Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop

Isao Goto (NICT)
Bin Lu (City Univ. of Hong Kong / Hong Kong Institute of Education)
Ka Po Chow (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
- Remarkable Findings at NTCIR-9
- PatentMT at NTCIR-9
- JE and EJ Subtasks
- CE Subtask
- Meta-Evaluation of Automatic Evaluation based on Human 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**.
  - 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 Japanese, English, and Chinese.

- To **compare** the effects of **different methods** of patent translation by applying them to the same test data.

- 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. ■ The translations were used as bag-of-words. ■ This means that **word selection** by SMT was better than that by RBMT.
NTCIR-8	Automatic evaluation	A hybrid system (RBMT with statistical post edit) achieved the best score for JE.

Comparison of NTCIR-7, 8, and 9

	NTCIR-7	NTCIR-8	**NTCIR-9** New
Language	**Japanese to English** **English to Japanese**	**Japanese to English** **English to Japanese**	**Chinese to English** **Japanese to English** **English to Japanese**
Human evaluation	**Adequacy** **Fluency**	No human evaluation	**Adequacy** New **Acceptability**
Extrinsic evaluation	CLIR	CLIR	No extrinsic evaluation
Number of participants	15	8	**21**

At NTCIR-9, participants can choose subtasks from three language directions, including **Chinese to English**.

Remarkable Findings at NTCIR-9

- **SMT** was the **best** system for **CE** and **EJ** patent translation.
  - This is the **first time** for **SMT** to be **demonstrated equal or better** quality than that of the top-level RBMT for **EJ** patent translation.

- **80%** of patent sentences could be understood in the best system for **CE** patent translation.

PatentMT at NTCIR-9

Features of PatentMT at NTCIR-9

- Provided data

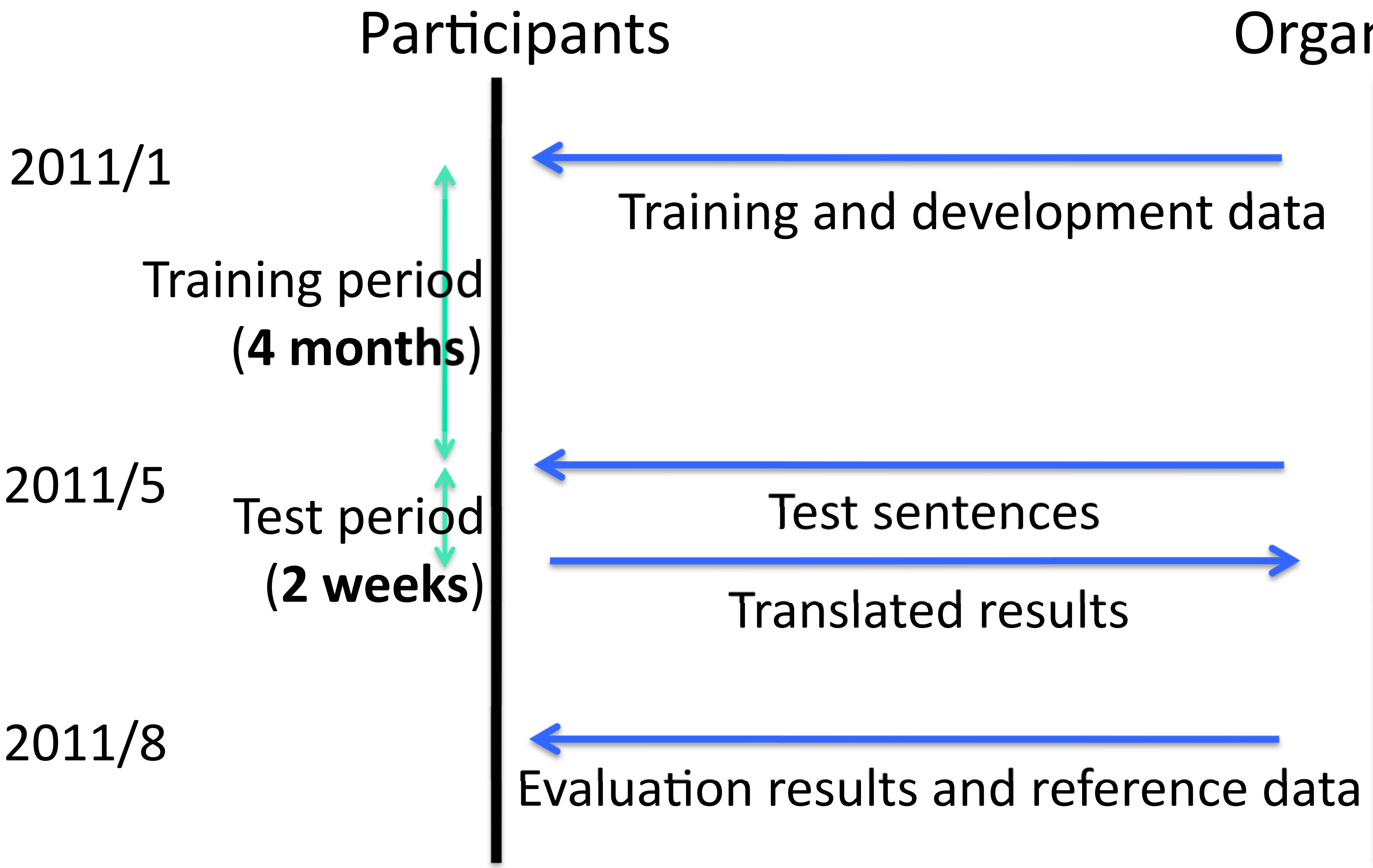
Training	CE	1 million patent parallel sentence pairs
		Over **300 million** patent **monolingual** sentences in English
	JE	Approximately 3.2 million patent parallel sentence pairs
		Over **300 million** patent **monolingual** sentences in English
	EJ	Approximately 3.2 million patent parallel sentence pairs
		Over **400 million** patent **monolingual** sentences in Japanese
Development	All	2,000 patent description parallel sentence pairs
Test	All	2,000 patent description sentences
		2,000 reference translations

  - The **periods** for the training and test data are **different**

    (Training data: 2005 or before, Test data: 2006 or later)

- **Human** evaluation ••• Primary evaluation
  - **Adequacy** and **Acceptability**

Flow and Schedule of the Task



Participants

Organizers

2011/1

Training and development data

Training period **(4 months)**

2011/5

Test period **(2 weeks)**

Test sentences

Translated results

2011/8

Evaluation results and reference data

The Hong Kong Institute of Education 香港教育學院

NICT

Participants

Group ID	Organization	Nationality	CE	JE	EJ
BJTUX	Beijing Jiaotong University	P.R. China	1		1
FRDC	Fujitsu R&D Center CO., LTD	P.R. China	1	1	1
ISTIC	Institute of Scientific and Technical Information of China	P.R. China	1		
ICT	Institute of Computing Technology, Chinese Academy of Sciences	P.R. China	1	1	1
BUAA	Institute of Intelligent Information Processing, Beihang University	P.R. China	1		
NEU	Northeastern University	P.R. China	1	1	
KECIR	Shenyang Aerospace University	P.R. China	1		
JAPIO	Japan Patent Information Organization	Japan		1	1
KYOTO	Kyoto University	Japan	1	1	1
NAIST	Nara Institute of Science and Technology	Japan		1	
NTT-UT	NTT Communication Science Labs. and the University of Tokyo	Japan	1	1	1
UOTTS	The University of Tokyo	Japan	1	1	1
TORI	Tottori University	Japan		1	1
EIWA	Yamanashi Eiwa College	Japan	1	1	
IDEAS	Institute for Information Industry, Chaoyang University of Technology and National Tsing Hua University	Taiwan	1		
NCW	NTNU, NCCU, and WebGenie Information Ltd.	Taiwan	1		
KLE	Pohang University of Science and Technology (POSTECH)	Korea	1	1	1
LIUM	University of Le Mans	France	1		
RWTH	RWTH Aachen University	Germany	1	1	
IBM	IBM Research	USA	1		
BBN	Raytheon BBN Technologies	USA	1		

Baseline Systems

SYSTEM-ID	System	Type	CE	JE	EJ
BASELINE1	Moses hierarchical phrase-based SMT system	SMT	1	1	1
BASELINE2	Moses phrase-based SMT system		1	1	1
RBMTx	SYSTRAN 7 Premium Translator	**RBMT**	1		
RBMTx	Huajian Multilingual EasyTrans version 3.0		1		
RBMTx	The Honyaku 2009 premium patent edition			1	1
RBMTx	ATLAS V14			1	1
RBMTx	PAT-Transer 2009			1	1
ONLINE1	Google online translation system	SMT	1	1	1

- 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.

Human Evaluation

- 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**.
    - **All of the first priority submissions** were evaluated at the least.
  - 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.

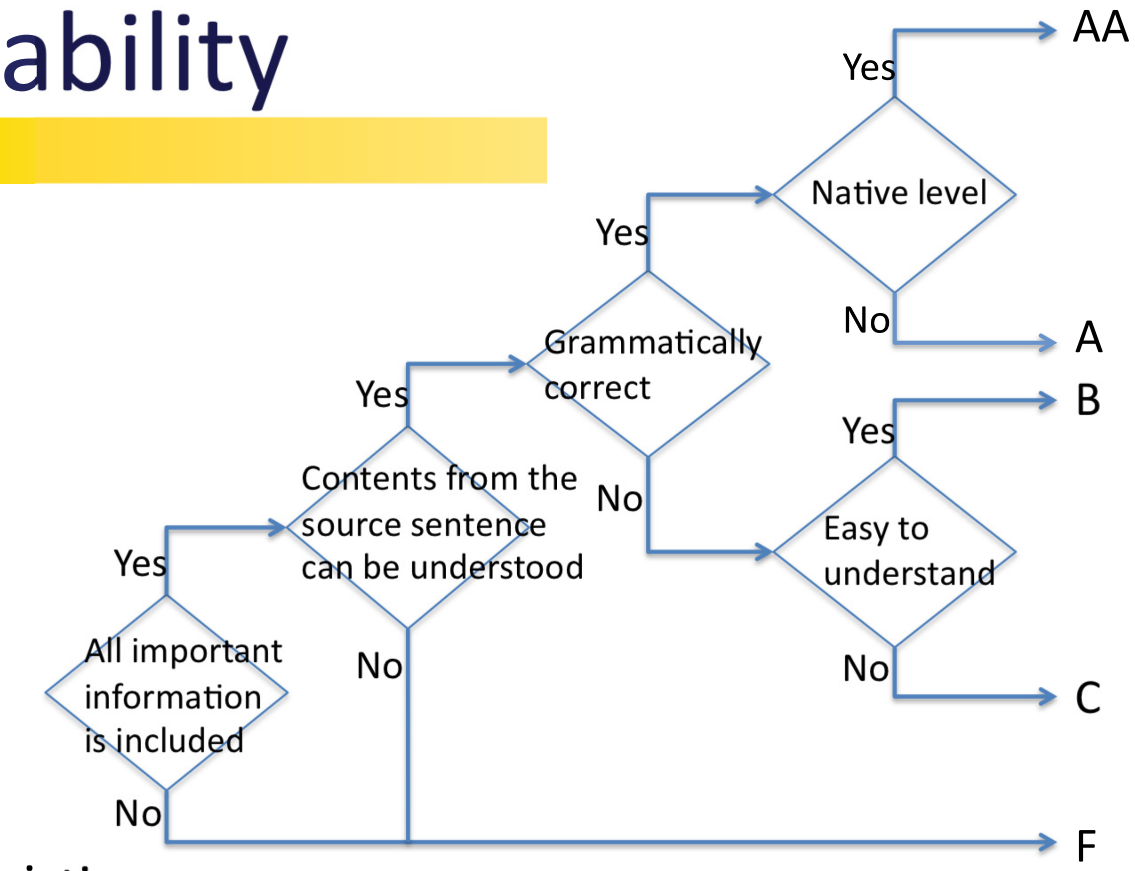
Adequacy

- The criterion of adequacy used 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.
  - It is **known** what **percentage** of the translated sentences express the **correct meaning of the source sentence**.
  - If a requirement for a translation system is that the source sentence meaning can be understood, then translations of **C or higher** are useful.

JE and EJ Subtasks

JE Patent Parallel Corpus

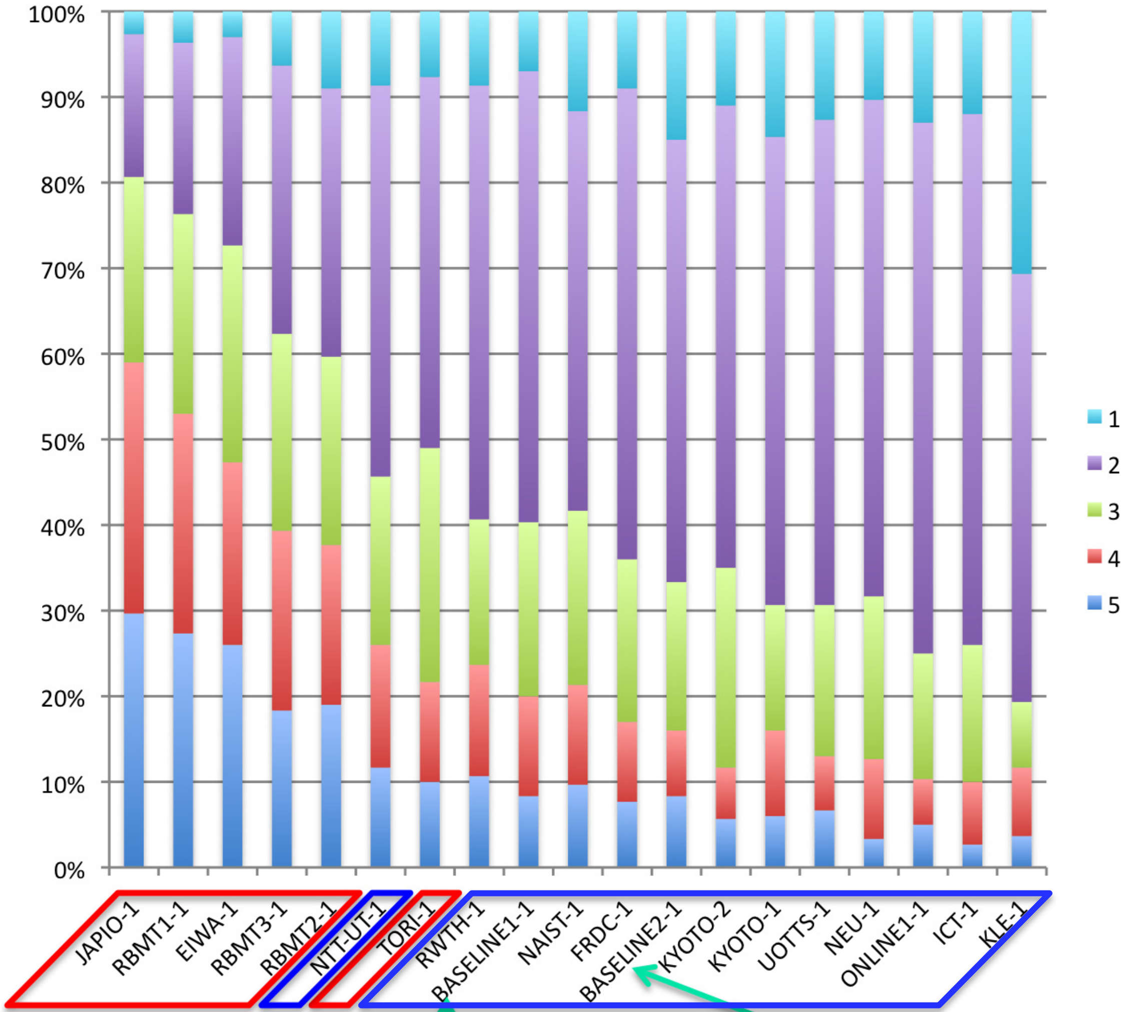
- How a corpus was built
  - Parallel patent documents in Japanese and English were extracted from **patent families**.
    - Patent families are one of the ways to apply for patents in more than one country.
  - The parallel sentences were automatically extracted from the parallel patent documents using bilingual dictionaries.

- Test data and reference translations
  - We **manually selected** 2,000 correct parallel sentence pairs from the automatically extracted pairs.

Explored Ideas for JE Subtask

Type	Ideas
Pre-ordering	POS-based reordering for dependency structure of Japanese (NTT-UT)
	Linear ordering problem based reordering (NAIST)
Hybrid decoder	RBMT and statistical post edit (EIWA, TORI)
Decoding	Hybrid reordering model (NEU)
	Example-based MT (KYOTO, NEU)
System combination	Generalized minimum Bayes risk system combination (NTT-UT)
Reranking	Bagging-based reranking (ICT)
Tokenization	Merging Japanese verb endings / splitting for katakana words (RWTH)
Preprocessing	Handling parentheses (FRDC)
Dictionary	Adding technical field dictionaries to RBMT (JAPIO)
Alignment	Bayesian subtree alignment (KYOTO)

JE Adequacy Results



RBMT or including RBMT

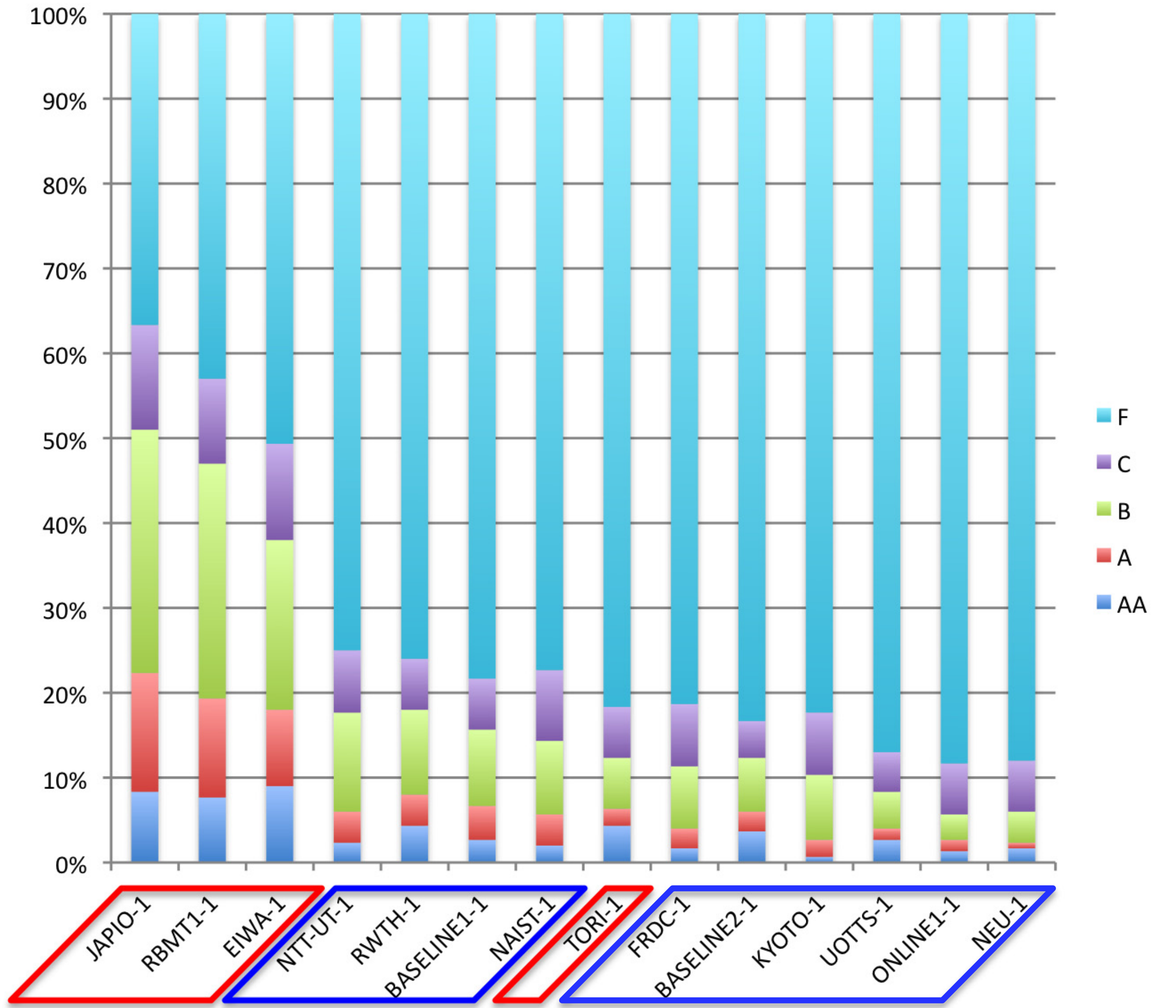
Not including RBMT

Baseline hierarchical SMT

Baseline non-hierarchical SMT

- The **RBMT** systems were **better** than the state-of-the-art SMT systems.
- The baseline **hierarchical** phrase-based SMT was slightly better than the baseline phrase-based SMT.

JE Acceptability Results



RBMT or including RBMT

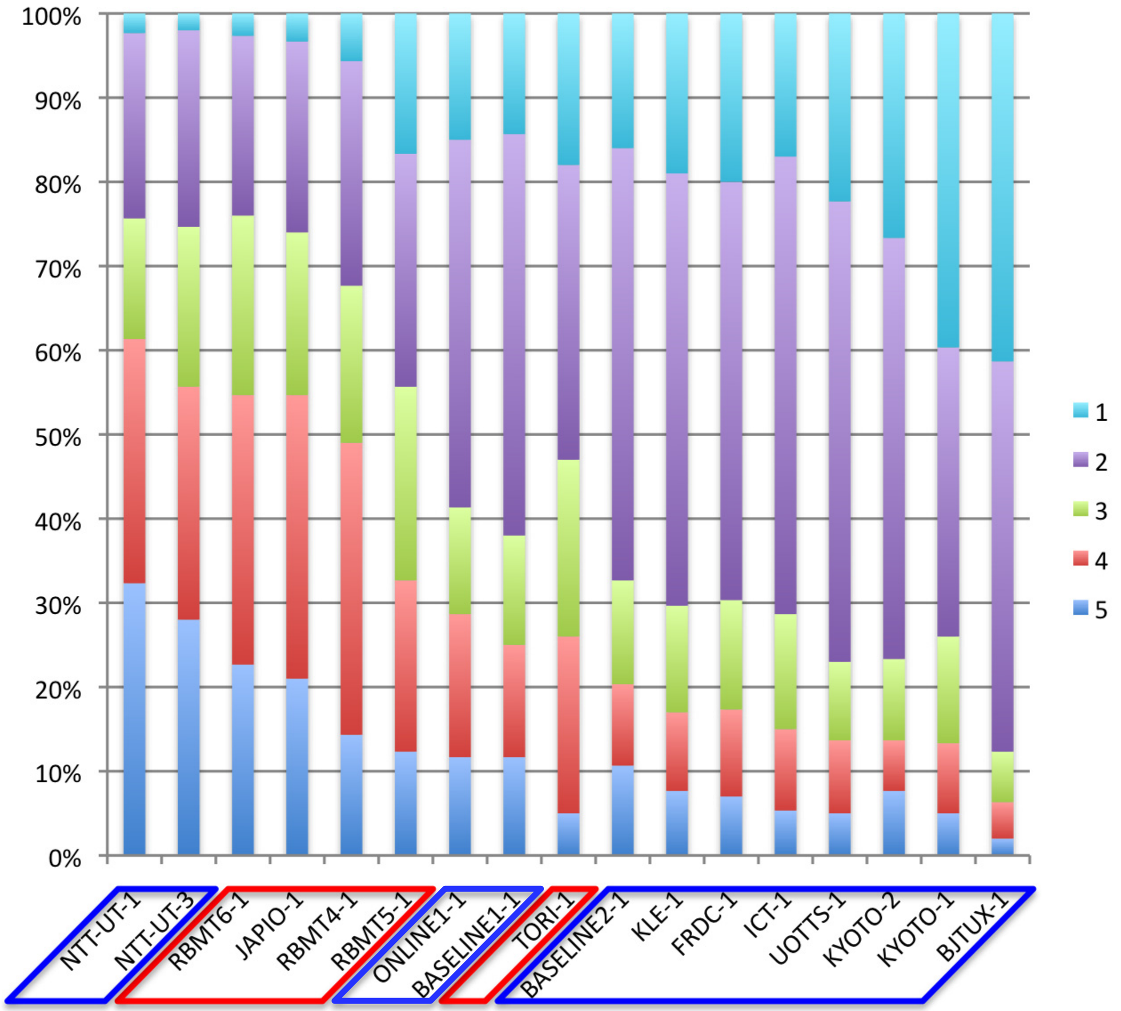
Not including RBMT

- **63%** sentences could be understood (C-rank and above) in the **best system** (JAPIO-1) using **RBMT**.
- **25%** sentences could be understood for the **best SMT** (NTT-UT1).
- There was a **large difference** in ability to **retain** the **sentence level meanings** between the **RBMT** systems and the **SMT** systems.

Explored Ideas for EJ Subtask

Type	Ideas
Pre-ordering	Head Finalization for English (NTT-UT)
	Syntactic reordering (KLE)
Preprocessing	Transferring syntactic roles (KLE)
	Inserting pseudo-particles (NTT-UT)
	Handling parentheses (FRDC)
Hybrid decoder	Cascading RBMT and SMT (TORI)
Decoding	HPSG forest-to-string MT (UOTTS)
	Example based MT (KYOTO)
	Factored translation model (BJTUX)
System combination	Generalized minimum Bayes risk system combination (NTT-UT)
Reranking	Bagging-based reranking (ICT)
Dictionary	Adding technical field dictionaries to RBMT (JAPIO)
Alignment	Bayesian subtree alignment (KYOTO)

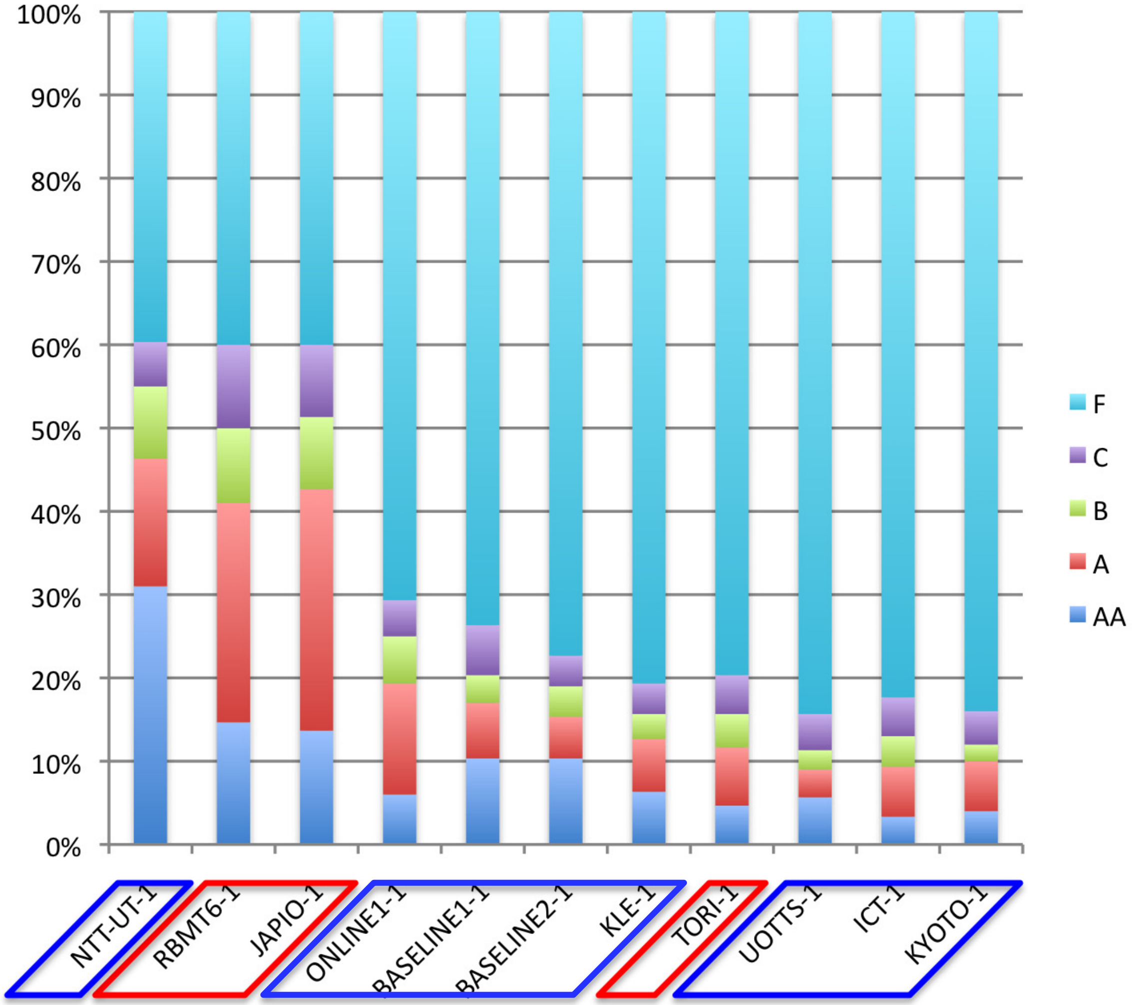
EJ Adequacy Results



Not including RBMT RBMT or including RBMT

- The top **SMT** systems (NTT-UT-1 and NTT-UT3) were **equal or better than** the top-level commercial **RBMT** systems.
- No SMT system did this at NTCIR-7, and it is the **first time** for this **achievement**.
- The baseline RBMT systems were better than those for SMT systems other than NTT-UT-1 and NTT-UT-3.

EJ Acceptability Results



Not including RBMT RBMT or including RBMT

- **60%** sentences could be understood (C-rank and above) for the top three systems (NTT-UT-1, RBMT6-1, and JAPIO-1).
- The translation quality of the top **SMT** system (NTT-UT-1) was **equal to or surpassing** that of the top-level RBMT systems **in retaining the sentence-level meanings**.
- This evaluation demonstrated the effectiveness of the NTT-UT system.

CE Subtask

CE Patent Parallel Corpus

- How the corpus was built
  - Comparable patent documents in Chinese and English were extracted from **PCT patents**.
    - PCT patents are one of the ways to apply for patents in more than one country.
  - The parallel sentences were automatically extracted from the comparable patent documents using length information, bilingual dictionaries and statistical translation probability.
- Test data and reference translations
  - **manually selected** 2,000 parallel sentence pairs from the automatically extracted pairs.

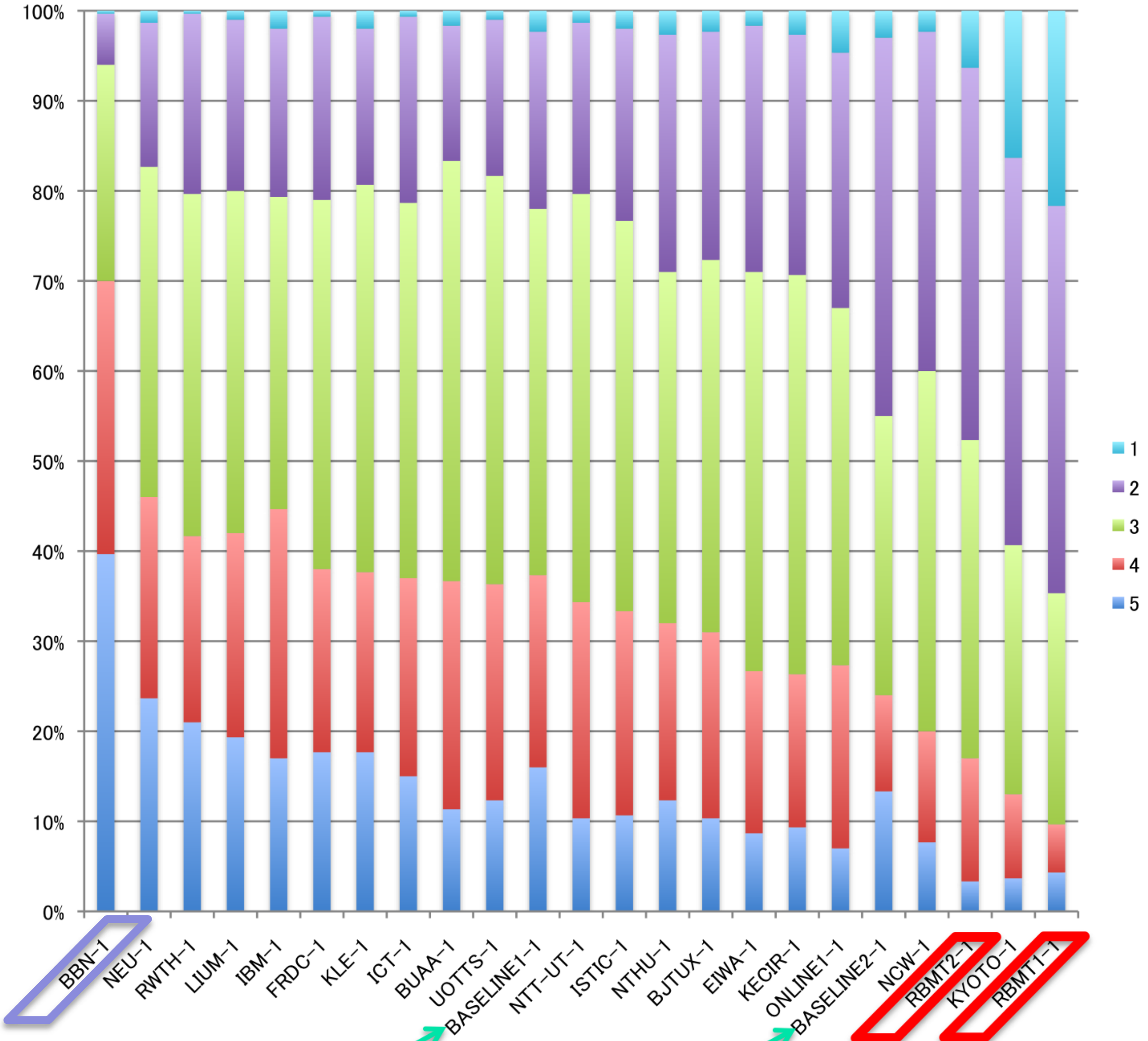
Explored Ideas for CE Subtask (1/2)

Type	Ideas
Tokenization	Optimizing the Chinese word segmenter based on MT performance (BBN)
	Tokenizing ASCII string in Chinese (BBN)
	Training the Chinese segmenter (NCW)
Preprocessing	Replacing infrequent special words to special tokens (BBN)
	Rule-based entity classing (IBM)
	Incorporating manually written templates (ICT)
	Chemical expression substitution (ICT)
	Chinese sentence paraphrasing (FRDC)
	Handling parentheses (FRDC)
	Prior Translation of unknown words and singletons (IDEAS)
Pre-ordering	Parsing-based pre-ordering (IBM)
Adaptation	LM adaptation for input sentences (BBN)
	TM adaptation using monolingual data (LIUM)
	Bayesian word alignment adaptation (NTT-UT)
	Domain adaptation using four domains (ICT)

Explored Ideas for CE Subtask (2/2)

Type	Ideas
Decoding	String-to-dependency MT (BBN)
	Using additional 8 features (BBN)
	Direct translation model (a special maximum entropy model) (IBM)
	Tree-to-string MT (IBM)
	Tree-to-tree MT (BUAA)
	BTG constraint into reordering model (BUAA)
	Example-based MT (KYOTO, NEU)
	SMT system using an example-based decorder (BUAA)
	Hybrid reodering model (NEU)
	Factored translation model (BJTUX)
Hybrid decoder	RBMT and statistical post edit (EIWA)
System combination	System combination of bidirectional translation systems (RWTH)
	System combination based on incremental alignment (IBM)
	Generalized minimum Bayes risk system combination (NTT-UT)
	System combination based on word and phrase (ISTIC)
Alignment	Bayesian subtree alignment (KYOTO)
Dictionary	Adding external bilingual dictionaries (NCW)

CE Adequacy Results



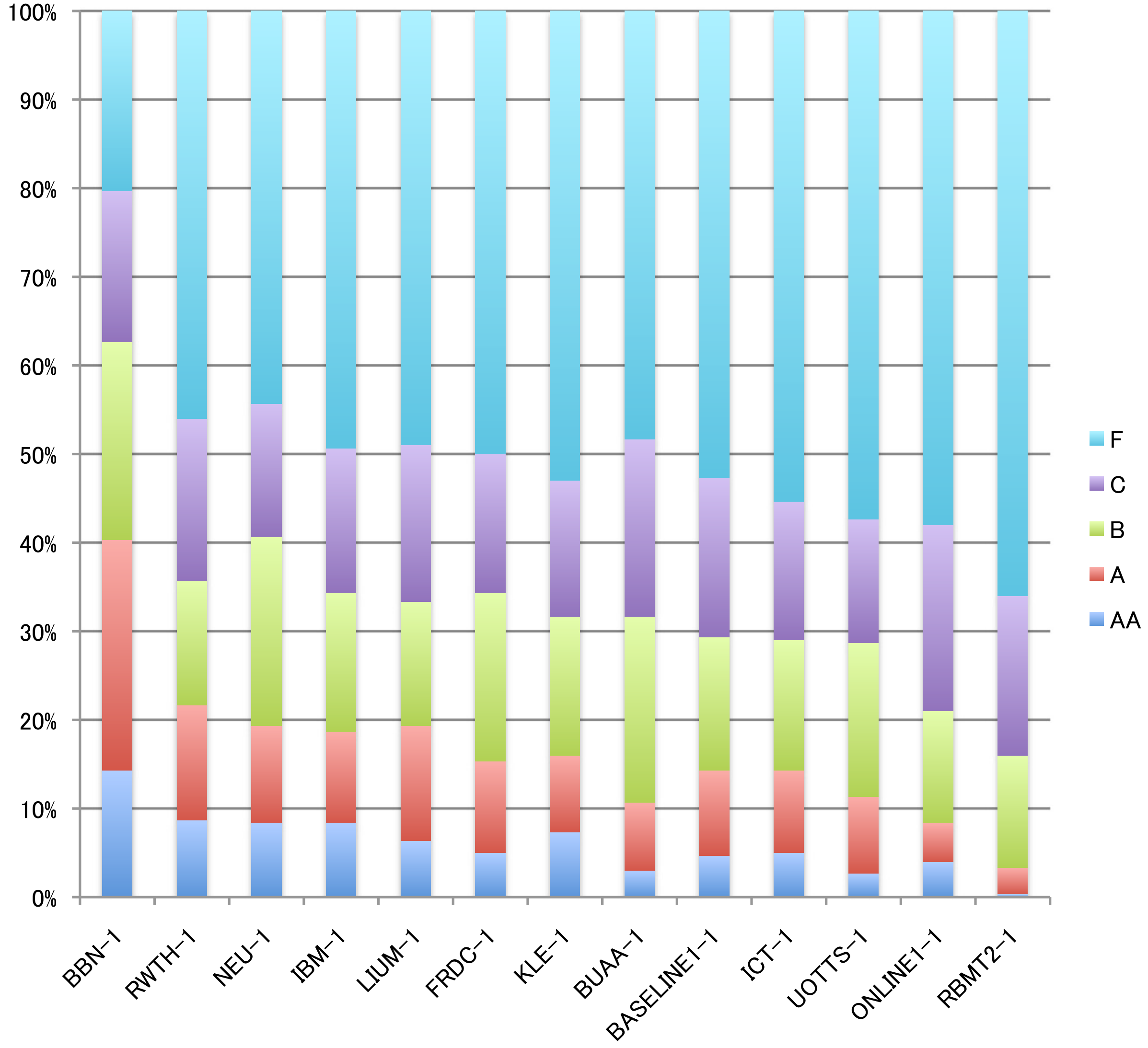
Baseline hierarchical SMT

Baseline non-hierarchical SMT

RBMT

- All the top systems were **SMT** systems.
- The top system (BBN-1) achieves **significantly better** scores than the other systems.
- The **hierarchical** phrase-based SMT baseline was **better** than the phrase-based SMT baseline.
- The SMT baseline systems were better than the baseline RBMT systems.

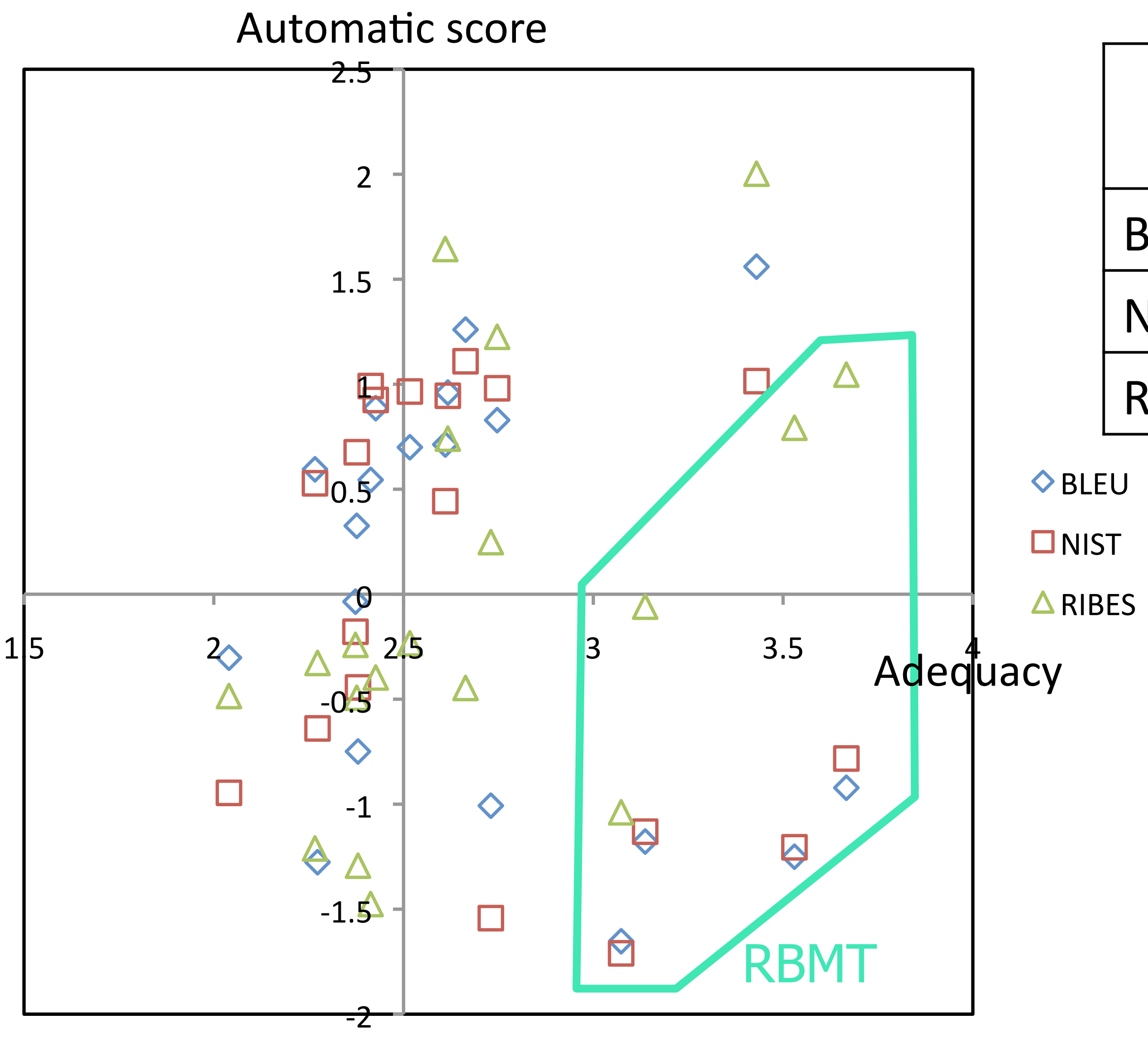
CE Acceptability Results



- **80%** sentences could be understood (C-rank and above) in the **best system** (BBN-1).
- This evaluation demonstrated the effectiveness of the BBN system.

Meta-Evaluation of Automatic Evaluation based on Human Evaluation

JE Correlations between Human and Auto

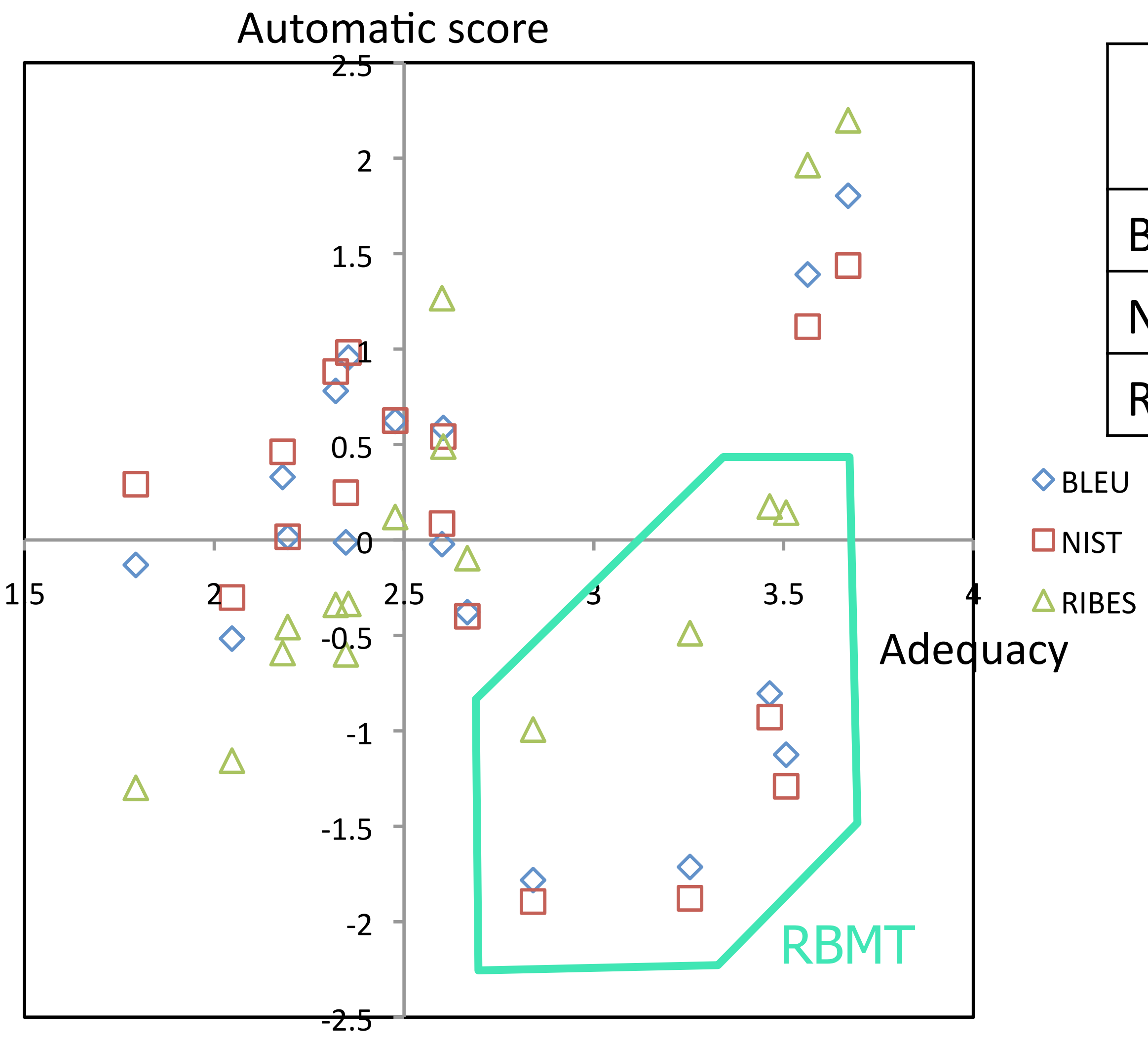


Spearman's ρ

	All	Excluding RBMT
BLEU	-0.042	0.618
NIST	-0.114	0.543
RIBES	0.632	0.679

- Reliability for RBMT was not high.
- RIBES was better than BLEU and NIST.

EJ Correlations between Human and Auto

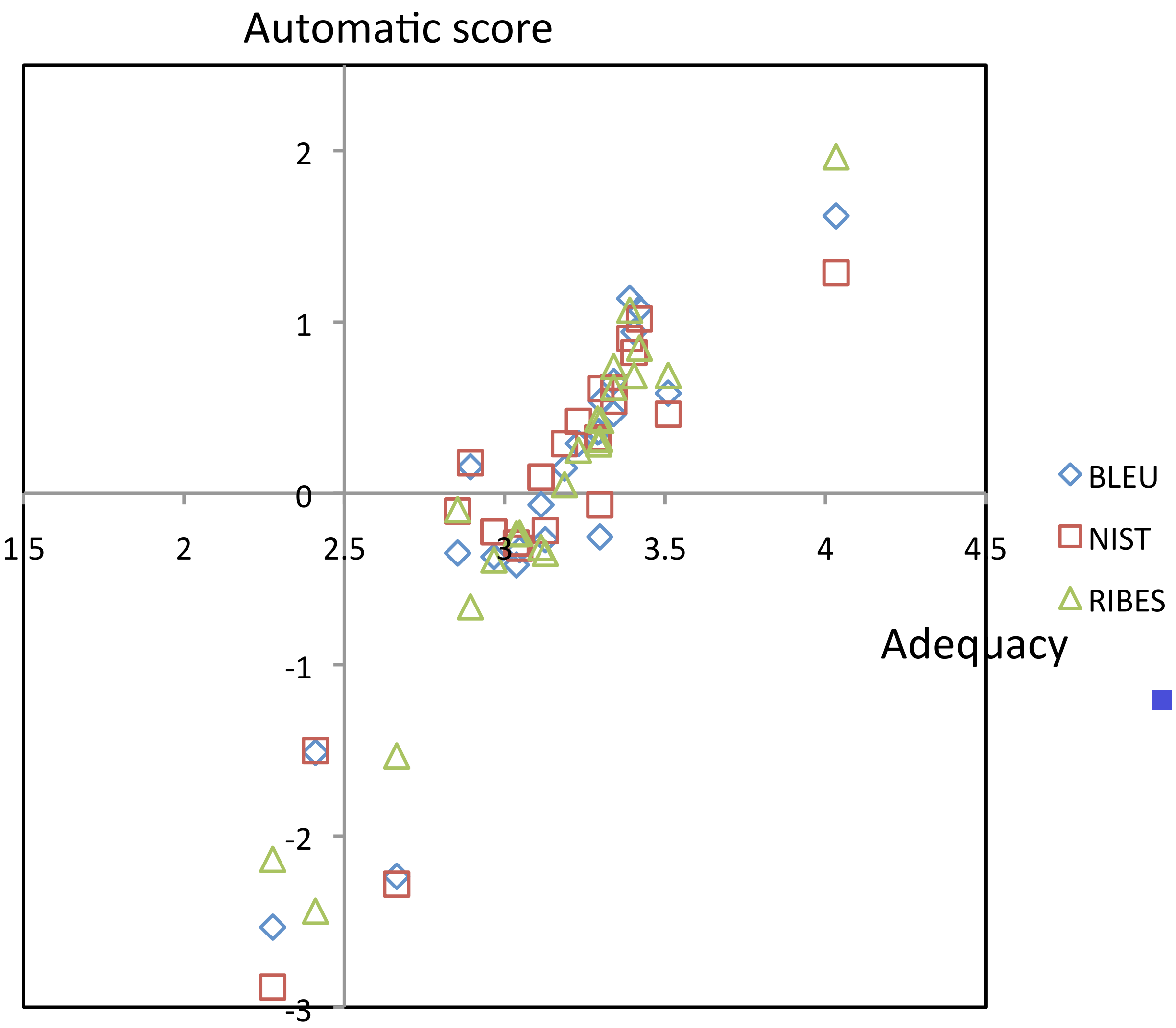


Spearman's ρ

	All	Excluding RBMT
BLEU	-0.029	0.511
NIST	-0.074	0.412
RIBES	0.716	0.929

- Reliability for RBMT was not high.
- RIBES was better than BLEU and NIST.

CE Correlations between Human and Auto



Spearman's ρ

	All
BLEU	0.931
NIST	0.911
RIBES	0.949

- Automatic scores had a high correlation with the human evaluation.

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.
- **8** baseline systems including 5 RBMT systems.
- **Human evaluations** were conducted.
- Various ideas were explored and the effectiveness of systems in patent translation was shown in evaluations.
- The effectiveness of each idea will be presented by the participants.

Oral Presentations of Participants

Group ID	Organization	Authors	Remarkable Results
BBN	*BBN Technologies, USA*	Jeff Ma and Spyros Matsoukas	The **best** system for the CE subtask
NTT-UT	*NTT Communication Science Labs., Japan and The University of Tokyo, Japan*	Katsuhito Sudoh et al.	The **best** system for the EJ subtask
NEU	*Northeastern University, P.R. China*	Tong Xiao et al.	Highly ranked system for the CE subtask
RWTH	*RWTH Aachen University, Germany*	Minwei Feng et al.	Highly ranked system for the CE subtask
IBM	*IBM T. J. Watson Research Center, USA*	Young-Suk Lee et al.	Highly ranked system for the CE subtask

(After this session, there will be a poster session that you are invited to attend.)

Thank you