Dependency Analysis *of* Scrambled References *for Better Evaluation of Japanese Translations*

Hideki ISOZAKI and Natsume KOUCHI

Okayama Prefectural University, Japan

WMT-2015

MAIN FOCUS OF THIS TALK

Isozaki+ 2014 proposed a method for regarding **SCRAMBLING** in **automatic evaluation of translation quality** with **RIBES**.

Here, we present its improvement.

What is **SCRAMBLING**?

What is **RIBES**?

OUTLINE

Background 1: SCRAMBLING

For instance, a Japanese sentence:

S1: John-**ga** Tokyo-**de** PC-**wo** (katta)。

(John (bought) a PC in Tokyo.)

can be reordered in the following ways. ( ) indicates a verb/adjective.

1. John-**ga** Tokyo-**de** PC-**wo** (katta)
2. John-**ga** PC-**wo** Tokyo-**de** (katta)
3. Tokyo-**de** John-**ga** PC-**wo** (katta)
4. Tokyo-**de** PC-**wo** John-**ga** (katta)
5. PC-**wo** John-**ga** Tokyo-**de** (katta)
6. PC-**wo** Tokyo-**de** John-**ga** (katta)

This is **SCRAMBLING** and some other languages such as German also have SCRAMBLING.

Background 1: SCRAMBLING

Japanese is known as a free word order language, but it is not completely free.

John-**ga** Tokyo-**de** PC-**wo** (katta)

Japanese Word Order Constraint 1:

Case markers (**ga**=subject, **de**=location, **wo**=object) should follow corresponding noun phrases.

Japanese Word Order Constraint 2:

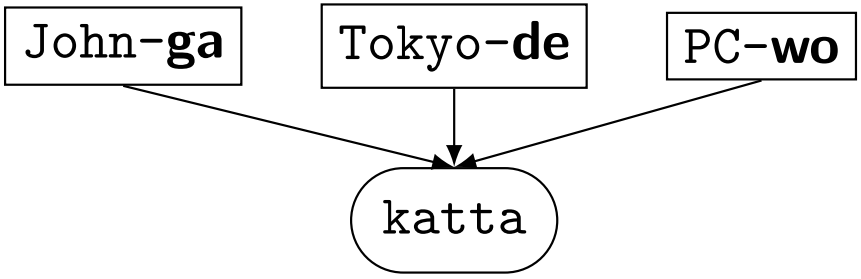
Japanese is a **head final** language.

A head should appear after all of its modifiers (dependents).

Here, the verb (katta) (bought) is the head.

Background 1: SCRAMBLING

S1 has this dependency tree:



The verb (katta) has three children.
The above scrambled sentences are **permutations of the three children** ($3! = 6$).

1. John-**ga** Tokyo-**de** PC-**wo** katta
2. John-**ga** PC-**wo** Tokyo-**de** katta
3. Tokyo-**de** John-**ga** PC-**wo** katta
4. Tokyo-**de** PC-**wo** John-**ga** katta
5. PC-**wo** John-**ga** Tokyo-**de** katta
6. PC-**wo** Tokyo-**de** John-**ga** katta

OUTLINE

Background 2: RIBES

RIBES is our new evaluation metric **designed for translation between distant language pairs such as Japanese and English**. (Isozaki+ EMNLP-2010, Hirao+ 2014)

RIBES measures **word order similarity** between an MT output and a reference translation.

RIBES shows a **strong correlation with human-judged adequacy** in EJ/JE translation.

Nowadays, most papers on JE/EJ translation use both **BLEU** and **RIBES** for evaluation.

Our meta-evaluation with NTCIR-7 JE data

System-level Spearman's ρ with adequacy, Single reference, 5 MT systems

BLEU	METEOR	ROUGE-L	IMPACT	**RIBES**
0.515	0.490	0.903	0.826	**0.947**

Meta-evaluation by NTCIR-9 PatentMT organizers.

System-level Spearman's ρ with adequacy, single reference, 17 MT systems

	BLEU	NIST	**RIBES**
NTCIR-9 JE	**−0.042**	−0.114	**0.632**
NTCIR-9 EJ	**−0.029**	−0.074	**0.716**
NTCIR-10 JE	**0.31**	0.36	**0.88**
NTCIR-10 EJ	**0.36**	0.22	**0.79**

Background 2: RIBES

SMT tends to follow the global word order given in the source.

In English ↔ Japanese translation, this tendency causes **swap of Cause and Effect**, but **BLEU** disregards the swap and overestimates SMT output.

Source: 彼は雨に濡れたので、風邪をひいた

Reference translation:

He caught a cold because he got soaked in the rain.

SMT output: **BLEU=0.74 very good!?**

He got soaked in the rain because he caught a cold.

Such an inadequate translation should be penalized much more.

Therefore, we designed **RIBES** to measure **word order**.

Background 2: RIBES

$$\textbf{RIBES} \stackrel{\text{def}}{=} \textbf{NKT} \times P^{\alpha} \times \textbf{BP}^{\beta}$$

where $\textbf{NKT} \stackrel{\text{def}}{=} \frac{\tau + 1}{2}$ is normalized Kendall's τ which **measures similarity of word order**.

P is unigram precision. **BP** is **BLEU**'s Brevity Penalty.

α and β are parameters for these penalties.

Default values are $\alpha = 0.25$, $\beta = 0.10$.

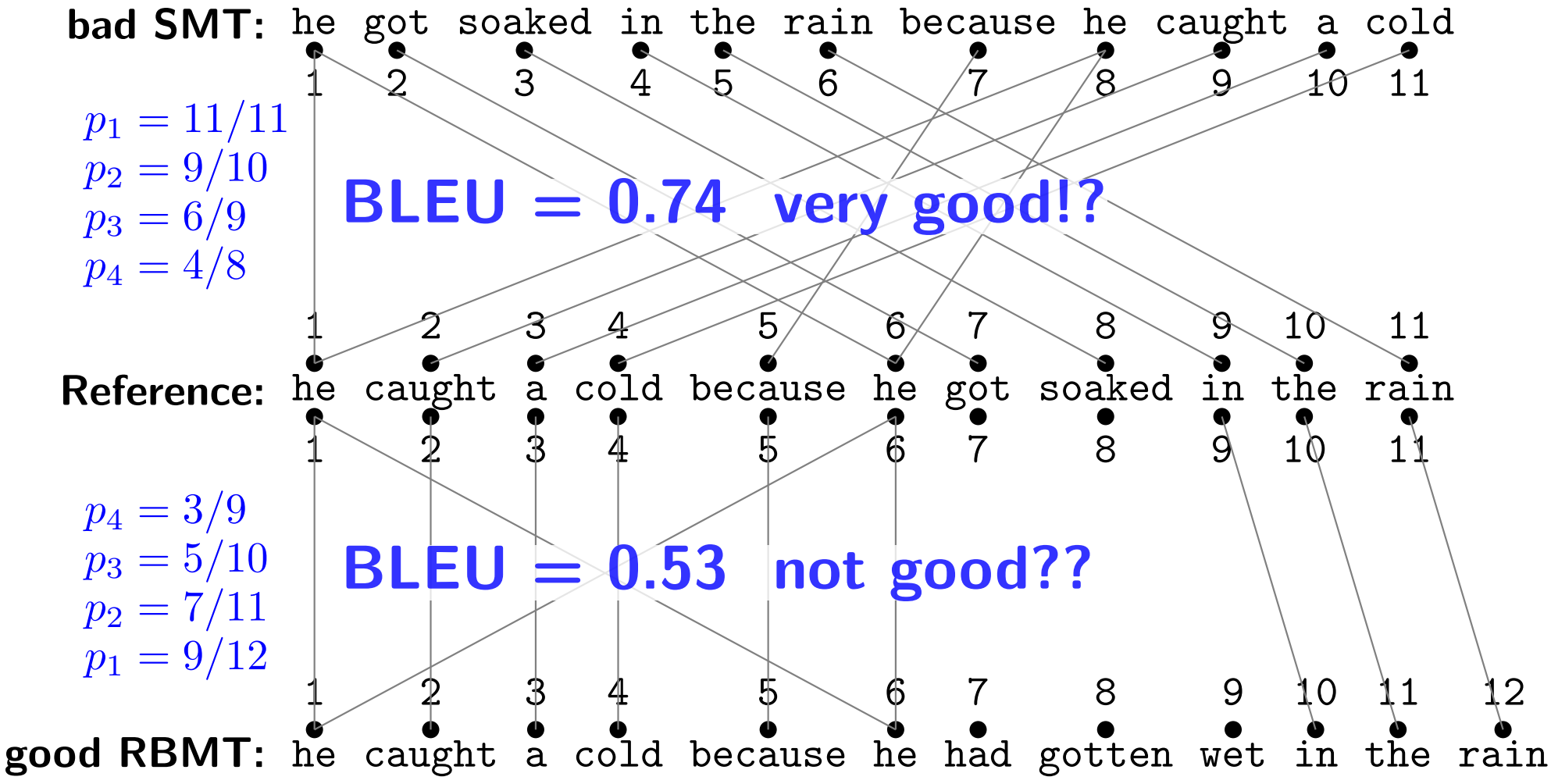
$$\text{(worst) } 0.0 \leq \textbf{RIBES} \leq 1.0 \text{ (best)}$$

`http://www.kecl.ntt.co.jp/icl/lirg/ribes/`

Hirao et al.: Evaluating Translation Quality with Word Order Correlations (in Japanese), Journal of Natural Language Processing, Vol. 21, No. 3, pp.421–444, 2014.

Background 2: RIBES

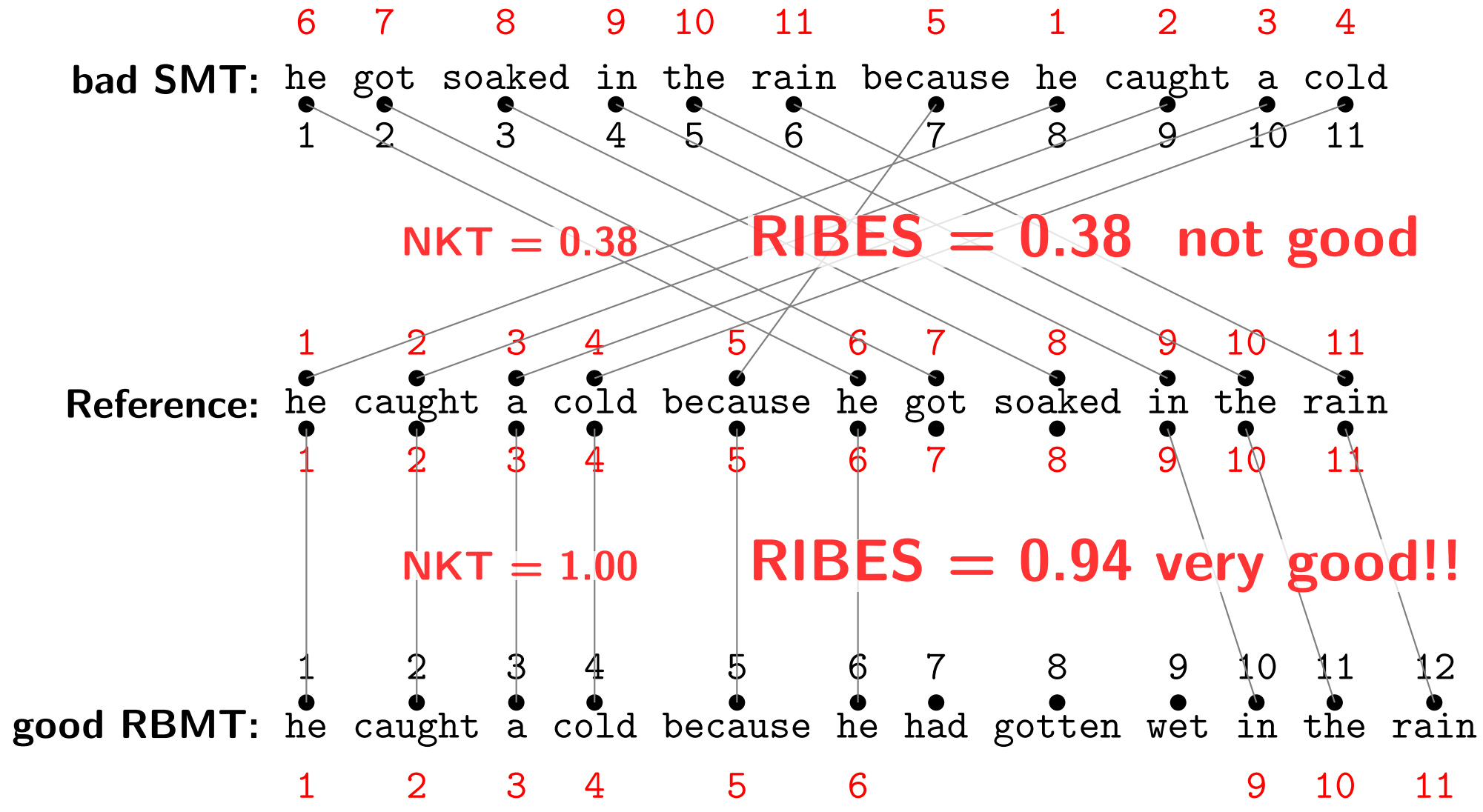
BLEU tends to prefer bad SMT output to good RBMT output.



BLUE is counterintuitive.

Background 2: RIBES

RIBES tends to prefer good RBMT output to bad SMT output.



RIBES is more intuitive.

RIBES versus SCRAMBLING

However, **RIBES** underestimates scrambled sentences.

Reference: `John-`**`ga`** `Tokyo-`**`de`** `PC-`**`wo`** (`katta`)

MT output: `PC-`**`wo`** `Tokyo-`**`de`** `John-`**`ga`** (`katta`)

This MT output is perfect for most Japanese speakers.

But its **RIBES** score is very low: 0.43.

Can we make the **RIBES** score higher?

OUTLINE

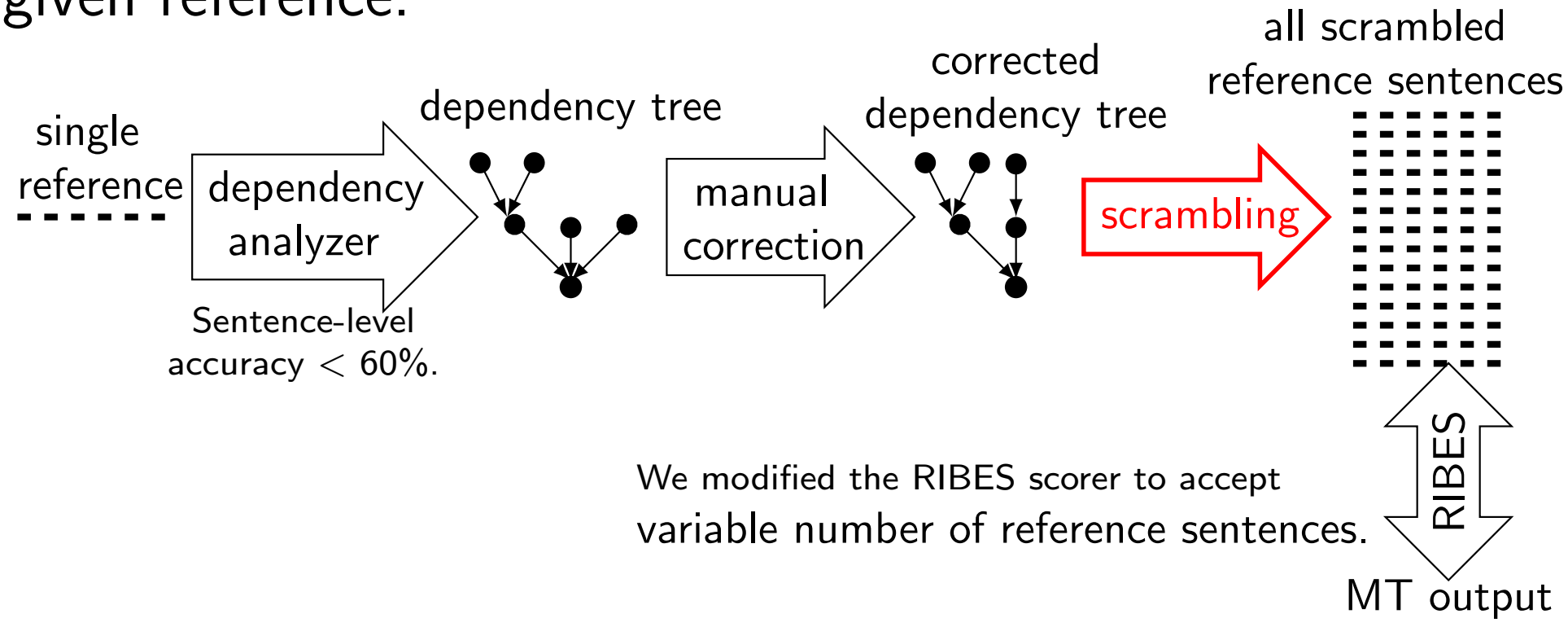
Our Idea in WMT-2014

Generate all scrambled sentences

from the given reference.

Then, use them as reference sentences.

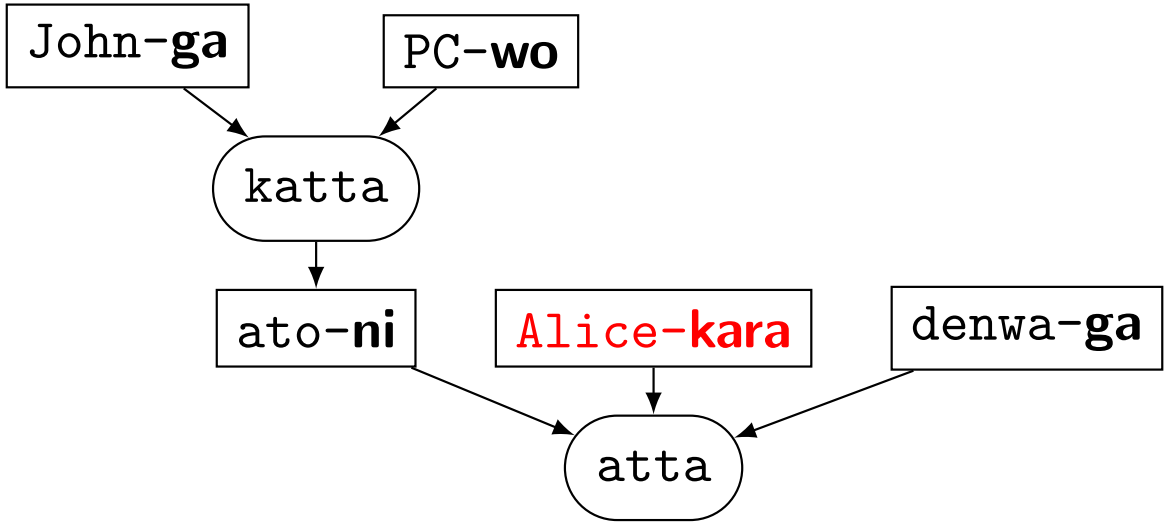
For this generation, we need the dependency tree of the given reference.



Scrambling by Post-Order traversal

S2: John-**ga** PC-**wo** katta ato-**ni** Alice-**kara** denwa-**ga** atta.
(After John bought a PC, there was a phone call from Alice.)

S2 has two verbs: katta (bought) and atta (was).



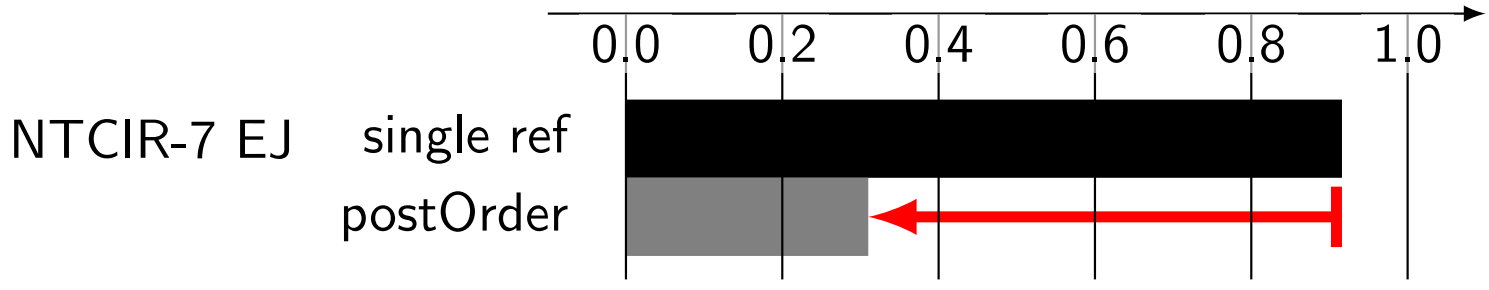
In order to generate Japanese-like **head final** sentences, we should output words in the dependency tree in **Post Order**.
But siblings can be output in any order.

In this case, we can generate $2! \times 3! = 12$ permutations.

Scrambling by Post-Order traversal

Now, we can generate scrambled references from the dependency tree of a reference sentence.

We used all scrambled sentences as references (postOrder).
But it damaged system-level correlation with adequacy.



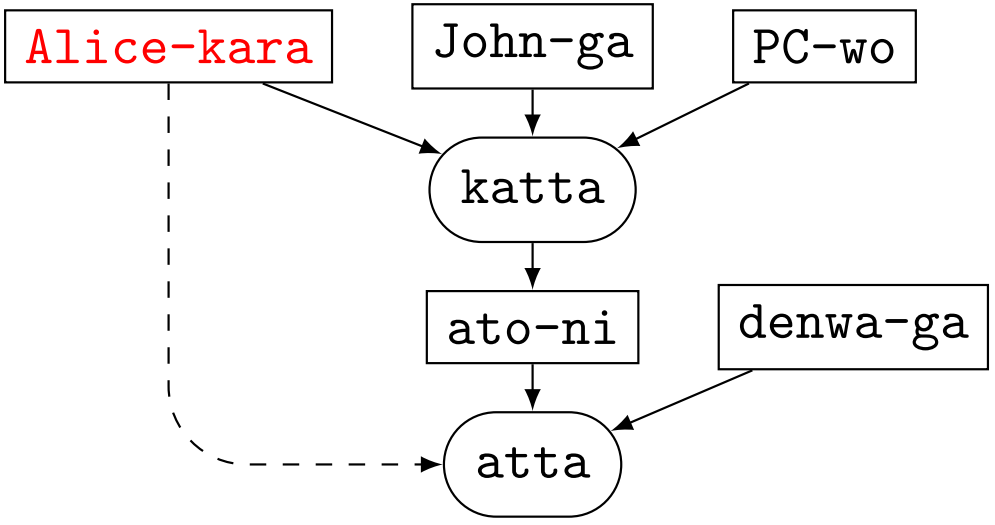
Perhaps, some scrambled sentences are not appropriate as references and they increases RIBES scores of bad MT outputs.

Scrambling of a complex sentence

S2: John-ga PC-wo (katta) ato-**ni** Alice-**kara** denwa-**ga** (atta).
(After John (bought) a PC, there (was) a phone call from Alice.)

One of S2's postOrder outputs is:

S2bad: Alice-**kara** John-**ga** PC-**wo** (katta) ato-**ni** denwa-**ga** (atta).
(After John (bought) a PC from Alice, there (was) a phone call.)



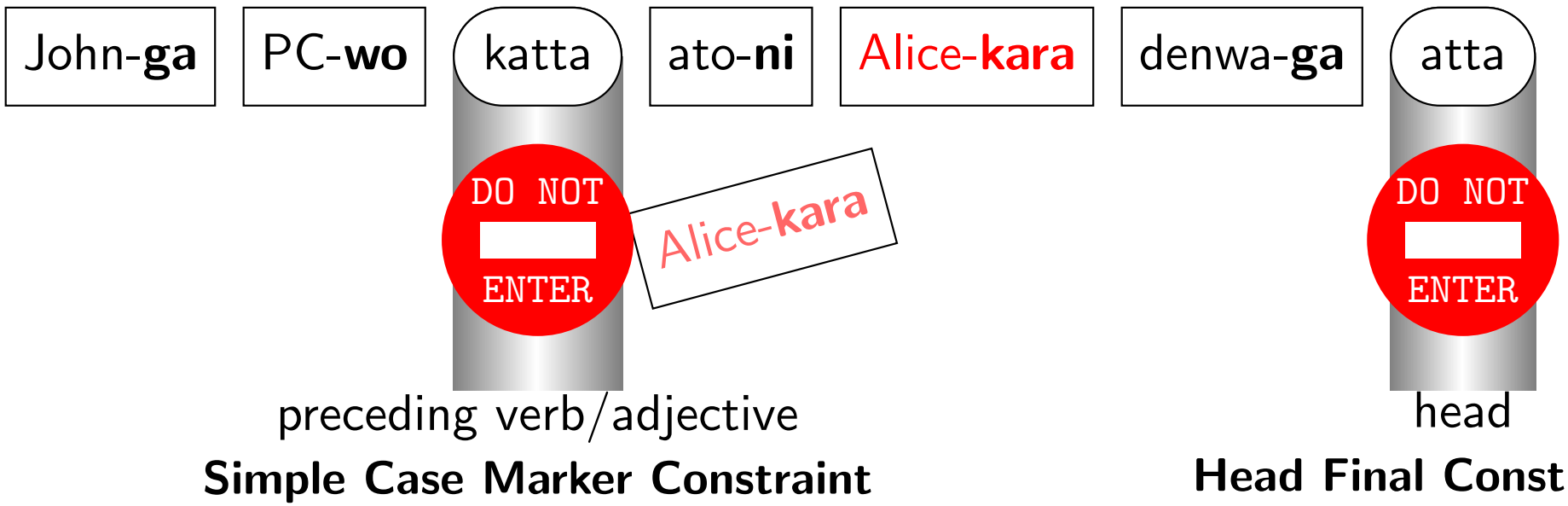
We should inhibit such misleading sentences.

Scrambling of a Complex Sentence

In order to inhibit such misleading sentences, Isozaki+ 2014 introduced

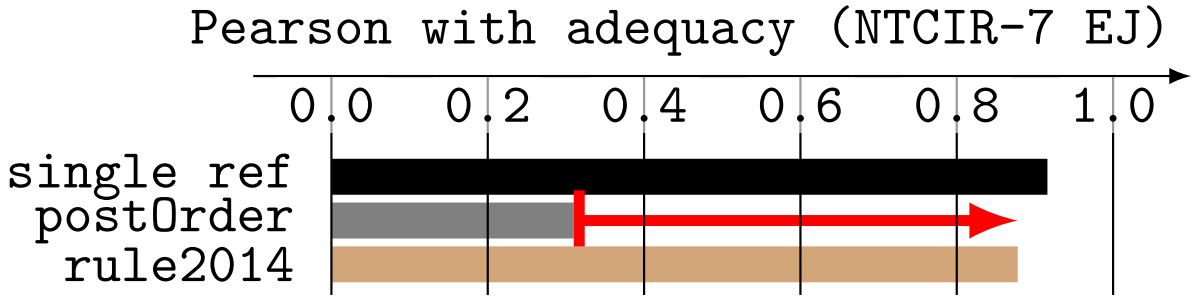
Simple Case Marker Constraint (rule2014)

You should not put case-marked modifiers of a verb/adjective before a preceding verb/adjective.

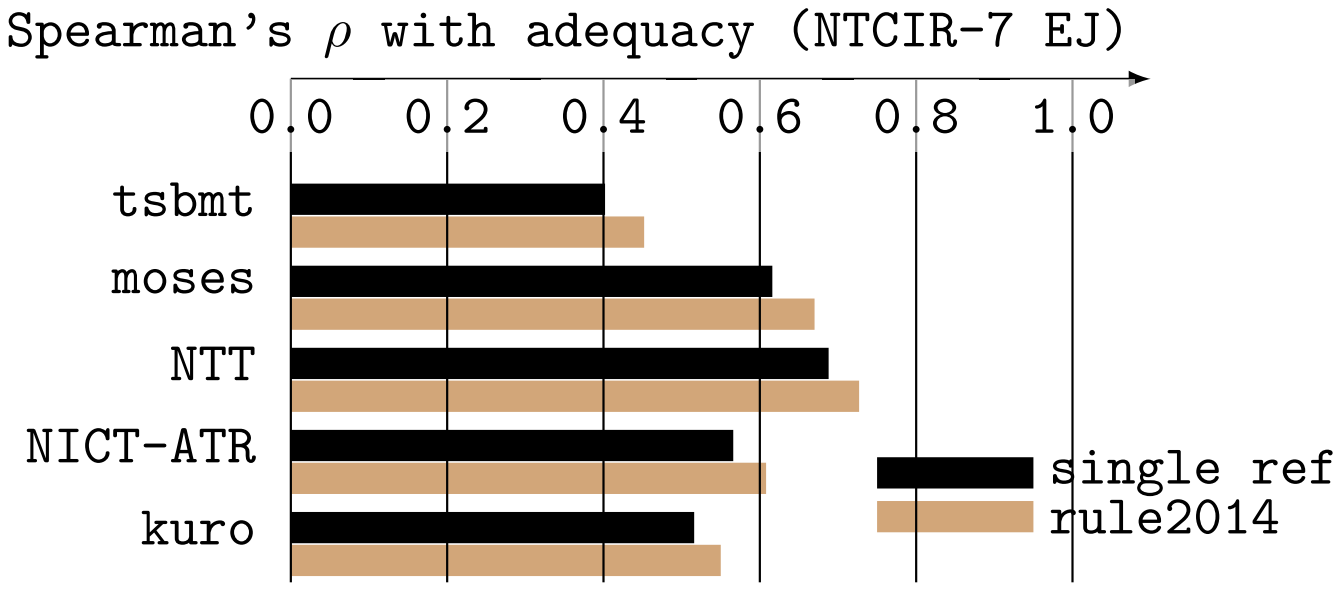


Effectiveness of rule2014

System-level correlation with adequacy was **recovered**.



Sentence-level correlation with adequacy was **improved**.



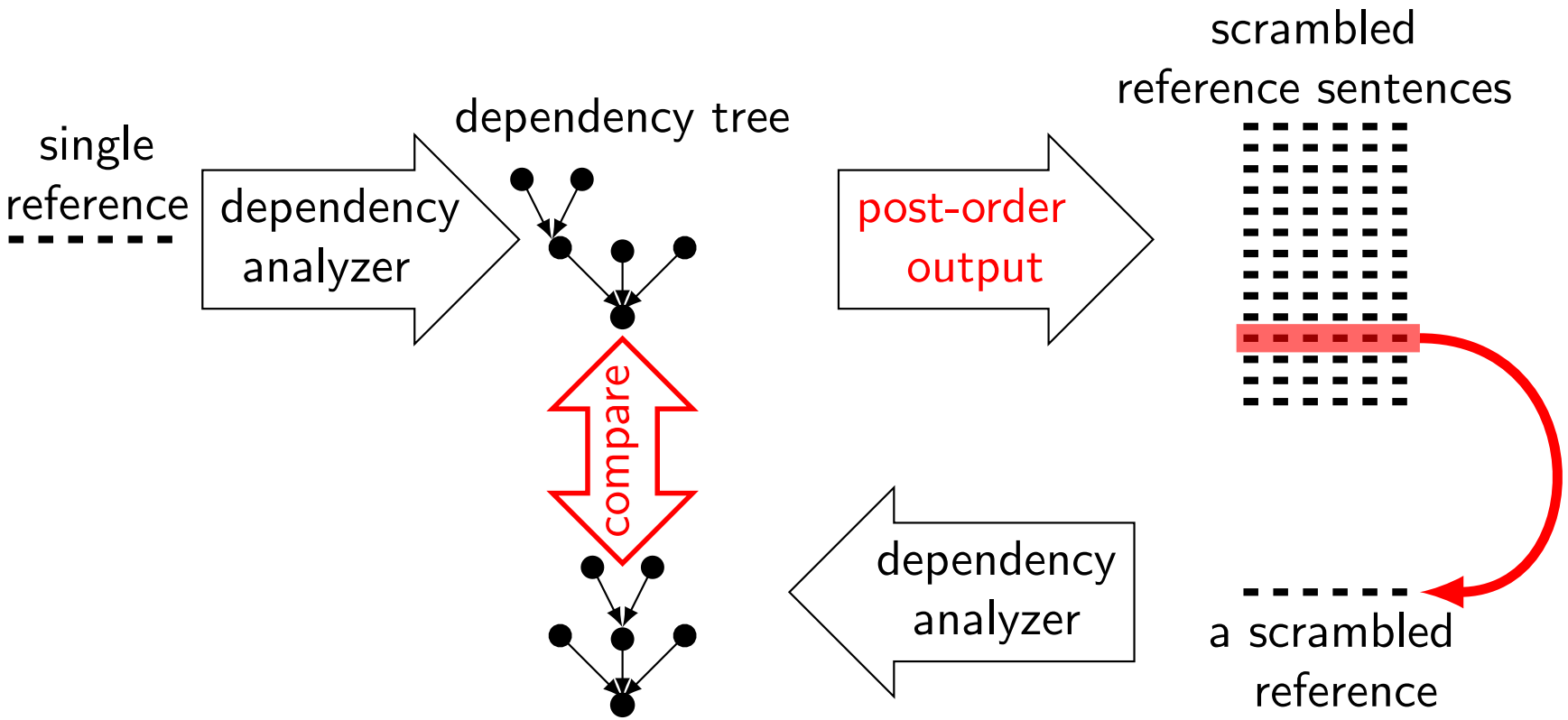
Problems of rule2014

- It covered **only 30% of NTCIR-7 EJ reference sentences**.
  (covered = generated alternative word orders for)
- In order to cover more sentences, **we will need more rules**.
- It requires **manual correction** of dependency trees.

OUTLINE

NEW IDEA for WMT-2015

If a sentence is misleading, parsers will be misled.



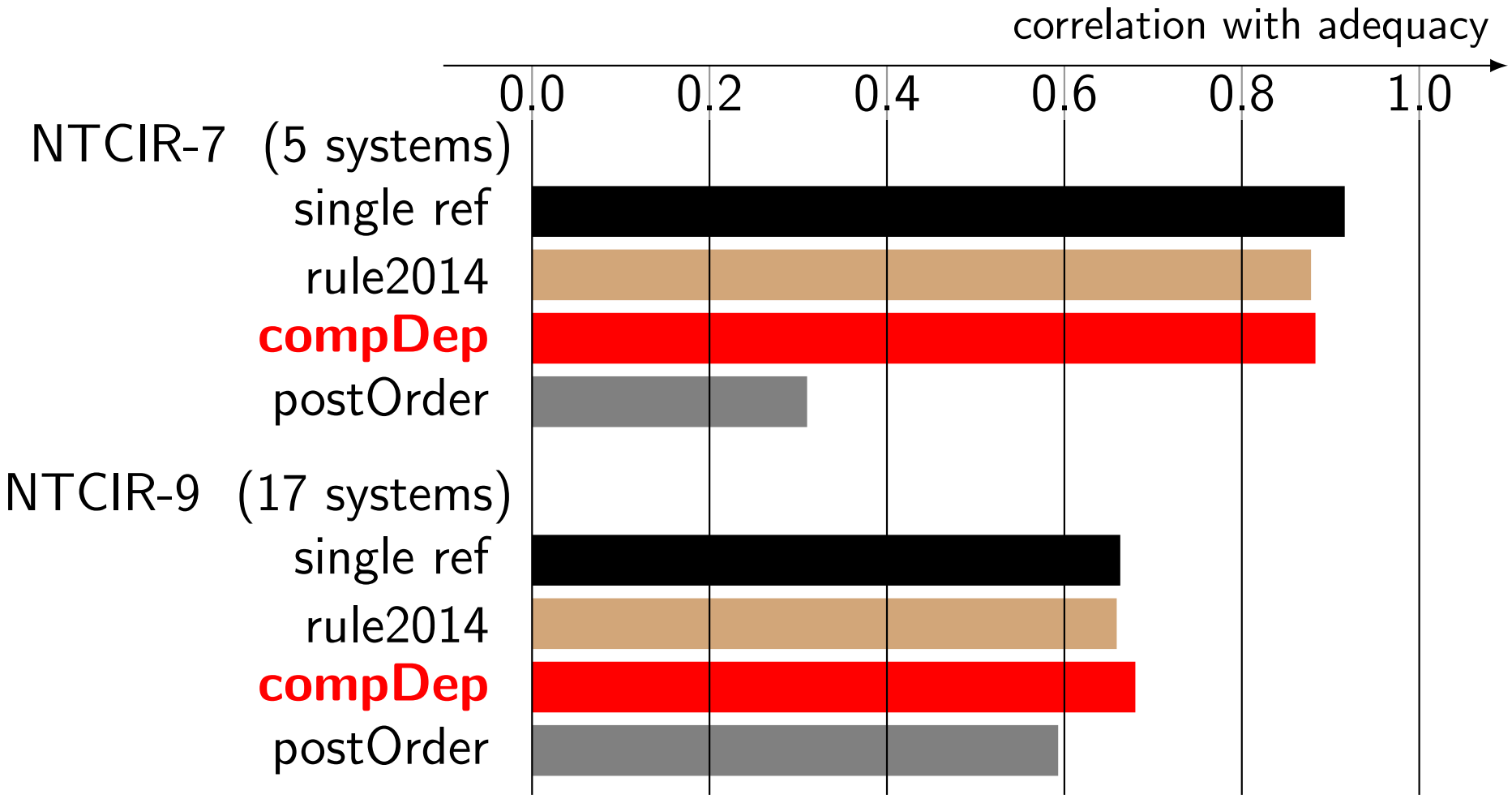
compDep (compare dependency trees):

If the two dependency trees are the same except sibling orders, we accept the new word order as a new reference.

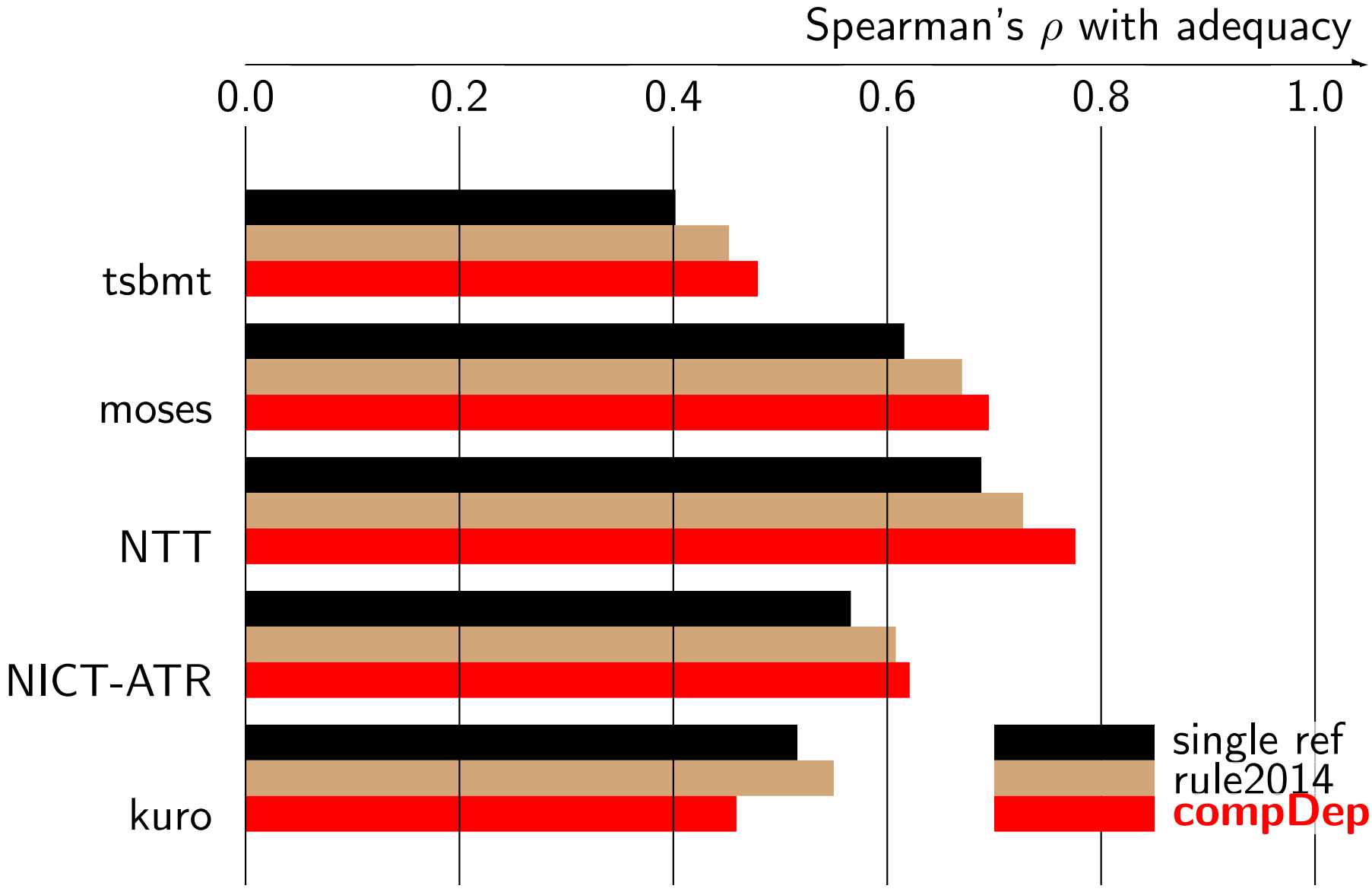
Otherwise, this word order is misleading and we reject it.

System-level correlation with adequacy

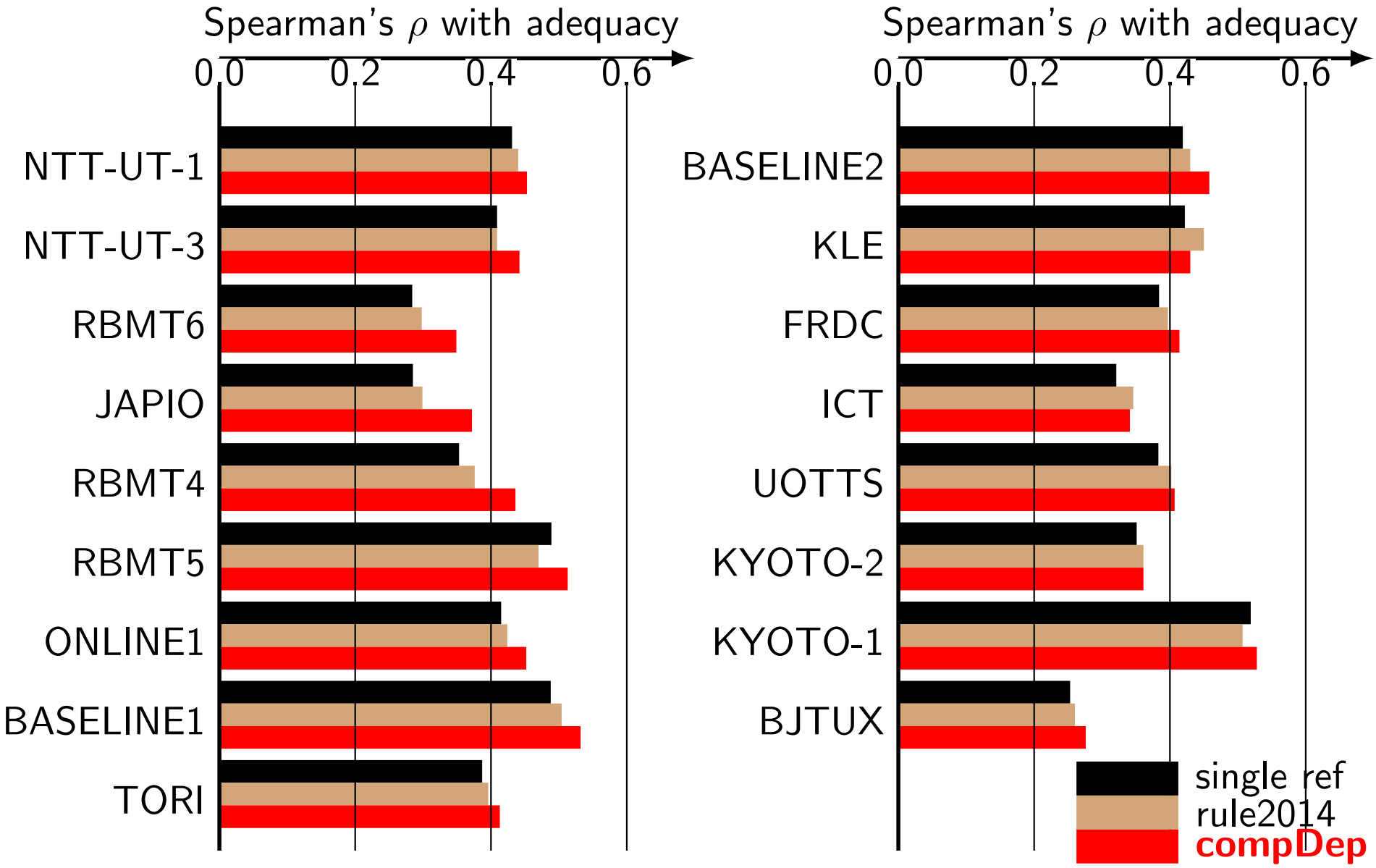
compDep's system-level correlation with adequacy is comparable to `single ref`'s and `rule2014`'s.



Improvement of sentence-level correlation with adequacy (NTCIR-7 JE)



Improvement of sentence-level correlation with adequacy (NTCIR-9 JE)



Number of generated word orders

compDep covers more reference sentences than rule2014.

	#perms	1	2–10	11–100	101–1000	>1000	total
NTCIR-7 EJ	single ref	100	0	0	0	0	100
	rule2014	**70**	30	0	0	0	100
	compDep	**20**	61	15	4	0	100
	postOrder	1	41	41	13	4	100

	#perms	1	2–10	11–100	101–1000	>1000	total
NTCIR-9 EJ	single ref	300	0	0	0	0	300
	rule2014	**267**	25	7	1	0	300
	compDep	**41**	189	63	5	2	300
	postOrder	0	100	124	58	18	300

compDep failed to generate alternative word orders for only (20+41)/(100+300)=15.3% of reference sentences while rule2014 failed for (70+267)/(100+300) = 84.3%.

Conclusions

We proposed **compDep** method to regard scrambling in automatic evaluation of translation quality with **RIBES**.

Experimental results show that

- **compDep** improved **sentence**-level correlation with human-judged adequacy.
- **compDep** does not damage the strong **system**-level correlation of **RIBES** very much.
- **compDep** covers $100\% - 15.3\% = 84.7\%$ of reference sentences.
- Manual correction does not change the results very much. (skipped in this talk).

Future work

- Application to other evaluaion measures such as **BLEU**.
- Application to other languages such as German.