# Simultaneous Speech Translation

Graham Neubig
Nara Institute of Science and Technology (NAIST)
10/16/2015

Joint Work With:
Satoshi Nakamura, Tomoki Toda, Sakriani Sakti, Tomoki Fujita, Hiroaki Shimizu, Yusuke Oda, Takashi Mieno, Quoc Truong Do

# Background

# Speech Translation

Source: NICT
http://www.nict.go.jp/press/2010/06/29-1.html


Source: Microsoft Research
http://research.microsoft.com/en-us/news/features/translator-052714.aspx

Source: Karlsruhe Institute of Technology
http://isl.anthropomatik.kit.edu/english/1520.php

# Traditional Speech Translation

ASR

Divide at
sentence boundaries

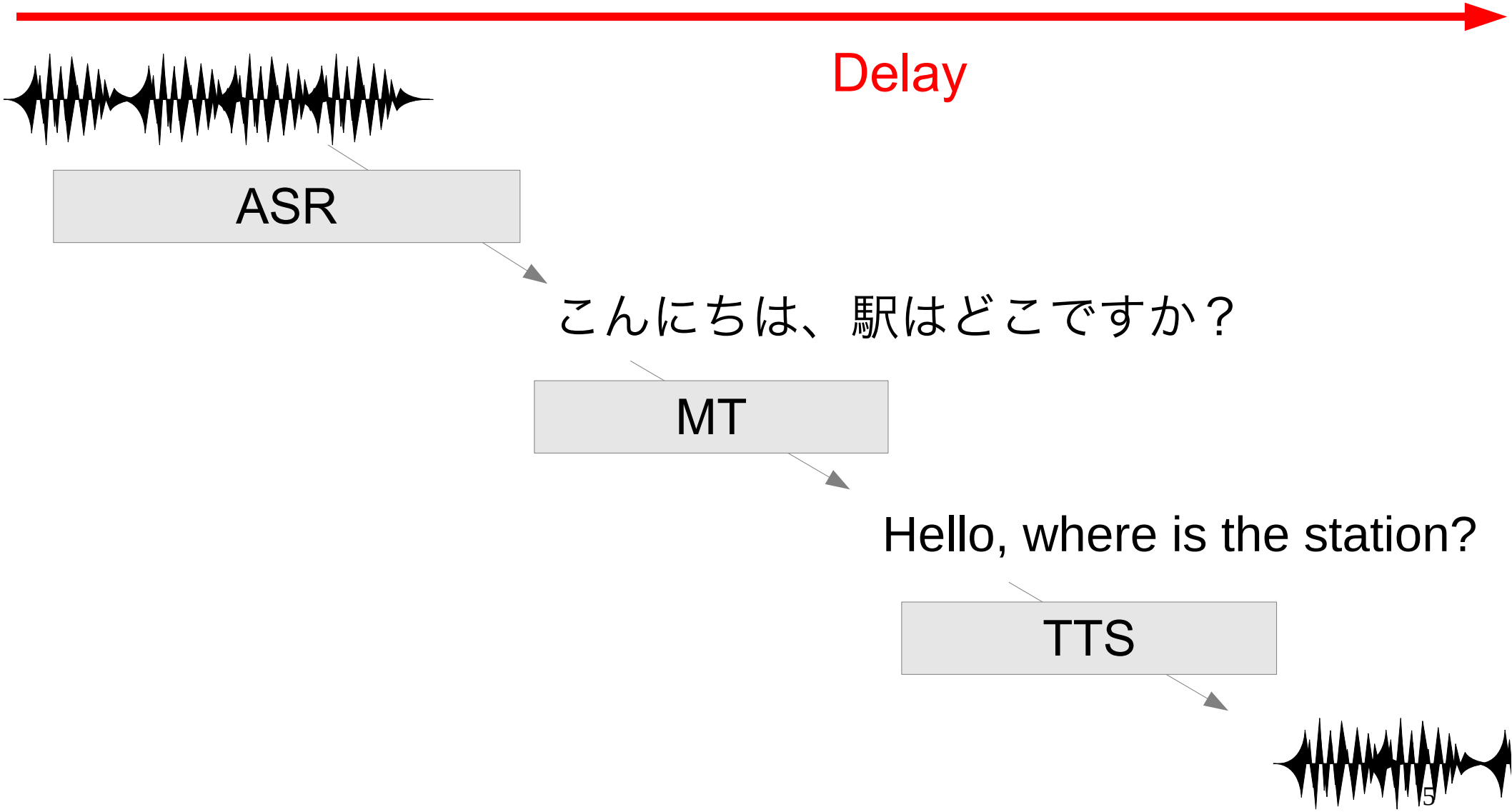
こんにちは、駅はどこですか？

MT

Hello, where is the station?

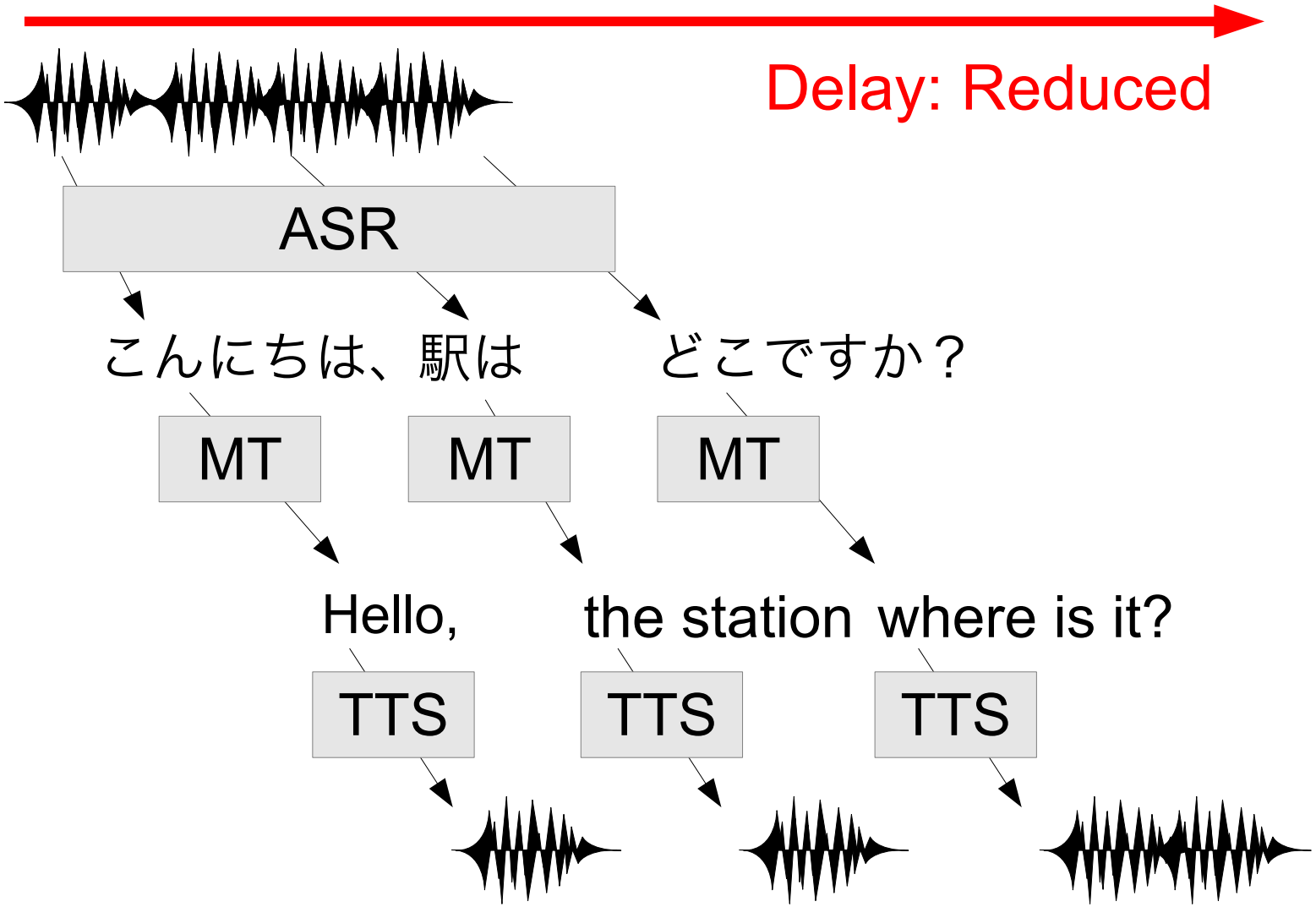
TTS

# Problem: Delay (Ear-Voice Span)



# Speech Translation Example

# Simultaneous Speech Translation




But, this is not easy!

# Professional Simultaneous Interpretation

Photo Credit:
https://www.flickr.com/photos/joi/2027679714
https://www.flickr.com/photos/european_parliament/4268490015

# Simultaneous Interpretation Data [Shimizu+ LREC14]

- Recorded data
  - About 10 Hours of TED Talks (English-Japanese, Japanese-English)

- Simultaneous interpreters
  - 3 pros with varying years of experience
  - Ranked S, A, and B

| Experience | Rank |
|---|---|
| 15 years | S rank |
| 4 years | A rank |
| 1 year | B rank |

Freely available for research purposes:
http://ahclab.naist.jp/resource/stc/

# Simultaneous Interpreter Example

# So How do Simultaneous Interpreters Do It?

## Source:

今ご覧いただいたこの映像は今から五年前、日本で世間を
賑わせていた裁判員制度が始まる一年前、大学四年生だった
私が模擬裁判用の資料として作った物です

## Translation:

Five years ago, as a college senior, I created the video that you just saw as a reference material for a mock trial, one year before the much-talked-about jury system commenced in Japan.

## Interpretation:

Predict NP

You just saw this video clip. Five years ago, at that time in Japan, the ordinary people's justice system, jury system, was very much talked about in Japan, and I created this video as a reference material for that.

Segmentation Prediction Rewording Summarization

# Can We Do the Same in Speech Translation Systems?

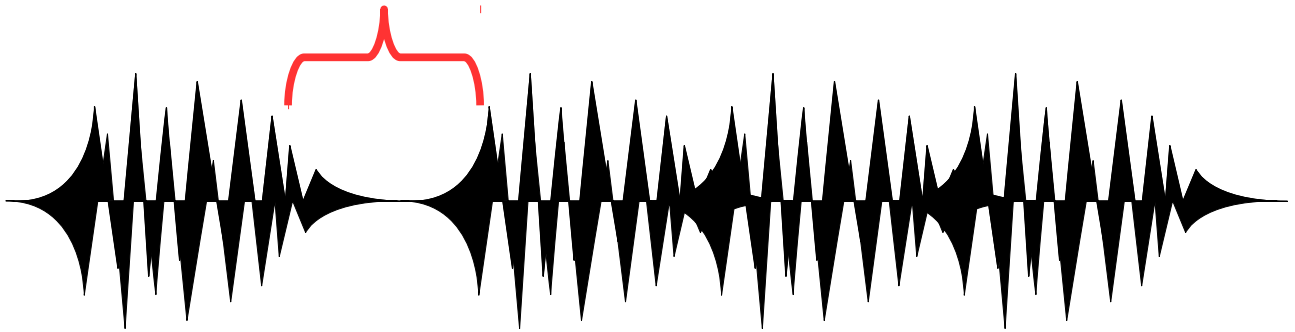
Four problems in this talk:

- Segmentation: When do we start translating?
- Prediction: Can we predict things that haven't been said?
- Rewording: Can we reword sentences to be conducive to simultaneous translation?
- Evaluation: How do we decide which results are better?

# Segmentation

# Heuristic Segmentation Strategies

Division on pauses [Fugen+ 07, Bangalore+ 12]



hello where is the station

comma no comma

Division on predicted commas [Sridhar+ 13]

Division based on reordering probabilities [Fujita+ 13]

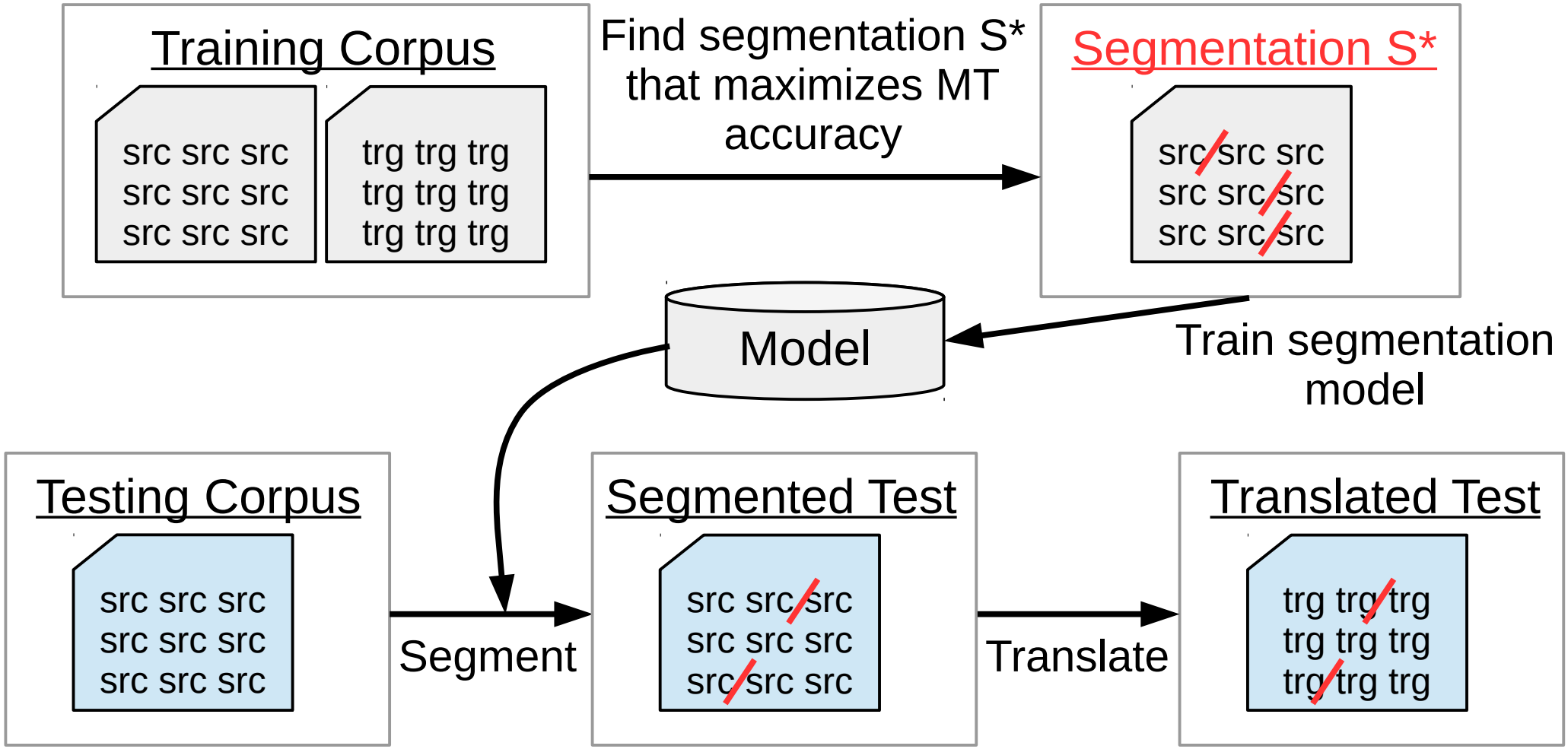
hello → probability of reordering 0.1
where → probability of reordering 0.8

# Optimizing Segmentation Strategies for Simultaneous Speech Translation [Oda+ ACL14]

- All previous segmentation strategies were based on heuristics
- Don't directly take into account effect on translation accuracy

What if we could directly optimize sentence segmentation for translation accuracy?

# Training/Testing Framework



# S* Search Method 1: Greedy Search

I ate lunch but she left 私は昼食を食べたが彼女は帰った

| | | |
|---|---|---|
| I/ate lunch but she left | 私/昼食を食べたが彼女は帰った | 0.7 |
| I ate/lunch but she left | 私は食べた/ランチ彼女は帰った | 0.4 |
| I ate lunch/but she left | 私は昼食を食べた/しかし彼女は帰った | 0.6 |
| I ate lunch but/she left | 私は昼食を食べたが/彼女は帰った | 1.0 |
| I ate lunch but she/left | 私は食べたが彼女/左 | 0.2 |

I ate lunch but/she left

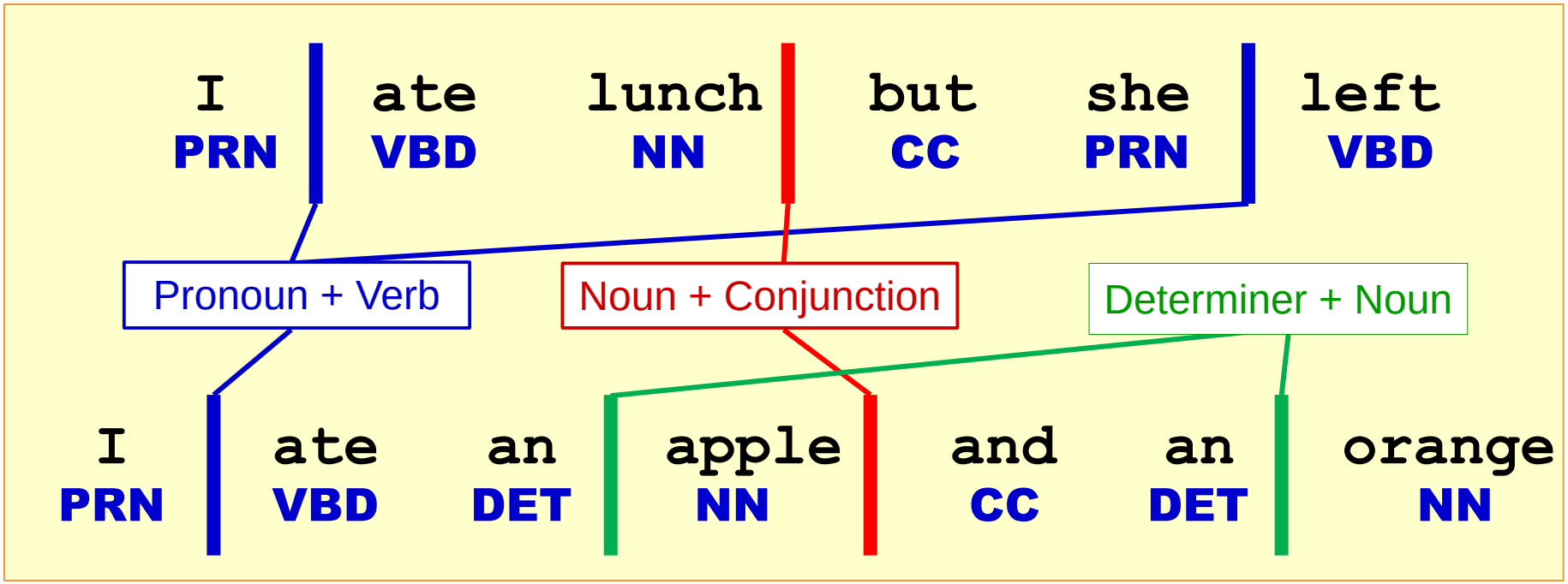
| | | |
|---|---|---|
| I/ate lunch but/she left | 私/昼食を食べたが/彼女は帰った | 0.9 |
| I ate/lunch but/she left | 私は食べた/昼食だが/彼女は帰った | 0.3 |
| I ate lunch/but/she left | 私は昼食を食べた/しかし/彼女は帰った | 0.6 |
| I ate lunch but/she/left | 私は昼食を食べたが/彼女/左 | 0.2 |

I/ate lunch but/she left

Train SVM classifier to recover / at test time

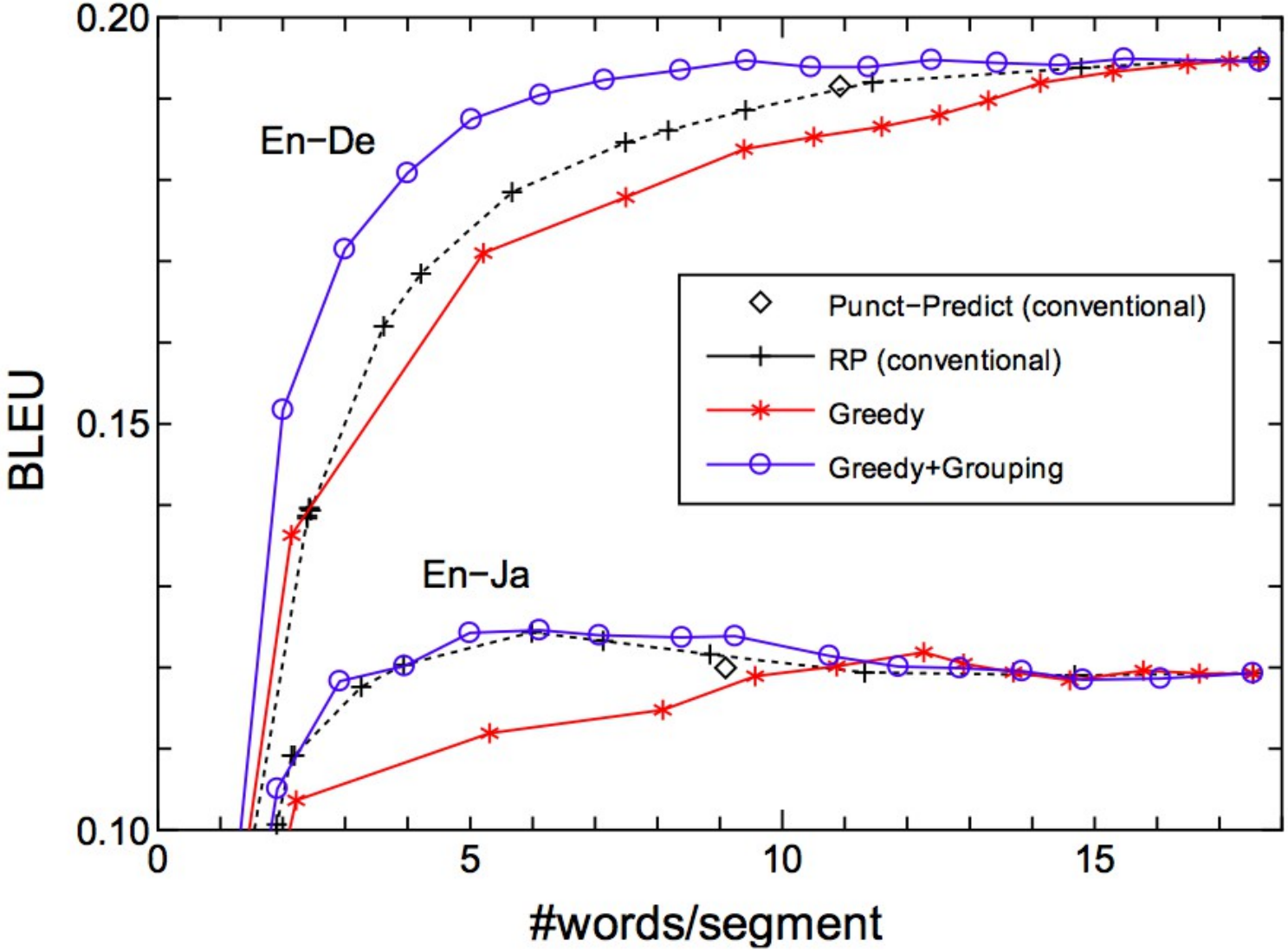
# S* Search Method 2: Grouping by Features

- Because MT/Evaluation is complicated, there is the potential to overfit
- Solution: group boundaries by features




Search can be performed using dynamic programming
Features for the model trivial, no learning is needed

# Results on TED Talks




→ 2-3 times faster with no loss in BLEU

# Simultaneous Translation Demo

- Greedy+Grouping at 10 words

# Future Contributions to Segmentation?

- **Speech:**
  Optimized models using acoustic features?
- **Parsing:**
  Incorporation with incremental parsing? e.g. [Ryu+ 06]
- **Machine Learning:**
  Smarter models: neural networks?
- **Algorithms:**
  Integration with incremental decoding? e.g. [Sankaran+ 10]

# Prediction

# What Kind of Prediction do Simultaneous Interpreters Do? [Wilss 78, Chernov+ 04]

- Lexical prediction

サイエンスを正しく楽しく、これを合い言葉にサイエンス CG
*science factual fun this keyword as science CG*

**then what I wanted to do is to**

クリエーターとして活動しています。
*creator as working*

**promote** fun and factual science, that's my keyword. I'm a …

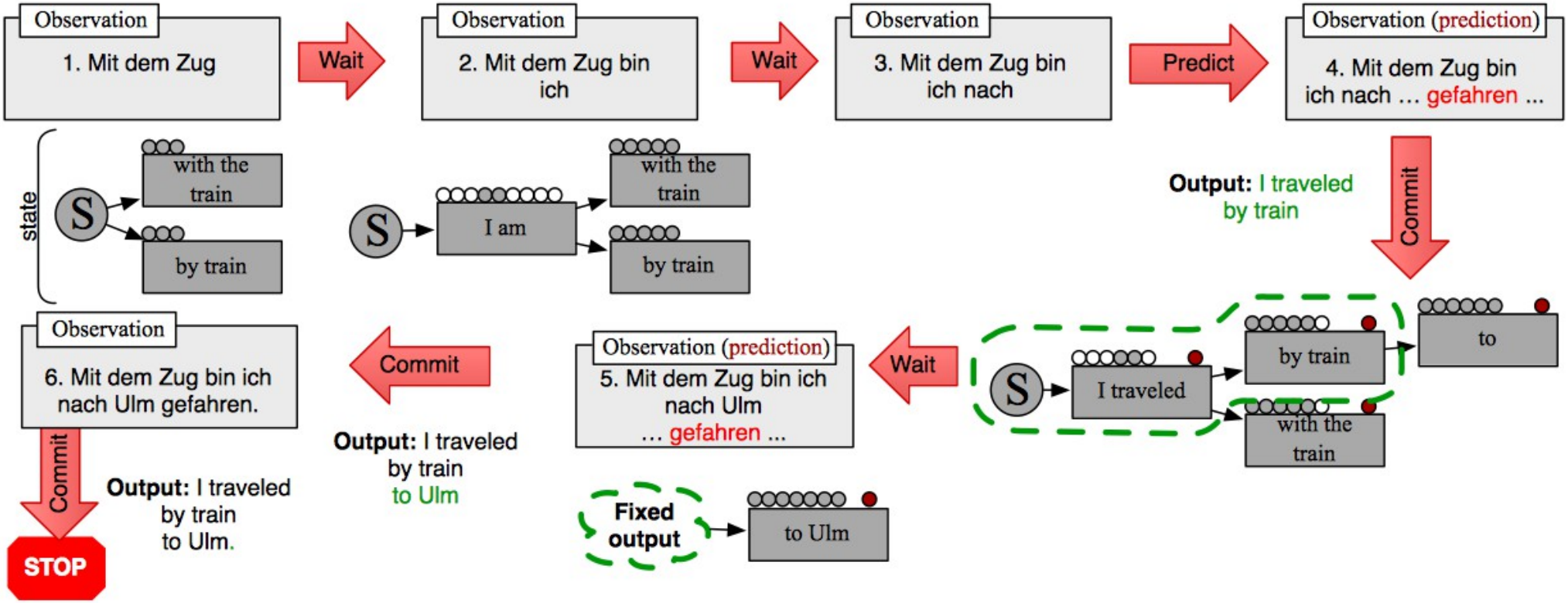
- Structural prediction

今 ご覧頂いた 映像
*now you saw video*

**you just saw** a video clip

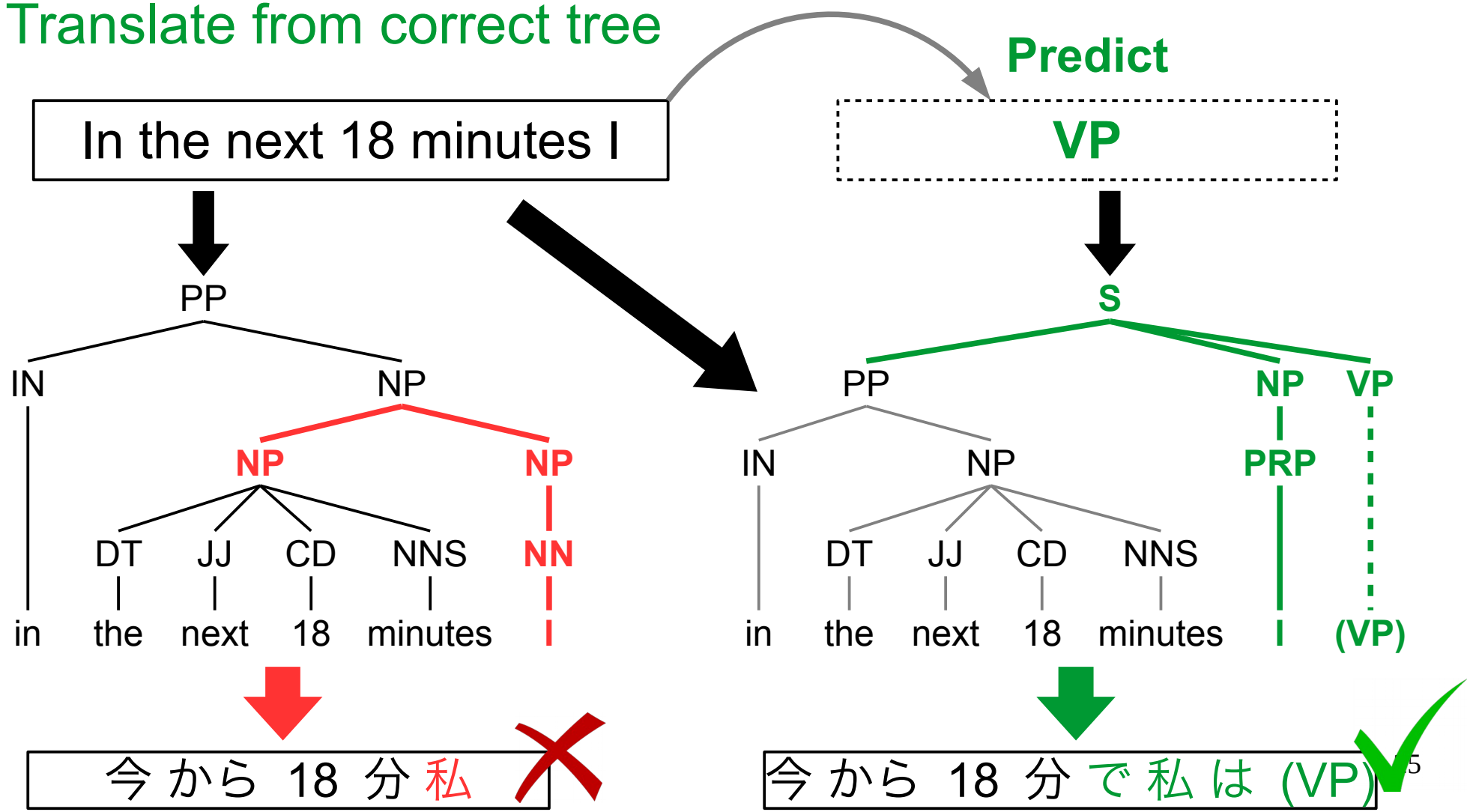
# Predicting Sentence-final Verbs [Grissom et al., EMNLP14]

- Method for translating from verb-final languages (e.g. German)
- Train a classifier to predict the sentence-final verb
- Use reinforcement learning to decide to “wait” “predict” or “commit”



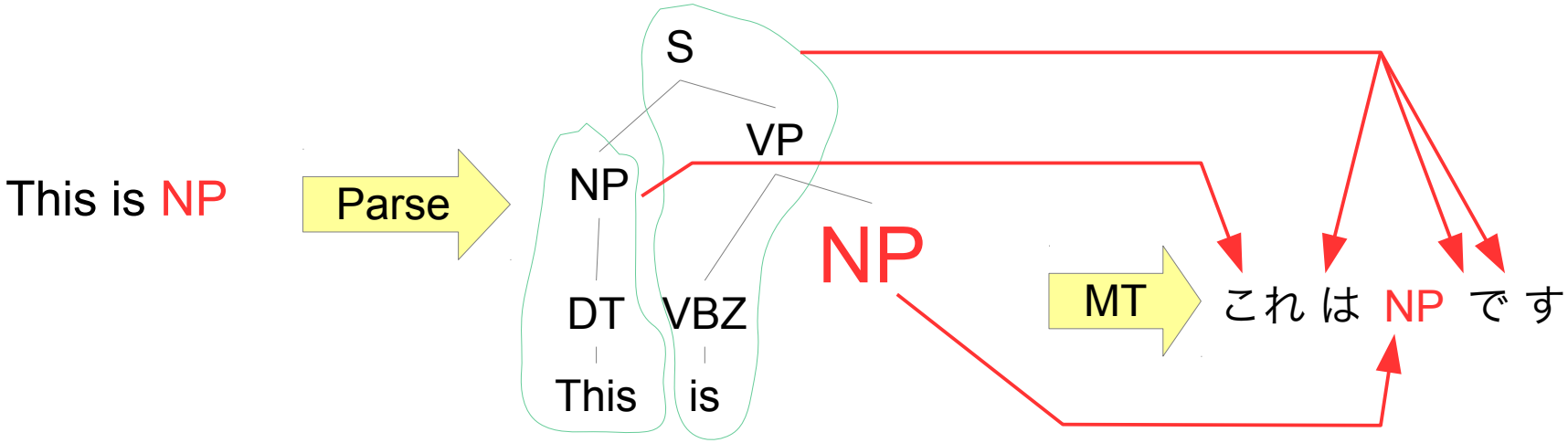
# Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents [Oda+ ACL15]

- Predict unseen syntax constituents
- Translate from correct tree



# Why is Syntax Necessary?

- Tree-to-string (T2S) MT framework
  - Obtains state-of-the-art results on syntactically distant language pairs (c.f. phrase-based translation; PBMT)
  - Possible to use additional syntactic constituents explicitly

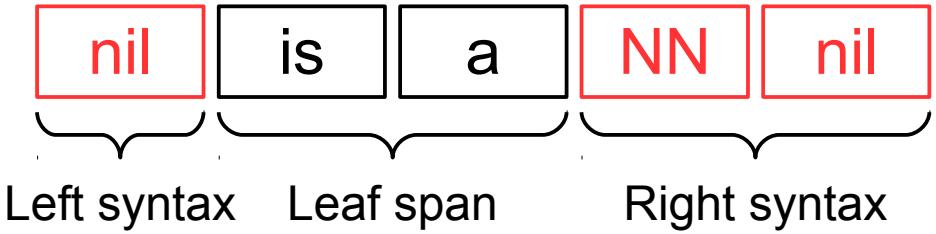


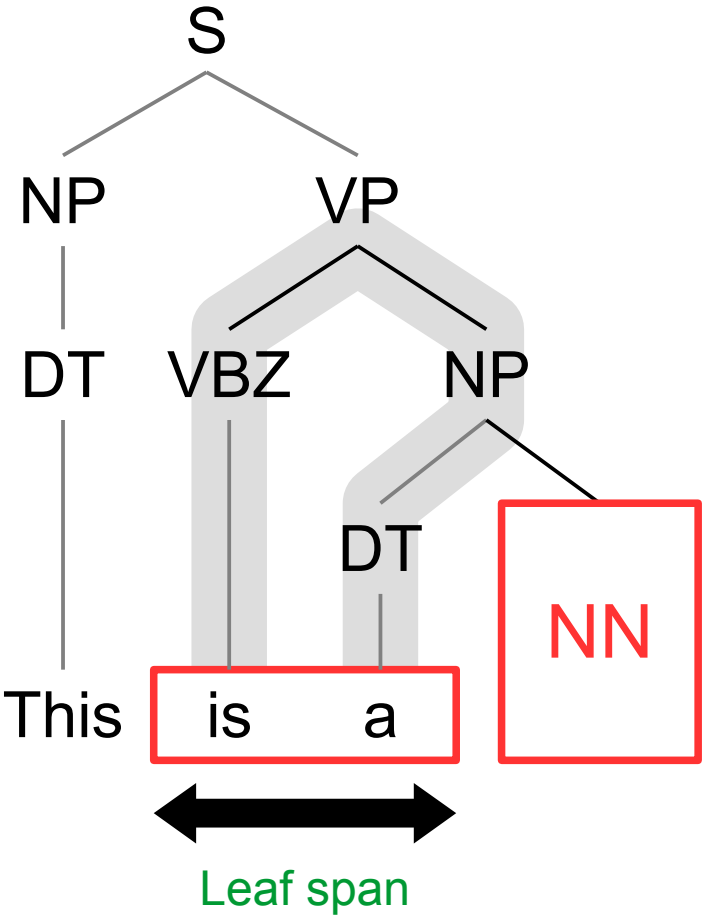

  - Additional heuristic to wait for more input based on when translation requires reordering

# Making Training Data for Syntax Prediction

- Decompose gold trees in the treebank

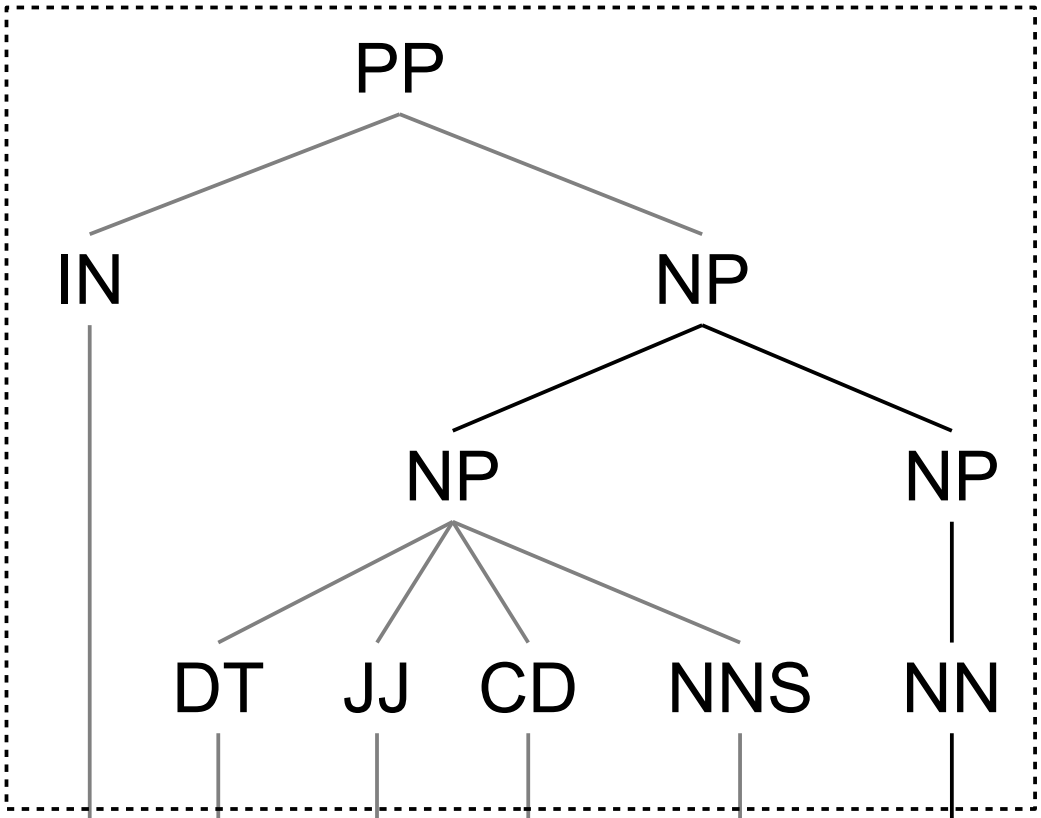
1. ***Select*** any leaf span in the tree
2. ***Find*** the path between leftmost/rightmost leaves
3. ***Delete*** the outside subtree
4. ***Replace*** inside subtrees with topmost phrase label
5. Finally we obtain:



# Syntax Prediction Process

1. ***Parse*** the input as-is




2. ***Extract*** features

| | |
|---|---|
| Word:R1=I | ROOT=PP |
| POS:R1=NN | ROOT-L=IN |
| Word:R1-2=I,minutes | ROOT-R=NP |
| POS:R1-2=NN,NNS | ... |
| ... | |

3. ***Predict*** the next tag (linear SVM)

VP ... 0.65
NP ... 0.28
nil ... 0.04
...

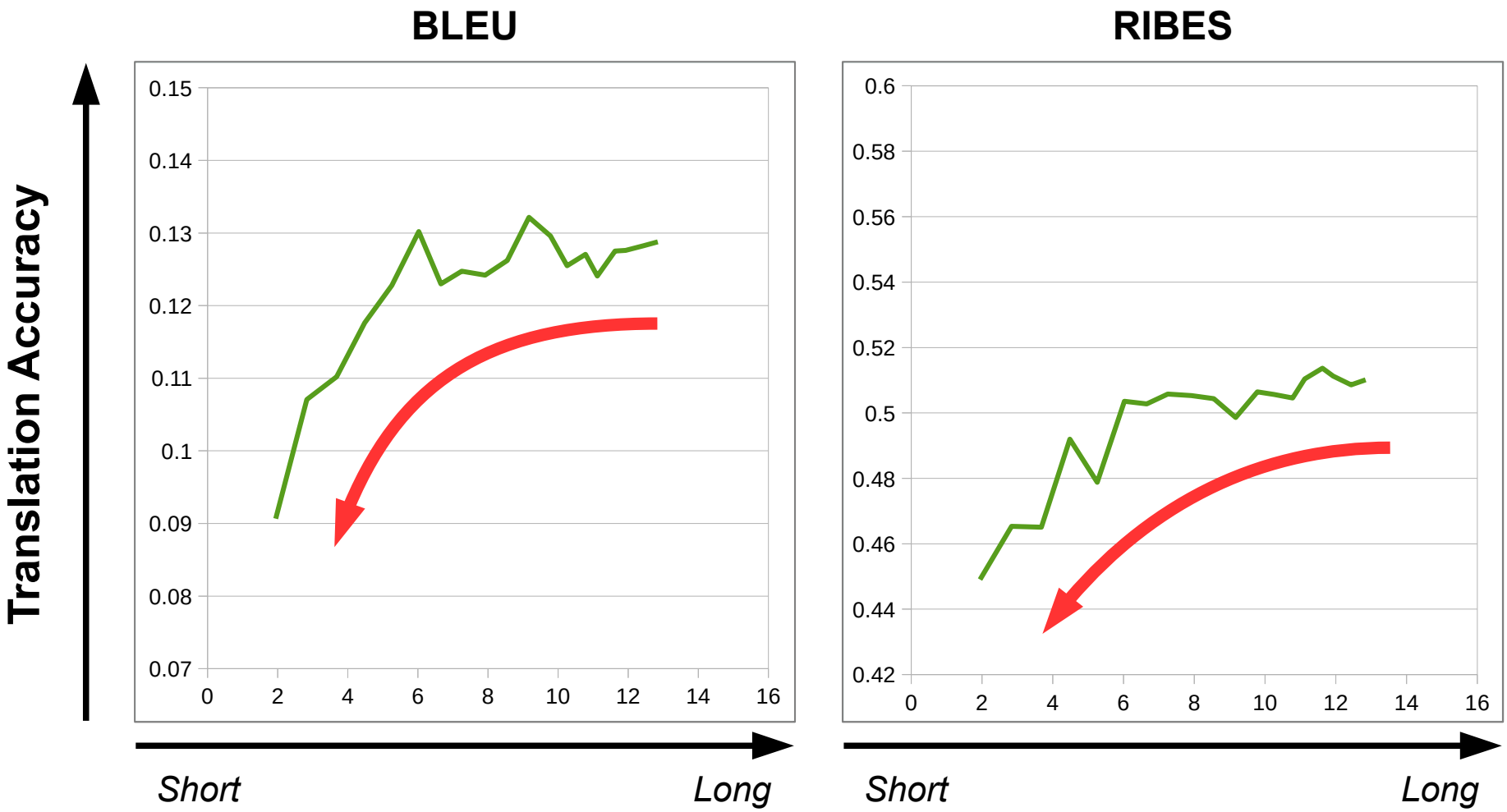
4. ***Append*** to sequence

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| in | the | next | 18 | minutes | I | **VP** | | | **nil** |

Input translation unit

5. ***Repeat*** until nil

# Results: Translation Trade-off (1)



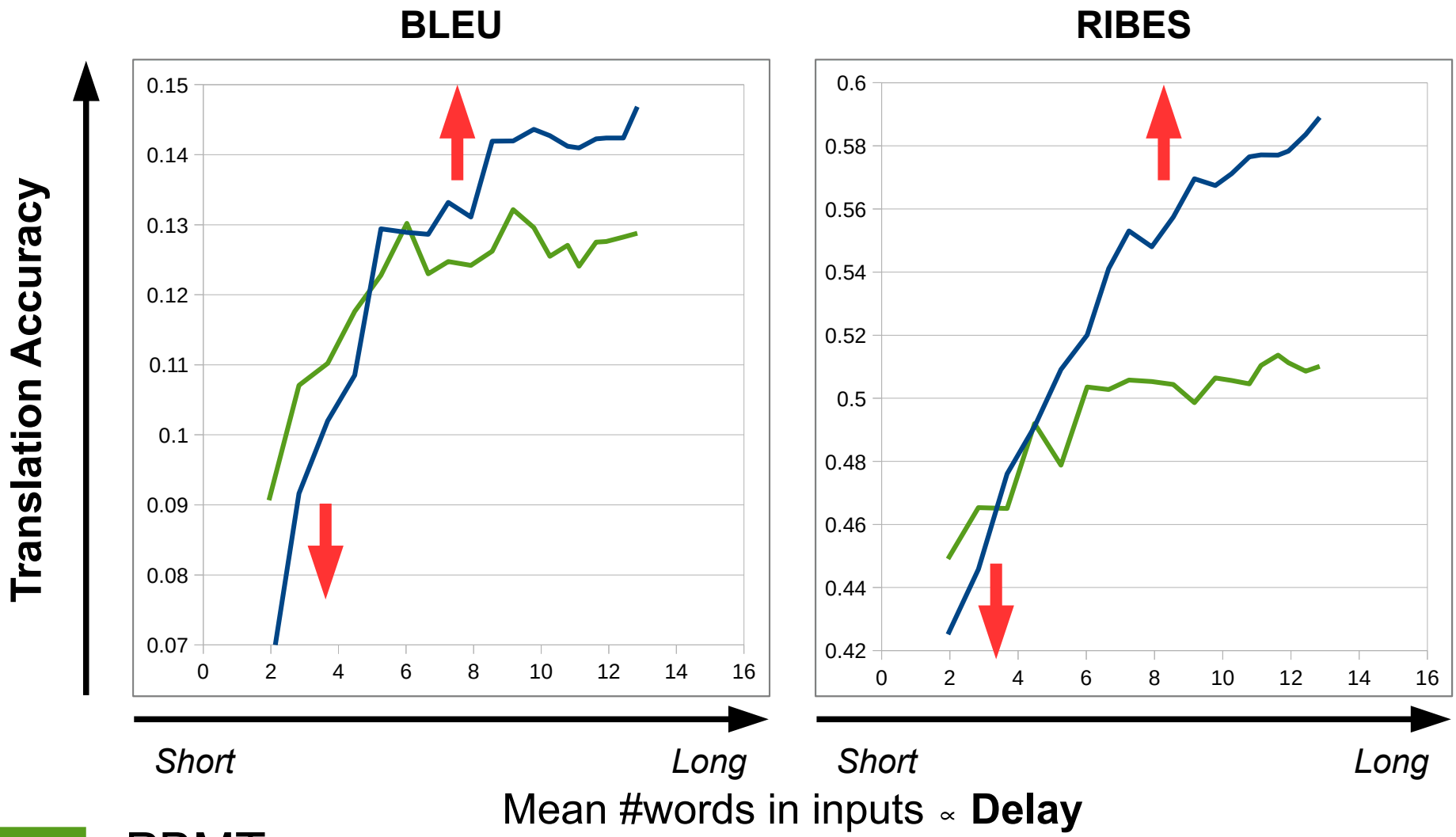

Mean #words in inputs ∝ **Delay**

Using N-words segmentation (not-optimized)

PBMT

- Short inputs reduce translation accuracies

# Results: Translation Trade-off (2)

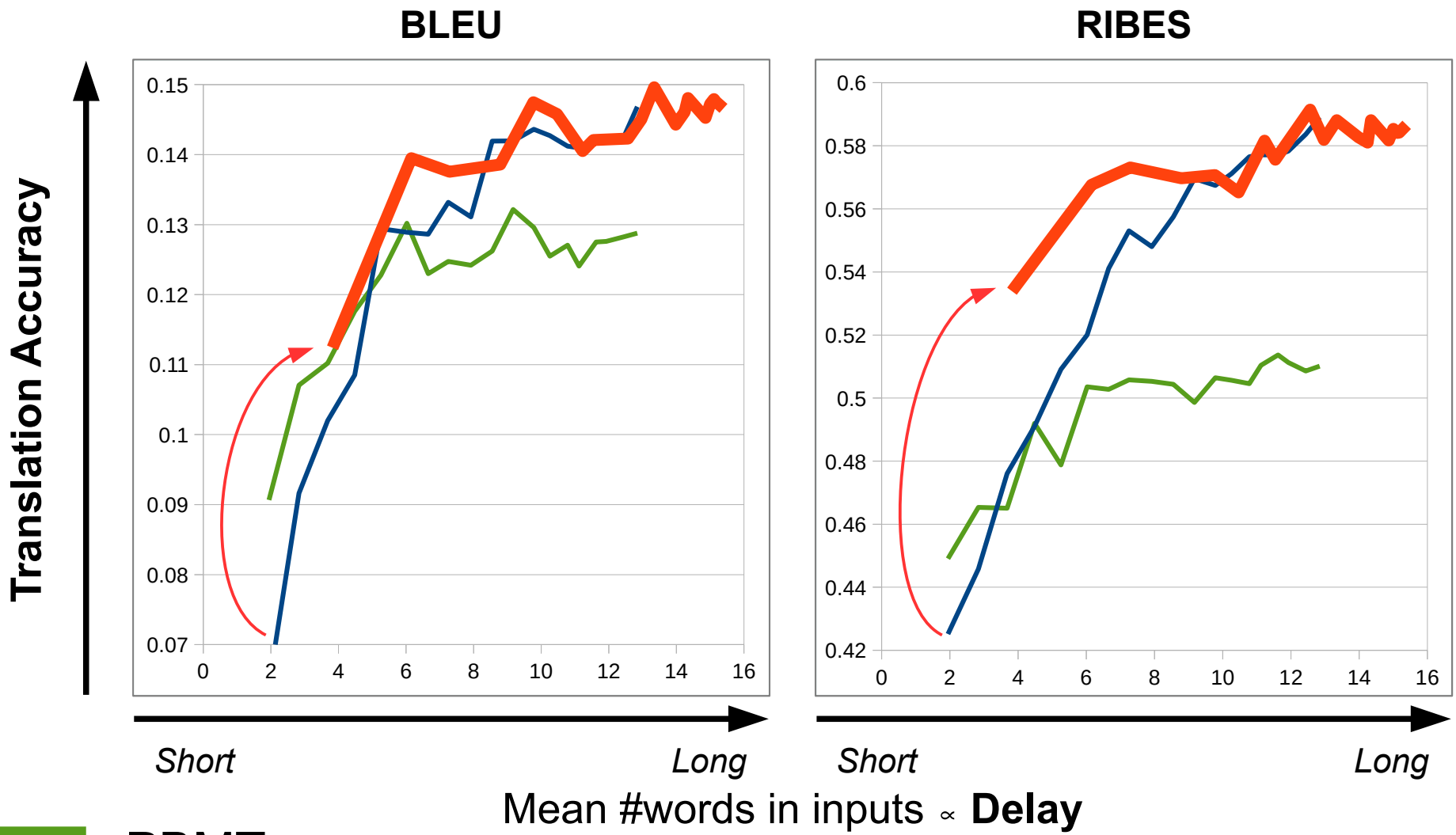



PBMT

T2S

- Long phrase ... **T2S** > PBMT
- Short phrase ... T2S < **PBMT**

# Results: Translation Trade-off (3)




PBMT

T2S

**Proposed**

- Prevent accuracy decreasing in short phrases
- More robustness for reordering

# Future Contributions to Prediction?

- **Language Modeling:**
  More sophisticated models for lexical prediction.
- **Lexical Simplification:**
  Predict a more general word, then replace it later?
- **Machine Learning:**
  End-to-end reinforcement learning of the whole system?
  Application of neural MT models?

# Rewording

# What Kinds of Rewording May Be Helpful?

- Conjunction Clauses [Shimizu+ 13]

X because Y

Y *dakara* X

X *nazenaraba* Y

- Passivization [He+ 15]

私は 昨日 安い 本を 買った

*I yesterday a cheap book bought*

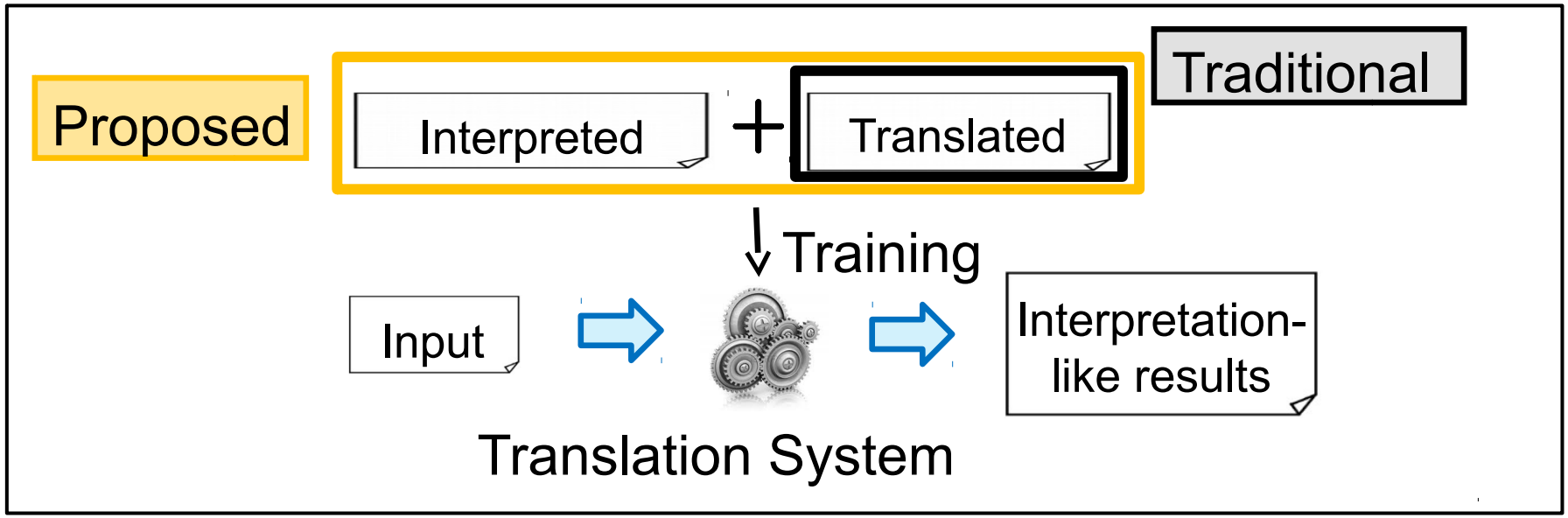
I bought a cheap book yesterday

yesterday a cheap book was bought by me

- etc.

# Constructing a Speech Translation System using Simultaneous Interpretation Data [Shimizu+ IWSLT13]

- Approach:
  - Incorporate simultaneous interpretation data in training the MT system




- [Paulik+ 08] use interpretation data, but to improve accuracy

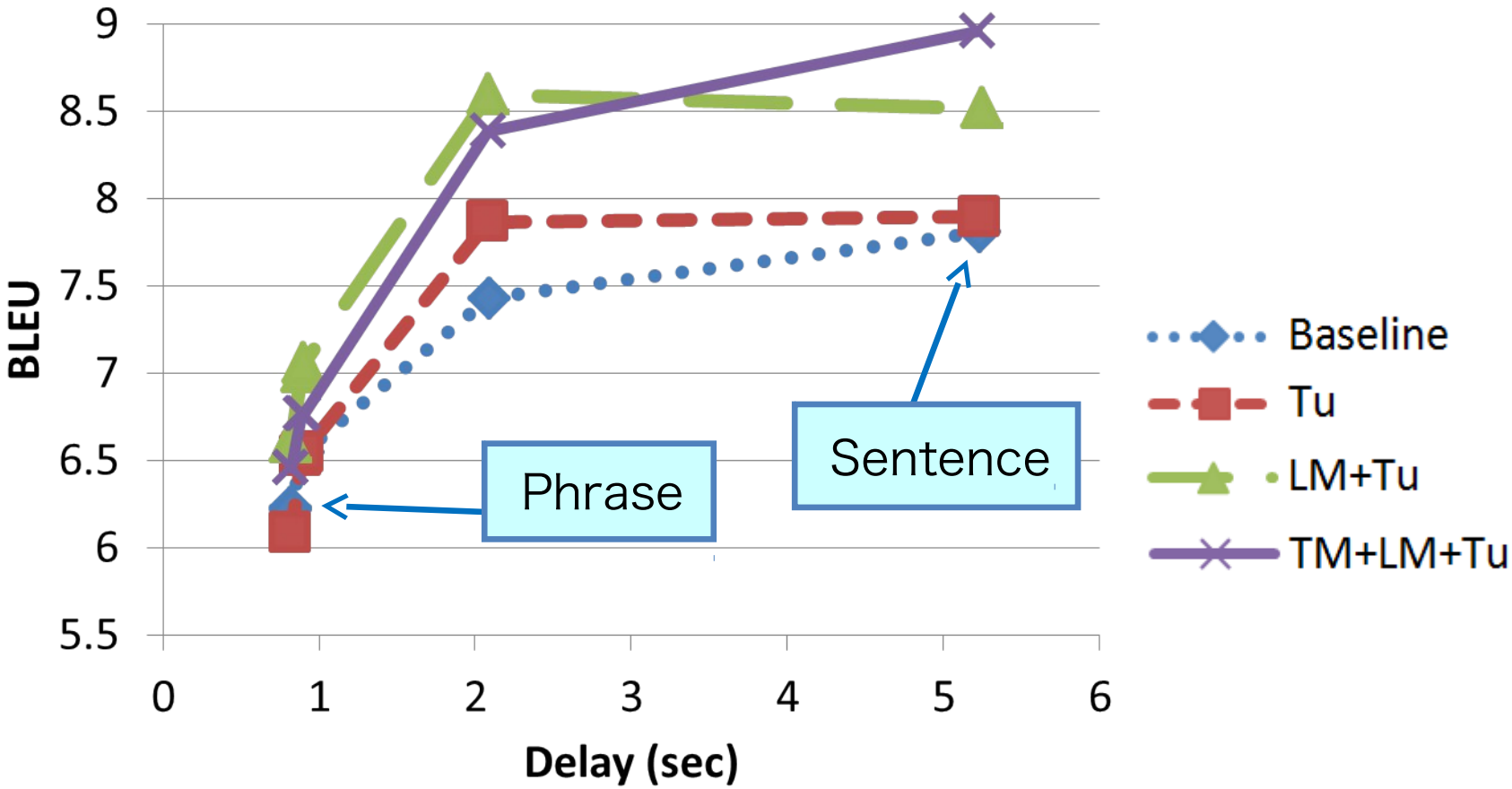
# Incorporating Interpretation Data

Interpretation data is small,
so use adaptation techniques

- Tuning (Tu)
  - Tune the parameters of the translation systems to match the interpretation data
- Language Model (LM): Linear Interpolation
  - Match the style of simultaneous interpreters
- Translation Model (TM): fill-up [Bisazza+ 11]
  - Like the LM, adapt the TM to match interpretation data

# Experimental Evaluation

Accuracy measured against simultaneous interpretation reference



# Examples of Learned Traits

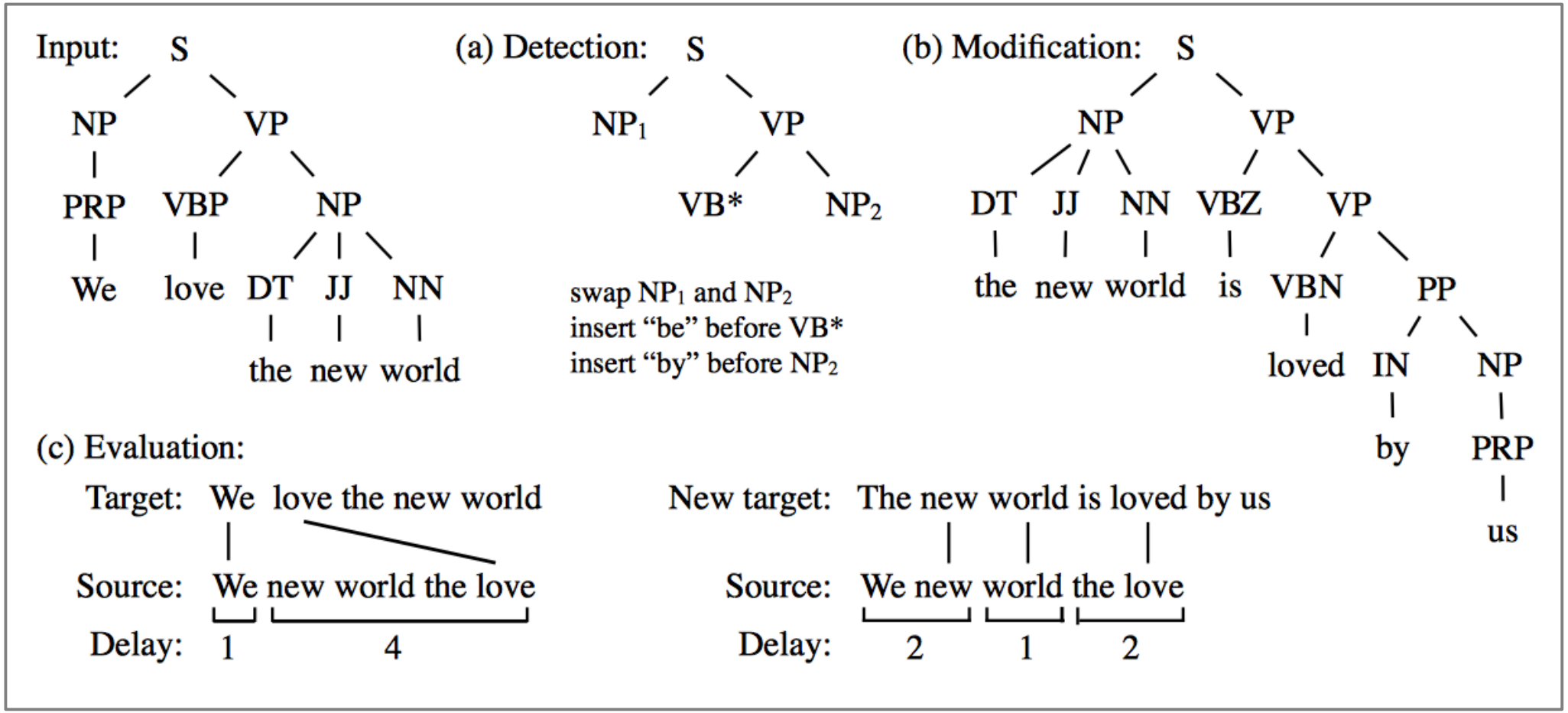
| | Sentence |
|---|---|
| Source | if you look at in the context of history you can see what this is doing |
| S Rank Reference | 過去から / 流れを見てみますと / 災害は / このように / 増えています<br>from the past / look at the context and / disasters are / like this / increasing |
| Baseline (RP 1.0) | 見てみると / 歴史の中で / 見ることができます / これがやっていること<br>looking at / in the history / you can see / what this is doing |
| TM+LM+Tu (RP 1.0) | では / 歴史の中で / 見ることができます / これがやっていること<br>ok / in the history / you can see / what this is doing |

Shortening

Starting sentences with “OK” or “And”
(Also done by interpreter in 25% of sentences)

# Syntax-based Rewriting for Simultaneous Machine Translation [He+ EMNLP15]

- Reword the target language to be closer to source
- Passivizing, changing order of clauses when beneficial



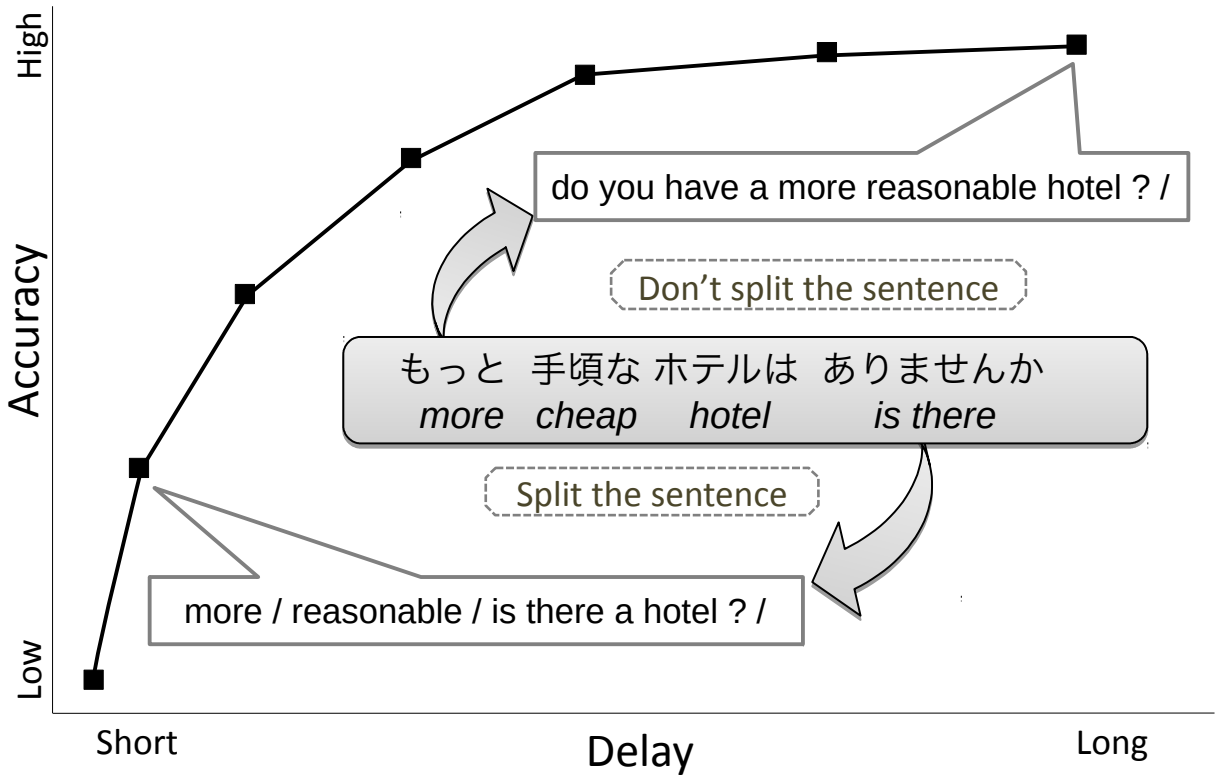
# Future Contributions in Rewording?

- **Paraphrasing:**
  More generalized models of structural paraphrasing?
- **Semantic Similarity:**
  How can we evaluate semantic similarity between sentences structurally different from the reference?

# Evaluation

# Speed vs. Accuracy

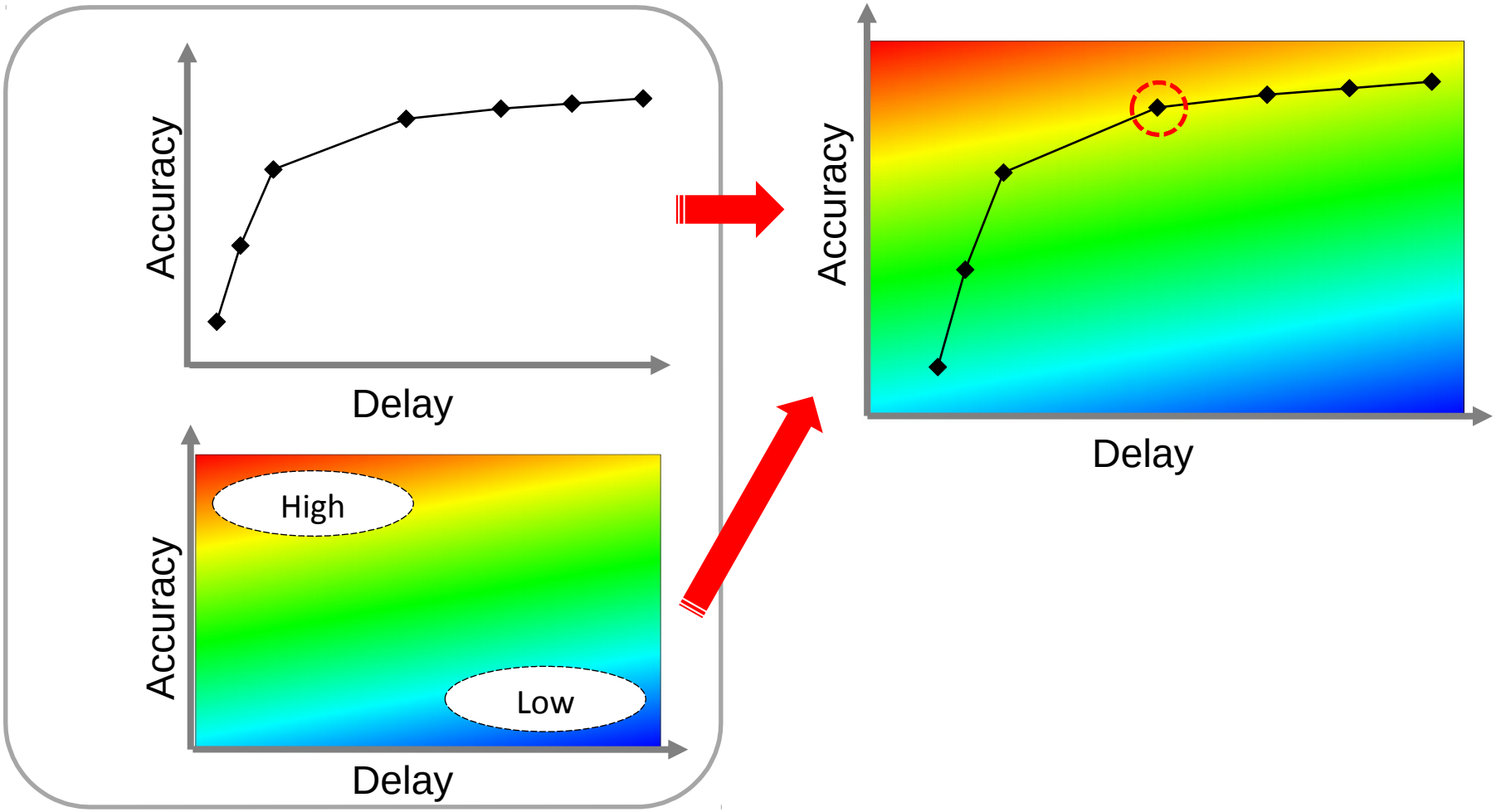
- Tradeoff between speed and accuracy.



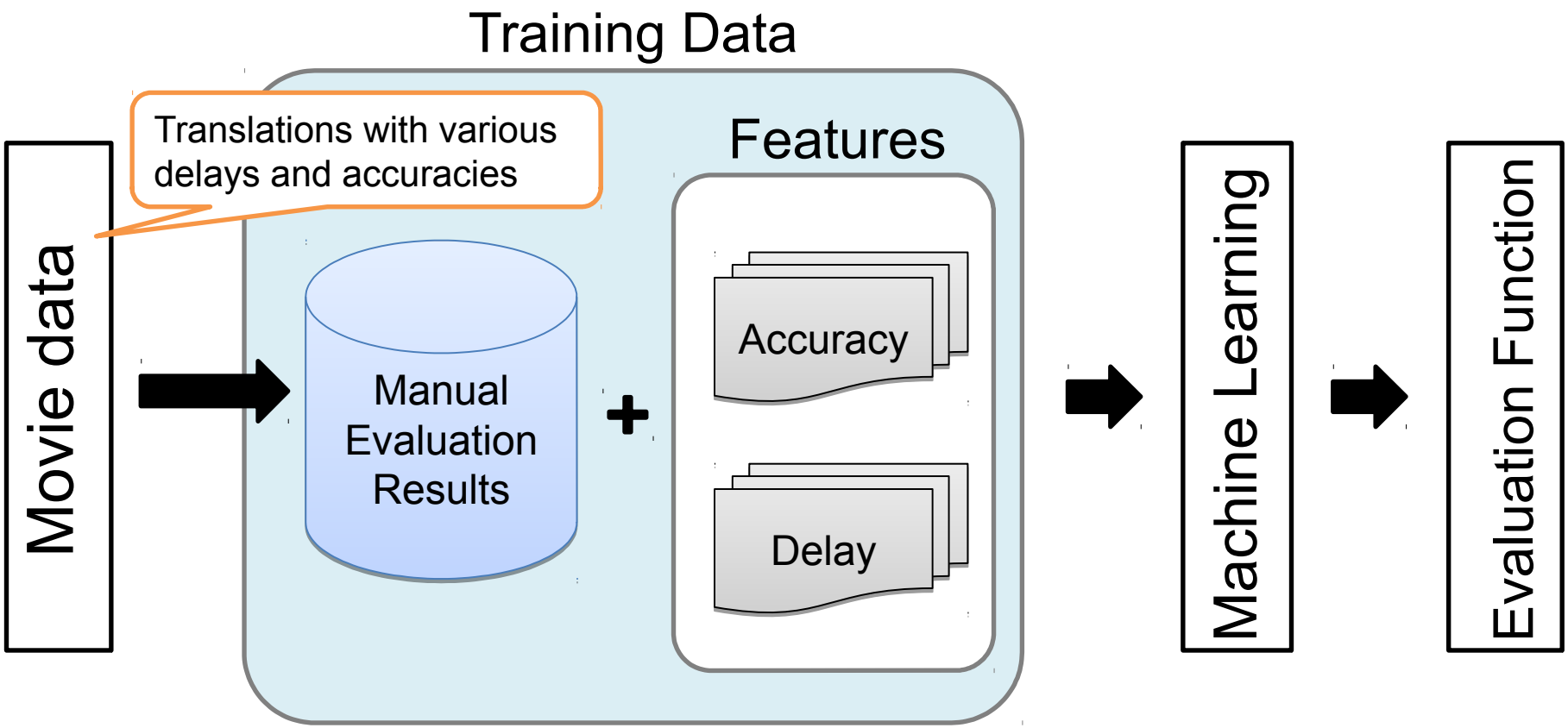

- Given two systems of different speed and accuracy, which is better?

# Speed or Accuracy? A Study in Evaluation of Simultaneous Speech Translation Systems [Mieno+ InterSpeech15]

- Based on speed and accuracy, determine which system is better

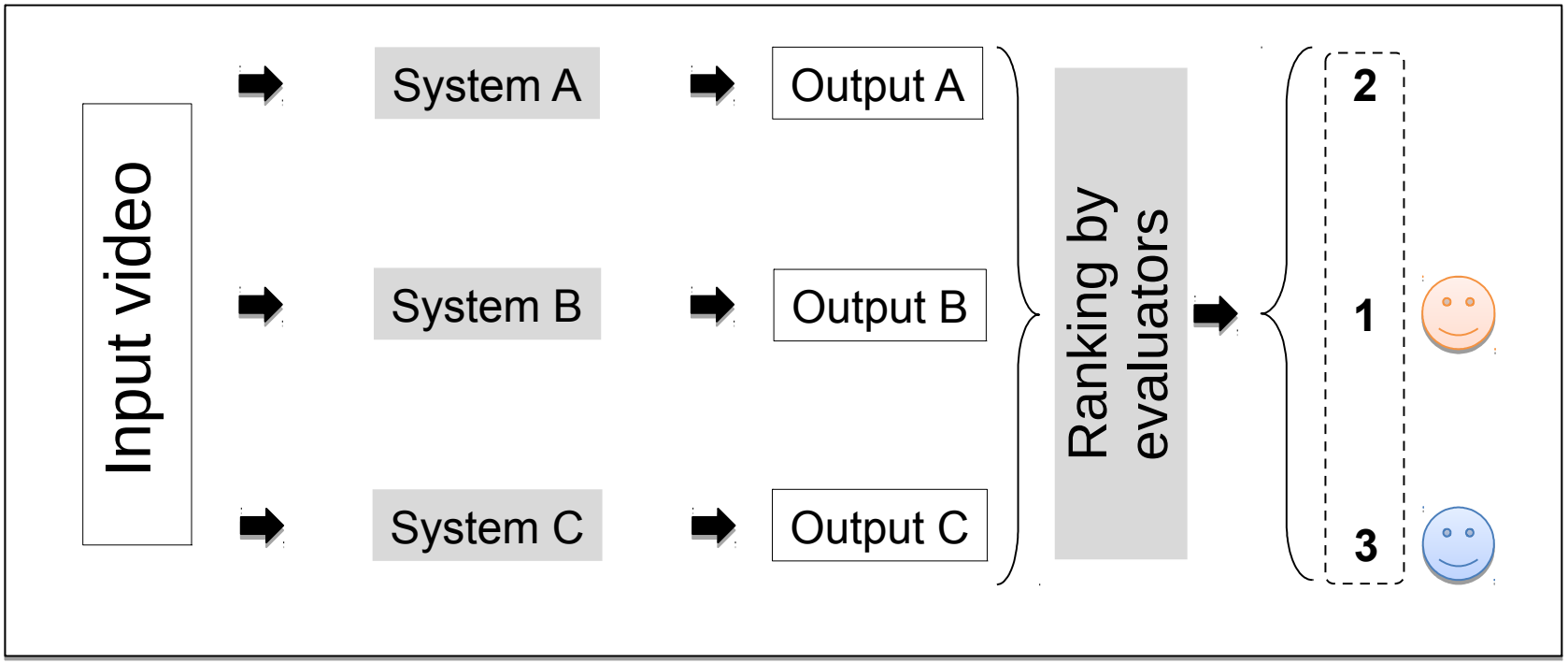


# How to Create an Evaluation Function? (Based on Data)



# Manual Evaluation Format

- Rank-based evaluation
  - Perform comparative evaluation of which output is “better”
  - Allows for consideration of both speed and accuracy



# Evaluation Sheet Example

# Learning an Evaluation Function

Define a linear function that takes a video as input and returns a score

$$s = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

Weight vector

Features useful in evaluation (i.e., delay and accuracy)

Displayed video

This function can be learned from ranked data using “learning to rank”

# Experimental Setup

- Target video

TED Talks
① Realtime trans. is important
② Often used in MT evaluation

- Gathered data

| | | |
|---|---|---|
| Video | 20 Types | 20-30 Seconds |
| Delay | 7 Types | 0,1,2,3,5,7,10 Seconds |
| Accuracy | 3 Types | Auto: BLEU/RIBES<br>Man: Adequacy |
| Subjects | 15 | Japanese speakers |
| Modalities | Subtitled | Dubbed |

- Translation data (5 varieties)

English → Japanese

**Translator**

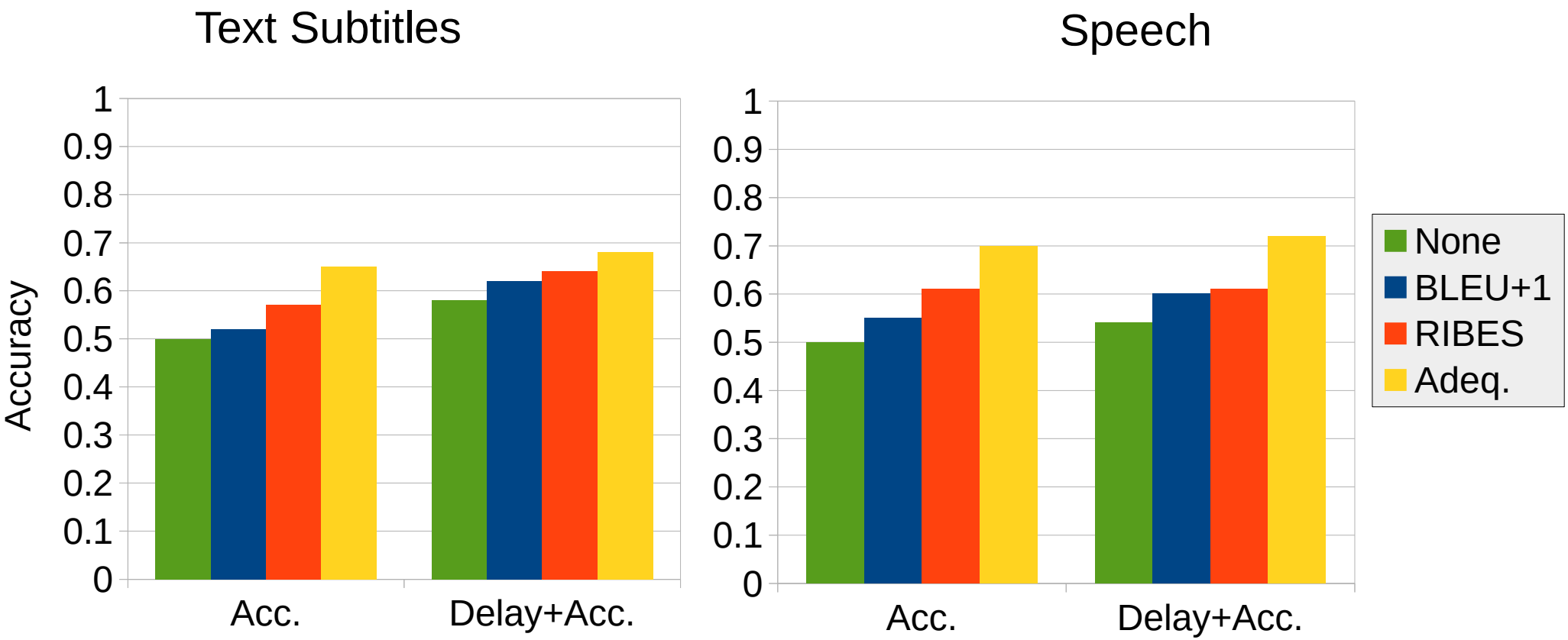
**Interpreter 1 (S Rank)**

**Interpreter 2 (A Rank)**

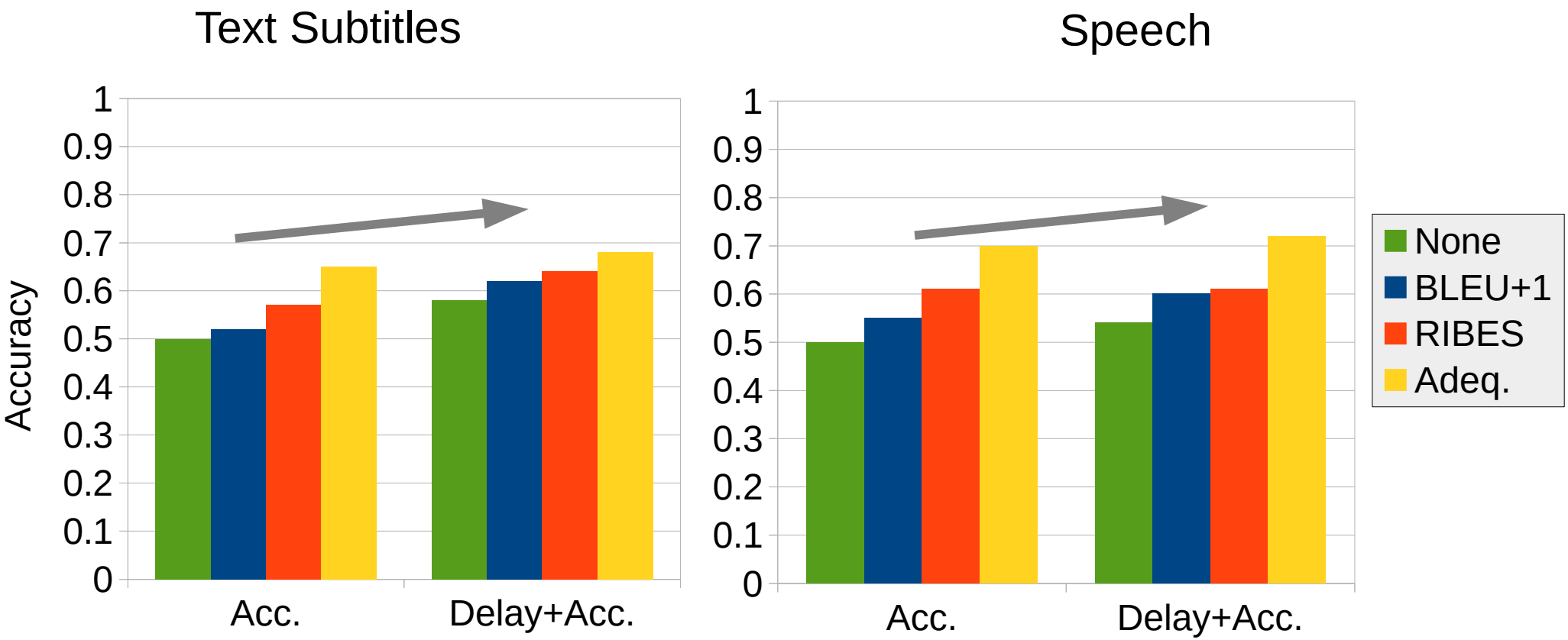
**Syntax-based MT**

**Phrase-based MT**

# Evaluation of Evaluation

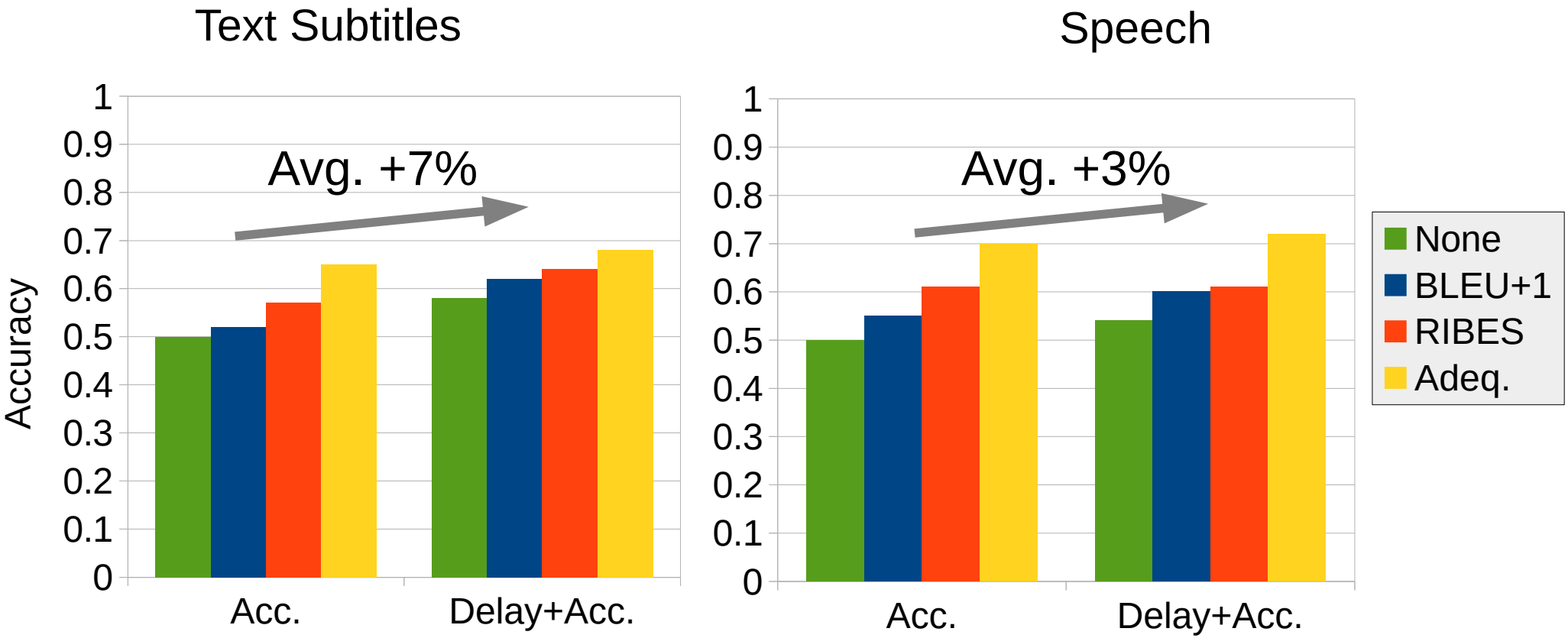


# Q1: Is Delay Important in S2S Translation?




**A: Yes!** In all cases, the scoring function considering delay did as good or better than just considering accuracy.

# Q2: Does Importance Depend on Modality of Presentation?




**A: Yes!** Considering delay was more useful when presenting results through subtitles.

**Why?:** Probably because when watching subtitles, it is possible to hear the original speech.

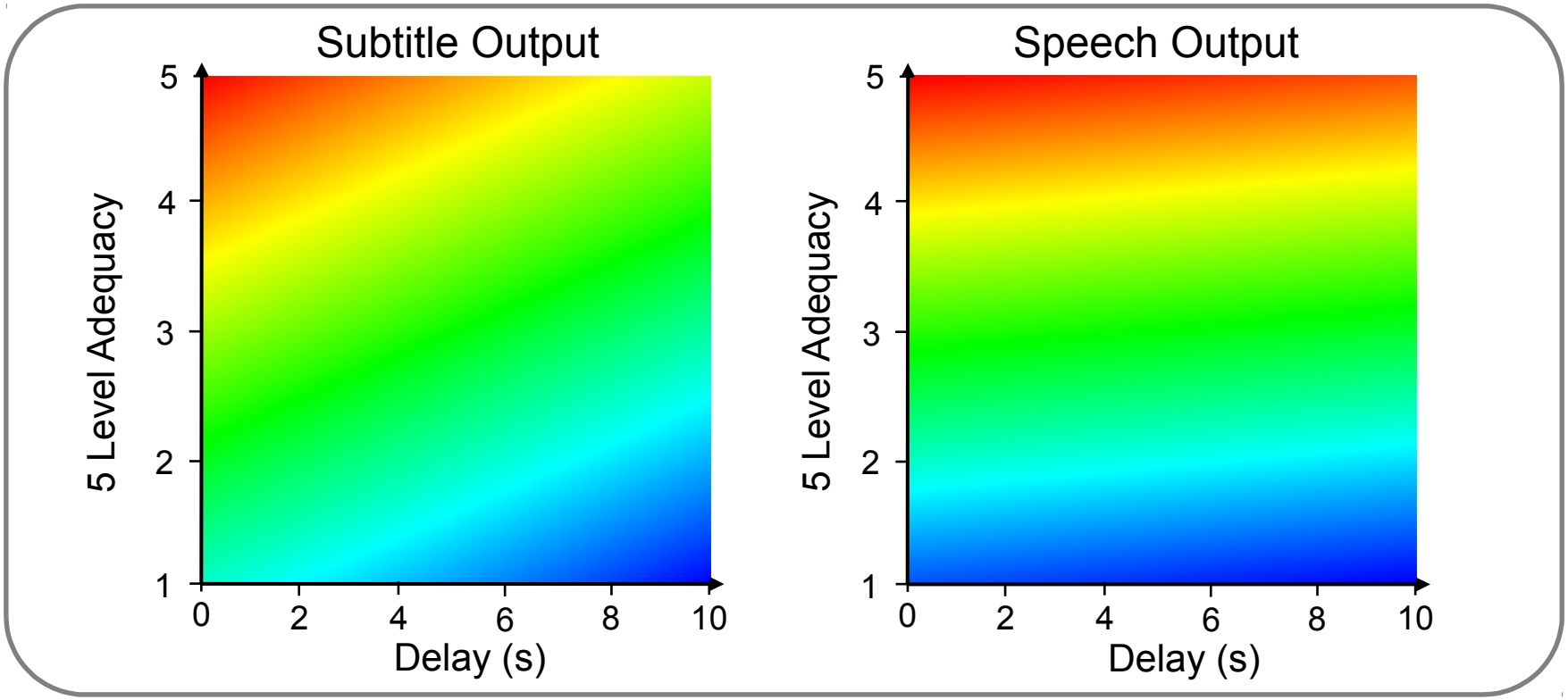
# Learned Evaluation Functions (for Adequacy)

| | Accuracy | Delay |
|---|---|---|
| Subtitle Output | 1.40 | -0.059 |
| Speech Output | 1.99 | -0.018 |

1 point of adequacy =

8.0 sec. of delay

28.5 sec. of delay



# Future Contributions in Evaluation?

- **Adaptation:**
  A more flexible evaluation measure that generalizes to many modalities, genres, tasks.
- **Machine Learning:**
  Non-linear regression functions?
- **Speech/UI:**
  Other factors including presentation modality (avatars?), synthesis quality play a large role.

# Conclusion

# Conclusion

- The problem of high-accuracy simultaneous translation covers many fields of NLP/Speech: parsing, machine learning, language modeling, prosody, paraphrasing.
- Still a new field, lots of opportunities for interesting applications of NLP tech!