

NTT-UT SMT System for NTCIR-9 PatentMT

Katsuhito Sudoh, Kevin Duh,
Hajime Tsukada, Masaaki Nagata
*NTT Communication Science
Laboratories
Kyoto, Japan*

Xianchao Wu, Takuya Matsuzaki,
Jun'ichi Tsujii
*The University of Tokyo
Tokyo, Japan*

Overview

| task | English-Japanese | Japanese-English | Chinese-English |
|------------------------|---|--|---|
| single system features | <ul style="list-style-type: none">• Pre-ordering• Big LM• WFST decode | <ul style="list-style-type: none">• Pre-ordering | <ul style="list-style-type: none">• WA Adaptation |
| additional feature | Sys. Comb. + U-Tokyo forest-to-tree | Sys. Comb. + U-Tokyo HPBMT | Sys. Comb. + U-Tokyo HPBMT |
| rank | 1st | 5th | 9th |

Overview

| task | English-Japanese | Japanese-English | Chinese-English |
|------------------------|---|--|---|
| single system features | <ul style="list-style-type: none"> • Pre-ordering • Big LM • WFST decode | <ul style="list-style-type: none"> • Pre-ordering | <ul style="list-style-type: none"> • WA Adaptation |
| additional feature | Sys. Comb. + U-Tokyo forest-to-tree | | Sys. Comb. + U-Tokyo RBMT |
| rank | 1st | 5th | 9th |

Better than RBMT even in Subjective Evaluation!!!

Overview

| task | English-Japanese | Japanese-English | Chinese-English |
|------------------------|---|--|---|
| single system features | <ul style="list-style-type: none"> • Pre-ordering • Big LM • WFST decode | <ul style="list-style-type: none"> • Pre-ordering | <ul style="list-style-type: none"> • WA Adaptation |
| additional feature | Sys. Comb. + U-Tokyo forest-to-tree | | Sys. Comb. + U-Tokyo RBMT |
| rank | 1st | 5th | 9th |

Better than RBMT even in Subjective Evaluation!!!

Today's Focus !

Head Finalization for En-Ja pre-ordering

Head Finalization for En-Ja pre-ordering

- Isozaki et al. (WMT 2010)

Head Finalization for En-Ja pre-ordering

- Isozaki et al. (WMT 2010)
- Moving heads to rhs on HPSG tree
 - English HPSG Parser “Enju” (U-Tokyo)

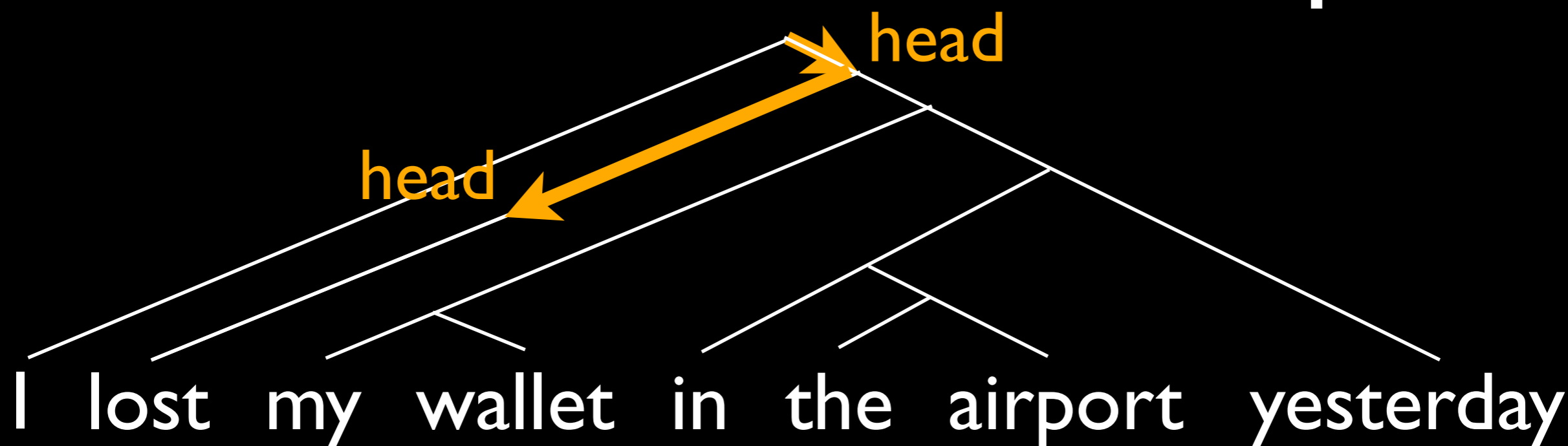
Head Finalization for En-Ja pre-ordering

- Isozaki et al. (WMT 2010)
- Moving heads to rhs on HPSG tree
 - English HPSG Parser “Enju” (U-Tokyo)
- Pseudo-word insertion for Ja particles
 - Predicate-argument structure by Enju

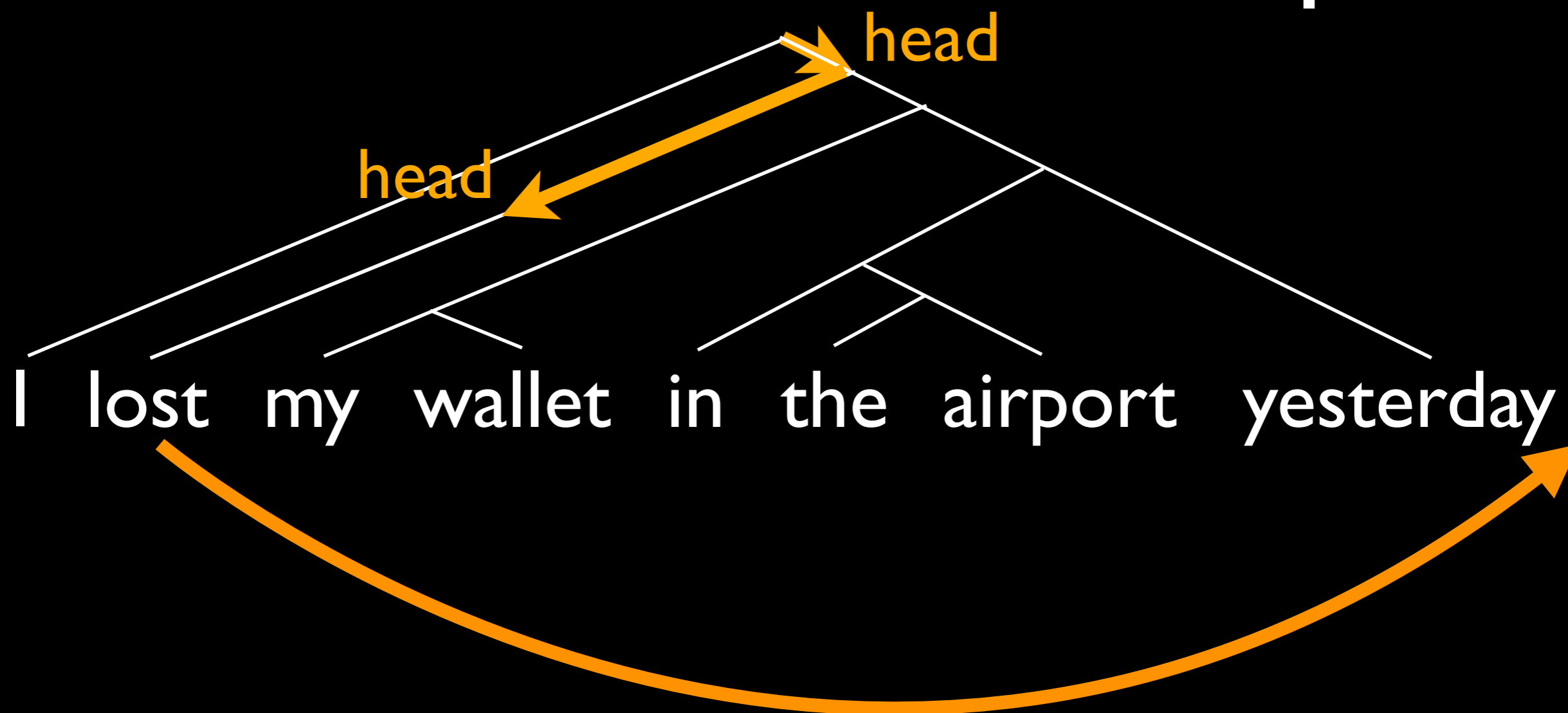
Head Finalization for En-Ja pre-ordering

- Isozaki et al. (WMT 2010)
- Moving heads to rhs on HPSG tree
 - English HPSG Parser “Enju” (U-Tokyo)
- Pseudo-word insertion for Ja particles
 - Predicate-argument structure by Enju
- Determiner (a/an/the) deletion

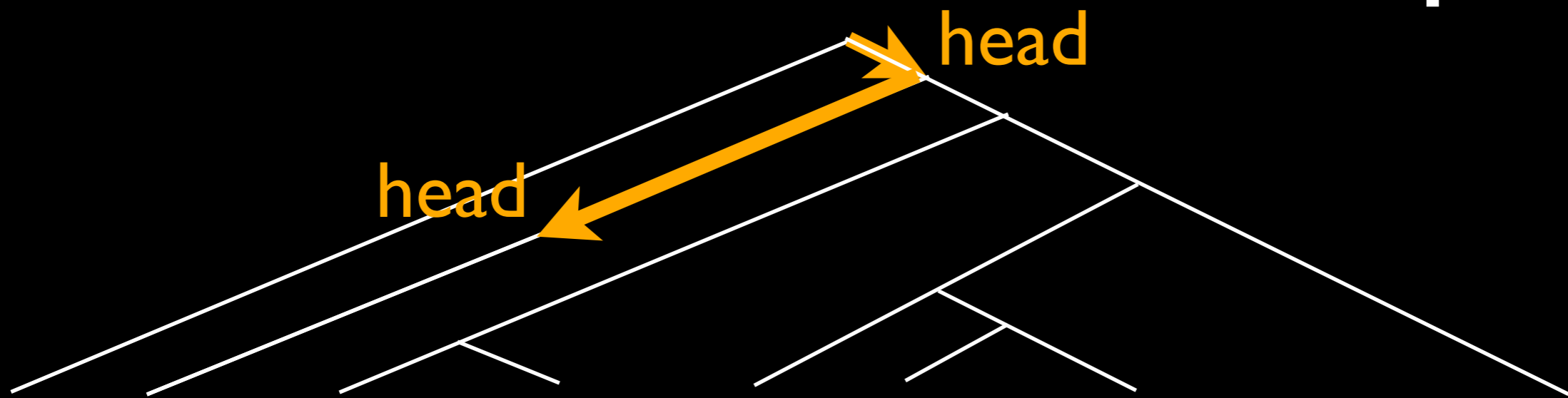
Head Finalization Example



Head Finalization Example

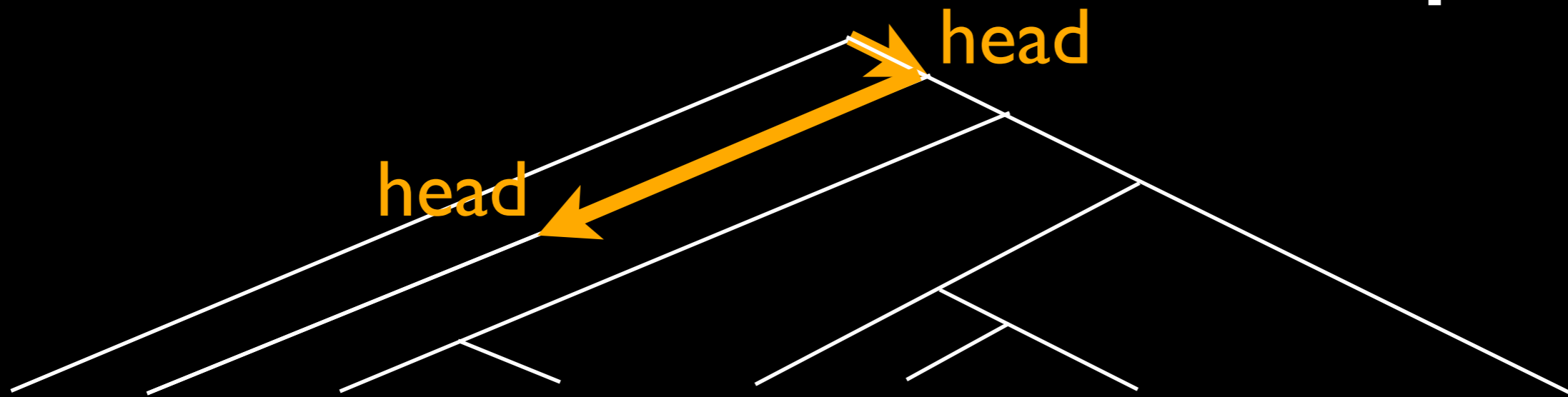


Head Finalization Example



I yesterday the airport in my wallet lost

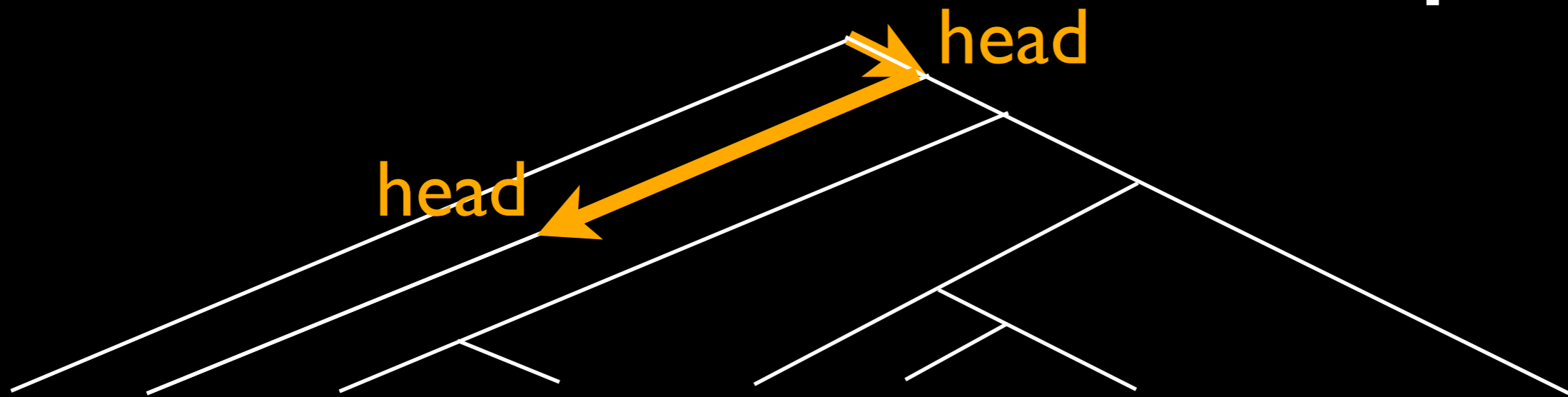
Head Finalization Example



I yesterday the airport in my wallet lost

- Move Heads

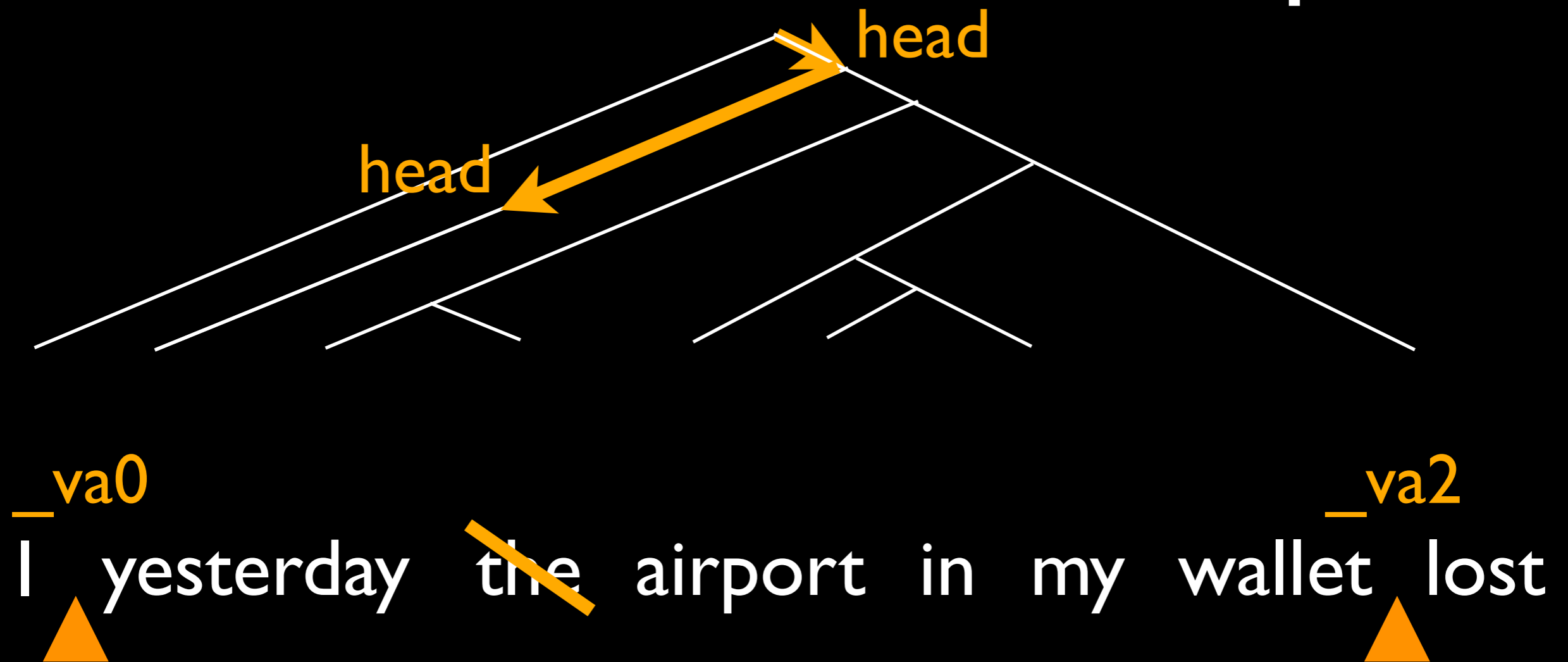
Head Finalization Example



I yesterday ~~the~~ airport in my wallet lost

- Move Heads
- Remove a, an, the

Head Finalization Example



- Move Heads
- Remove a, an, the
- Insert pseudo-particles for subjects & objects

Head Finalization Example

I **_va0** yesterday airport in my wallet **_va2** lost

Head Finalization Example

I **_va0** yesterday airport in my wallet **_va2** lost

私は昨日空港で私の財布をなくした

Head Finalization Example

I **_va0** yesterday airport in my wallet **_va2** lost

私は昨日空港で私の財布をなくした

Monotone Translation !!

Japanese Big LM

- Word 5-gram LM from **300M** Ja sentences

WFST-based Monotone Decoding

WFST-based Monotone Decoding

- MT becomes monotone by pre-ordering

WFST-based Monotone Decoding

- MT becomes monotone by pre-ordering
- Efficient decoding by WFST

WFST-based Monotone Decoding

- MT becomes monotone by pre-ordering
- Efficient decoding by WFST
 - phrase segmentation > phrase translation
> word segmentation > LM

WFST-based Monotone Decoding

- MT becomes monotone by pre-ordering
- Efficient decoding by WFST
 - phrase segmentation > phrase translation
> word segmentation > LM
- Efficient on-the-fly composition

WFST-based Monotone Decoding

- MT becomes monotone by pre-ordering
- Efficient decoding by WFST
 - phrase segmentation > phrase translation
> word segmentation > LM
- Efficient on-the-fly composition
- ~3x faster than Moses PBMT

Generalized MBR-based System Combination

Generalized MBR-based System Combination

- Duh et al. (IJCNLP 2011)

Generalized MBR-based System Combination

- Duh et al. (IJCNLP 2011)
- Hyp. selection on N -bests on M systems

Generalized MBR-based System Combination

- Duh et al. (IJCNLP 2011)
- Hyp. selection on N -bests on M systems
- Optimization in RIBES+BLEU

Generalized MBR-based System Combination

- Duh et al. (IJCNLP 2011)
- Hyp. selection on N -bests on M systems
- Optimization in RIBES+BLEU
- System-independent “agreement” features
 - Sub-components on RIBES & BLEU

Generalized MBR-based System Combination

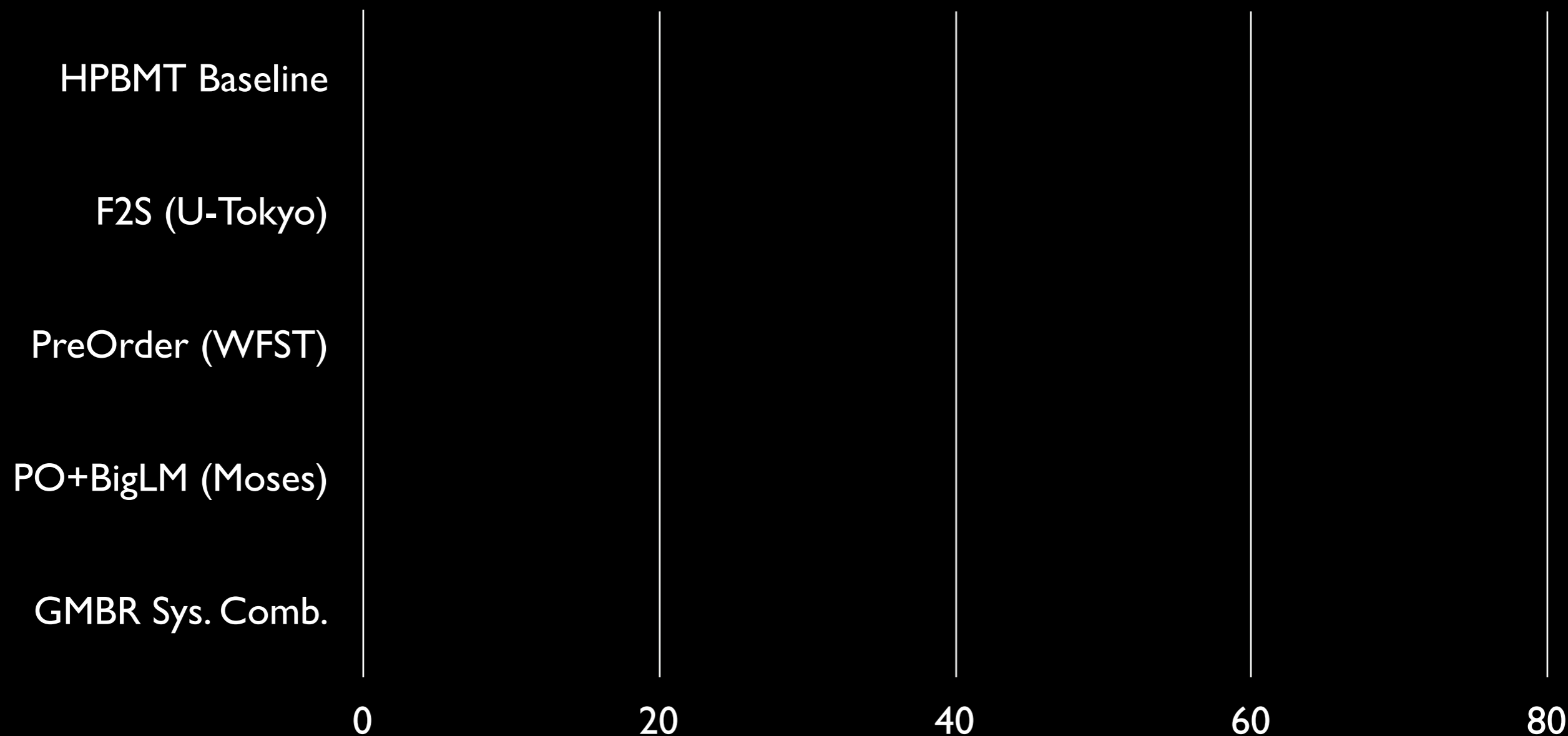
- Duh et al. (IJCNLP 2011)
- Hyp. selection on N -bests on M systems
- Optimization in RIBES+BLEU
- System-independent “agreement” features
 - Sub-components on RIBES & BLEU
- Ranking SVM-like pairwise training

EJ Auto-Eval Results

■ BLEU (%) ■ RIBES (%)

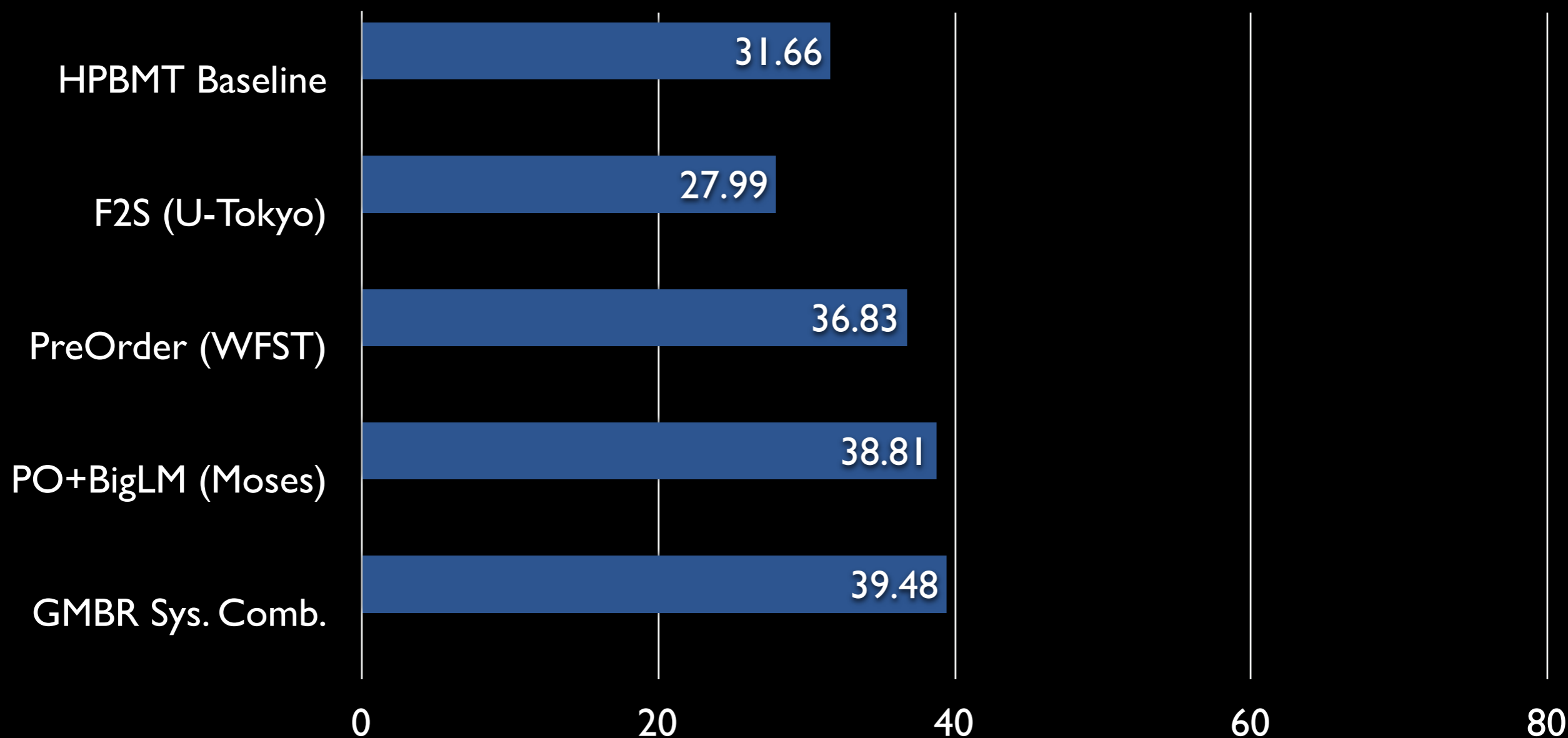
EJ Auto-Eval Results

■ BLEU (%) ■ RIBES (%)



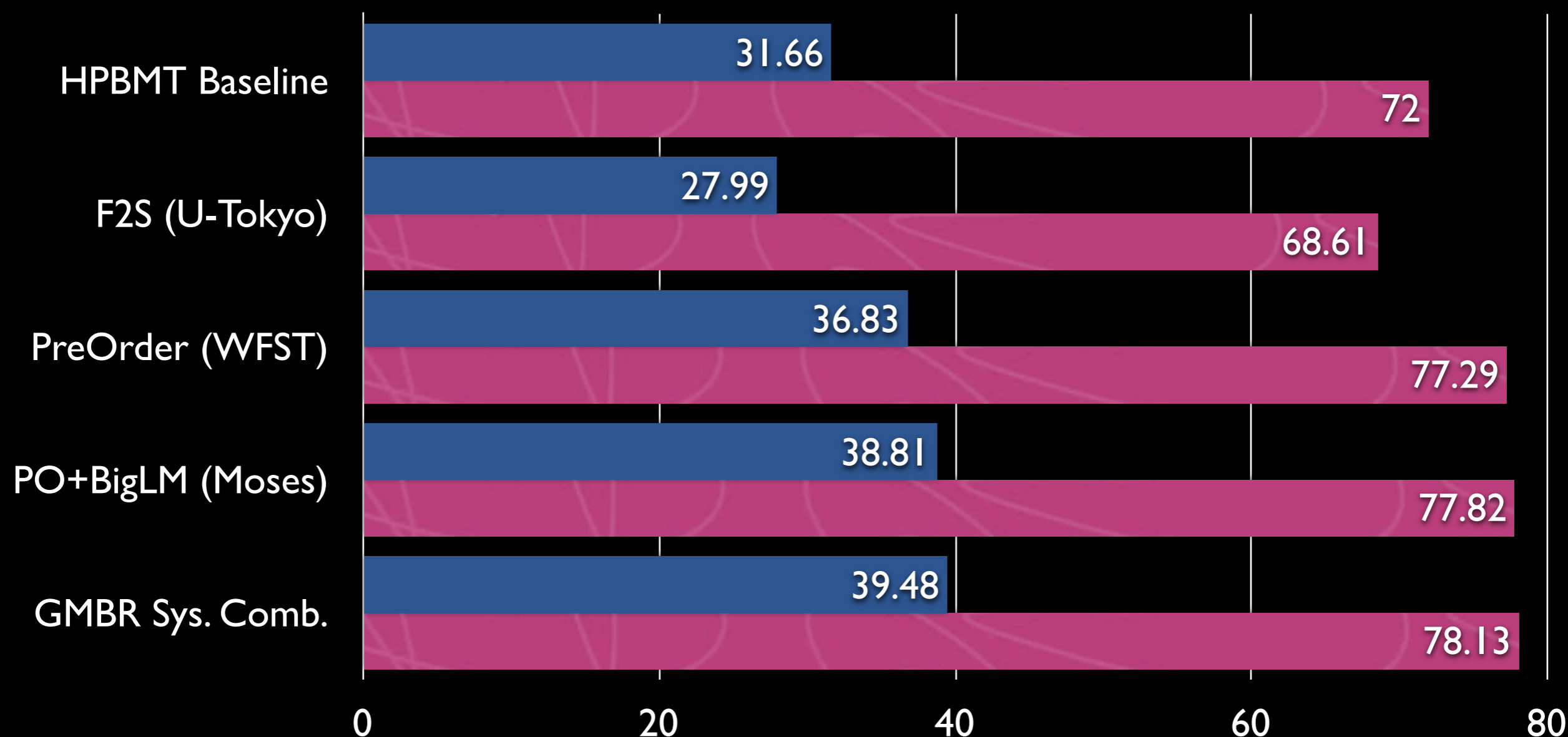
EJ Auto-Eval Results

■ BLEU (%) ■ RIBES (%)



EJ Auto-Eval Results

■ BLEU (%) ■ RIBES (%)

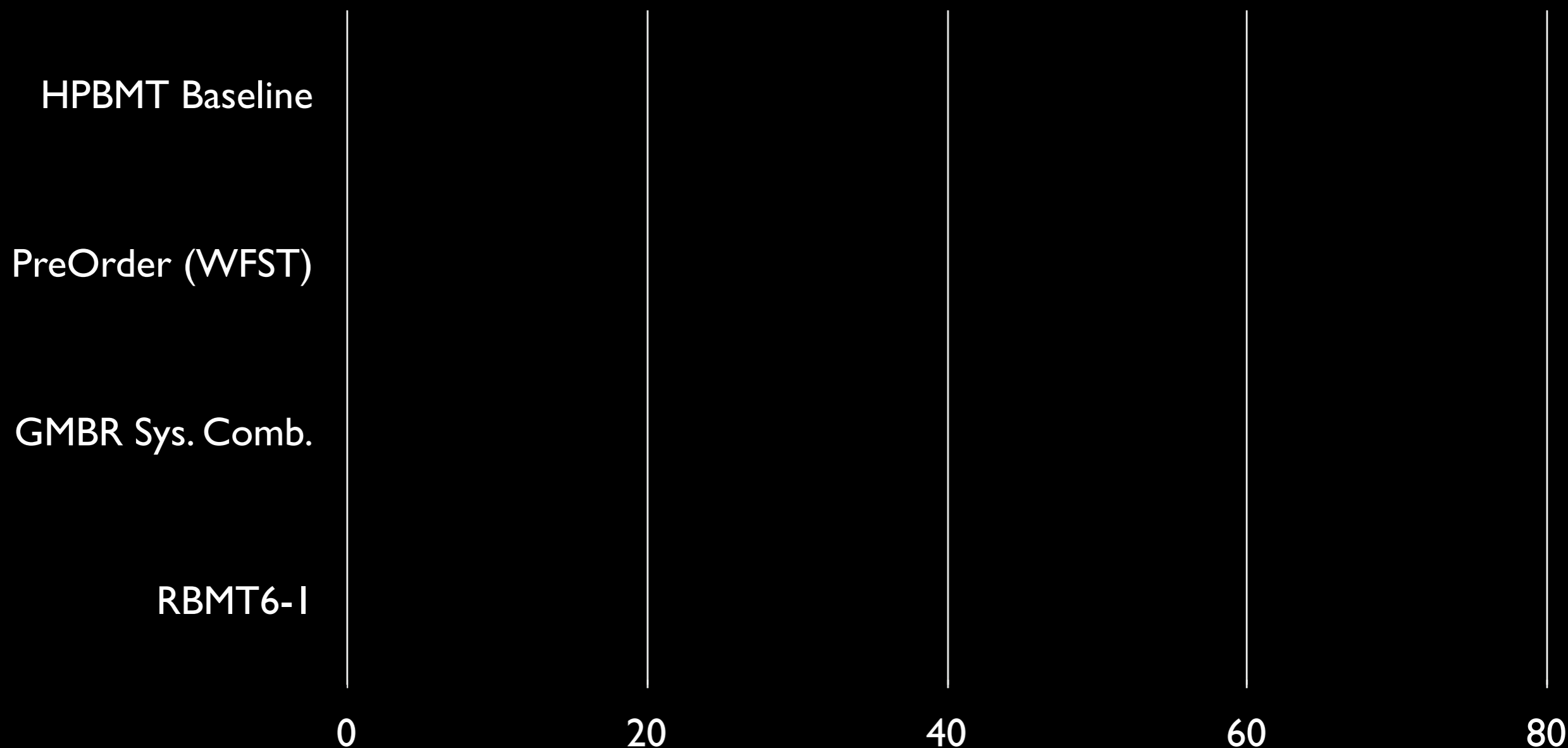


EJ Subj.-Eval Results

■ Adequacy ■ Acceptability (%)

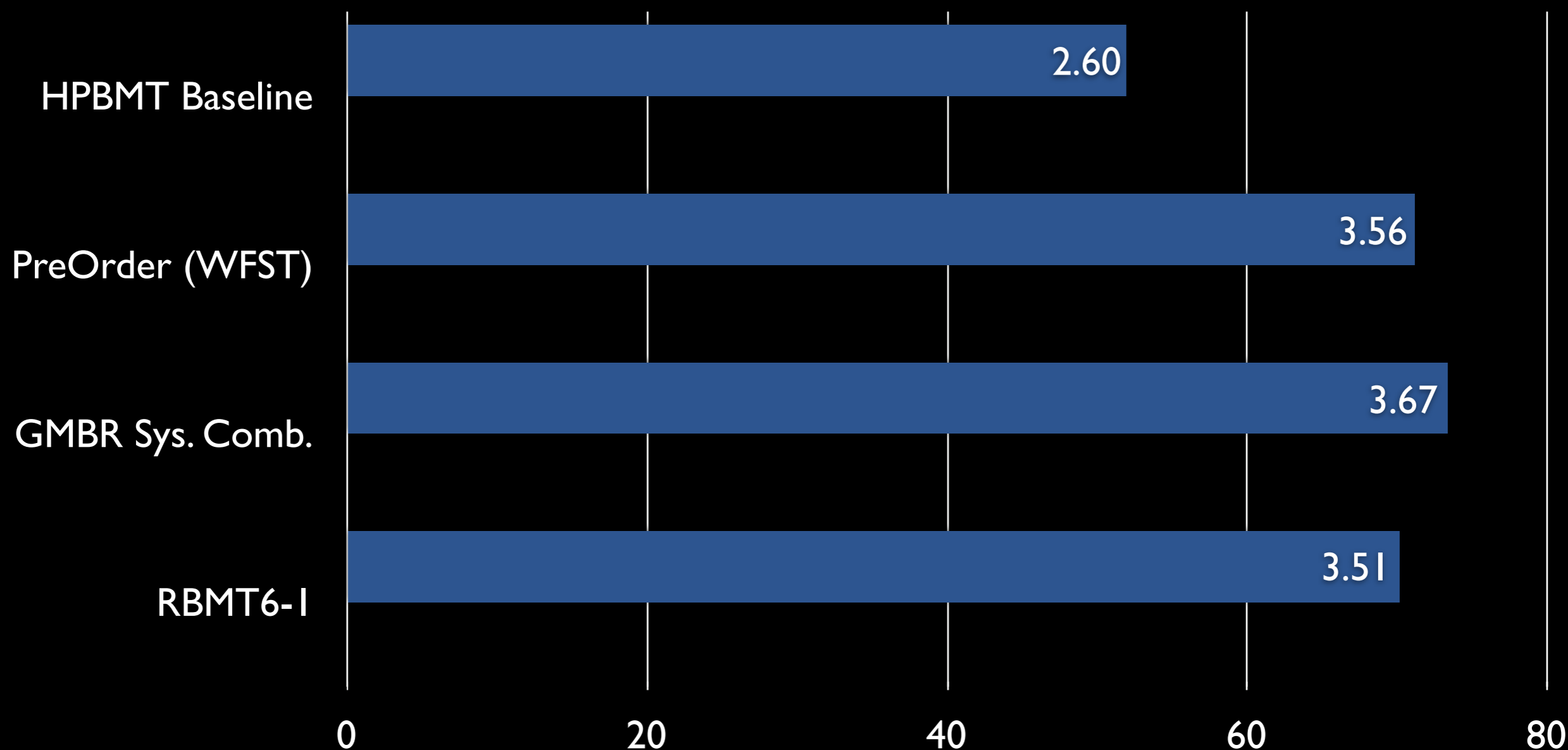
EJ Subj.-Eval Results

■ Adequacy ■ Acceptability (%)



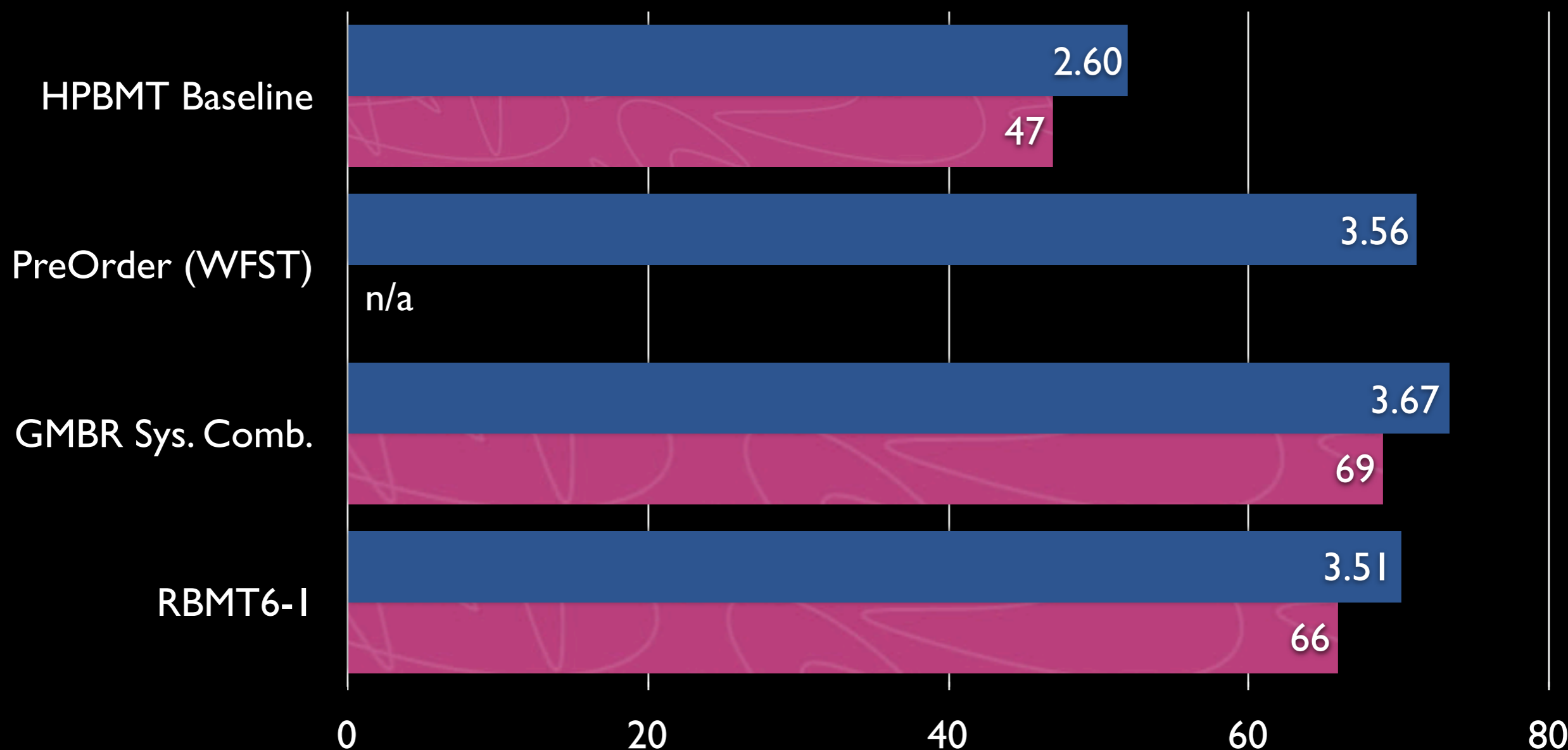
EJ Subj.-Eval Results

■ Adequacy ■ Acceptability (%)



EJ Subj.-Eval Results

■ Adequacy ■ Acceptability (%)



What we found...

What we found...

- Head Finalization worked **QUITE** well !
 - Simple but effective way for EJ translation
 - Monotone translation is relatively easy?

What we found...

- Head Finalization worked **QUITE** well !
 - Simple but effective way for E_J translation
 - Monotone translation is relatively easy?
- Further improved by GMBR Sys. Comb.
 - System variance (diversity) is important?

Conclusion

Conclusion

- State-of-the-art EJ translation
 - even better than RBMT !

Conclusion

- State-of-the-art EJ translation
 - even better than RBMT !
- ... moderate in JE/CE
 - JE pre-ordering, CE adaptation

That's It!

Acknowledgments

- PatentMT organizers, for all of this great task!
- Prof. Hideki Isozaki, for Head Finalization
- Dr. Takaaki Hori, Dr. Shinji Watanabe, for WFST decoding