patent examination.
JE	JAPIO-1	**Consistent in its translation quality**. The system seemed to try to **translate complicated input sentences depending on context** and I would like to **applaud this**.
	EIWA-1	There were fragmental translations. To understand, sentences before or after, or common knowledge of technology were needed for many parts.
	NTITI-1	There were **good results** and **not good results**. Impression was inconsistent. **If** this problem were **improved**, it **would be a good** system.

JAPIO-1 and BBN-1 were highly evaluated.

Comprehensive Comments (Evaluator 2)

CE	BBN-1	**A little inconsistent**. There were some English grammatical problems.
	RWSYS-1	There were **good results** and also **not good results**.
	SRI-1	The translations were hard to read.
JE	JAPIO-1	Even if the input Japanese sentences were **abstruse**, it **sometimes could translate**. Not only were the English **translations good**, but **so was analyzing input Japanese sentences**.
	EIWA-1	It was similar to JAPIO-1. It would be **better** than BBN-1.
	NTITI-1	There were **good results** and also **not good results**. It would be **slightly better** than RWSYS-1.

JAPIO-1, EIWA-1, and BBN-1 were highly evaluated.

Summary of PatentMT

- Goal: To foster **challenging** and **practical** research into patent machine translation
- Large-scale **CE** and **JE patent parallel corpora** were provided.
- **21** research groups participated.
- **Human evaluations** were conducted.
- The top MT systems for JE and CE were **useful** for **patent examination**.
- Various ideas were explored and the effectiveness of the systems for patent translation was shown in evaluations.
- The effectiveness of each idea will be presented by the participants.

Thank you

Oral Presentations of Participants

Group ID	Organization	Authors	Notable points
BBN	*BBN Technologies*	Zhongqiang Huang et al.	The **best** system for CE
NTITI	*NTT Corporation / National Institute of Informatics*	Katsuhito Sudoh et al.	The **best SMT** system for JE and the **best** system for EJ
RWSYS/ RWTH	*RWTH Aachen University / Systran*	Minwei Feng et al.	Highly ranked systems for CE and JE
SRI	*SRI International*	Bing Zhao et al.	Highly ranked system for CE
HDU	*Institute for Computational Linguistics, Heidelberg University*	Patrick Simianer et al.	Highly ranked systems for CE and JE
FUN-NRC	*Future University Hakodate / National Research Council Canada*	Atsushi Fujita and Marine Carpuat	Exploring paraphrasing
EIWA	*Yamanashi Eiwa College*	Terumasa Ehara	Exploring hybrid decoder and system combination
TRGTK	*Torangetek Inc.*	Hao Xiong and Weihua Luo	Exploring document-level decoding and utilizing GPU