

# Automatic Evaluation of Translation Quality for Distant Language Pairs

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh,  
Hajime Tsukada

NTT Communication Science Laboratories, NTT Corporation

October 10, 2010

## Motivation

**Word order** is essential in **Japanese-to-English** translation.

---

But popular metrics do not regard global word order very much.

---

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

*kare wa ame ni nureta* node *kaze wo hiita*

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

*kare wa ame ni nureta* *node* *kaze wo hiita*

SMT output: (performs only local reordering)

*he got soaked in the rain* because *he caught a cold*

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

*kare wa ame ni nureta* *node* *kaze wo hiita*

SMT output: (performs only local reordering)

*he got soaked in the rain* because *he caught a cold* ??

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

**kare wa ame ni nureta** *node* **kaze wo hiita**

SMT output: (performs only local reordering)

**he got soaked in the rain** because **he caught a cold** ??

Reference (requires global reordering):

**he caught a cold** because **he got soaked in the rain**

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

**kare wa ame ni nureta** node **kaze wo hiita**

SMT output: (performs only local reordering)

**he got soaked in the rain** because **he caught a cold** ??

Reference (requires global reordering):

**he caught a cold** because **he got soaked in the rain**

This inadequate SMT output obtains very good scores:

BLEU= 0.771, PER = 0.00, TER = 0.167

# Motivation

**Word order** is essential in **Japanese-to-English** translation.

But popular metrics do not regard global word order very much.

Most SMT systems make mistakes like this.

Japanese: (input)

**kare wa ame ni nureta** node **kaze wo hiita**

SMT output: (performs only local reordering)

**he got soaked in the rain** because **he caught a cold** ??

Reference (requires global reordering):

**he caught a cold** because **he got soaked in the rain**

This inadequate SMT output obtains very good scores:

BLEU= 0.771, PER = 0.00, TER = 0.167

We cannot believe these evaluation metrics at all.

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: *meari* wa *jon* wo koroshita.

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: *meari wa jon wo koroshita.*

SMT output: *John killed Mary.*

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: *meari wa jon wo koroshita.*

SMT output: *John killed Mary.*

Reference: *Mary killed John.*

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: *meari wa jon wo koroshita.*

SMT output: *John killed Mary.*

Reference: *Mary killed John.*

What will happen if you swap these two names?

Japanese: *jon wa meari wo koroshita.*

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: *meari wa jon wo koroshita.*

SMT output: John killed Mary .

Reference: Mary killed John .

What will happen if you swap these two names?

Japanese: *jon wa meari wo koroshita.*

SMT output: John killed Mary .

Reference: john killed Mary .

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: **meari** wa **jon** wo koroshita.

SMT output: **John** killed **Mary** .

Reference: **Mary** killed **John** .

What will happen if you swap these two names?

Japanese: **jon** wa **meari** wo koroshita.

SMT output: **John** killed **Mary** .

Reference: **john** killed **Mary** .

**This SMT service has a PREJUDICE!**

## Word order is essential: More examples

Now, we show real outputs of a famous state-of-the-art SMT service.

Japanese: **meari** wa **jon** wo koroshita.

SMT output: **John** killed **Mary** .

Reference: **Mary** killed **John** .

What will happen if you swap these two names?

Japanese: **jon** wa **meari** wo koroshita.

SMT output: **John** killed **Mary** .

Reference: **john** killed **Mary** .

**This SMT service has a PREJUDICE!** ... Language Model?

## Yet another example

Japanese:

*bobu wa meari ni yubiwa wo kau tameni, jon no mise ni itta.*

## Yet another example

Japanese:

*bobu wa meari ni yubiwa wo kau tameni, jon no mise ni itta.*

SMT output:

Bob Mary to buy the ring, John went to the store.

## Yet another example

Japanese:

**bobu** wa **meari** ni yubiwa wo kau tameni, **jon** no mise ni itta.

SMT output:

**Bob** **Mary** to buy the ring, **John** went to the store.

????

## Yet another example

Japanese:

**bobu** wa **meari** ni yubiwa wo kau tameni, **jon** no mise ni itta.

SMT output:

**Bob** **Mary** to buy the ring, **John** went to the store.

????

Reference:

**Bob** went to **John**'s store to buy a ring for **Mary**.

## Yet another example

Japanese:

**bobu** wa **meari** ni yubiwa wo kau tameni, **jon** no mise ni itta.

SMT output:

**Bob** **Mary** to buy the ring, **John** went to the store.

????

Reference:

**Bob** went to **John**'s store to buy a ring for **Mary**.

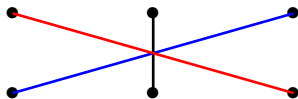
In this way, wrong word order leads to misunderstanding or incomprehensibility.

# Our Proposal

We propose new metrics based on

**rank correlation coefficients** (Kendall's  $\tau$  and Spearman's  $\rho$ ) to measure the difference of word order.

Reference: **Mary**<sub>1</sub> killed<sub>2</sub> **John**<sub>3</sub>



MT output: **John**<sub>3</sub> killed<sub>2</sub> **Mary**<sub>1</sub>

In this case,  $\tau = \rho = -1$ .

## Main Result (Meta-Evaluation of NTCIR-8 PATMT JE task, 15 systems)

Proposed metrics (single reference) have strong correlation with human judgements (adequacy and fluency)

Proposed metric	based on $\tau$	0.940
Proposed metric	based on $\rho$	0.947

## Main Result (Meta-Evaluation of NTCIR-8 PATMT JE task, 15 systems)

Proposed metrics (single reference) have strong correlation with human judgements (adequacy and fluency)

<b>Proposed metric</b>	based on $\tau$	<b>0.940</b>
<b>Proposed metric</b>	based on $\rho$	<b>0.947</b>
<b>BLEU</b>		<b>0.515</b>

## Main Result (Meta-Evaluation of NTCIR-8 PATMT JE task, 15 systems)

Proposed metrics (single reference) have strong correlation with human judgements (adequacy and fluency)

Proposed metric	based on $\tau$	0.940
Proposed metric	based on $\rho$	0.947
BLEU		0.515

You should not use **BLEU** for Japanese-to-English translation!

## Main Result (Meta-Evaluation of NTCIR-8 PATMT JE task, 15 systems)

Proposed metrics (single reference) have strong correlation with human judgements (adequacy and fluency)

Proposed metric	based on $\tau$	0.940
Proposed metric	based on $\rho$	0.947
BLEU		0.515

You should not use **BLEU** for Japanese-to-English translation!

Thank you!