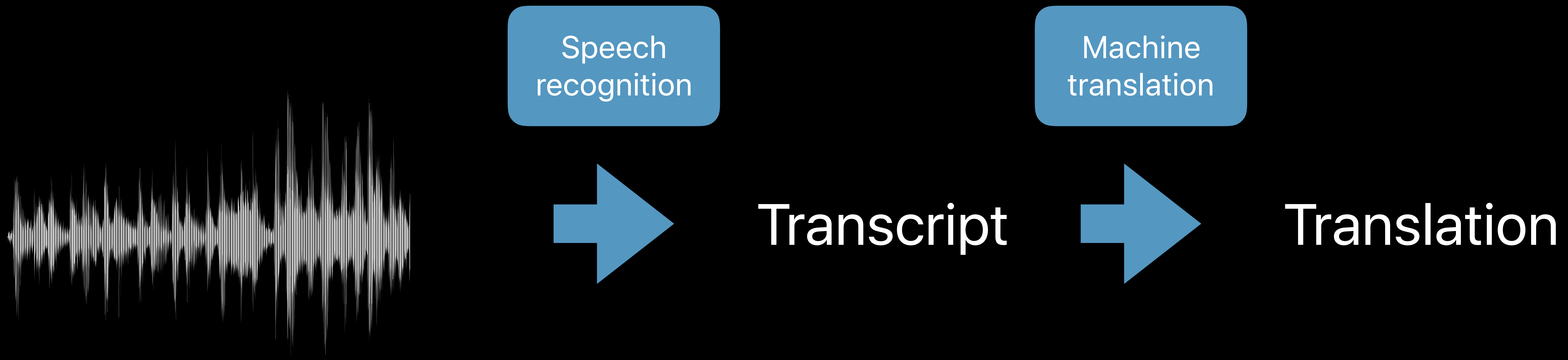
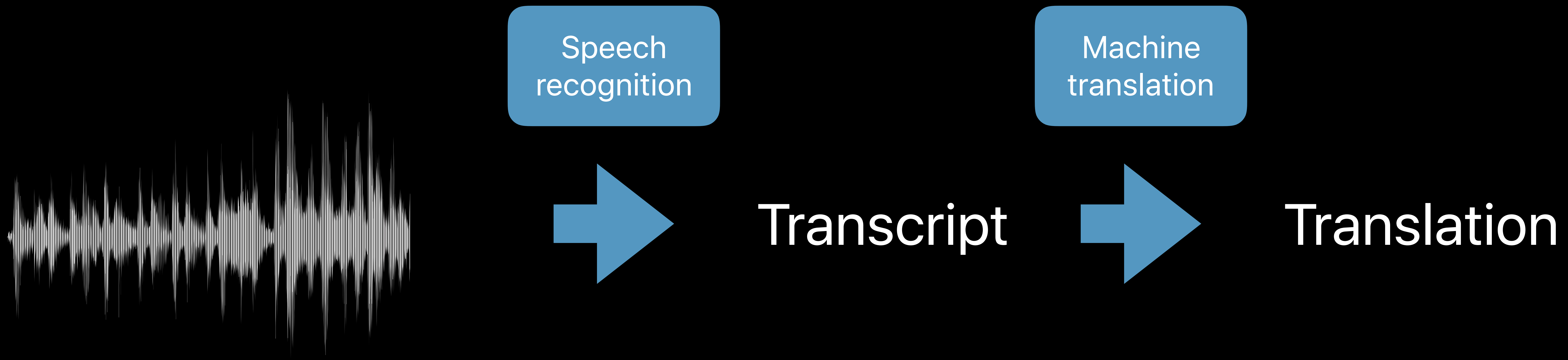
Speech translation

Matthias Sperber


Speech recognition
Transcript
Machine translation
Translation



Problem solved?

Agenda

Agenda

Challenges & applications

Agenda

Challenges & applications

Cascaded models

Agenda

Challenges & applications

Cascaded models

Simultaneous translation

Agenda

Challenges & applications
Cascaded models
Simultaneous translation
End-to-end models

Agenda

Challenges & applications

Cascaded models

Simultaneous translation

End-to-end models

Challenges & applications

How does speech differ from text?

Data representations

Speech

Continuous signal

Written Language

word1 word13 word5621 word15625
word256 word13 …

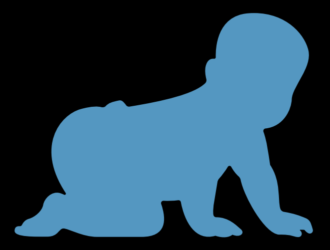
Discrete sequence

★ Modeling approaches (used to) differ

Acquisition

Speech

As infants, naturally



Written Language

Needs a writing system

Needs to be taught

★ Speech-enabled services reach new users

Information content

Written language approximates speech

Information content

Written language approximates speech

- *“Will you have marmalade or jam?”*

Information content

Written language approximates speech

- *"Will you have marmalade or jam?"*
- 🎧

Information content

Written language approximates speech

- *"Will you have marmalade or jam?"*
- 🎧
- *"Will you have marmalade, jam, or something else?"*

Information content

Written language approximates speech

- *"Will you have marmalade or jam?"*
- 🎧
- *"Will you have marmalade, jam, or something else?"*

★ Prosody (non-verbal parts) are partly lost

★ Semantics can become ambiguous

Fluency

Speech

Often spontaneous

"Hi um yeah I'd like to talk about how you dress for work and and um what do you normally what type of outfit do you normally have to wear"

Written Language

Often fluent, grammatical sentences

Fluency

Speech

Often spontaneous

“Hi ~~um yeah~~ I’d like to talk about how you dress for work and ~~and um what do you normally~~ what type of outfit ~~do~~ you normally have to wear”

Written Language

Often fluent, grammatical sentences

Fluency

Speech

Often spontaneous

"Hi ~~um yeah~~ I'd like to talk about how you dress for work and ~~and um what do you normally~~ what type of outfit ~~do~~ you normally have to wear"

★ Usability: literal speech hard to read

★ Data: hard to find textual training data

★ Translatability: clean before translating

Written Language

Often fluent, grammatical sentences

Applications

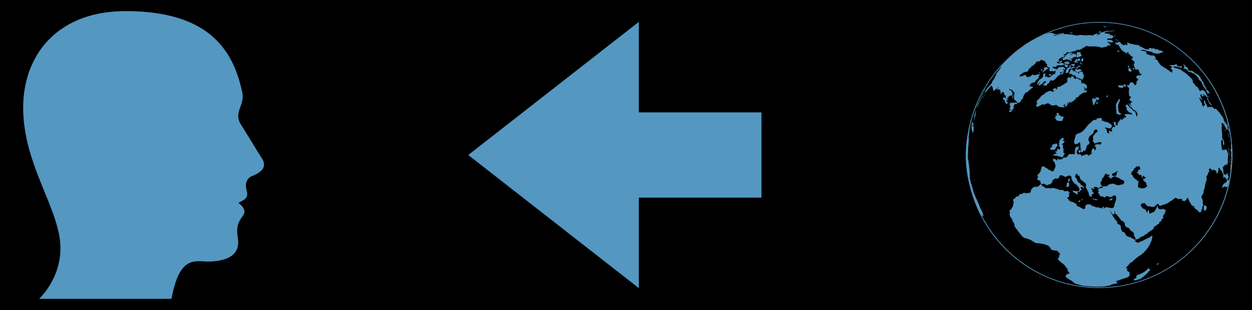
Applications

Information flow

Applications

Information flow

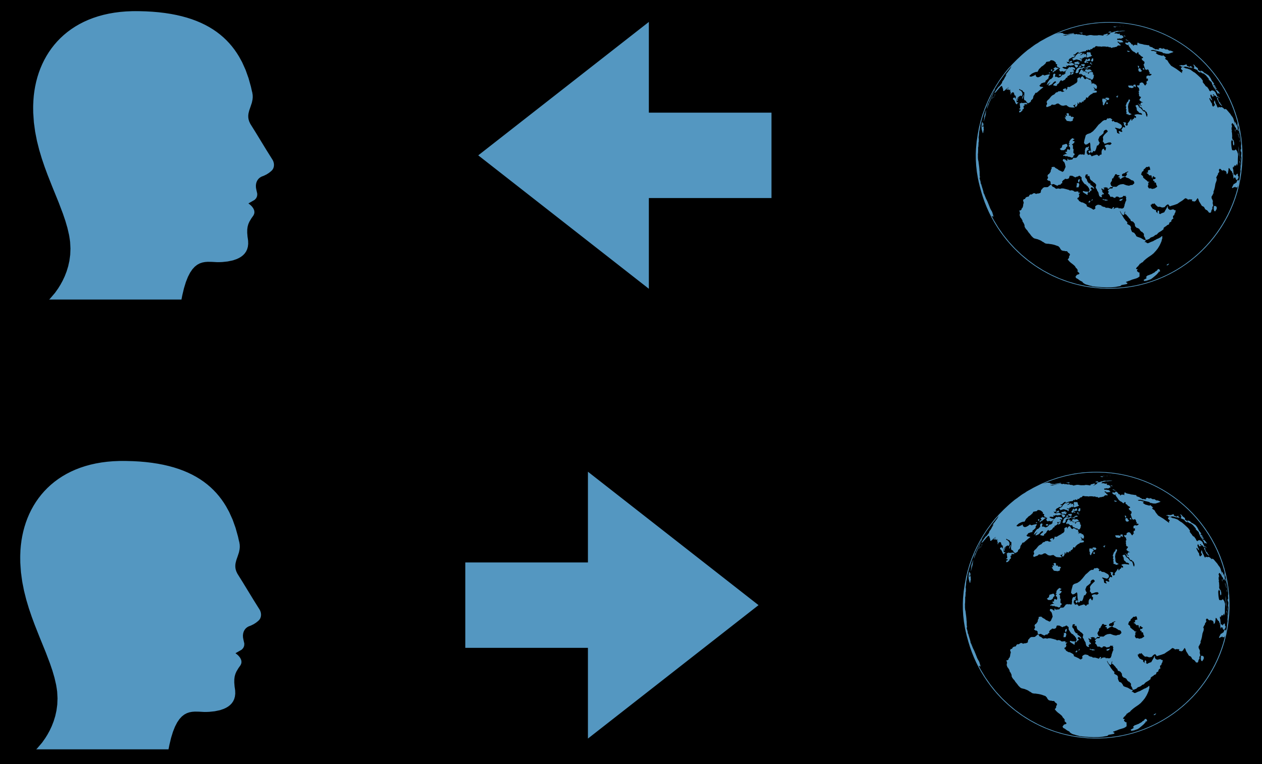
- Assimilation / information access



Applications

Information flow

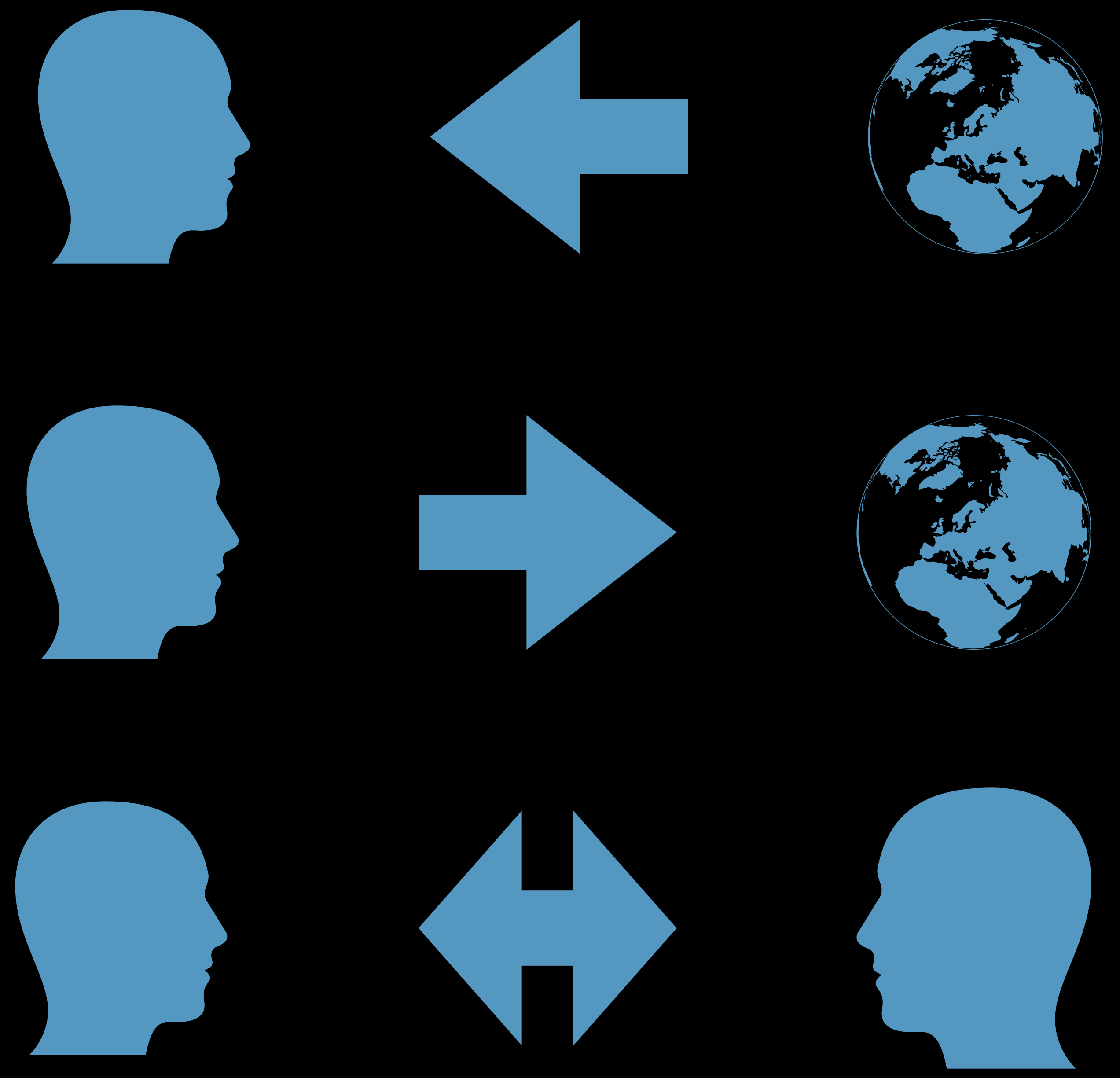
- Assimilation / information access
- Dissemination / broadcasting



Applications

Information flow

- Assimilation / information access
- Dissemination / broadcasting
- Interactive communication



Applications

Simultaneous translation

- No segments
- No pauses
- Translation delivered simultaneously
- Additive latency

Applications

Consecutive translation

- Fixed, short, natural segments
- Multiplicative latency
- Examples:
  - Voice commands
  - Consecutively translated speeches

Applications

Online vs. offline

- Online case: speed is important
  - Latency
  - Throughput
- Offline case: speed is less critical

Applications

Output modality

Applications

Output modality

- Text

Applications

Output modality

- Text
- Speech (e.g. text + TTS)

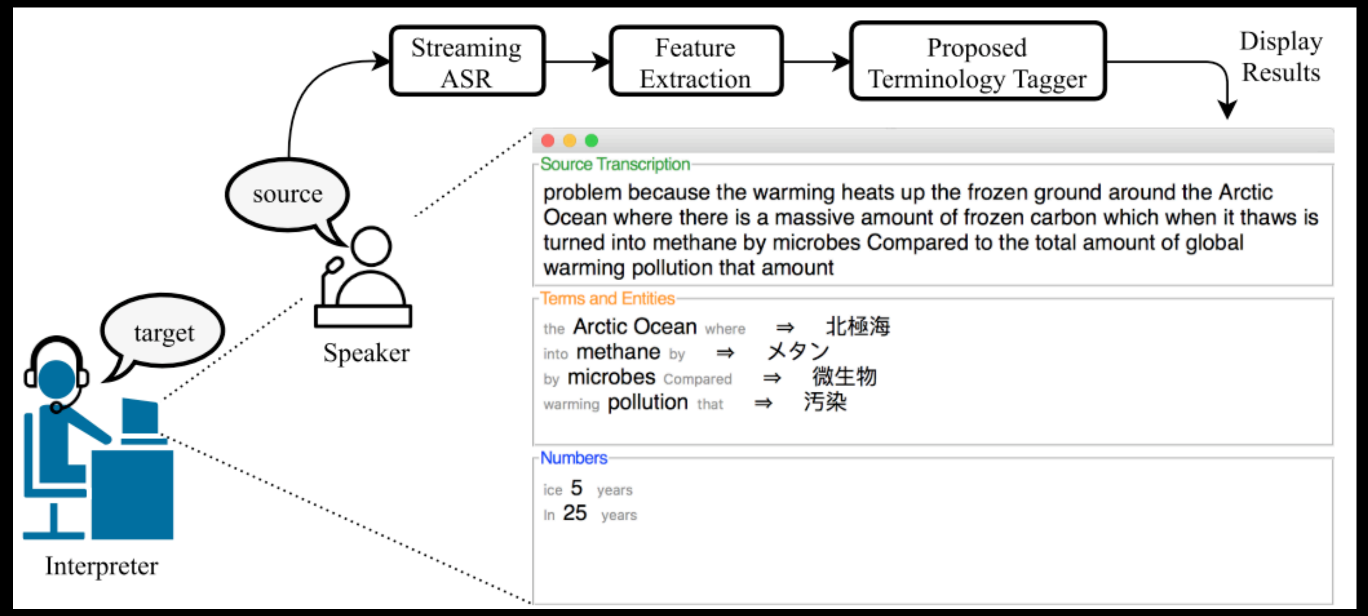
Applications

Output modality

- Text

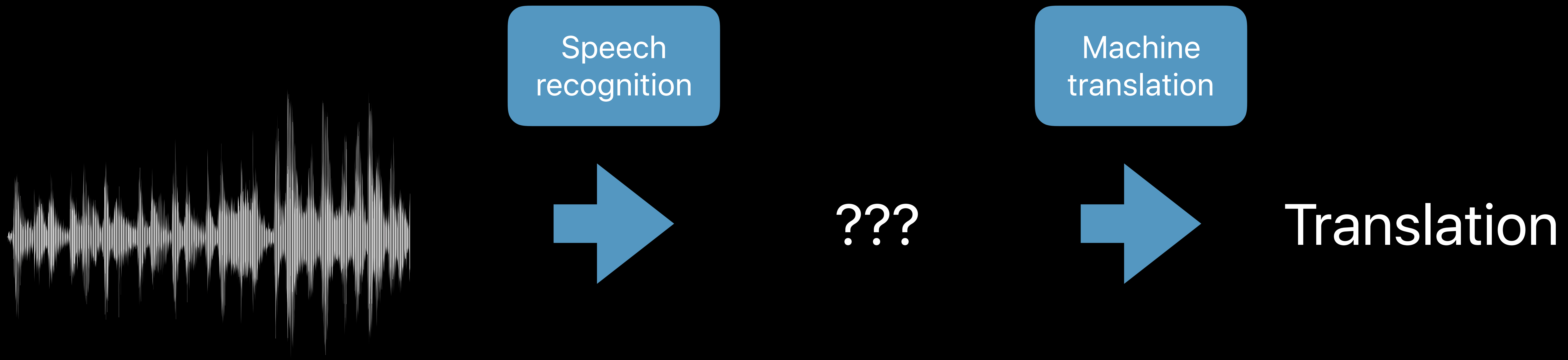
- Speech (e.g. text + TTS)

- Condensed information (e.g. only named entities)

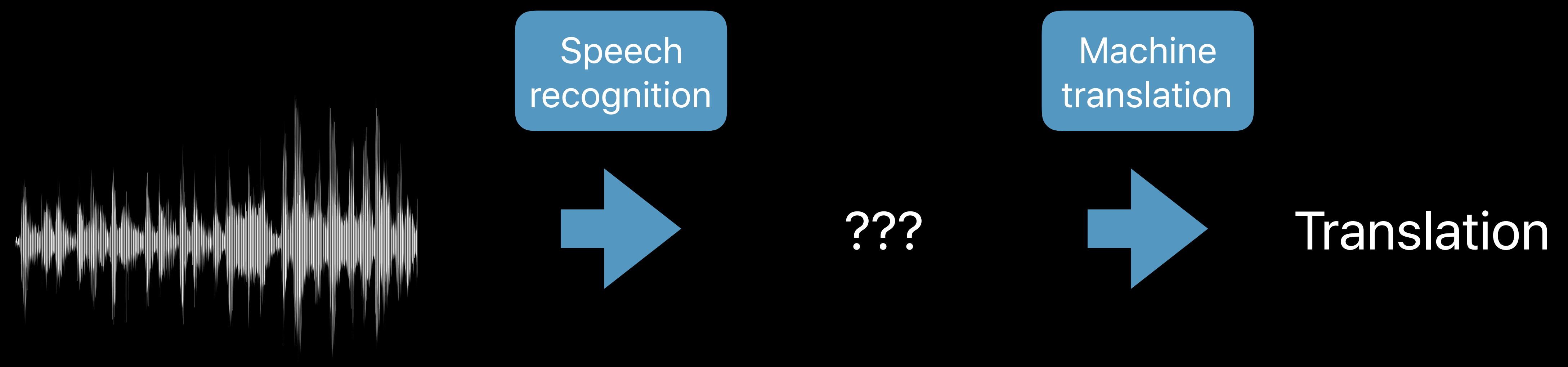


Cascaded Models

Cascaded Approach

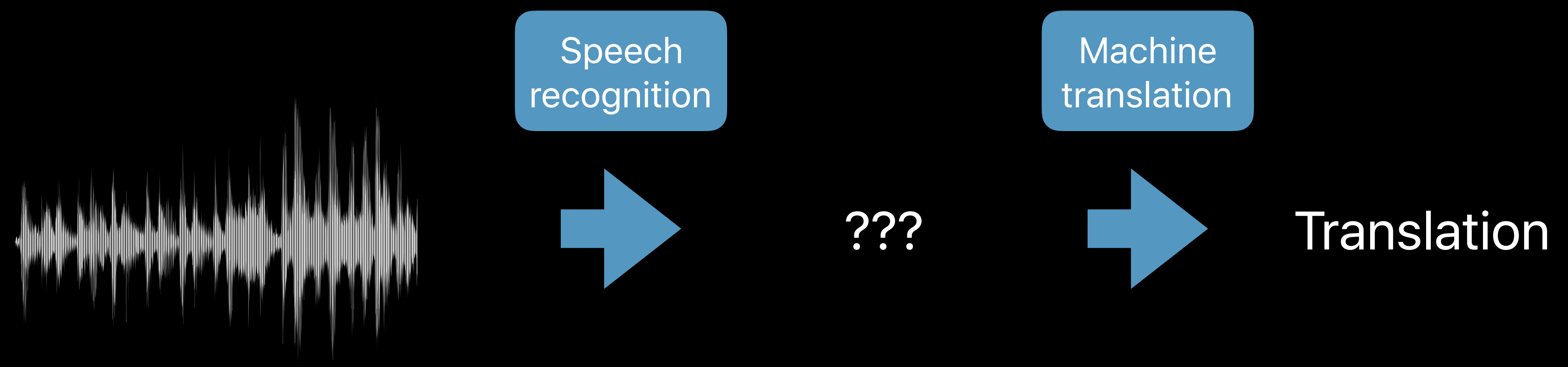


Cascaded Approach



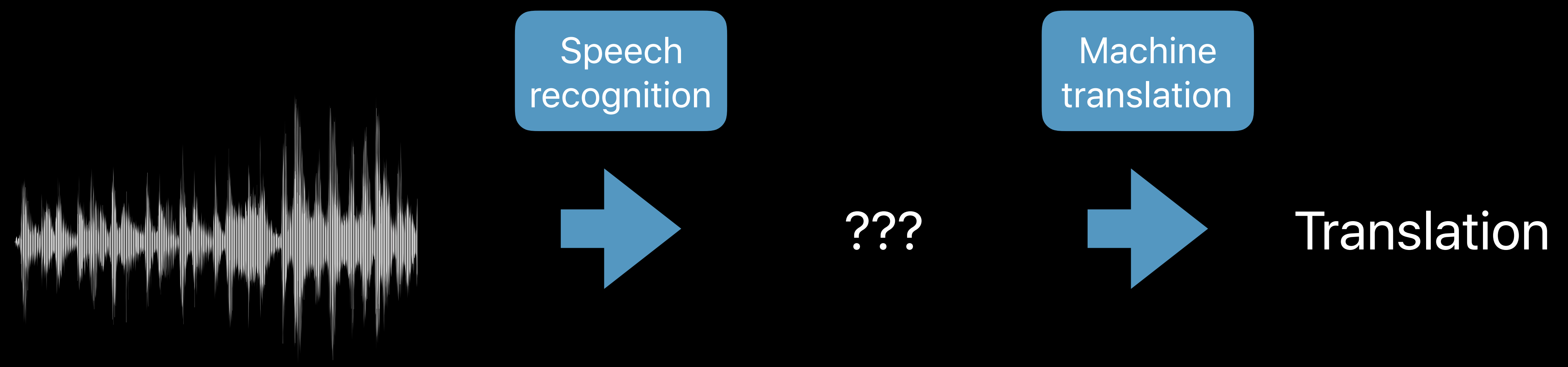
- Problem 1: Error propagation
- Problem 2: Domain Mismatch
- Problem 3: Information Loss

Cascaded Approach



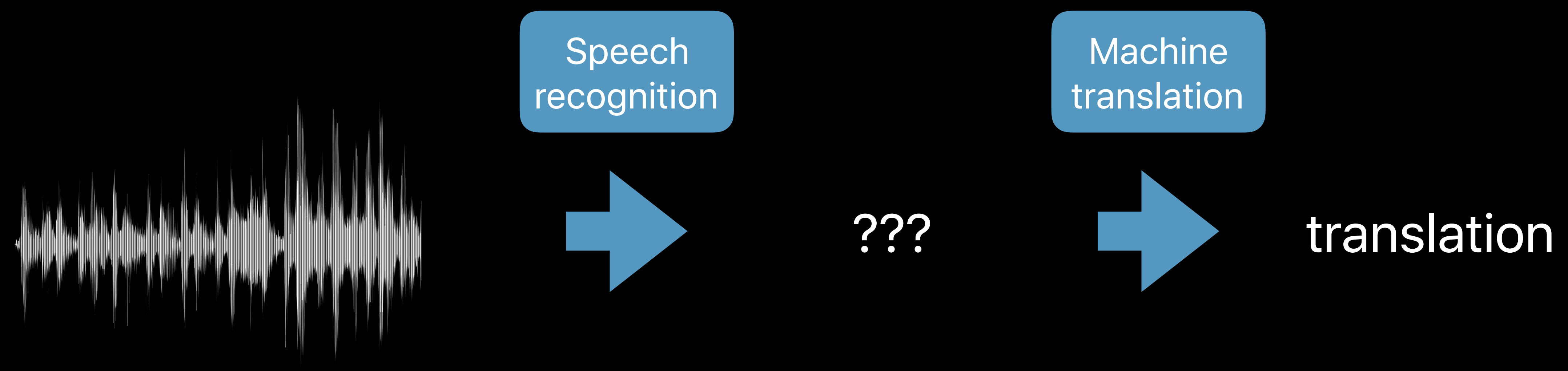
- Problem 1: Error propagation
  - All models make mistakes
  - How to translate ASR mistakes?
    - Avoid error propagation & compounding
- Problem 2: Domain Mismatch
- Problem 3: Information Loss

Cascaded Approach



- Problem 1: Error propagation
- Problem 2: Domain mismatch
  - Speech recognizer outputs verbatim, spontaneous language
    - Possibly disfluent, no punctuation, no capitalization
  - MT trained on written-style data
- Problem 3: Information Loss

Cascaded Approach



- Problem 1: Error propagation
- Problem 2: Domain mismatch
- Problem 3: Information loss
  - Transcript discards prosody

Cascaded Approach

Target text

Source speech

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \text{argmax}_T Pr\left(T \mid X\right)$$

$$= \text{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

$$\approx \text{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr(T \mid X)$$

$$= \operatorname{argmax}_T \sum_S Pr(T \mid S, X)\, Pr(S \mid X)$$

$$\approx \operatorname{argmax}_T \sum_S Pr(T \mid S)\, Pr(S \mid X)$$

Assume cond. independence: $(T \perp\!\!\!\perp X) \mid S$

machine translation model

speech recognition model

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \mathrm{argmax}_T Pr\left(T \mid X\right)$$

$$= \mathrm{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

$$\approx \mathrm{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

$$\approx \mathrm{argmax}_T \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Early decision: consider only e.g. 1-best, *n*-best list, lattice

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Problem 1: error propagation

$$\approx \operatorname{argmax}_T \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Early decision: consider only e.g. 1-best, *n*-best list, lattice

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

Problem 3: information loss

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Problem 1: error propagation

Early decision: consider only e.g. 1-best, *n*-best list, lattice

$$\approx \operatorname{argmax}_T \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Cascaded Approach

Target text

Source speech

Source text

$$\hat{T} = \operatorname{argmax}_T Pr\left(T \mid X\right)$$

$$= \operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Problem 3: information loss

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

Problem 1: error propagation

Early decision: consider only e.g. 1-best, *n*-best list, lattice

Problem 2: domain mismatch

Different assumptions on $Pr\left(S\right)$

Addressing Error Propagation

$$\mathrm{argmax}_T \sum_{S \in \mathcal{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Early decision: consider only
e.g. 1-best, *n*-best list, lattice

Addressing Error Propagation

n-best lists

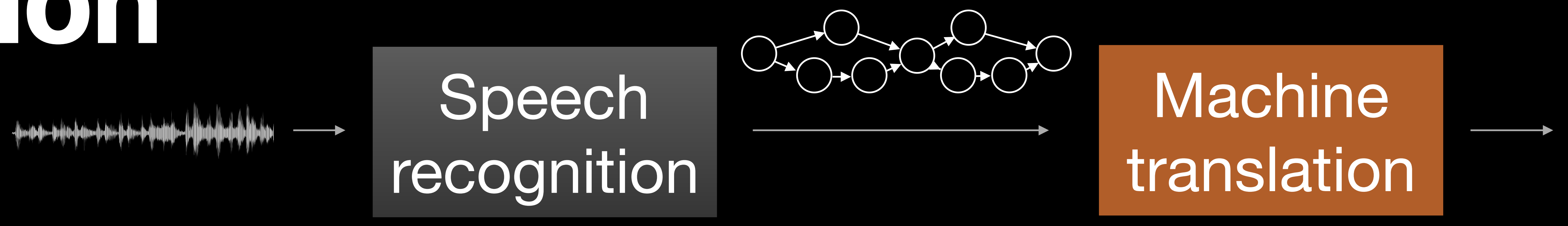
[Lavie+1995; Quan+2005; Lee+2007]

- Idea:
  - Speech recognizer outputs *n* best recognitions, including scores
  - Translate each, pick option with best combined score
- Problem:
  - Computationally inefficient

<s> hay qué bueno </s>	0.48
<s> ah qué bueno </s>	0.4
<s> hay que buena </s>	0.12

Addressing Error Propagation

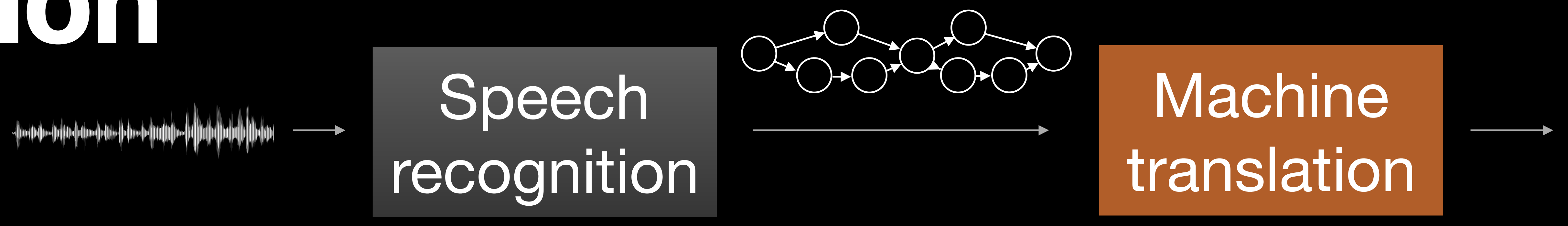
Lattices



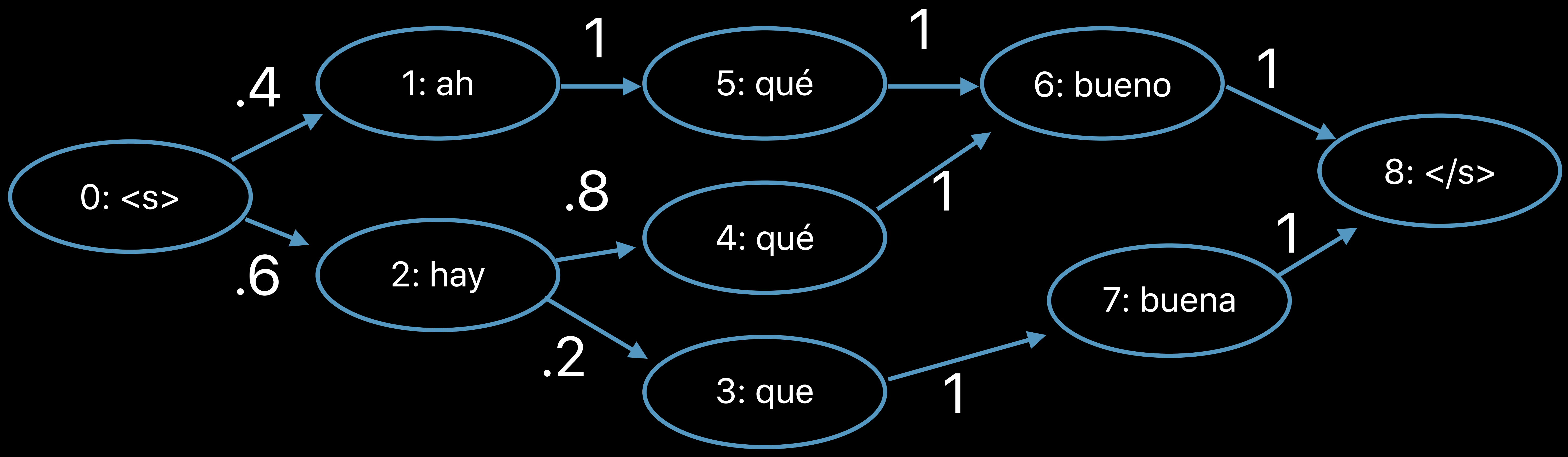
<s> ah qué bueno </s>	0.4
<s> hay qué bueno </s>	0.48
<s> hay que buena </s>	0.12

Addressing Error Propagation

Lattices

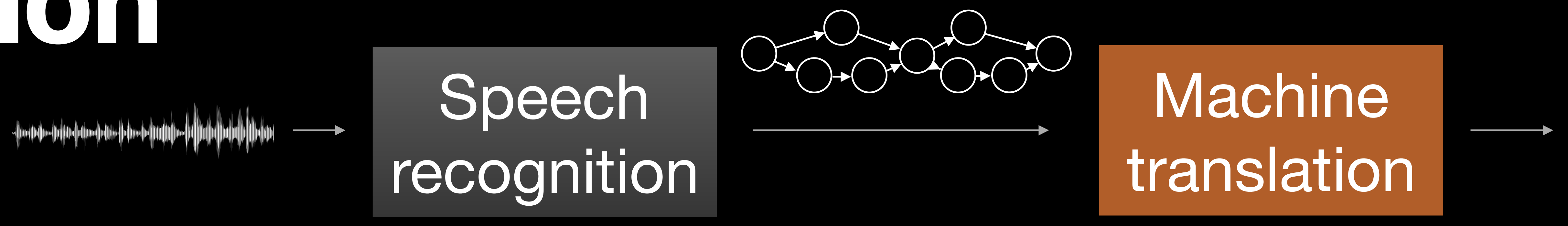


<s> ah qué bueno </s>	0.4
<s> hay qué bueno </s>	0.48
<s> hay que buena </s>	0.12

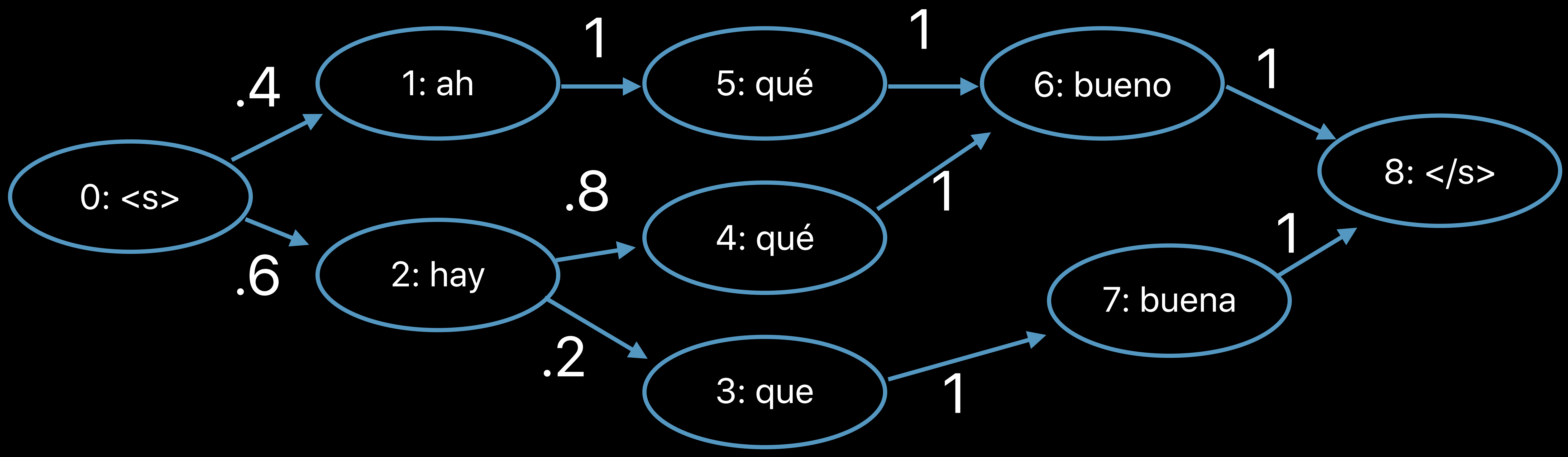


Addressing Error Propagation

Lattices



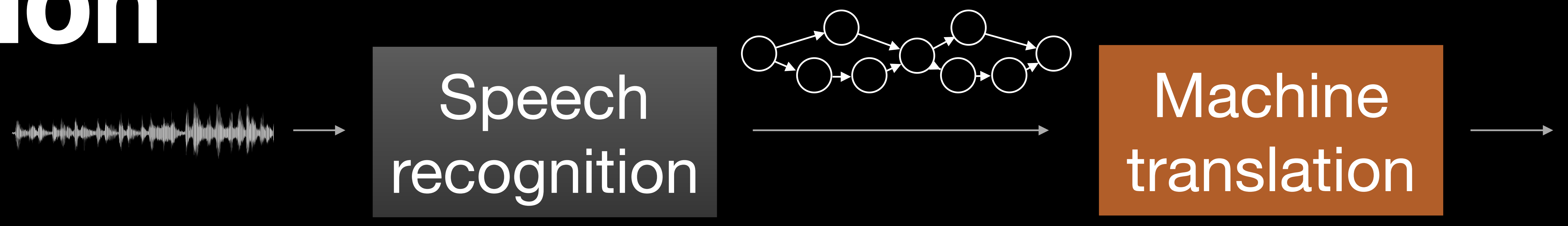
<s> ah qué bueno </s>	0.4
<s> hay qué bueno </s>	0.48
<s> hay que buena </s>	0.12



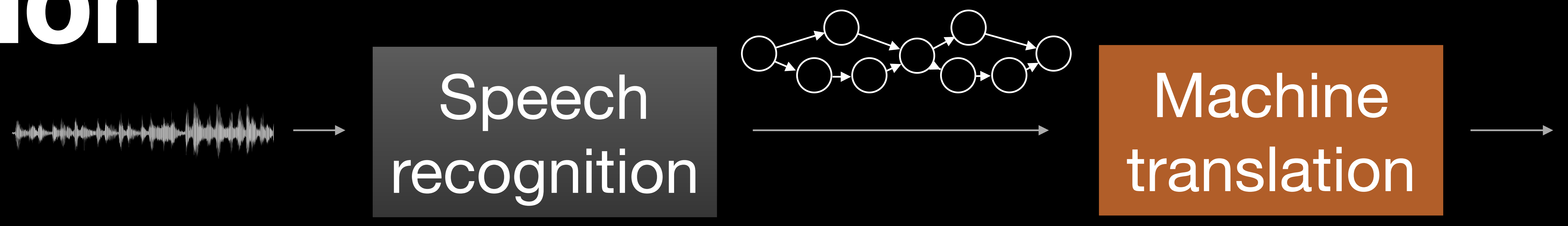
- Lattices: a compact representation of *n*-best lists

Addressing Error Propagation

Lattice Translation

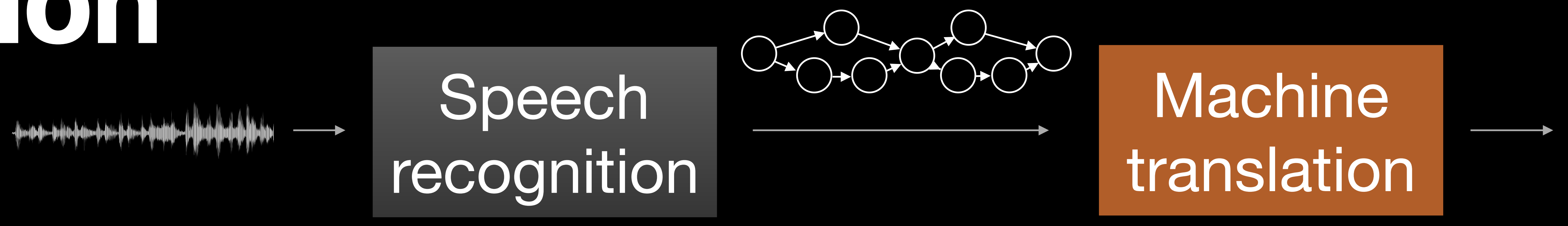


Addressing Error Propagation
Lattice Translation



- SMT: lattice decoding
  [Saleem+2004; Zhang+2005; Bertoldi+2007; Matusov+2008; …]

Addressing Error Propagation
Lattice Translation



- SMT: lattice decoding
  [Saleem+2004; Zhang+2005; Bertoldi+2007; Matusov+2008; ...]
- Lattice-to-sequence NMT
  [Su+2017; Sperber+2017; Sperber+2019; Xiao+2019; Zhang+2019]

Addressing Error Propagation

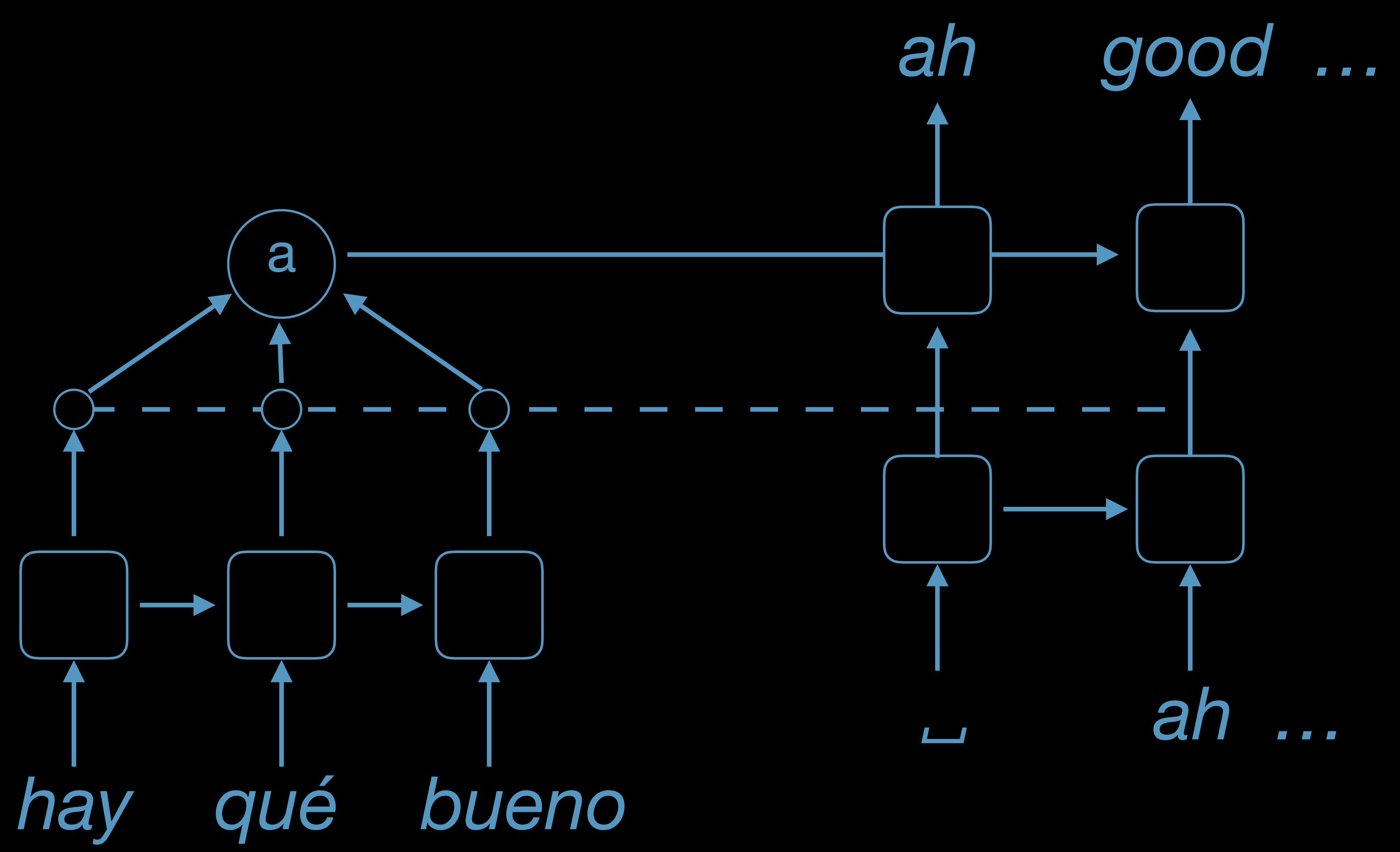
Lattices LSTM encoders

[Sperber+2017]

Addressing Error Propagation

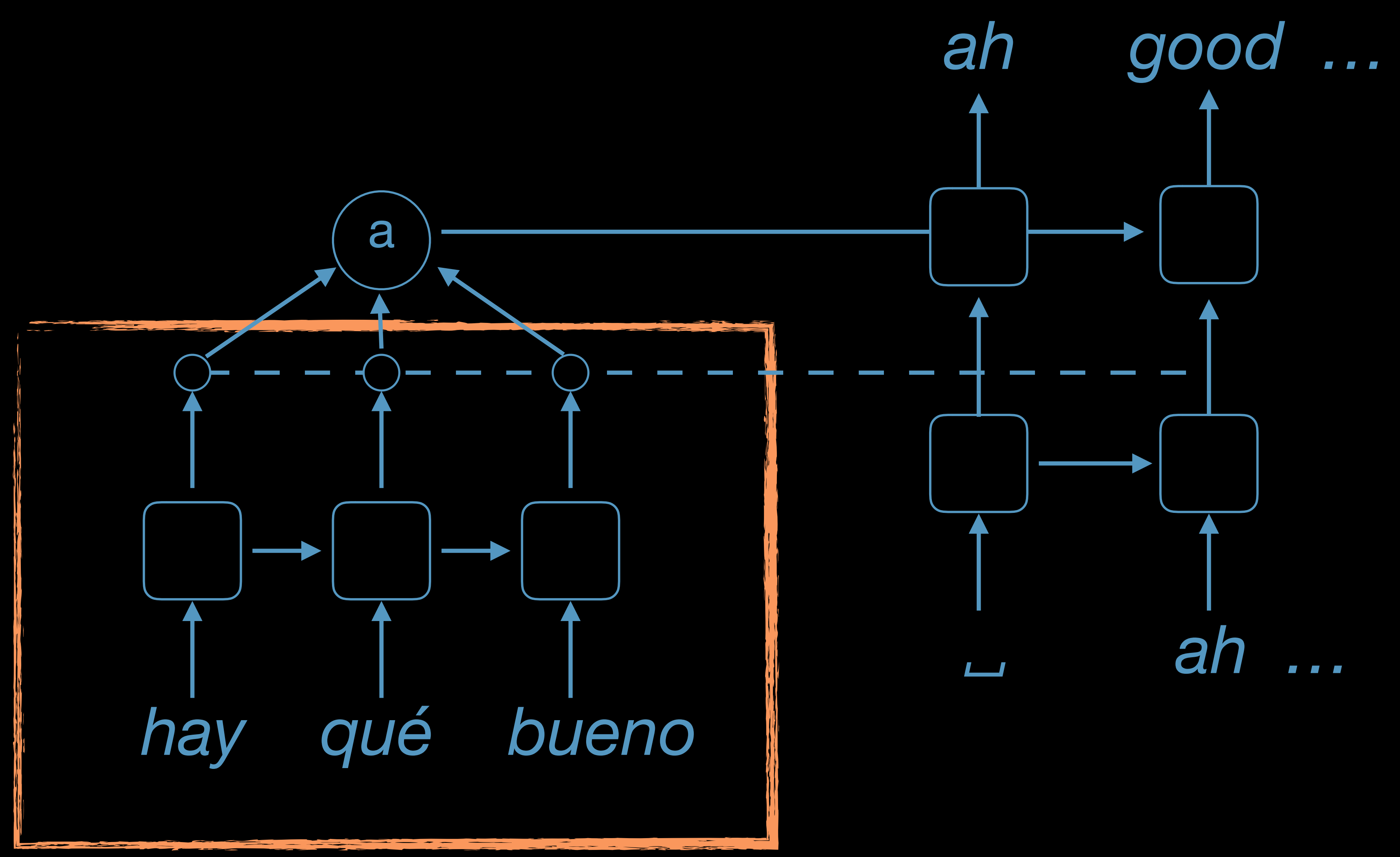
Lattices LSTM encoders

[Sperber+2017]



Addressing Error Propagation
Lattices LSTM encoders

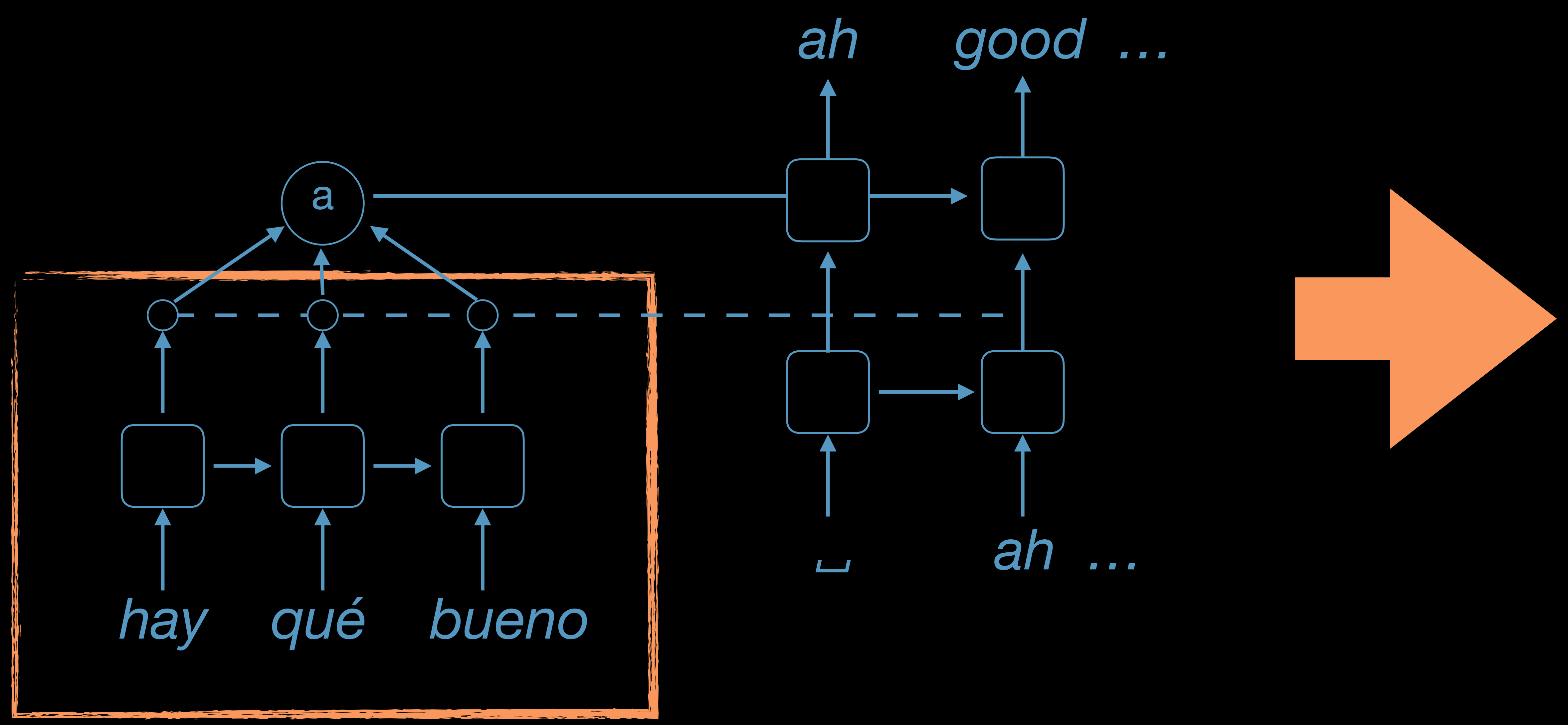
[Sperber+2017]



Addressing Error Propagation

Lattices LSTM encoders

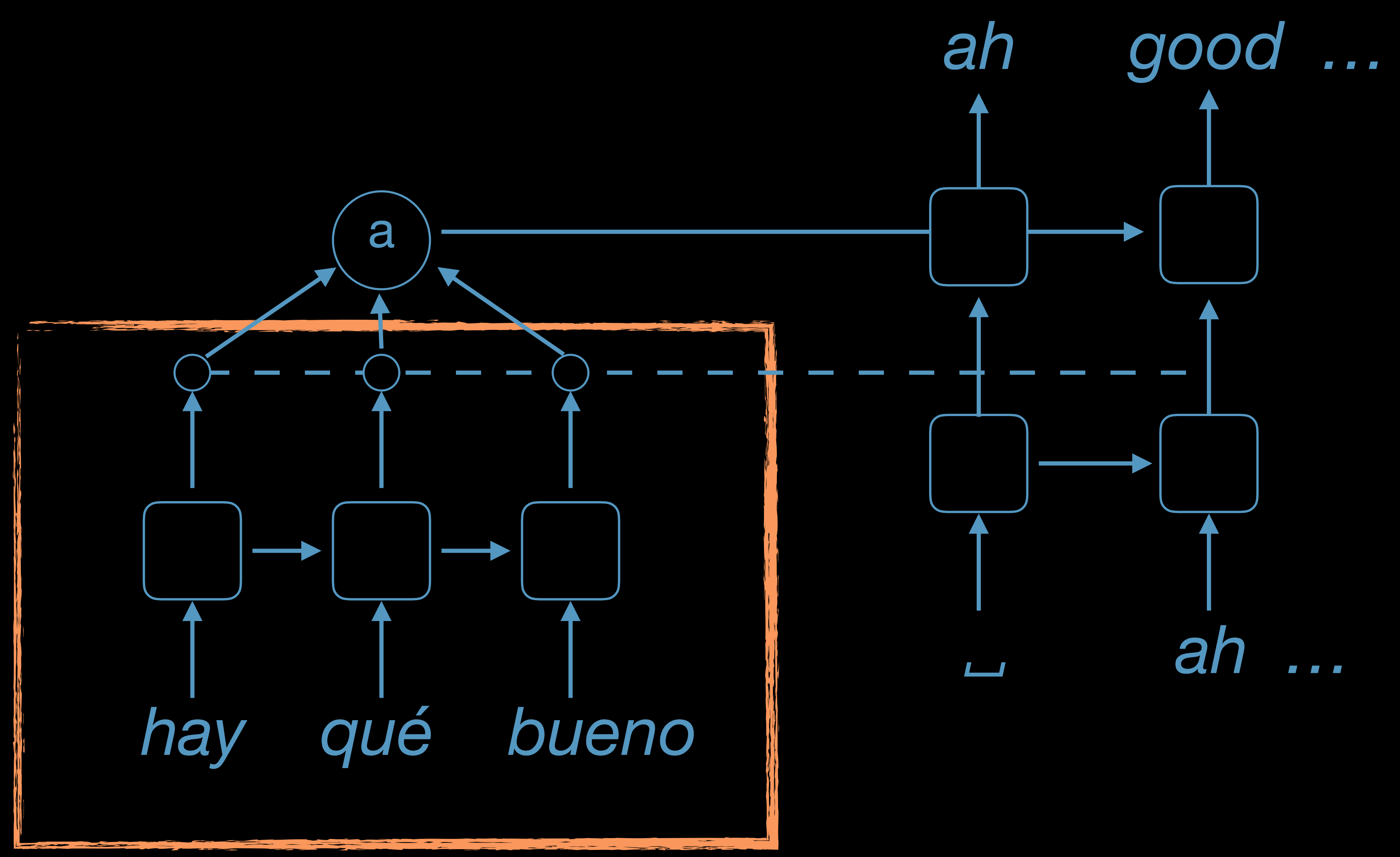
[Sperber+2017]

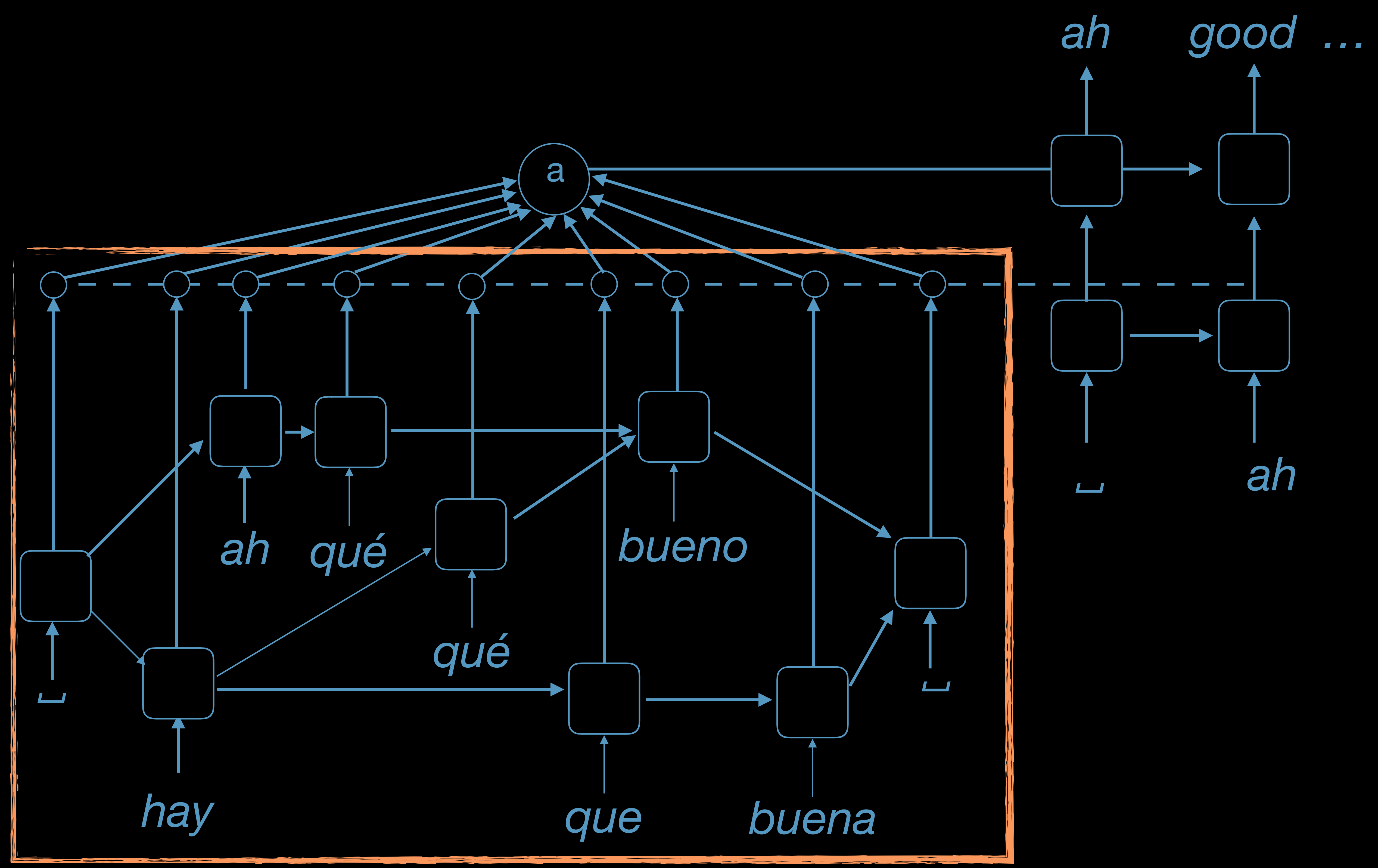


Addressing Error Propagation

Lattices LSTM encoders

[Sperber+2017]





Addressing Error Propagation

Lattices LSTM encoders

[Sperber+2017]

+ bidirectional

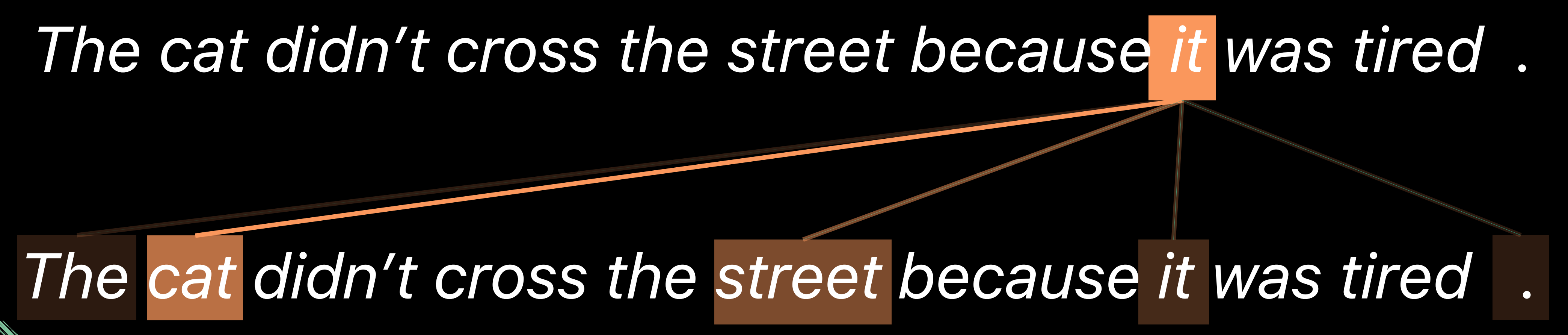
+ layer stacking

Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.

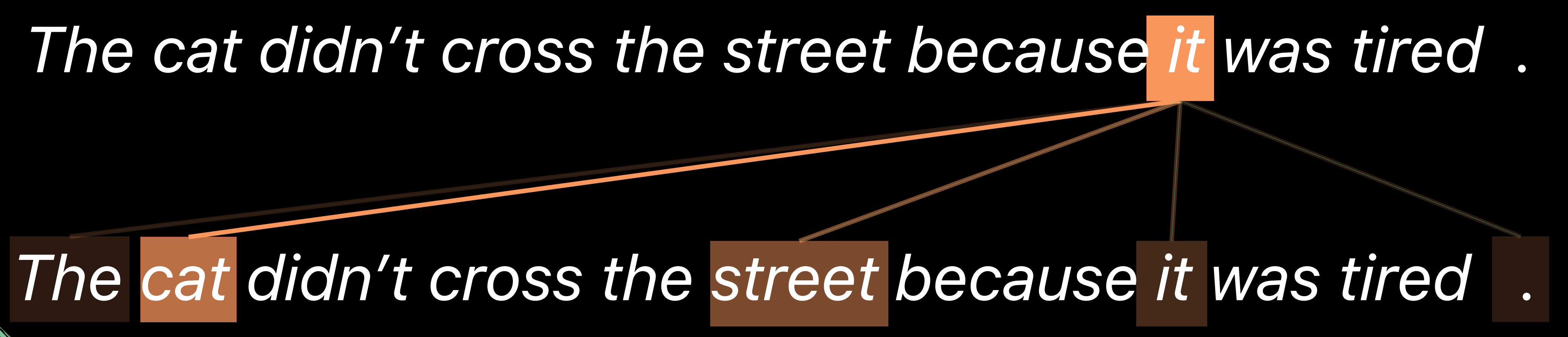


Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.

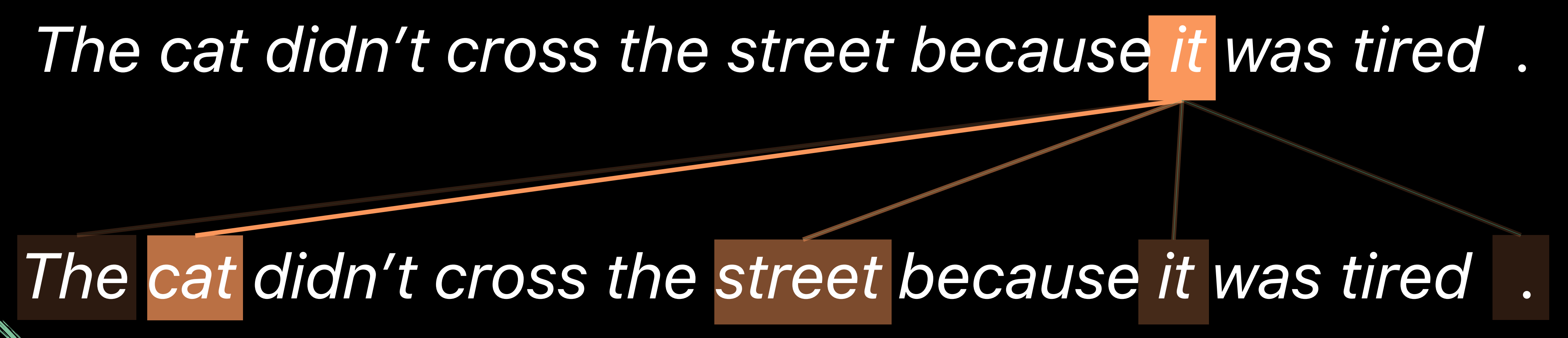


Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.

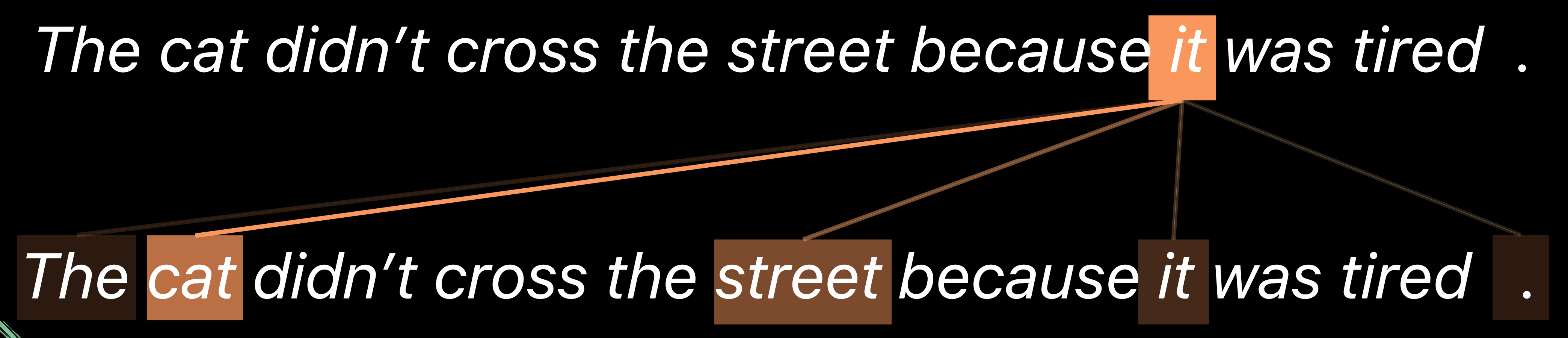


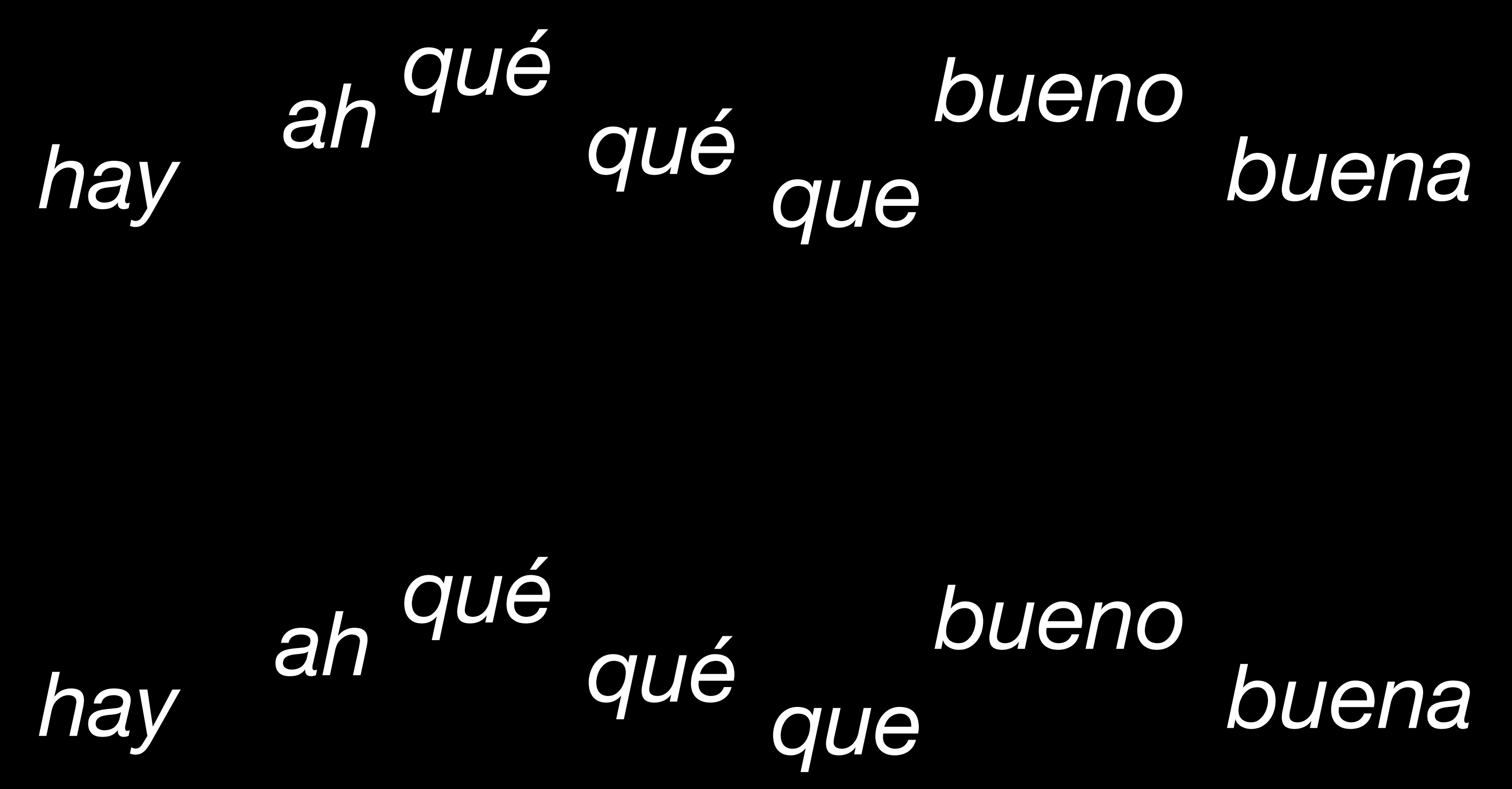
Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.



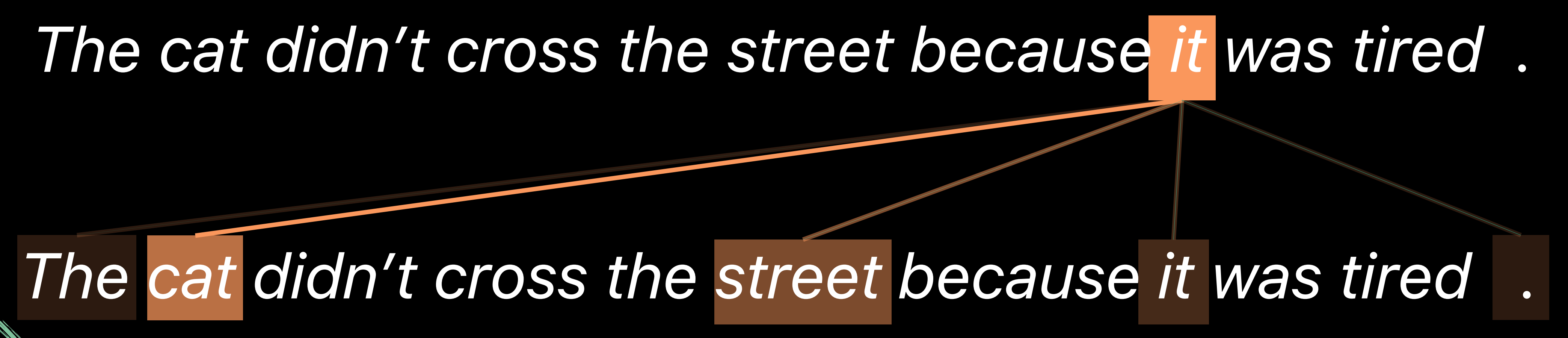


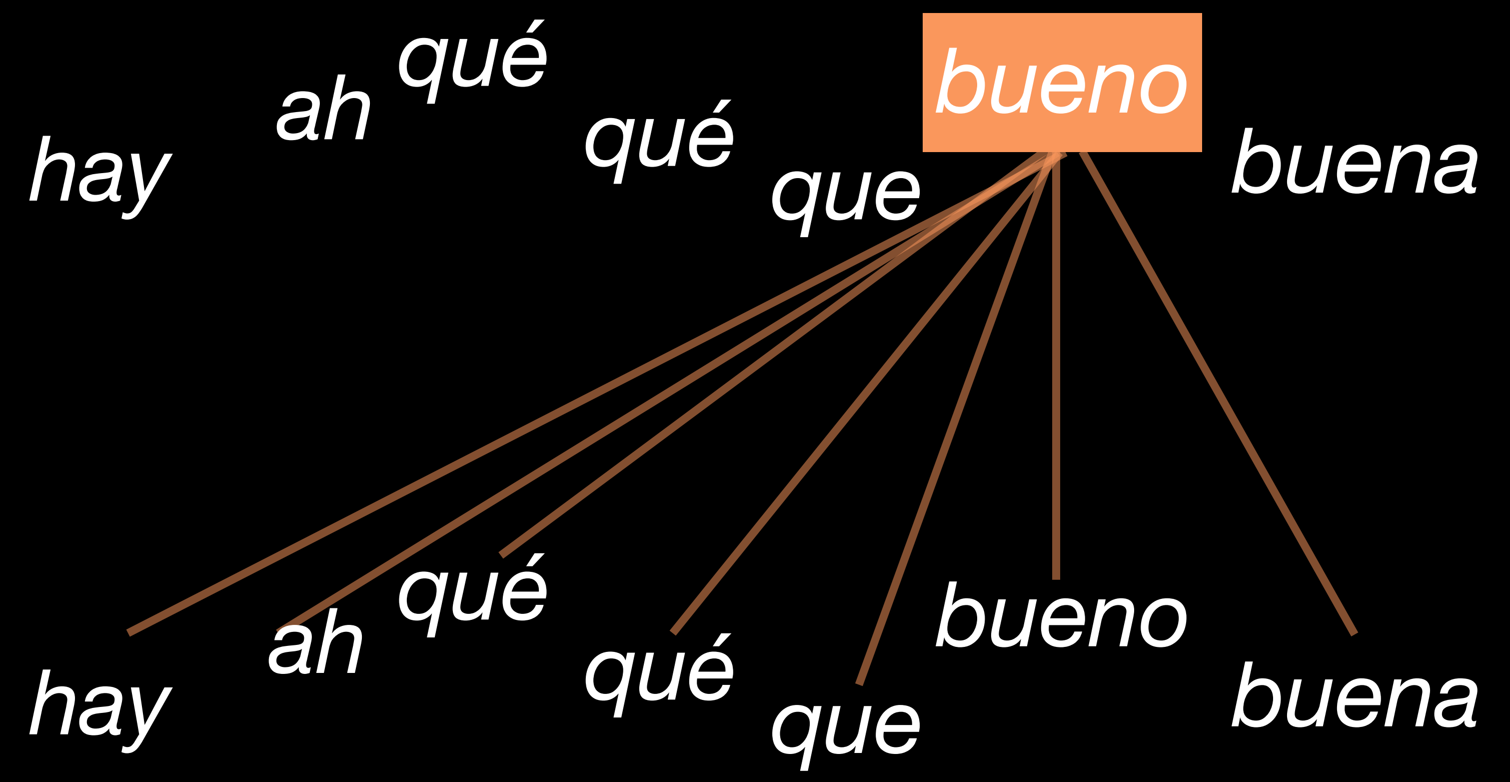
Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.



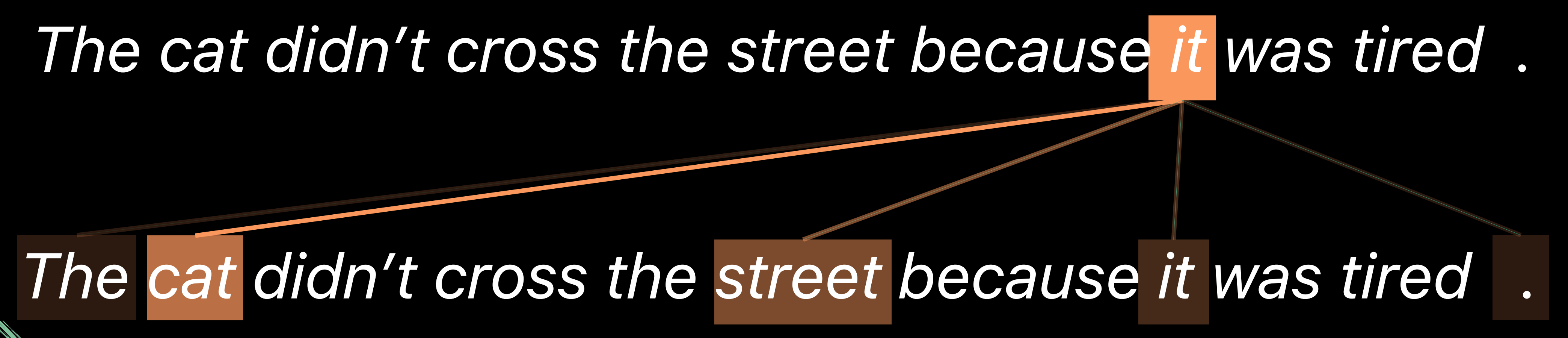


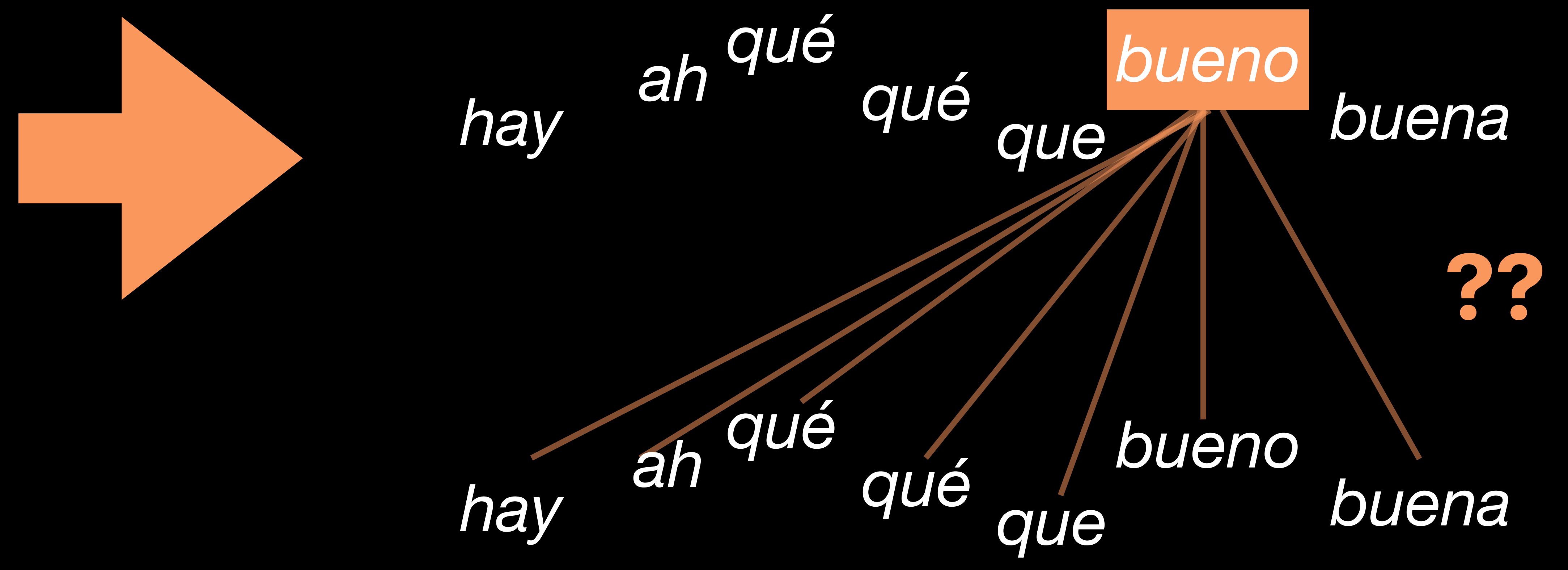
Addressing Error Propagation

Lattice Self-Attention

[Sperber+2019]

Self-attention encodes sequences of vectors by relating these vectors to each-other based on pairwise similarities.

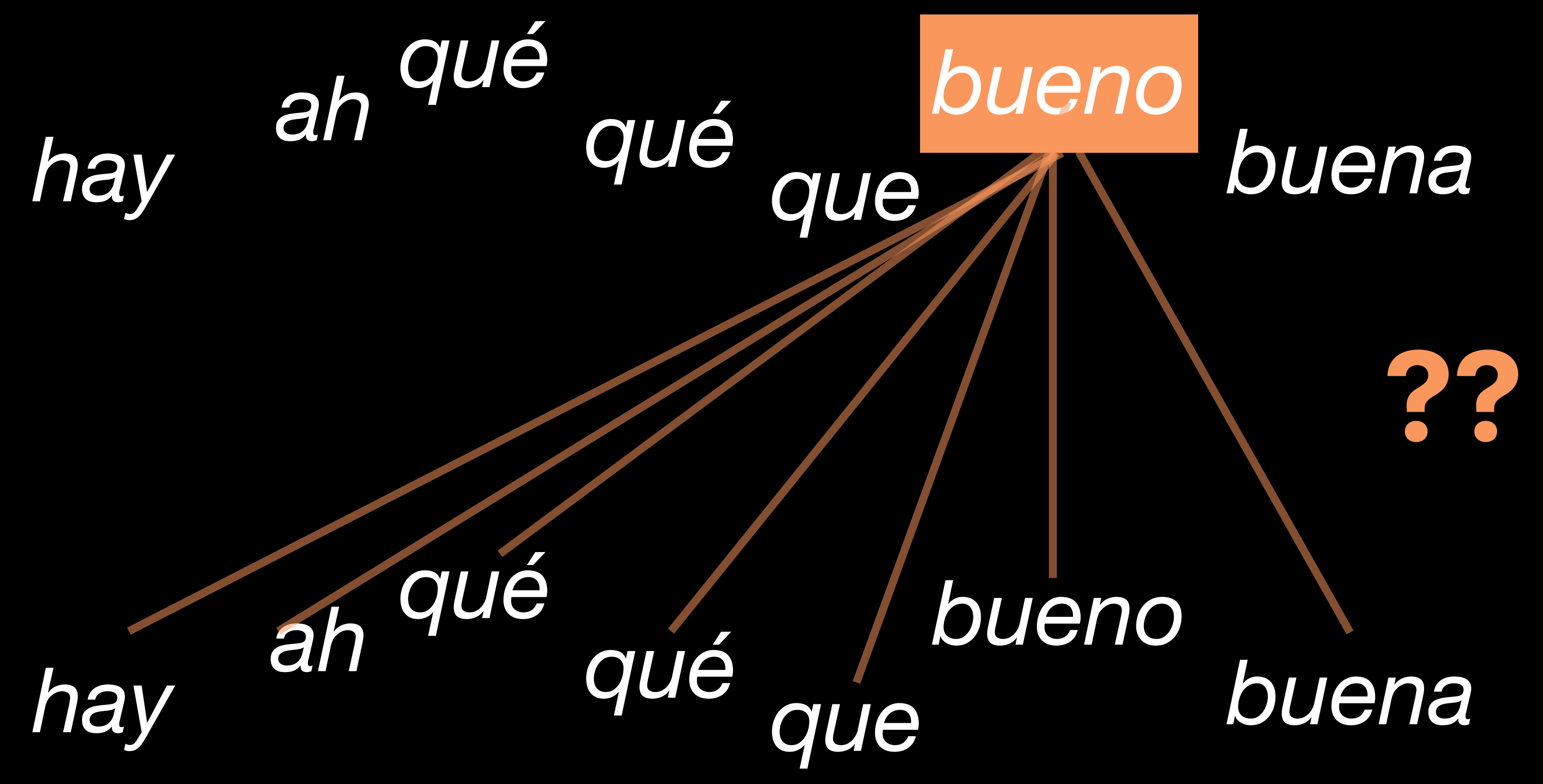




Addressing Error Propagation

Lattice Self-Attention: Positional Representation

[Sperber+2019]

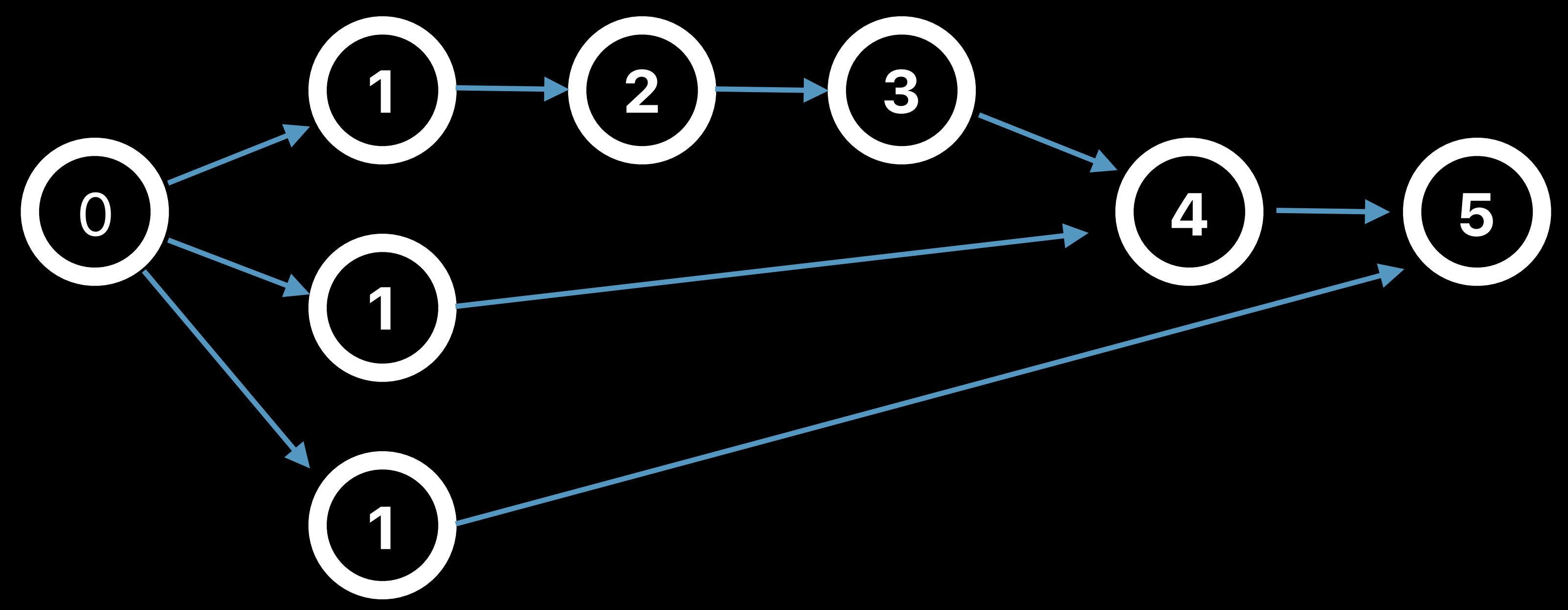


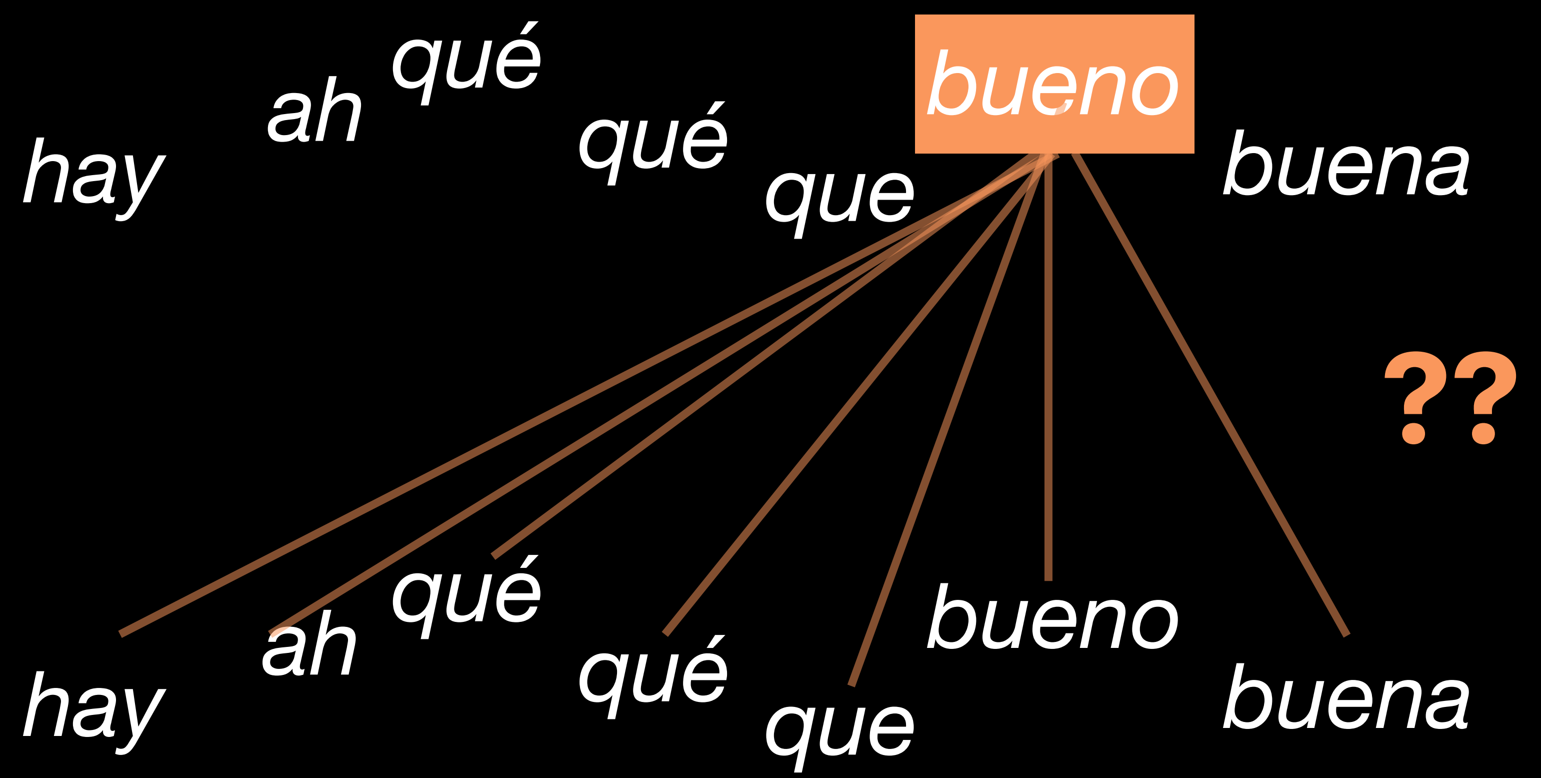
Addressing Error Propagation

Lattice Self-Attention: Positional Representation

[Sperber+2019]

Longest distance



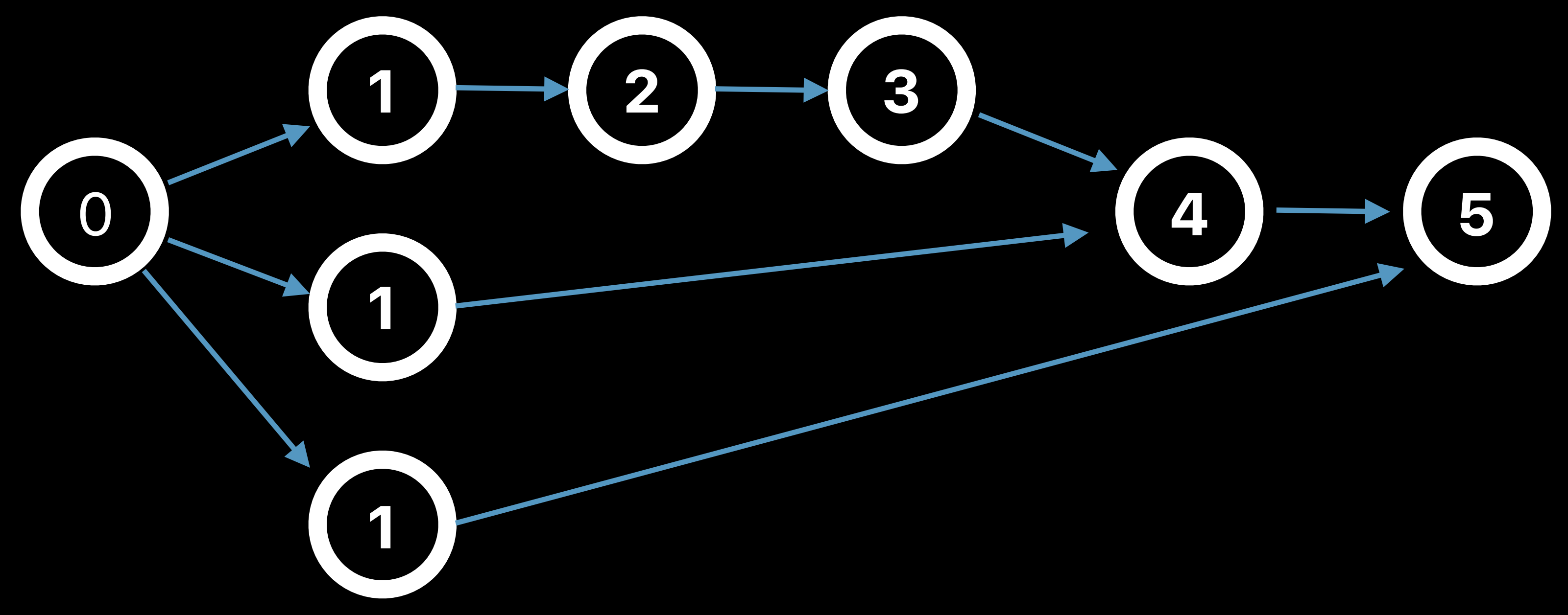


Addressing Error Propagation

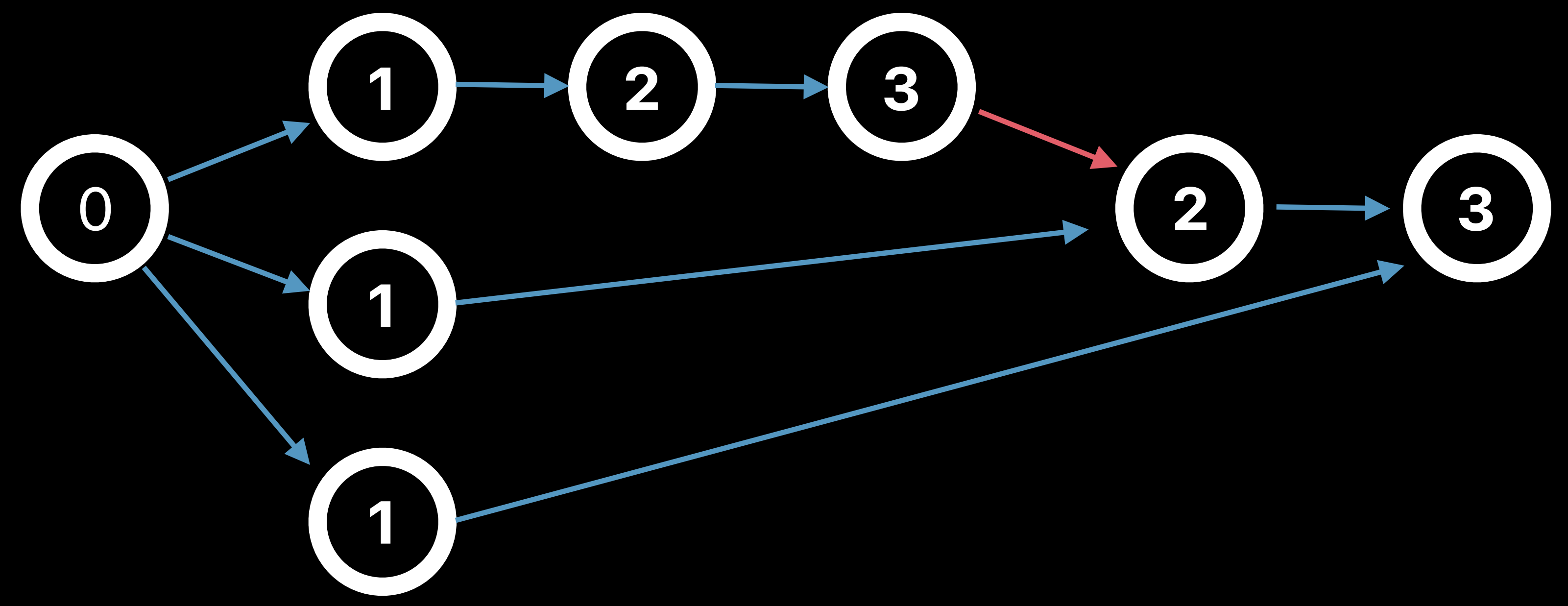
Lattice Self-Attention: Positional Representation

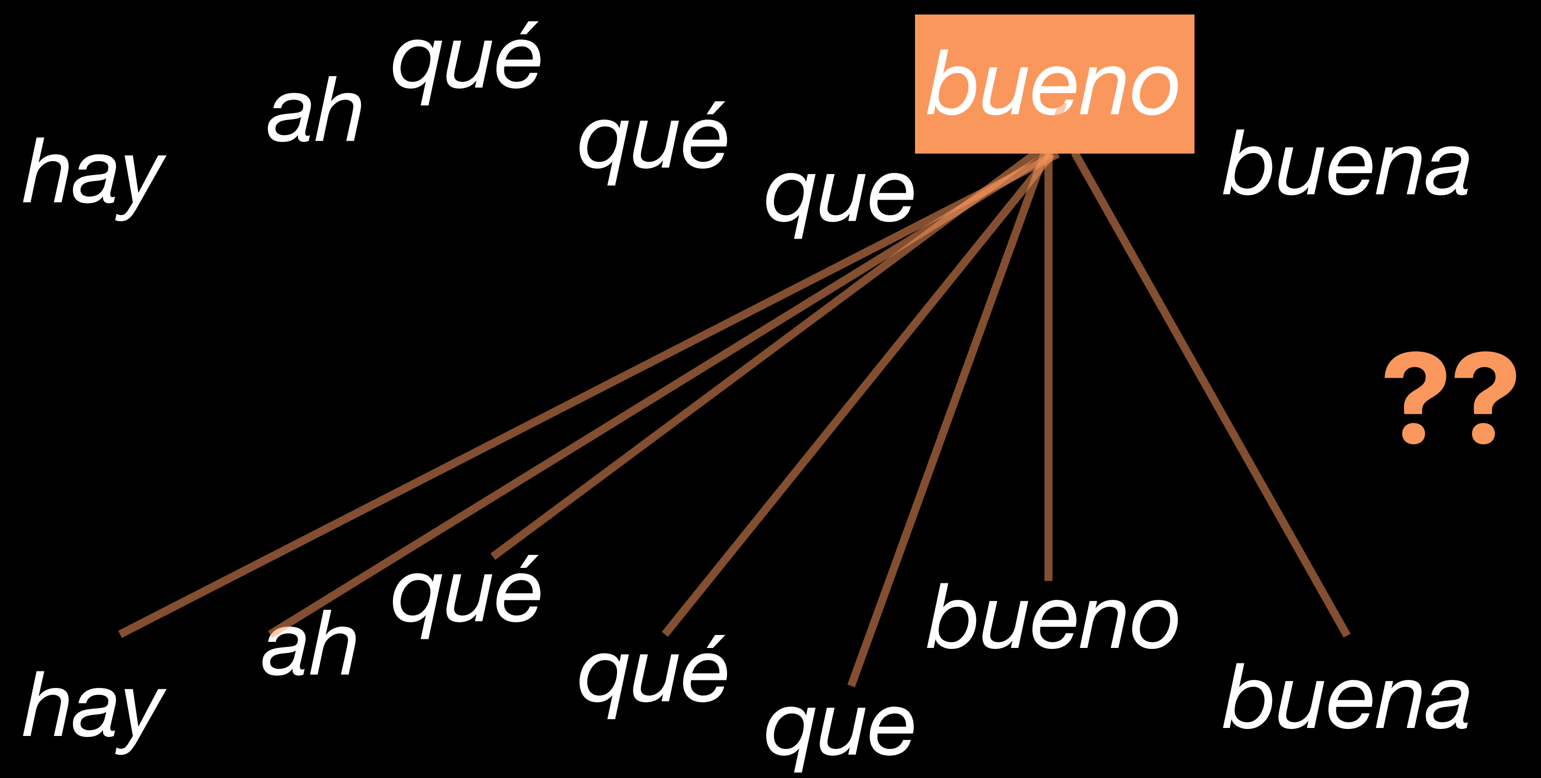
[Sperber+2019]

Longest distance



Shortest distance



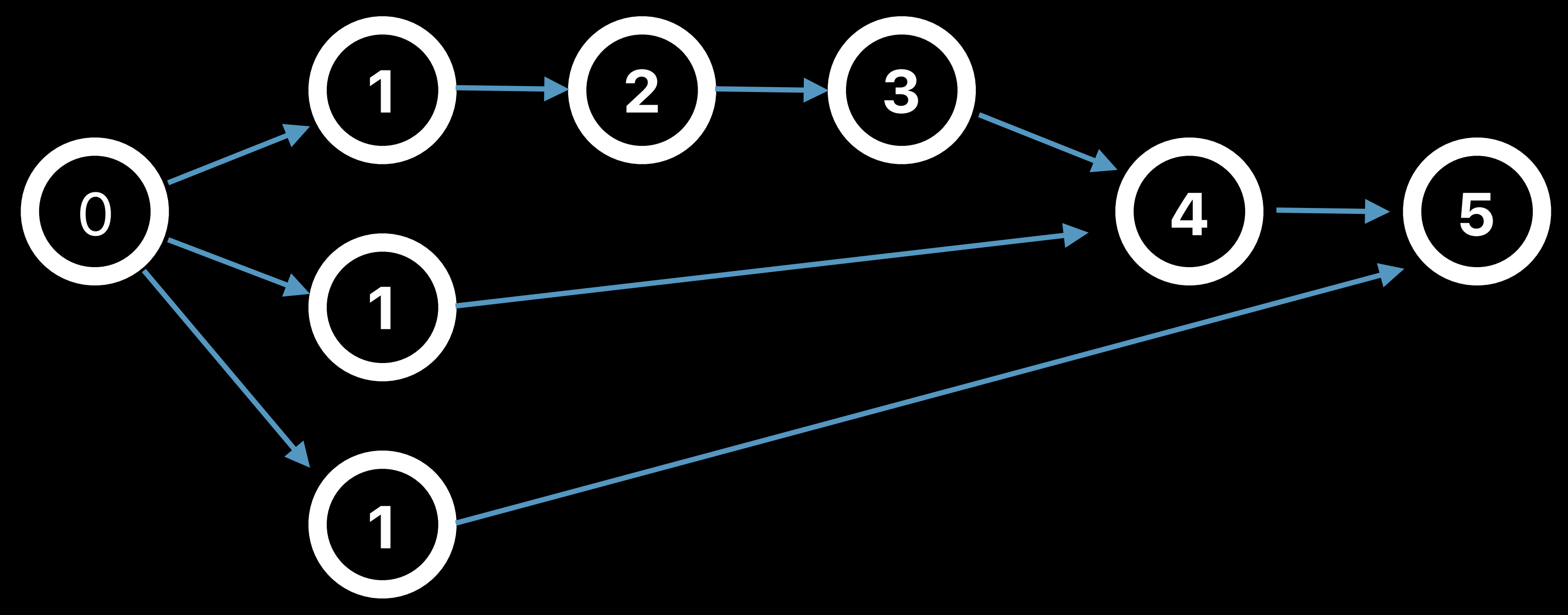


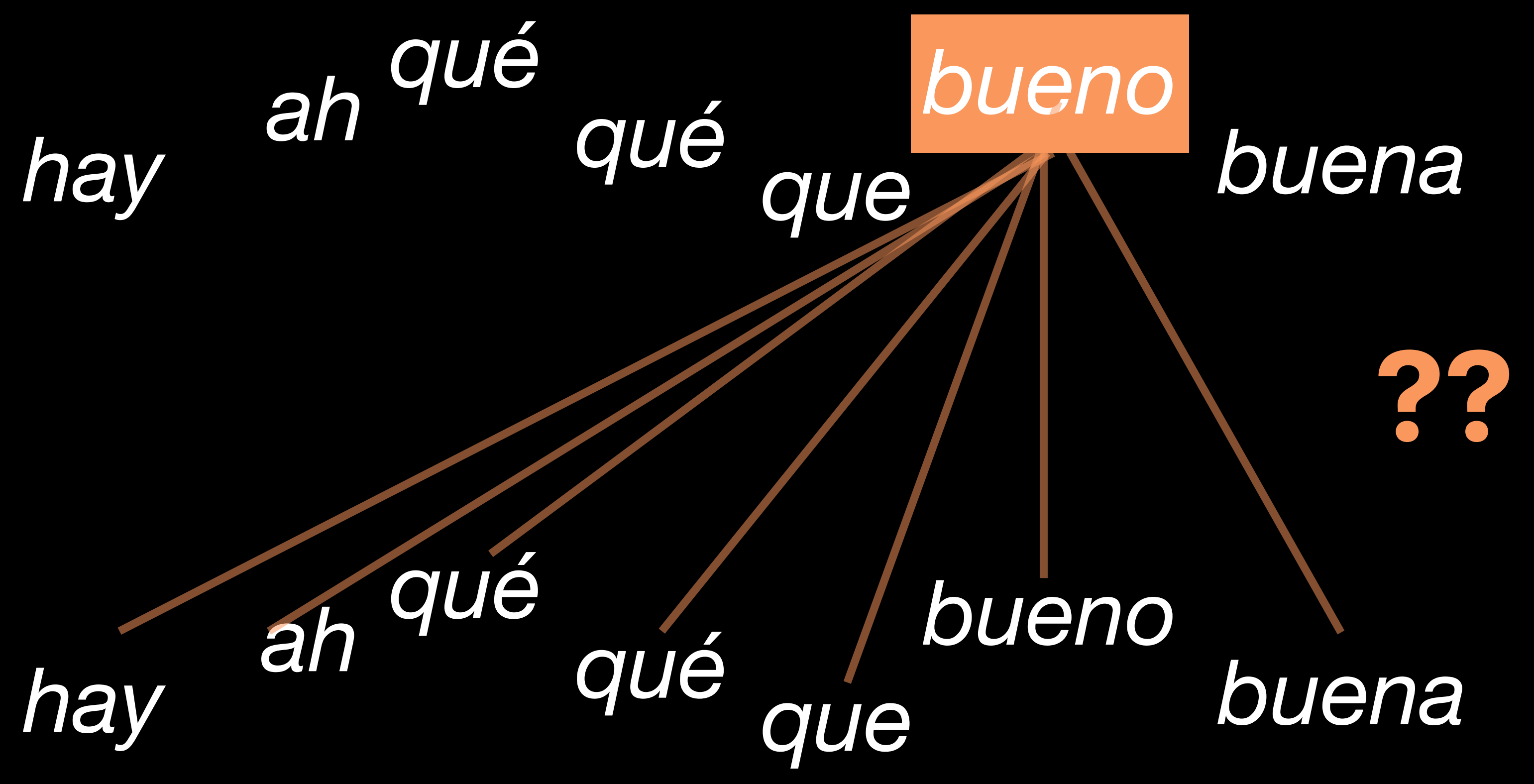
Addressing Error Propagation

Lattice Self-Attention: Positional Representation

[Sperber+2019]

Longest distance





Addressing Error Propagation

Lattice Self-Attention: Reachability Masks

[Sperber+2019]

Addressing Error Propagation

Lattice Self-Attention: Reachability Masks

[Sperber+2019]

query key

$$e_{ij} = f\left(\mathbf{x}_i, \mathbf{x}_j\right) + \vec{m}_{ij}$$

$$\alpha_i = \mathrm{softmax}\left(\mathbf{e}_i\right)$$

$$\mathbf{y}_i = \sum_{j=1}^{l} \alpha_{ij}\mathbf{x}_j$$

Addressing Error Propagation

Lattice Self-Attention: Reachability Masks

[Sperber+2019]

query key

$$e_{ij} = f\left(\mathbf{x}_i, \mathbf{x}_j\right) + \vec{m}_{ij}$$

$$\alpha_i = \mathrm{softmax}\left(\mathbf{e}_i\right)$$

$$\mathbf{y}_i = \sum_{j=1}^{l} \alpha_{ij}\mathbf{x}_j$$

- Binary $\vec{m}_{ij} = \begin{cases} 0 & \text{if } j \text{ successor of } i \\ -\infty & \text{else} \end{cases}$

Addressing Error Propagation

Lattice Self-Attention: Reachability Masks

[Sperber+2019]

query key

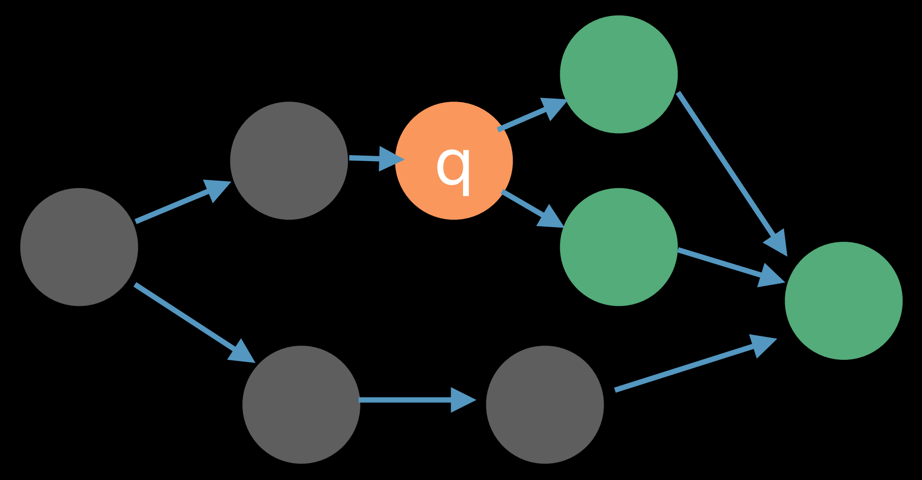
$$e_{ij} = f\left(\mathbf{x}_i, \mathbf{x}_j\right) + \overrightarrow{m}_{ij}$$

$$\alpha_i = \mathrm{softmax}\left(\mathbf{e}_i\right)$$

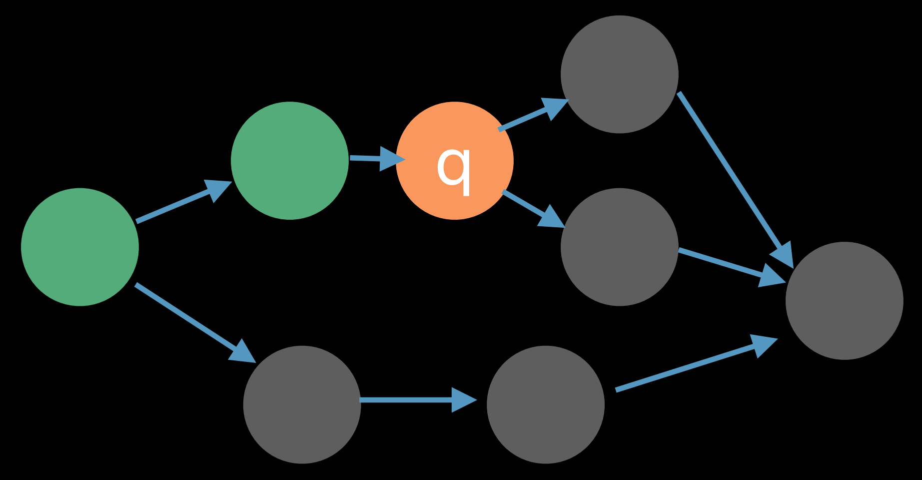
$$\mathbf{y}_i = \sum_{j=1}^{l} \alpha_{ij}\mathbf{x}_j$$

- Binary $\overrightarrow{m}_{ij} = \begin{cases} 0 & \text{if } j \text{ successor of } i \\ -\infty & \text{else} \end{cases}$

$\overrightarrow{m}_{ij}$



$\overleftarrow{m}_{ij}$



Addressing Error Propagation

Lattice Self-Attention: Reachability Masks

[Sperber+2019]

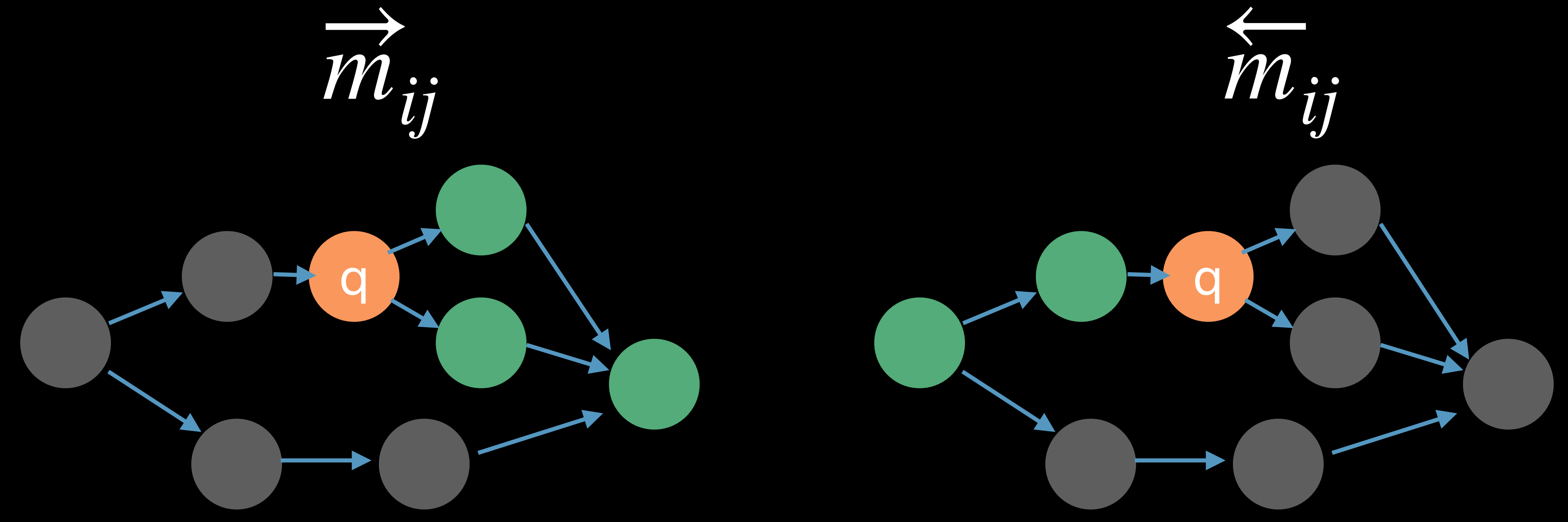
query key

$$e_{ij} = f\left(\mathbf{x}_i, \mathbf{x}_j\right) + \overrightarrow{m}_{ij}$$

$$\alpha_i = \mathrm{softmax}\left(\mathbf{e}_i\right)$$

$$\mathbf{y}_i = \sum_{j=1}^{l} \alpha_{ij} \mathbf{x}_j$$

- Binary $\overrightarrow{m}_{ij} = \begin{cases} 0 & \text{if } j \text{ successor of } i \\ -\infty & \text{else} \end{cases}$



- Probabilistic $\overrightarrow{m}_{ij} = \log P\left(j \textbf{ successor of } i\right)$

Addressing Error Propagation

Lattice-to-Sequence Results

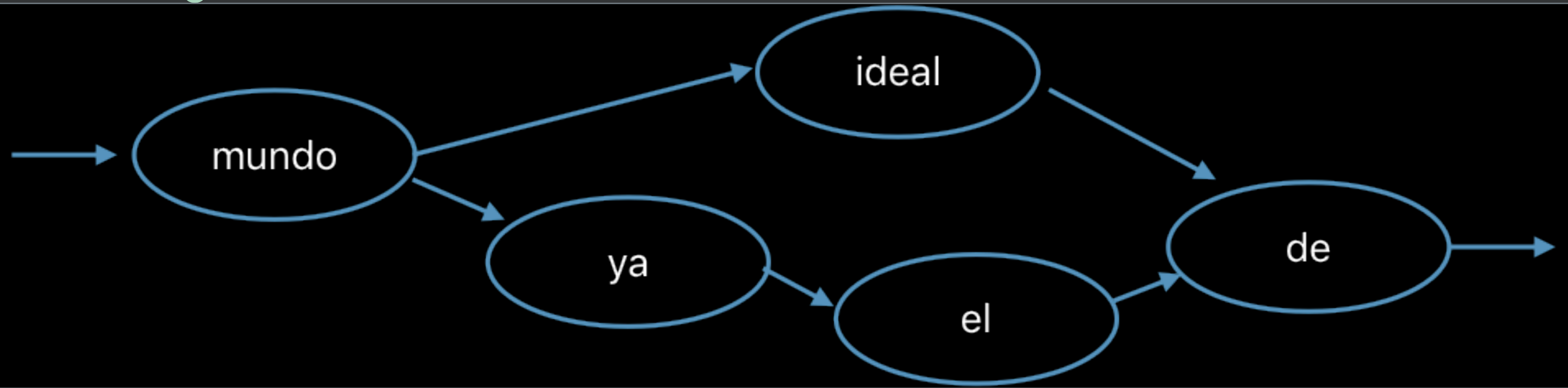
[Sperber+2019]

Encoder model	Inputs	BLEU (Fisher)	BLEU (Callhome)
LSTM	1-best	35.9	11.8
SA (self-attention)	1-best	35.7	12.3
directional SA	1-best	37.4	13.0
SA	linearized lattice (topo.)	30.6	9.4
LatticeLSTM	lattice	38.0	14.1
Lattice SA	lattice	**38.7**	**14.7**

Addressing Error Propagation

Lat2seq example - error in 1best

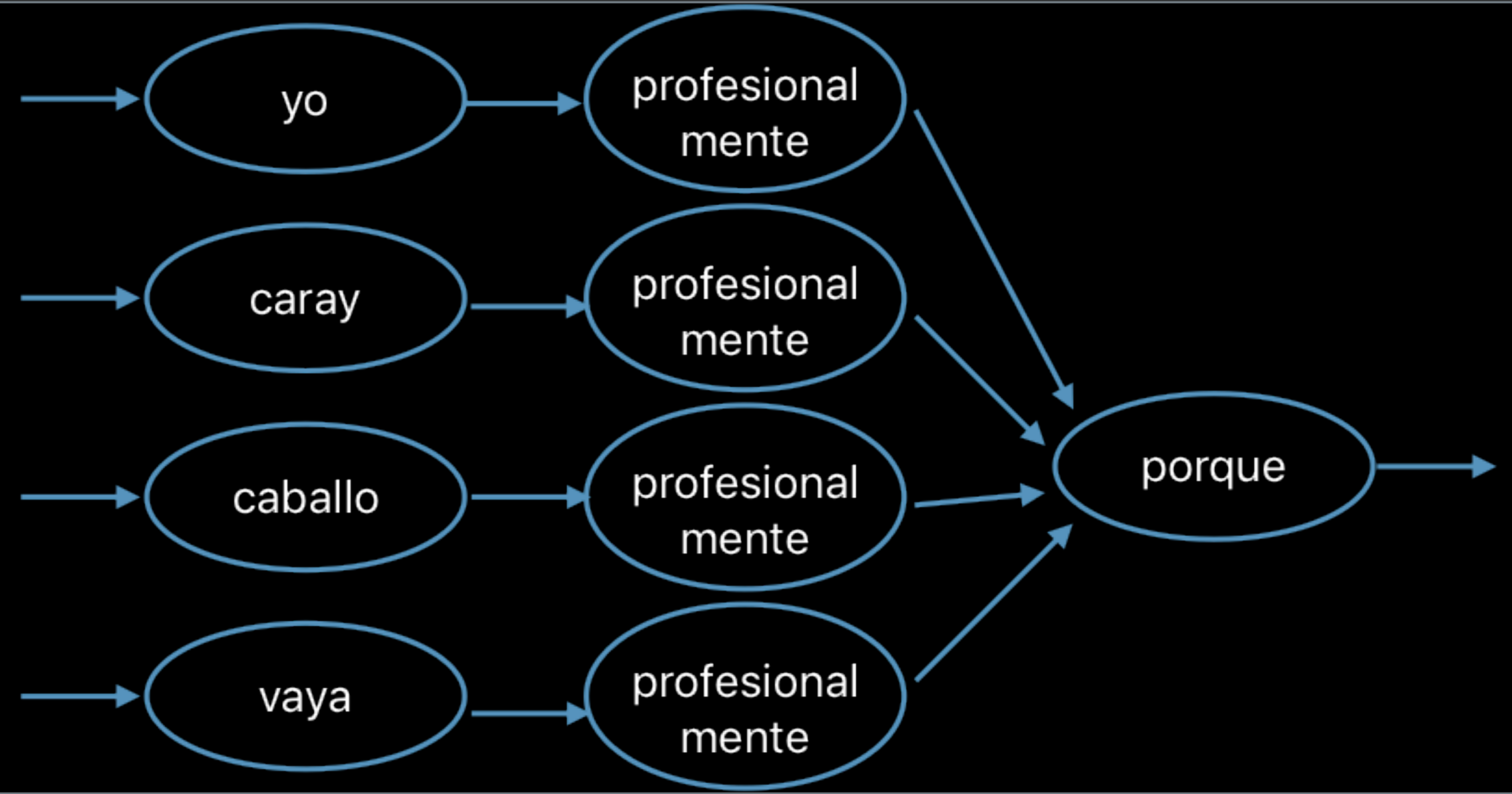
Reference: *and and that is something that i think is counterproductive right because one think that when everything is done one would like maybe a ideal world that one with power everyone will work for the good of everyone*

1-best recognition:	*y y eso es algo que a mi me parece contraproducente verdad porque uno piensa y cuando ya a todos uno quisiera tal vez **un mundo** ya el de que una vez que cadena cuerpos trabajarán por el bienestar de de todos*
Seq2seq output:	*and , and that 's something that seems to me , right ? because one thinks , and when you think , and when everyone would like perhaps **a world** already , the one time that the chain changes for the*
Recognition lattice:	
Lat2seq output:	*and , and that 's something that seems to me , right ? because one thinks , and when you see , when you go to **a ideal world** , you see that they are illegals for the , well , they are all foreigners*

Addressing Error Propagation

Lat2seq example -redundant content

Reference: *the ones who go to have fun for a day those who go because they don ' t have an addiction and they need to play and those who dedicate themselves professionally because there are certain games i think that you hear a game blackjack*

1-best recognition:	*los que van porque que es un día los que van porque no tiene alicia derrita jugar y los que sí caray **profesionalmente** porque hay ciertos counselor bueno creo que soy josé playa que*
Seq2seq output:	*the ones that go , because it 's a day that they go , because they don 't have alicia , play and the ones that are italian , because there are some **<unk>** , well , i think i 'm jose*
Recognition lattice:	
Lat2seq output:	*the ones that go , because it 's a day that they go because you don 't want to play and play , and the ones that influenced **professionally** , because there are certain things , well , i think that i 'm jose*

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- General-purpose regularization

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- General-purpose regularization
  - dropout, word dropout, L2 decay, ...

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- General-purpose regularization
  - dropout, word dropout, L2 decay, …
- Data augmentation/noising

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- General-purpose regularization
  - dropout, word dropout, L2 decay, …
- Data augmentation/noising
  - Idea: introduce “recognition errors” into the MT training data

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- General-purpose regularization
  - dropout, word dropout, L2 decay, …
- Data augmentation/noising
  - Idea: introduce "recognition errors" into the MT training data
  - Models learns how to translate these (ignore errors, or even correct common error patterns)

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- Data augmentation/noising

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- Data augmentation/noising
  - Random insertions/ substitutions/deletions

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- Data augmentation/noising
  - Random insertions/substitutions/deletions
  - Acoustic confusability

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

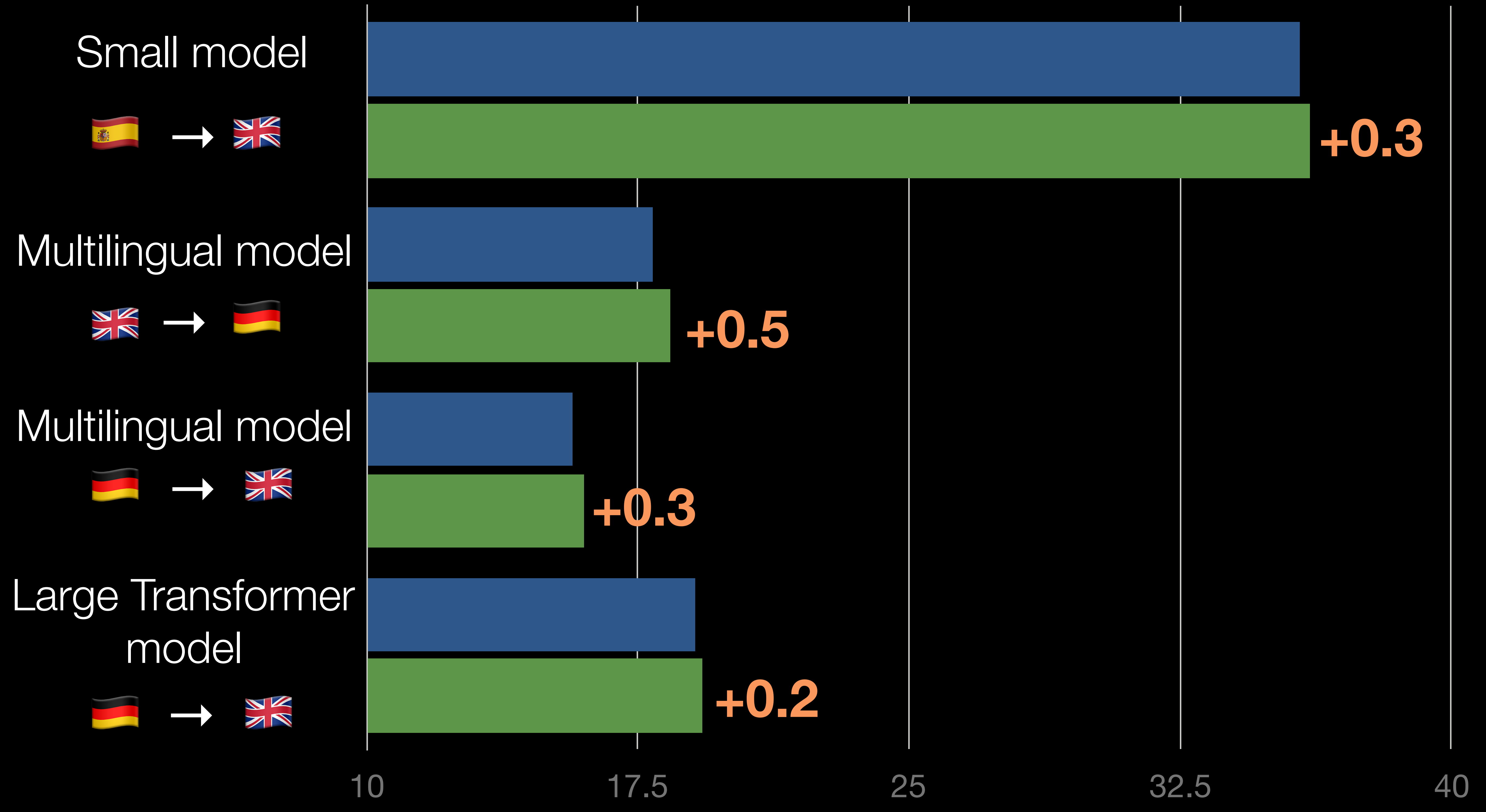
- Data augmentation/noising
  - Random insertions/substitutions/deletions
  - Acoustic confusability
  - Linguistic confusability

Addressing Error Propagation

Robust models

[Tsvetkov+2014; Ruiz+2015; Sperber+2017]

- Data augmentation/noising
  - Random insertions/ substitutions/deletions
  - Acoustic confusability
  - Linguistic confusability



Addressing Domain Mismatch

$$\text{argmax}_T \sum_{S \in \mathscr{H}} Pr(T \mid S)\, Pr(S \mid X)$$

Different assumptions on $Pr(S)$

Addressing Domain Mismatch

$$\text{argmax}_T \sum_{S \in \mathcal{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Different assumptions on $Pr\left(S\right)$

- Spoken vs. written style

Addressing Domain Mismatch

$$\text{argmax}_{T} \sum_{S \in \mathscr{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Different assumptions on $Pr\left(S\right)$

- Spoken vs. written style
- Punctuation

Addressing Domain Mismatch

$$\operatorname{argmax}_T \sum_{S \in \mathcal{H}} Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Different assumptions on $Pr\left(S\right)$

- Spoken vs. written style
- Punctuation
- Capitalisation

Addressing Domain Mismatch

End-to-end corpora

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

- Case study: "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus"

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

- Case study: “Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus”
  - Starting point: conversational ASR corpus 🇪🇸📞

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

- Case study: “Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus”
  - Starting point: conversational ASR corpus 🇪🇸📞
  - Crowd-source translations

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

- Case study: “Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus”
  - Starting point: conversational ASR corpus 🇪🇸📞
  - Crowd-source translations
    - $16k for 193 hours / 170k utterances

Addressing Domain Mismatch

End-to-end corpora

[Post+2013]

- Case study: "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus"
  - Starting point: conversational ASR corpus 🇪🇸📞
  - Crowd-source translations
    - $16k for 193 hours / 170k utterances
  - MT trained on this in-domain data much better than MT trained on 20x larger out-of-domain corpus

Interface	Euro	LDC
Transcript	41.8	58.7
1-best	24.3	35.4

Addressing Domain Mismatch

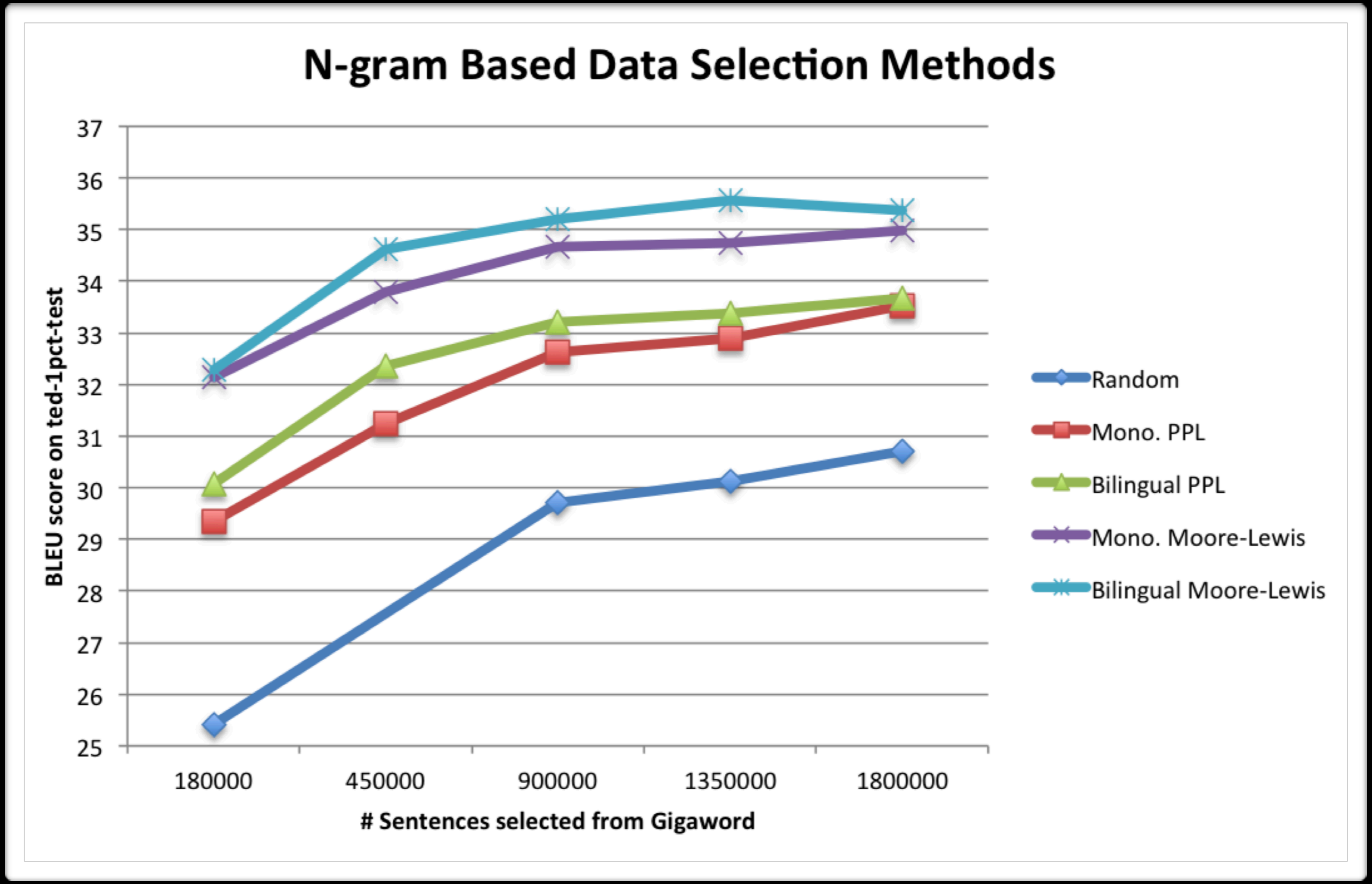
General-purpose domain adaptation

- Common situation:
  - Small amount of in-domain (spoken style) text data
  - Large amount of general-domain MT data
- Data filtering:
  - Select sentences from general-domain data that are "most similar" to the in-domain data

Addressing Domain Mismatch

General-purpose domain adaptation

- Common situation:
  - Small amount of in-domain (spoken style) text data
  - Large amount of general-domain MT data
- Data filtering:
  - Select sentences from general-domain data that are "most similar" to the in-domain data



[Axelrod, 2014]

Addressing Domain Mismatch

Segmentation

Raw ASR

we see here is an example from the european parliament the european parliament twenty languages

and you try simultaneously by help human translator translators the

talk to each of the speaker in other languages to translate it is possible to build computers

the similar to provide translation services

Segmented text

- Sent. boundaries
- Punctuation
- Capitalization
- Number format

We see here is an example from the European Parliament.

The European Parliament 20 languages are spoken, and you try by help human translator to translate simultaneously translators the speeches of the speaker in each case in other languages.

It is possible to build computers that are similar to provide translation services?

Addressing Domain Mismatch
Disfluencies

- Disfluency removal is hard:

  - Highly context dependent
  - Almost no training data

Addressing Domain Mismatch
Disfluencies

- Disfluency removal is hard:

Hesitation	*eh, eh, eh, um, yo pienso que es así.* *uh, uh, uh, um, i think it's like that.*
Repetition	*Y, y no cree que, que, que,* *And, and I don't believe that, that, that*
Correction	*no, no puede, no puedo irme para …* *no, it cannot, I cannot go there …*
False start	*porque qué va, mja ya te acuerda que …* *because what is, mhm do you recall now that …*

[Salesky+2018]

  - Highly context dependent
  - Almost no training data

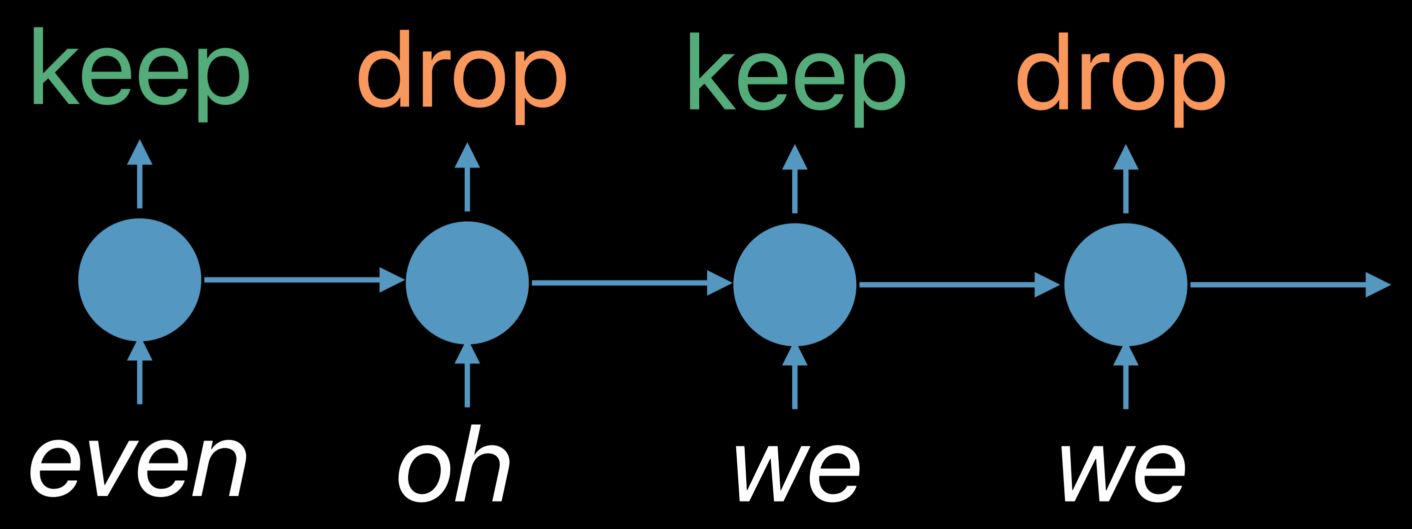
Addressing Domain Mismatch

Disfluencies

Addressing Domain Mismatch

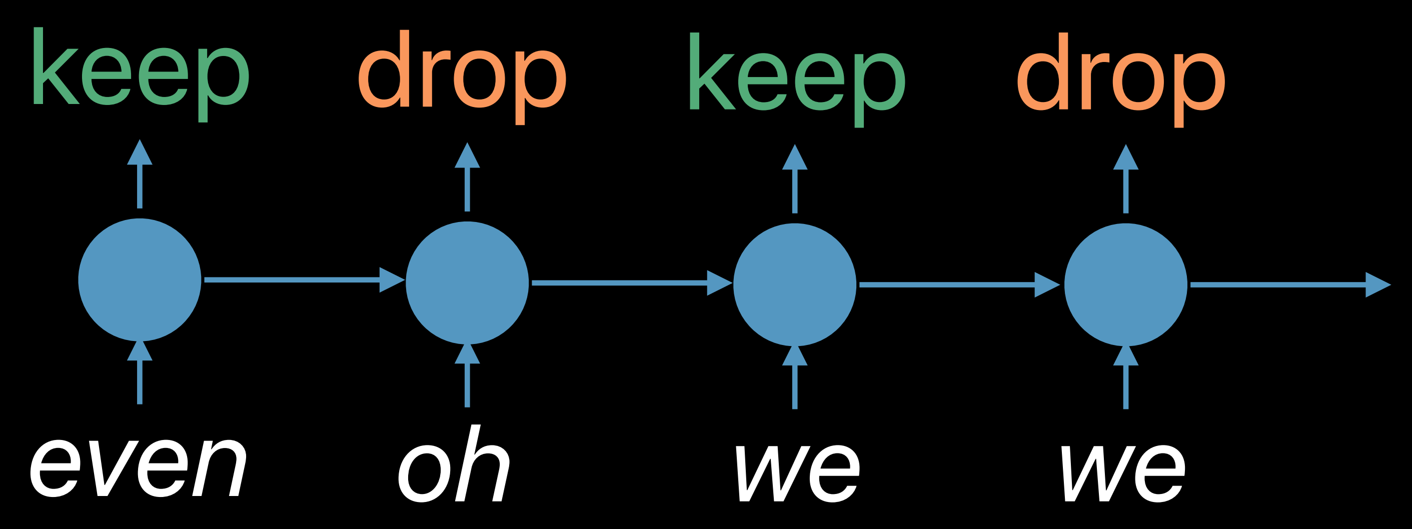
Disfluencies

- Disfluency as preprocessing



Addressing Domain Mismatch
Disfluencies

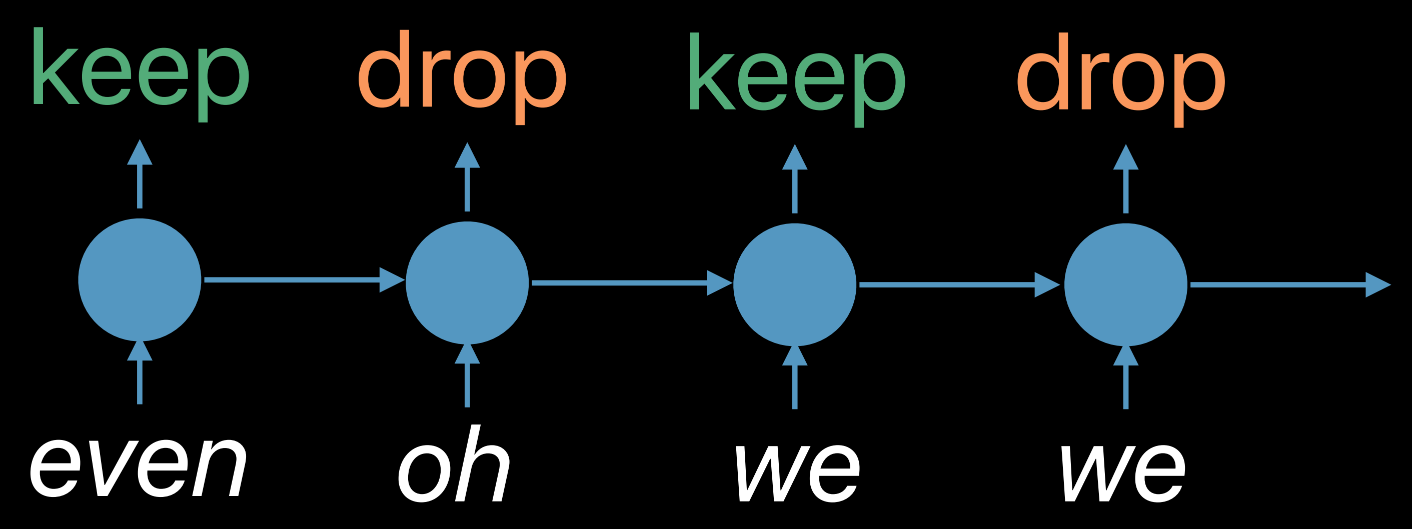
- Disfluency as preprocessing



- Joint translation and disfluency removal

Addressing Domain Mismatch
Disfluencies

- Disfluency as preprocessing

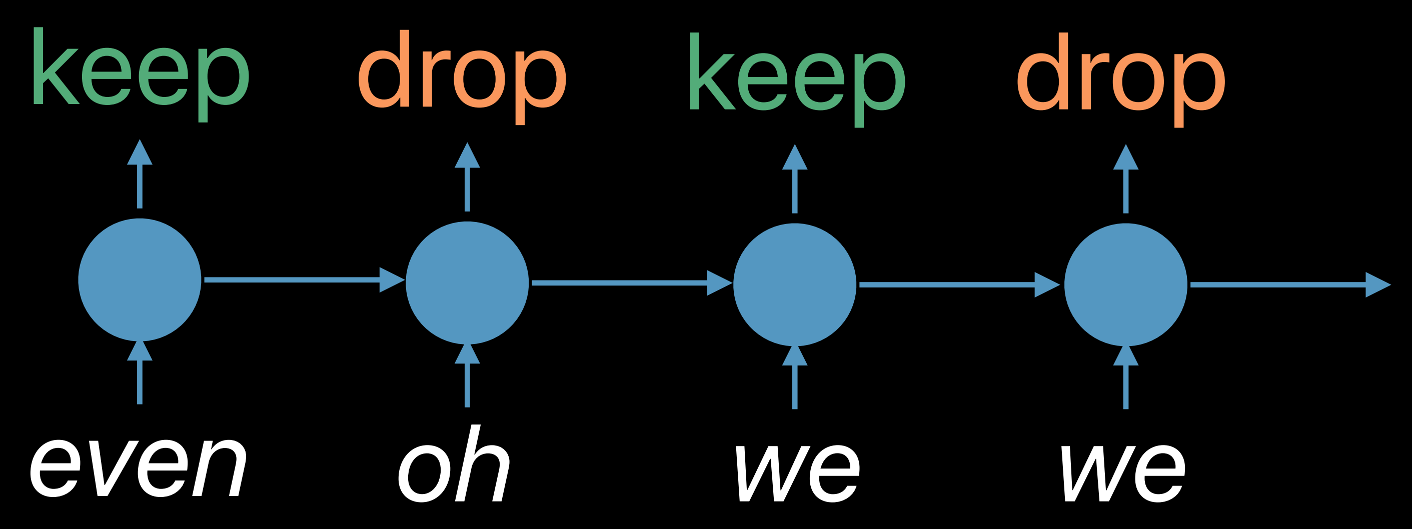


- Joint translation and disfluency removal
  - Train on disfluent source text → fluent target text

Addressing Domain Mismatch

Disfluencies

- Disfluency as preprocessing



- Joint translation and disfluency removal
  - Train on disfluent source text → fluent target text

SRC	*también tengo um eh estoy tomando una clase ...*
REF	*i also have um eh im taking a marketing class ...*
NMT	*im taking a class of marketing*

[Salesky+2019]

Addressing Information Loss

Addressing Information Loss

$$\operatorname{argmax}_T \sum_S Pr(T \mid S, X)\, Pr(S \mid X)$$

Addressing Information Loss

$$\operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Addressing Information Loss

$$\operatorname{argmax}_T \sum_S Pr\left(T \mid S, X\right) Pr\left(S \mid X\right)$$

$$\approx \operatorname{argmax}_T \sum_S Pr\left(T \mid S\right) Pr\left(S \mid X\right)$$

Assume cond. independence: $\left(T \perp\!\!\!\perp X\right) \mid S$

Addressing Information Loss

Prosody

- Speech ≈ phones + prosody ≈ verbal + non-verbal
- Prosody features:
  - Rhythm (time)
  - Melody (pitch)
  - Dynamics (energy)

Addressing Information Loss

Prosody

Addressing Information Loss

Prosody

- Functions:

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation

“THIS is my niece, Lucy.”

“THIS is my NIECE, LUCY.”

Addressing Information Loss
Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information

“I’ve lost an umBRELLa”
“a LAdy’s umbrella?”
“Yes, with STARS on it. GREEN stars.”

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information
    - Emphatic stress

"I'm NEVer eating clams again"

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information
    - Emphatic stress
    - Contrastive

“is this a LOW or a HIGH impact aerobics class”?

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information
    - Emphatic stress
    - Contrastive
  - Discourse

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information
    - Emphatic stress
    - Contrastive
  - Discourse
    - Speech act

Statement / question / acknowledgment / appreciation / agreement / abandonment / ...

Addressing Information Loss

Prosody

- Functions:
  - Distinctive: semantic disambiguation
  - Prominence
    - New information
    - Emphatic stress
    - Contrastive
  - Discourse
    - Speech act
  - ...

Statement / question / acknowledgment / appreciation / agreement / abandonment / ...

Addressing Information Loss

Prosody-aware translation

Addressing Information Loss

Prosody-aware translation

- The alignment approach

Addressing Information Loss

Prosody-aware translation

- The alignment approach
  - assume prosody does not change surface form

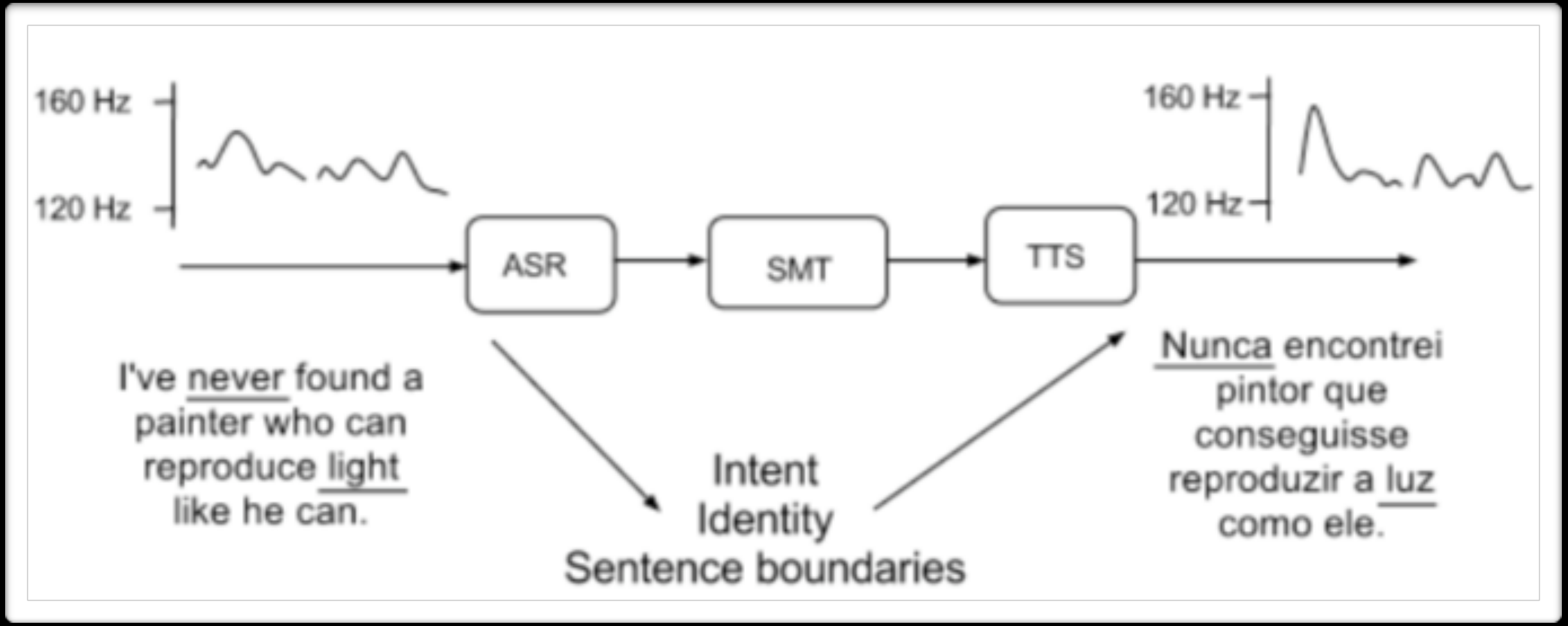
Addressing Information Loss
Prosody-aware translation

- The alignment approach
  - assume prosody does not change surface form
  - transfer prosody to aligned target words

Addressing Information Loss

Prosody-aware translation

- The alignment approach
  - assume prosody does not change surface form
  - transfer prosody to aligned target words

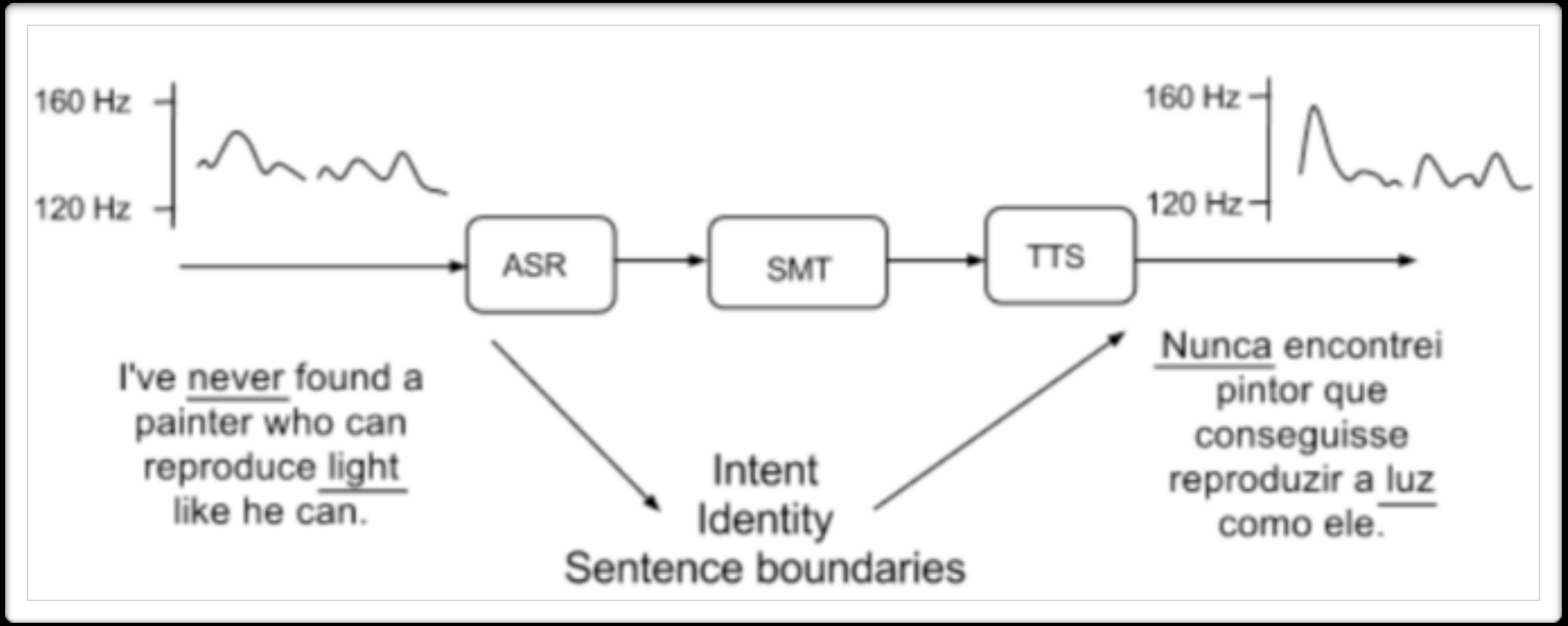


[Anumanchipalli et al., 2012]

Addressing Information Loss

Prosody-aware translation

- The alignment approach
  - assume prosody does not change surface form
  - transfer prosody to aligned target words
- Problem: works only for closely related languages, and not for text outputs



[Anumanchipalli et al., 2012]

Addressing Information Loss

Prosody-aware translation

Addressing Information Loss

Prosody-aware translation

- The markup approach

Addressing Information Loss

Prosody-aware translation

- The markup approach

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

Addressing Information Loss

Prosody-aware translation

- The markup approach
- Observations:

*this is my *niece , *lucy	こちら は 姪っ子 の ルーシー です 。
*this is my niece , lucy	ルーシー 、 こちら が 姪っ子 です 。
*this is my *niece , lucy ?	ルーシー 、 こちら が 姪っ子 かな 。

Addressing Information Loss

Prosody-aware translation

- The markup approach
- Observations:
  - English: emphasis required for disambiguation

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

Addressing Information Loss

Prosody-aware translation

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

Addressing Information Loss

Prosody-aware translation

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough
  - Japanese: disambiguation through sentence structure, no emphasis needed

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

Addressing Information Loss

Prosody-aware translation

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough
  - Japanese: disambiguation through sentence structure, no emphasis needed
- Translate (annotated-text → annotated text) MT?

Addressing Information Loss

Prosody-aware translation

*this is my *niece , *lucy	こちら　は　姪っ子　の　ルーシー　です　。
*this is my niece , lucy	ルーシー　、　こちら　が　姪っ子　です　。
*this is my *niece , lucy ?	ルーシー　、　こちら　が　姪っ子　かな　。

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough
  - Japanese: disambiguation through sentence structure, no emphasis needed
- Translate (annotated-text → annotated text) MT?
- Problems:

Addressing Information Loss

Prosody-aware translation

*this is my *niece , *lucy	こちら は 姪っ子 の ルーシー です 。
*this is my niece , lucy	ルーシー 、 こちら が 姪っ子 です 。
*this is my *niece , lucy ?	ルーシー 、 こちら が 姪っ子 かな 。

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough
  - Japanese: disambiguation through sentence structure, no emphasis needed
- Translate (annotated-text → annotated text) MT?
- Problems:
  - No such training data

Addressing Information Loss

Prosody-aware translation

*this is my *niece , *lucy	こちら は 姪っ子 の ルーシー です 。
*this is my niece , lucy	ルーシー 、 こちら が 姪っ子 です 。
*this is my *niece , lucy ?	ルーシー 、 こちら が 姪っ子 かな 。

- The markup approach
- Observations:
  - English: emphasis required for disambiguation
    - Punctuation helps, but not enough
  - Japanese: disambiguation through sentence structure, no emphasis needed
- Translate (annotated-text → annotated text) MT?
- Problems:
  - No such training data
  - Markup does not capture all phenomena

Simultaneous Translation

Simultaneous Translation

European parliament

Simultaneous Translation

European parliament

- 24 official languages, 552 language combinations!

Simultaneous Translation

European parliament

- 24 official languages, 552 language combinations!
- Employs ~800 interpreters

Simultaneous Translation

European parliament

- 24 official languages, 552 language combinations!
- Employs ~800 interpreters
- Active language / passive language / relay / retour

Simultaneous Translation

European parliament

- 24 official languages, 552 language combinations!
- Employs ~800 interpreters
- Active language / passive language / relay / retour
- Including translation: 460 million Euros / year

Simultaneous Translation

Interpreting vs. Translation

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints
- Interpreting (consecutive or simultaneous)

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints
- Interpreting (consecutive or simultaneous)
  - Direct spoken communication between people

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints
- Interpreting (consecutive or simultaneous)
  - Direct spoken communication between people
  - Enable natural communication

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints
- Interpreting (consecutive or simultaneous)
  - Direct spoken communication between people
  - Enable natural communication
  - Real-time constraints

Simultaneous Translation

Interpreting vs. Translation

- Both: carry meaning across languages
- Translation:
  - Offline, access to dictionary & other resources, no hard time constraints
- Interpreting (consecutive or simultaneous)
  - Direct spoken communication between people
  - Enable natural communication
  - Real-time constraints
- Here: “simultaneous translation” = “simultaneous interpretation”

Simultaneous Translation

Humans vs. machines

Simultaneous Translation

Humans vs. machines

- Text translation:

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations
- Interpretation: not the case

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations
- Interpretation: not the case
  - Cognitive limitations

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations
- Interpretation: not the case
  - Cognitive limitations
  - Time pressure

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations
- Interpretation: not the case
  - Cognitive limitations
  - Time pressure
  - Fatigue

Simultaneous Translation

Humans vs. machines

- Text translation:
  - With enough effort, humans can achieve near-perfect translations
- Interpretation: not the case
  - Cognitive limitations
  - Time pressure
  - Fatigue
- Realistic chance to outperform humans in simultaneous translation

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
+ computational overhead
+ network overhead

German	
Gloss	
Translation	

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich
Gloss	I
Translation	I

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde
Gloss	I	a) sign up b) sign off
Translation	I	???

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich
Gloss	I	a) sign up b) sign off	myself
Translation	I	???	

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich	zur
Gloss	I	a) sign up b) sign off	myself	to
Translation	I	???		

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich	zur	Summer
Gloss	I	a) sign up b) sign off	myself	to	summer
Translation	I	???			

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich	zur	Summer	School
Gloss	I	a) sign up b) sign off	myself	to	summer	school
Translation	I	???				

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich	zur	Summer	School	an
Gloss	I	a) sign up b) sign off	myself	to	summer	school	(up)
Translation	I	???					

Simultaneous Translation

Latency vs. accuracy

- Latency = waiting for linguistic context
  + computational overhead
  + network overhead

German	Ich	melde	mich	zur	Summer	School	an	.
Gloss	I	a) sign up b) sign off	myself	to	summer	school	(up)	.
Translation	I	???						

Simultaneous Translation

Strategies

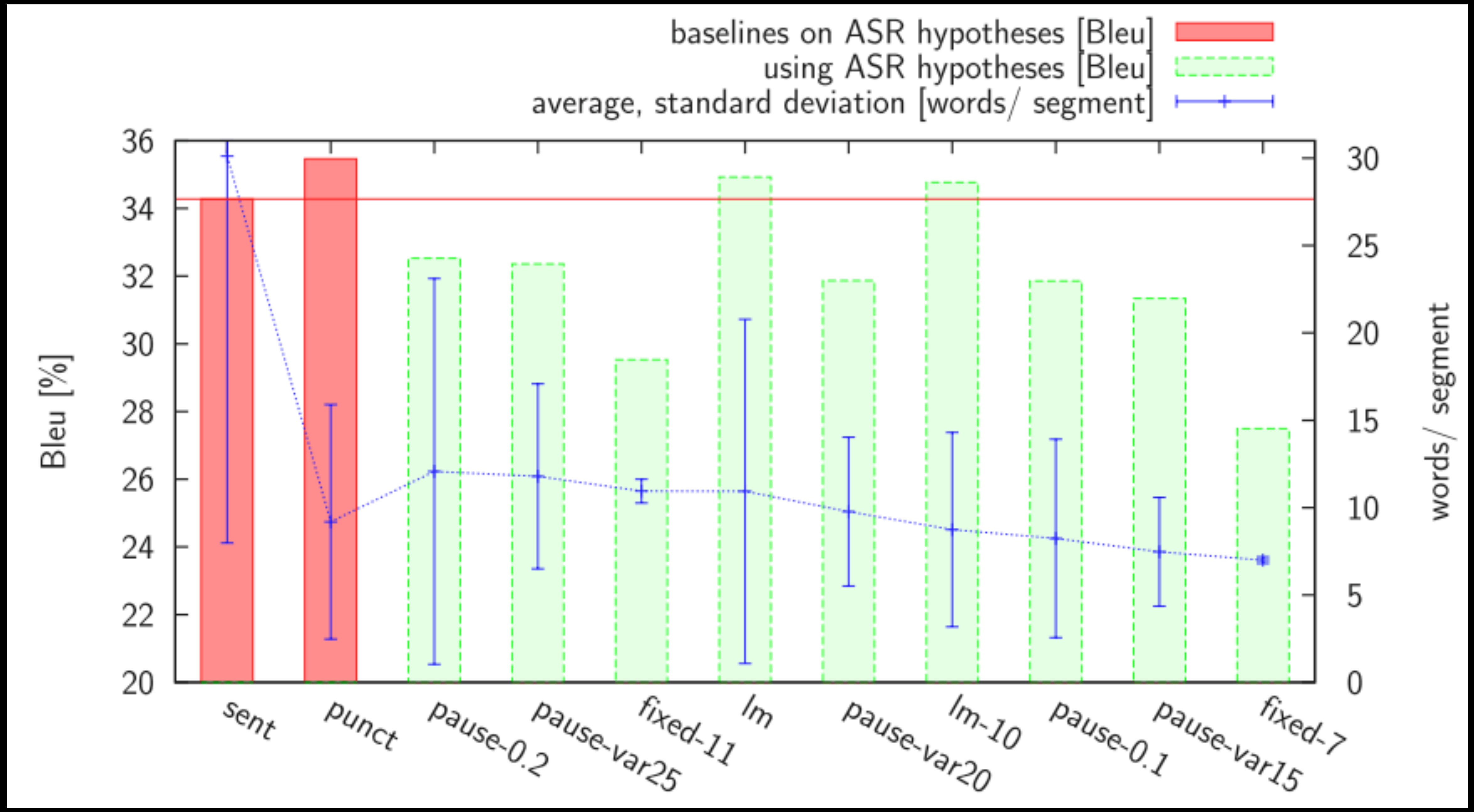
1. Segmented translation
2. Streaming models
3. Translate & revise

Simultaneous Translation

1. Segmented translation

[Fügen 2008]

- Find naturally occurring sentence breaks
  - Prosodic breaks
  - Predict sentence boundaries



Simultaneous Translation

1. Segmented translation

[Oda+2014]

I | ate lunch but | she left

I signed up to | the summer school

Simultaneous Translation

1. Segmented translation

[Oda+2014]

- Find “smallest translatable units”

I | ate lunch but | she left

I signed up to | the summer school

Simultaneous Translation

1. Segmented translation

[Oda+2014]

I | ate lunch but | she left

I signed up to | the summer school

- Find “smallest translatable units”
- Optimization problem:

Simultaneous Translation

1. Segmented translation

[Oda+2014]

I | ate lunch but | she left

I signed up to | the summer school

- Find “smallest translatable units”
- Optimization problem:
  - find segmentation that

Simultaneous Translation

1. Segmented translation

[Oda+2014]

I | ate lunch but | she left

I signed up to | the summer school

- Find "smallest translatable units"
- Optimization problem:
  - find segmentation that
  - maximizes BLEU

Simultaneous Translation

1. Segmented translation

[Oda+2014]

- Find "smallest translatable units"
- Optimization problem:
  - find segmentation that
  - maximizes BLEU
  - at given avg. segment length

I | ate lunch but | she left

I signed up to | the summer school

Simultaneous Translation

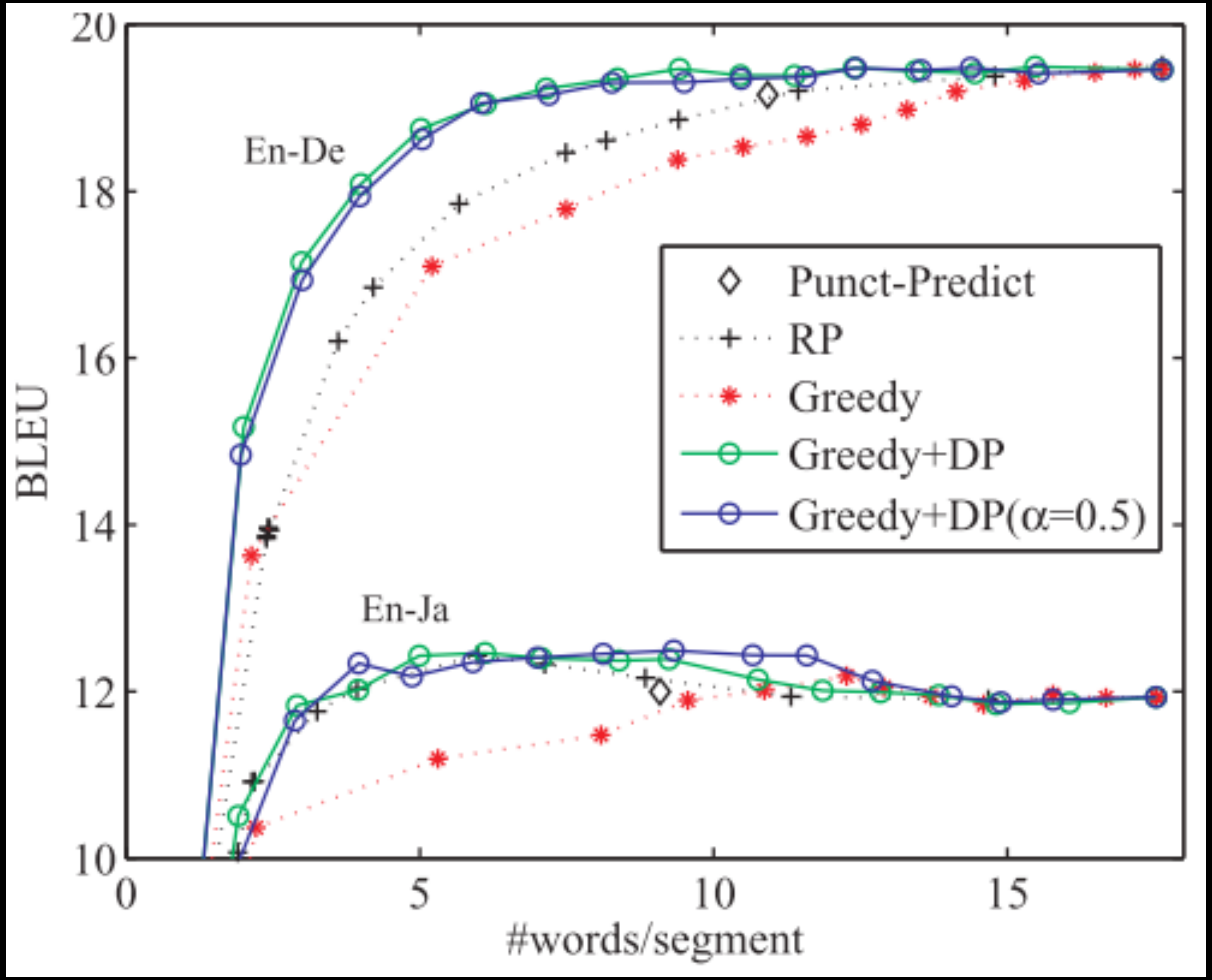
1. Segmented translation

[Oda+2014]

I | ate lunch but | she left

I signed up to | the summer school

- Find "smallest translatable units"
- Optimization problem:
  - find segmentation that
  - maximizes BLEU
  - at given avg. segment length



Simultaneous Translation

2. Streaming models

Simultaneous Translation

2. Streaming models

- MT model takes stream as input

Simultaneous Translation

2. Streaming models

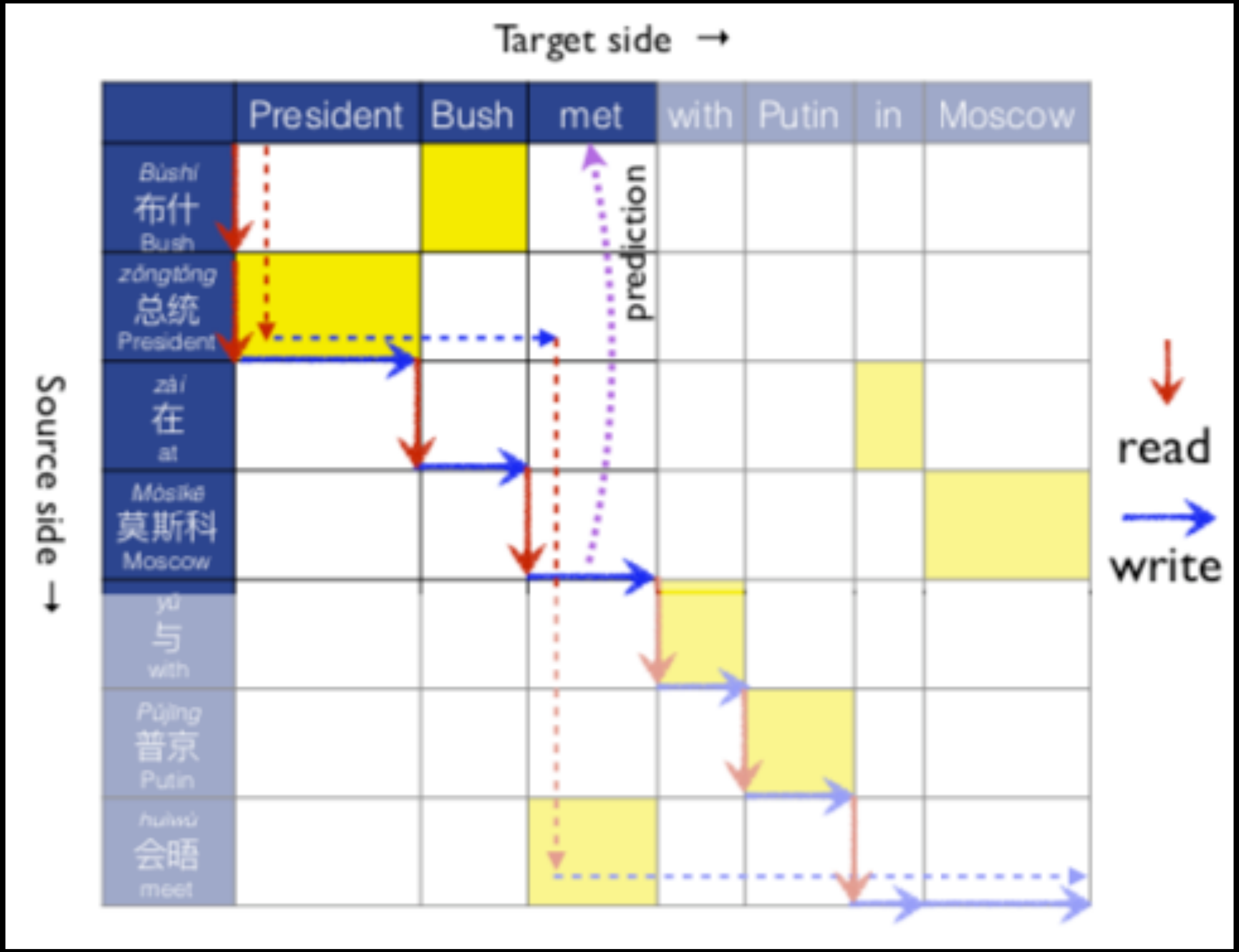
- MT model takes stream as input
- For each incoming word:
  - Do nothing
  - Or, produce one or more output words

Simultaneous Translation

2. Streaming models - Static delay

[Ma+2019]

- "wait-*k*" strategy
- initially, read *k* words
- then: read 1, write 1, ...

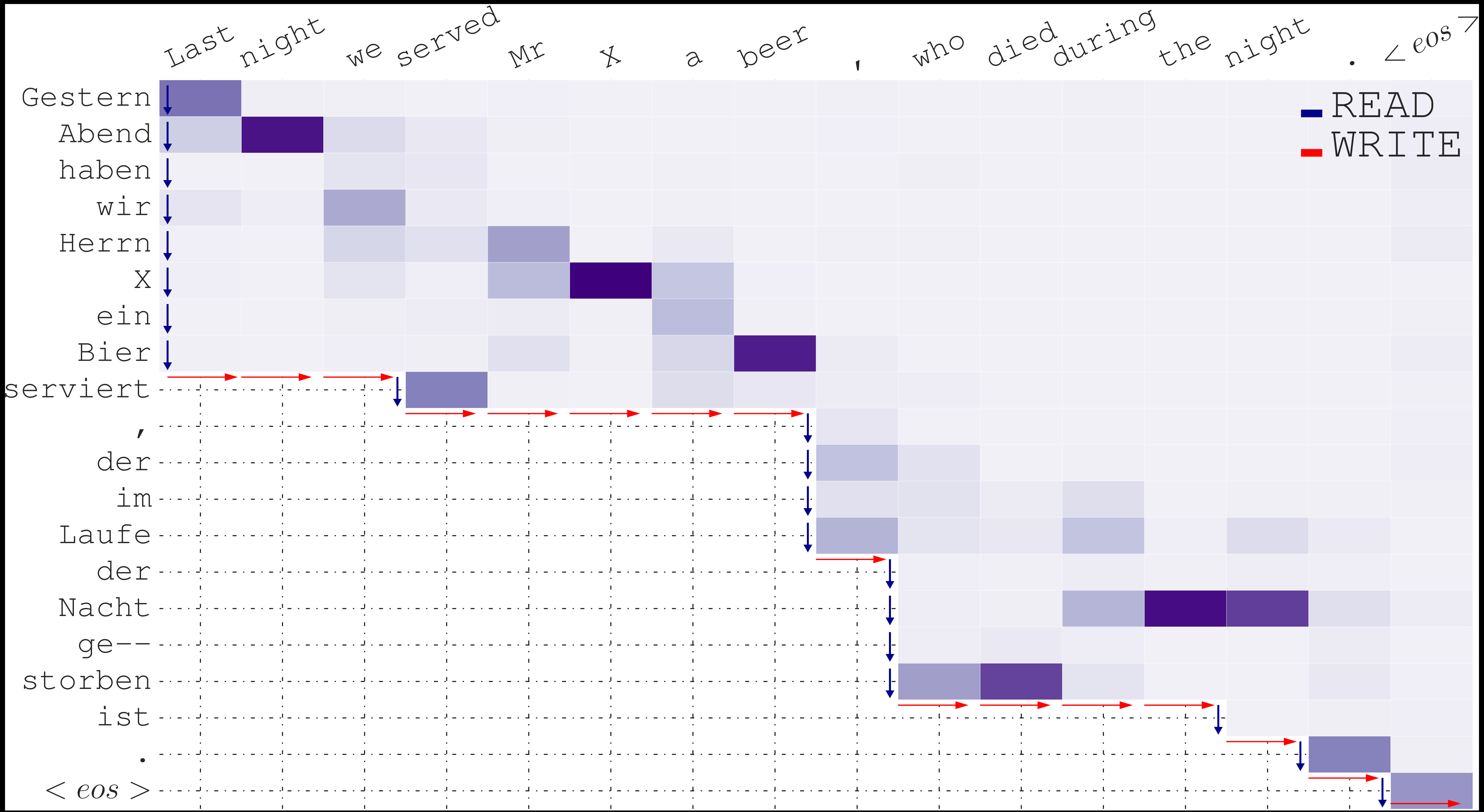


Simultaneous Translation

2. Streaming models - Dynamic delay

[Gu+2017; Xiong+2019; Arivazhagan+2019]

- Delay depending on current context



Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich

I

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Ich melde mich
I sign off

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Ich melde mich
I sign off

Ich melde mich zur
I sign off from

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Ich melde mich
I sign off

Ich melde mich zur
I sign off from

Ich melde mich zur Summerschool
I sign off from summer school

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Ich melde mich
I sign off

Ich melde mich zur
I sign off from

Ich melde mich zur Summerschool
I sign off from summer school

Ich melde mich zur Summerschool an
I sign up for summer school

Simultaneous Translation

3. Translate & revise

[Niehues+2018]

- Translate immediately, revise if necessary
- Usability goal: minimize number of revisions
- Needs appropriate text-based user interface

Ich
I

Ich melde
I notify

Ich melde mich
I sign off

Ich melde mich zur
I sign off from

Ich melde mich zur Summerschool
I sign off from summer school

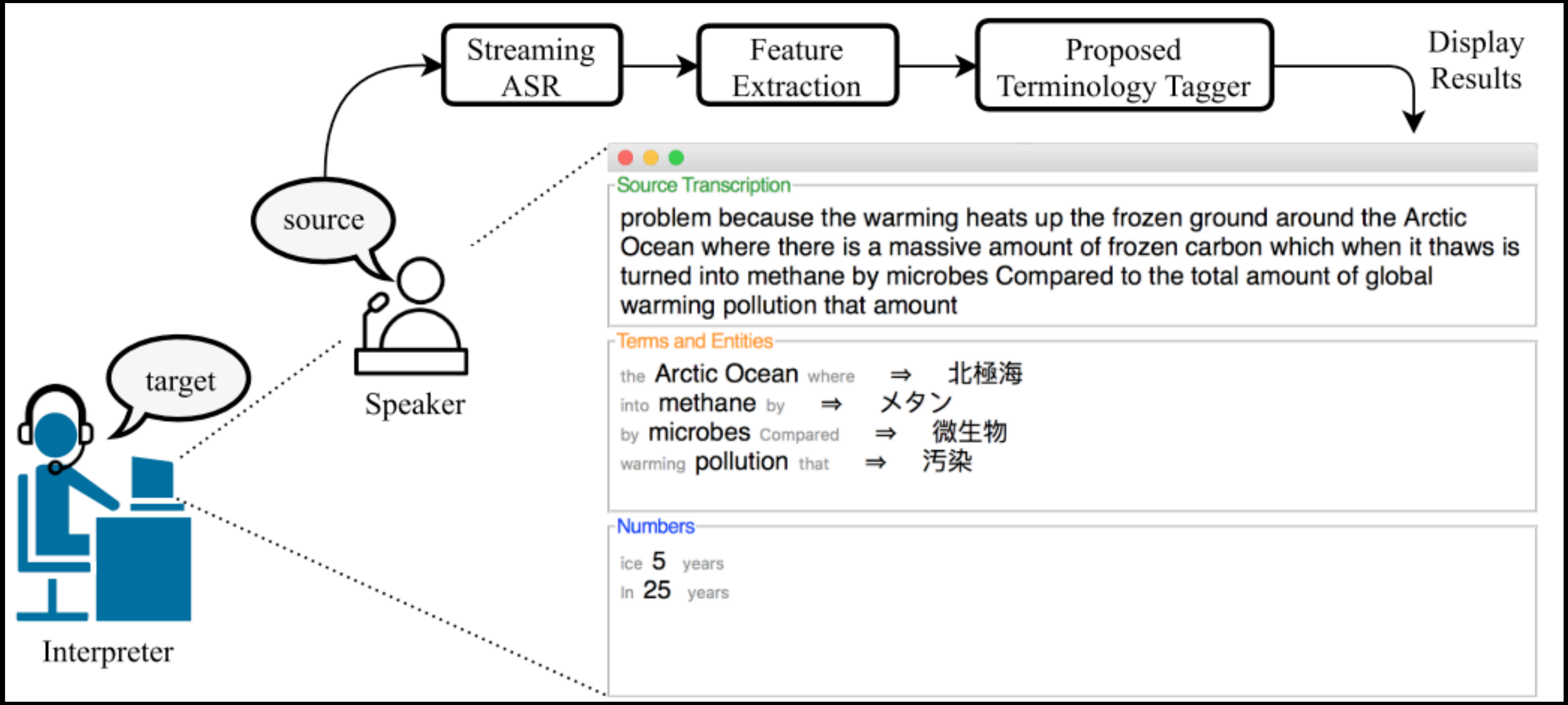
Ich melde mich zur Summerschool an
I sign up for summer school

Simultaneous Translation

Computer-assisted simultaneous translation

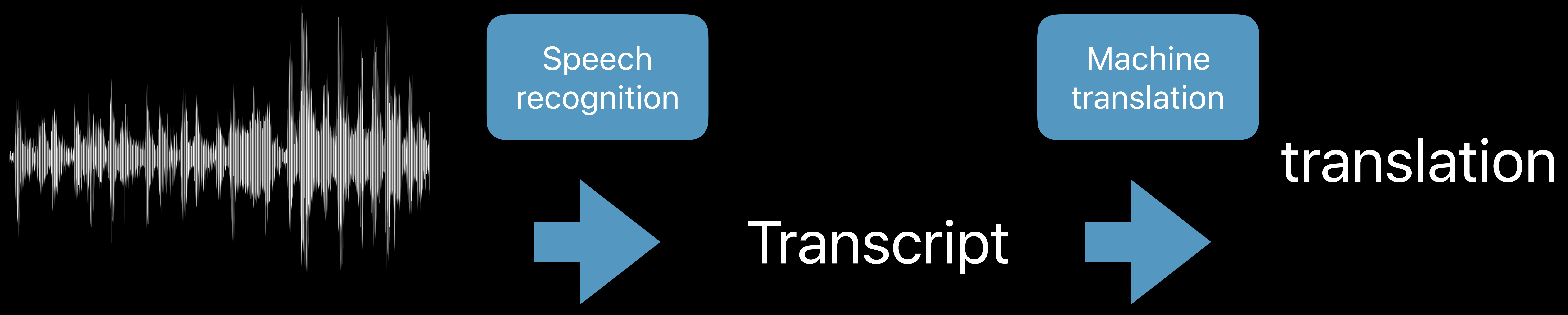
[Vogler+2019]

- Interpreters often work in pairs: One interprets, one writes down dates, lists, names, numbers
- Can we automize the second task?



End-to-end models

Motivation


Speech recognition
Transcript
Machine translation
translation

Motivation

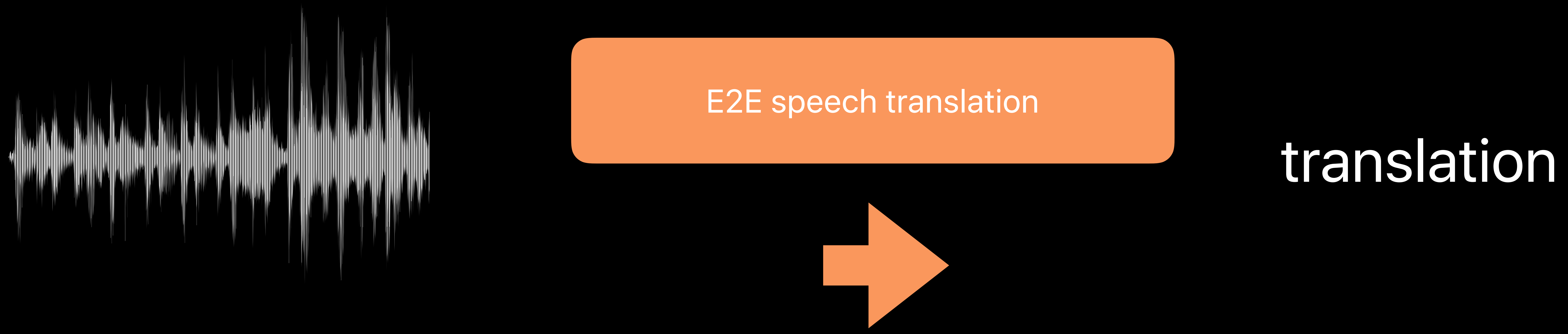
translation

Motivation

E2E speech translation

translation

Motivation



✓ Avoid cascade's problems: *error propagation, ASR/MT data mismatch, information loss*

Motivation

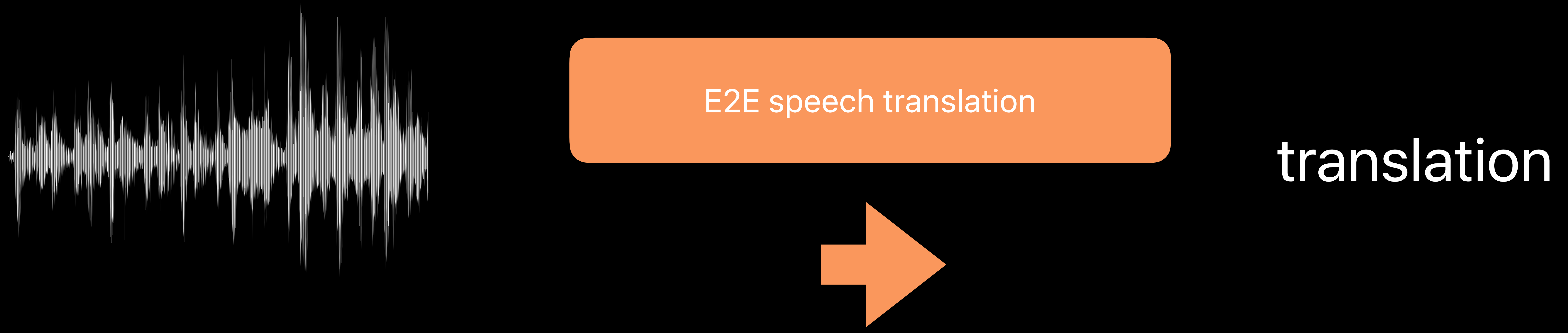
E2E speech translation

translation

✓Simplicity

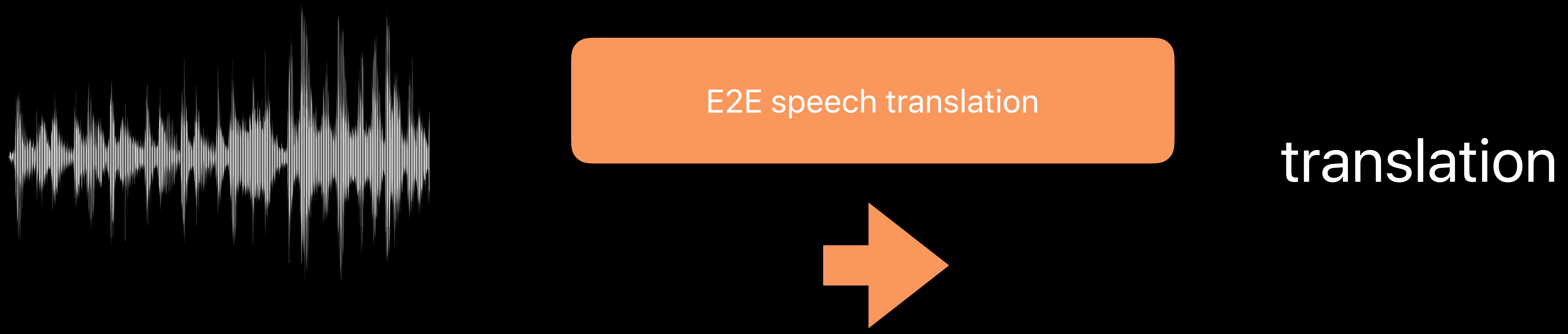
✓Avoid cascade’s problems:
error propagation, ASR/MT data mismatch, information loss

Motivation



- ✓ Avoid cascade's problems: *error propagation, ASR/MT data mismatch, information loss*
- ✓ Simplicity
- ✓ Joint parameter optimisation

Motivation



✓Avoid cascade's problems:
error propagation, ASR/MT data mismatch, information loss

✓Simplicity

✓Joint parameter optimisation

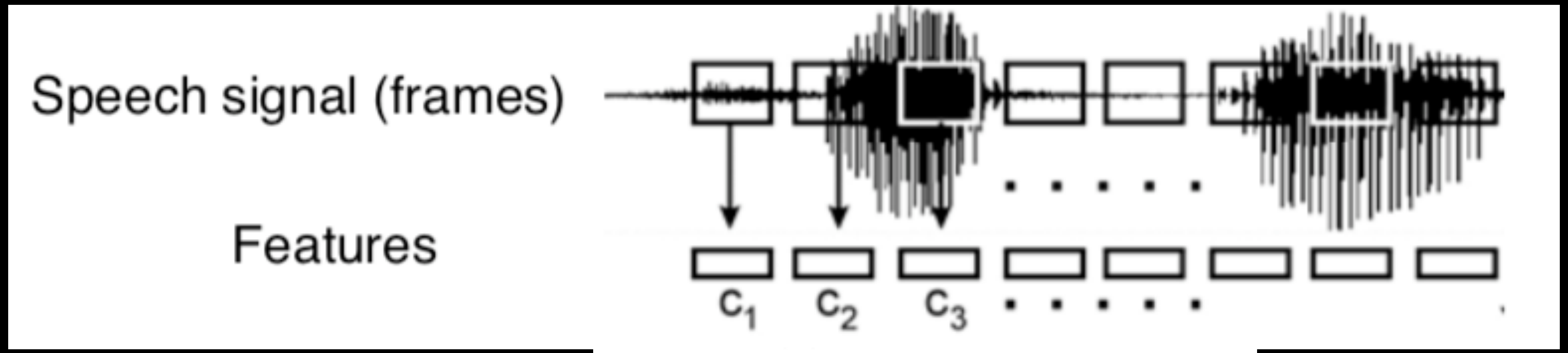
✓Computationally cheaper

End-to-end models

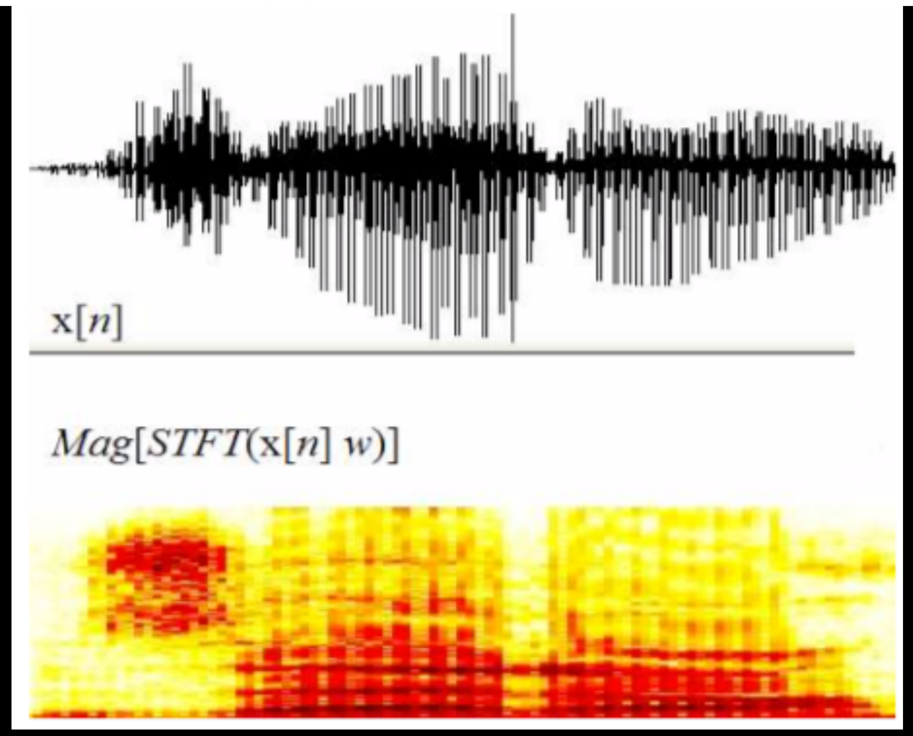
Preliminaries: Listen, attend, and spell

[Chan+2016]

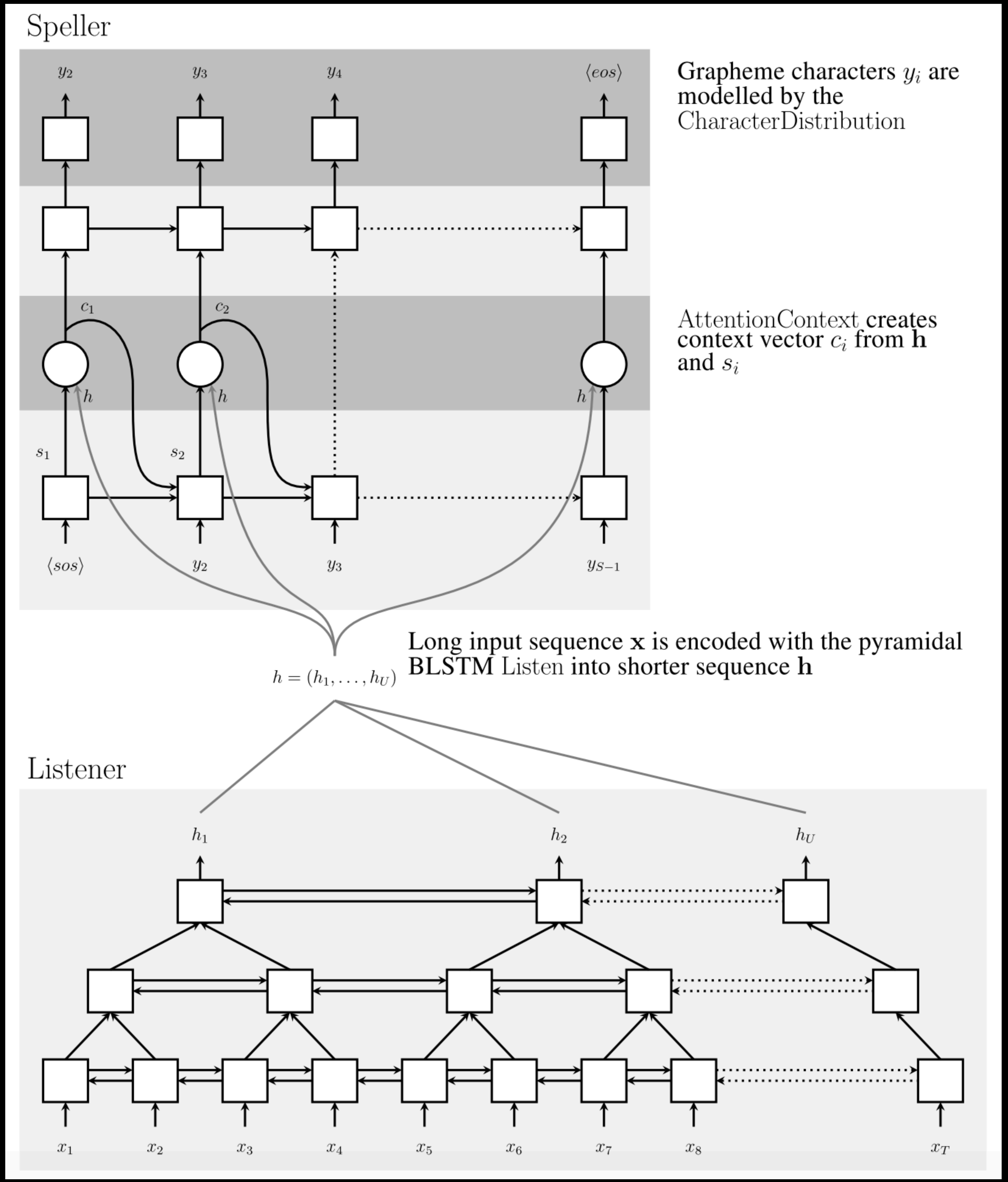
- Sequence-to-sequence models can do speech recognition, too
- Input: feature vectors



[Kasprzak]

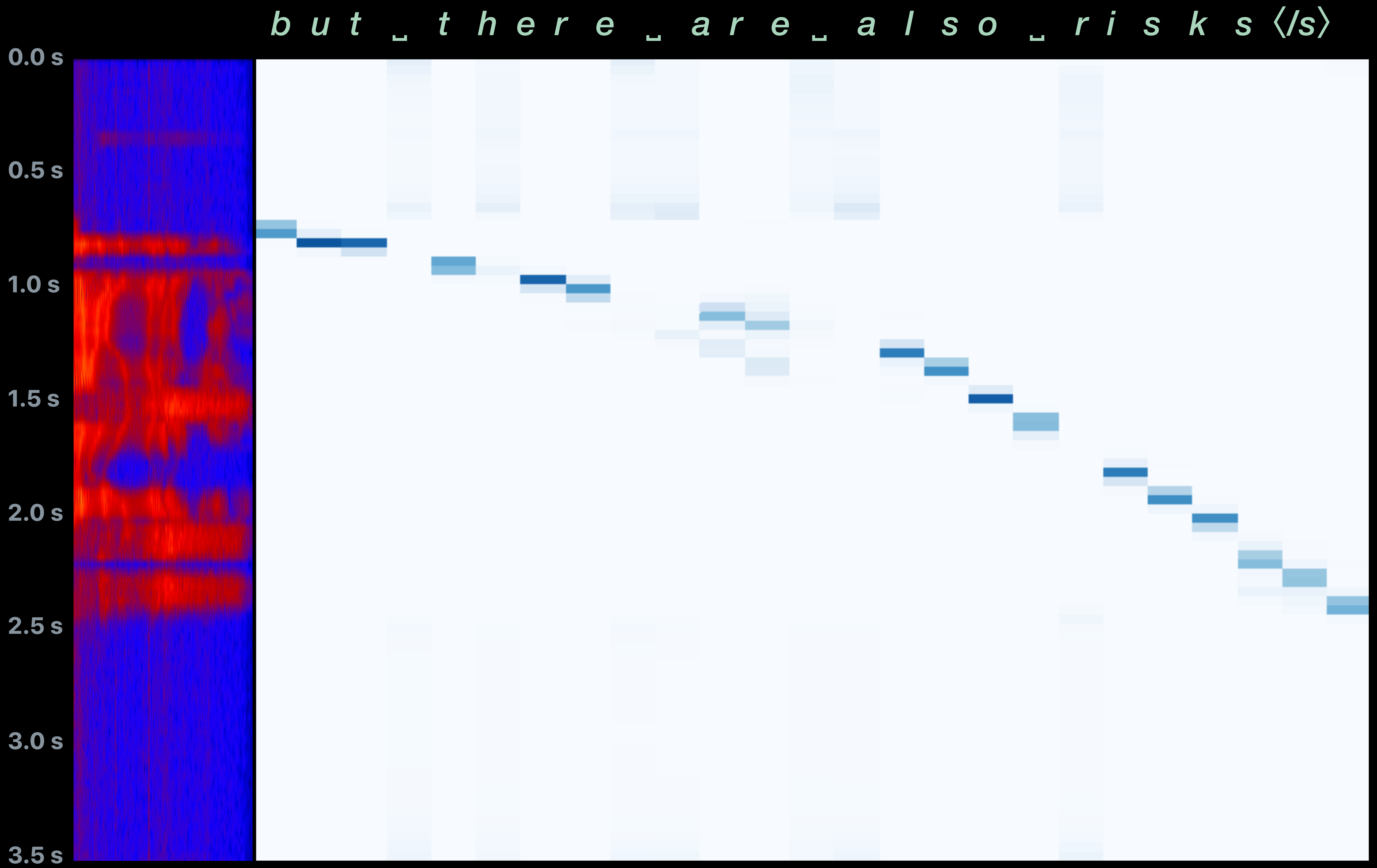


- Output: characters

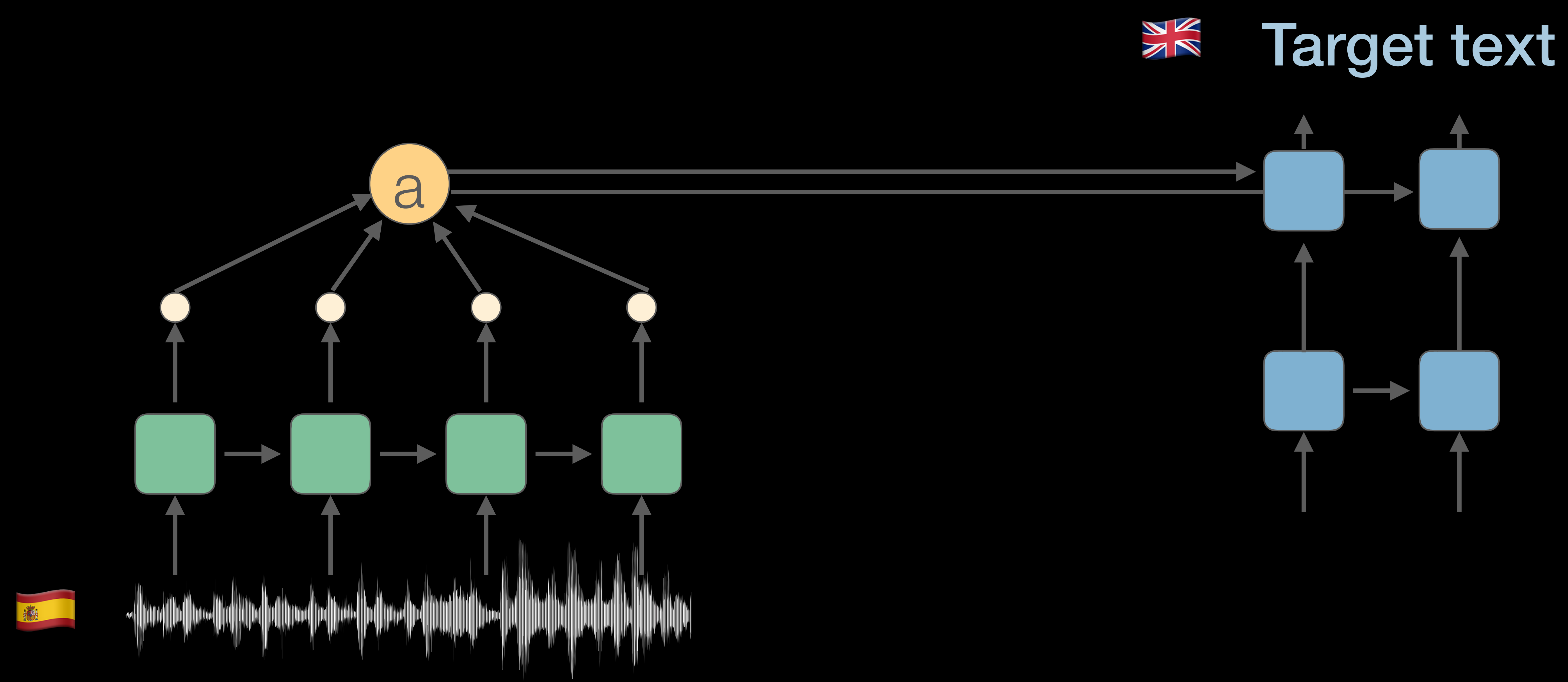


End-to-end models

Preliminaries: Listen, attend, and spell



Direct model



End-to-end models

Data

Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	
Machine translation		✓	✓
End-to-end	✓	(✓)	✓

End-to-end models

Data

Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	
Machine translation		✓	✓
End-to-end	✓	(✓)	✓

Public corpora	Language pairs	Domain	Size
Fisher *[Post+2013]*	🇪🇸 → 🇬🇧	Telephone (strangers)	162h
Callhome *[Post+2013]*	🇪🇸 → 🇬🇧	Telephone (family)	13h

End-to-end models

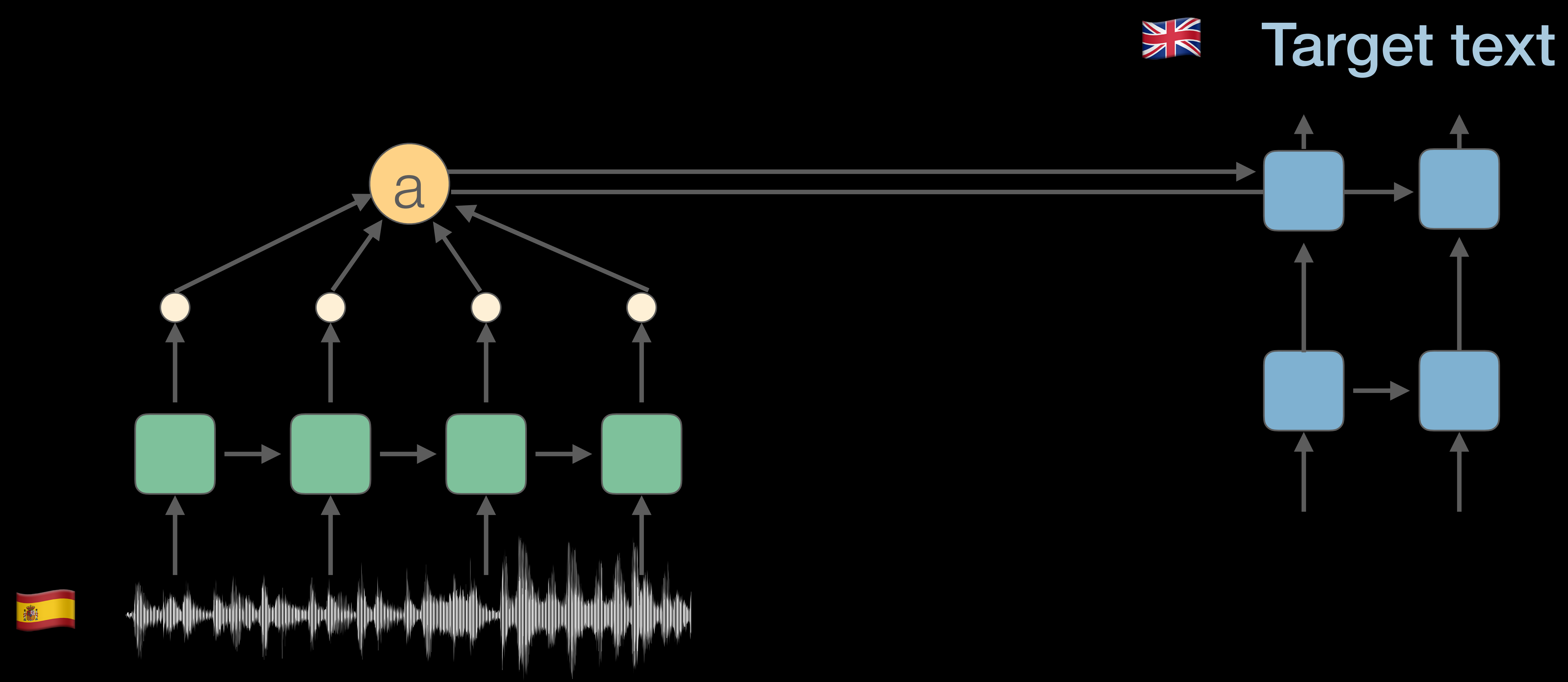
Data

Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	
Machine translation		✓	✓
End-to-end	✓	(✓)	✓

Public corpora	Language pairs	Domain	Size
Fisher *[Post+2013]*	🇪🇸 → 🇬🇧	Telephone (strangers)	162h
Callhome *[Post+2013]*	🇪🇸 → 🇬🇧	Telephone (family)	13h
LibriTrans *[Kocabiyikoglu+2018]*	🇬🇧 → 🇫🇷	Audio books	100h
MuST-C *[Di Gangi+2019]*	🇬🇧 → {🇳🇱,🇫🇷,🇩🇪,🇮🇹,🇵🇹,🇷🇴,🇷🇺,🇪🇸}	TED talks	~400h per language
MaSS *[Boito+2019]*	All directions: {🇬🇧,🇪🇸,🏴,🇫🇮,🇫🇷,🇭🇺,🇷🇴,🇷🇺}	Bible	~20h per language

Direct model

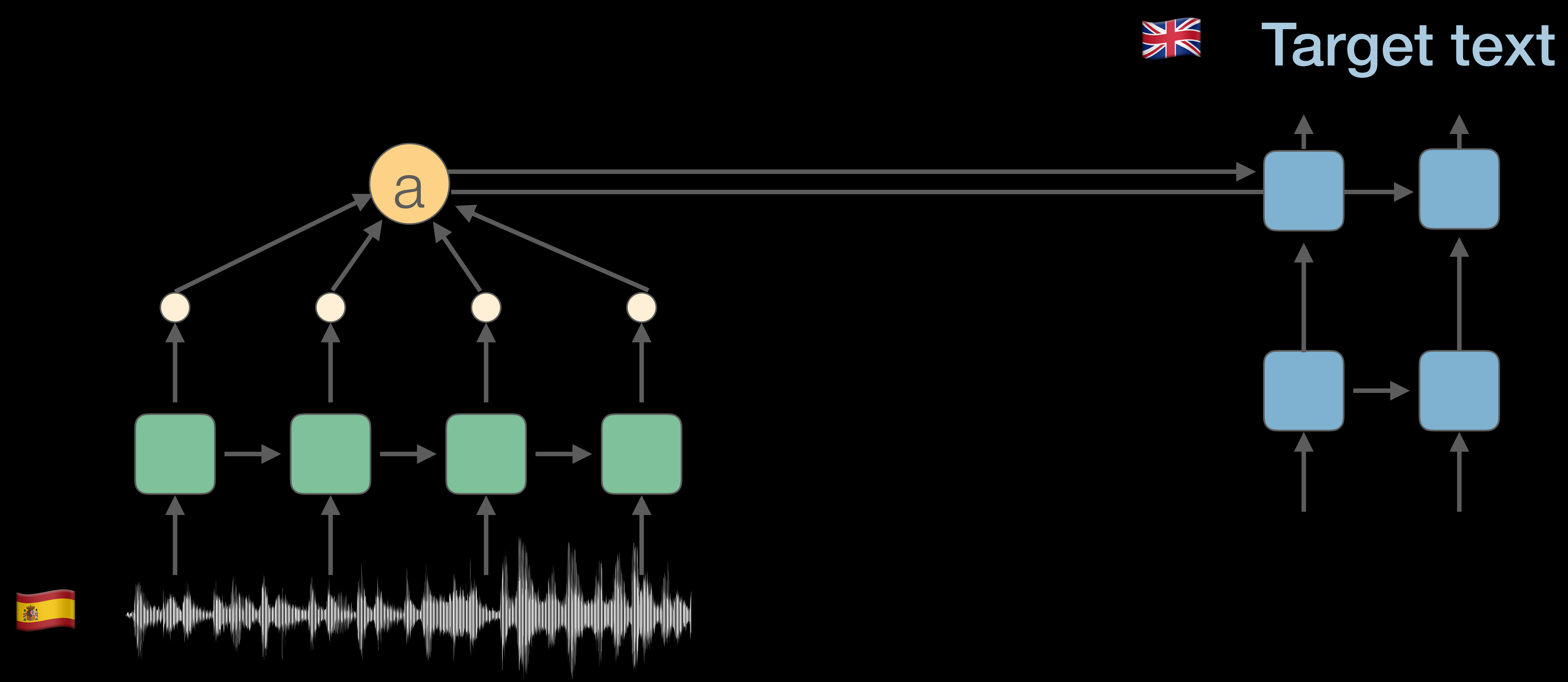
[Duong+2015]



Direct model

[Duong+2015]

- Endangered language documentation/ preservation



Language documentation

Transcript-free speech translation

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation
  - Data collection is time-consuming

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation
  - Data collection is time-consuming
  - Often no writing system

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation
  - Data collection is time-consuming
  - Often no writing system
  - Few or no expert linguists available

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation
  - Data collection is time-consuming
  - Often no writing system
  - Few or no expert linguists available
- Transcript-free speech translation can help

Language documentation

Transcript-free speech translation

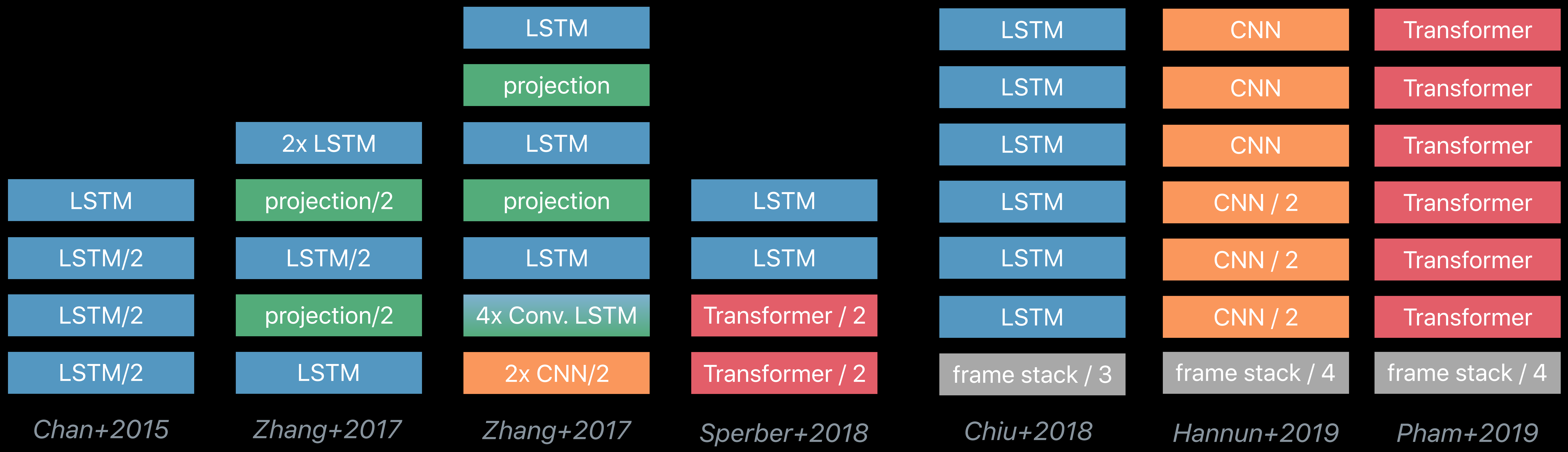
- Endangered language documentation/ preservation
  - Data collection is time-consuming
  - Often no writing system
  - Few or no expert linguists available
- Transcript-free speech translation can help
  - Need to cope with very small data

Language documentation

Transcript-free speech translation

- Endangered language documentation/ preservation
  - Data collection is time-consuming
  - Often no writing system
  - Few or no expert linguists available
- Transcript-free speech translation can help
  - Need to cope with very small data
  - Even if accuracy is bad: attention / alignment scores are already useful

Encoder architectures

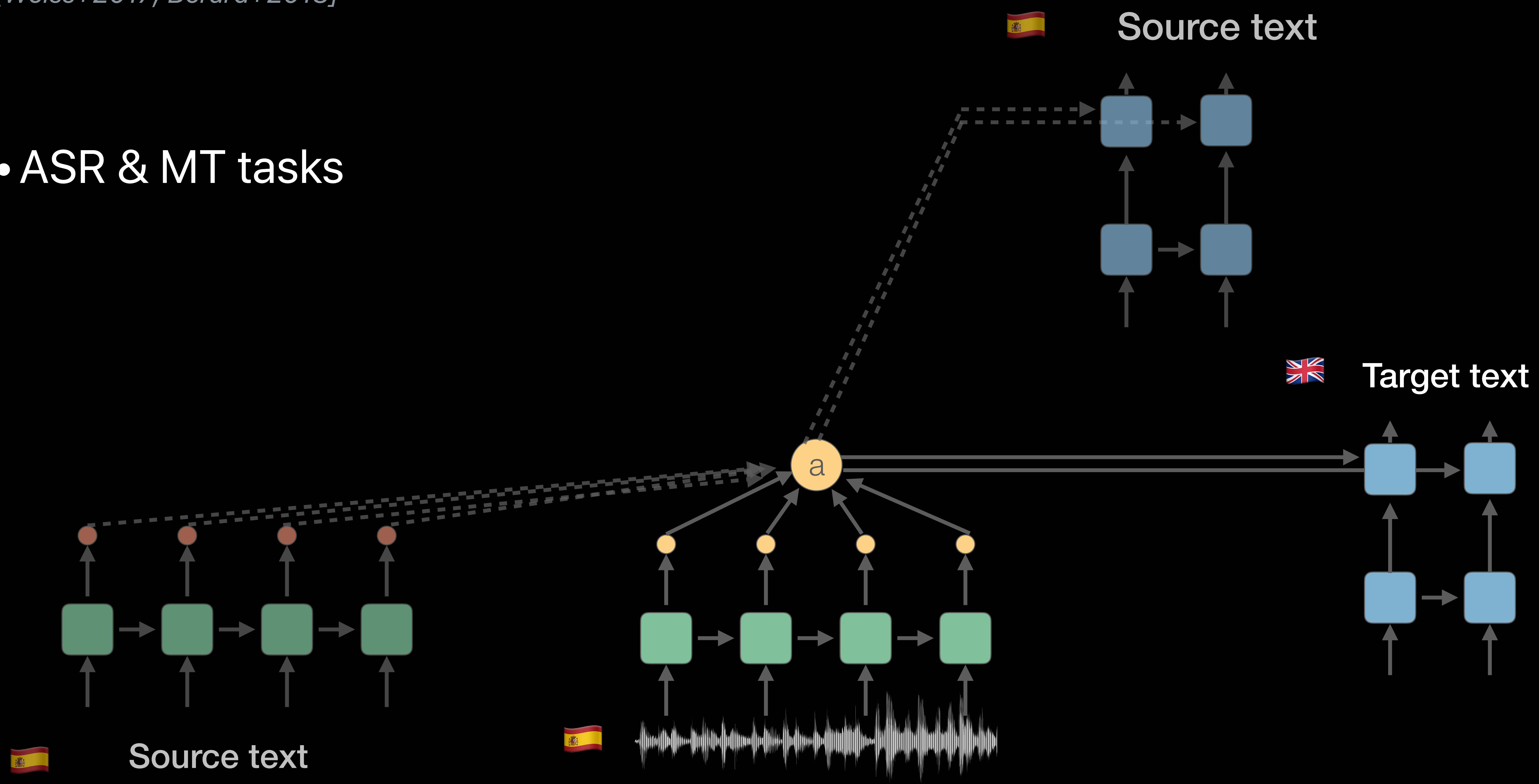


- Lots of choices for encoder architectures (mainly from ASR literature)
- Considerable differences in accuracy
- No consensus on "what works best for everyone" (yet)

Multi-task training

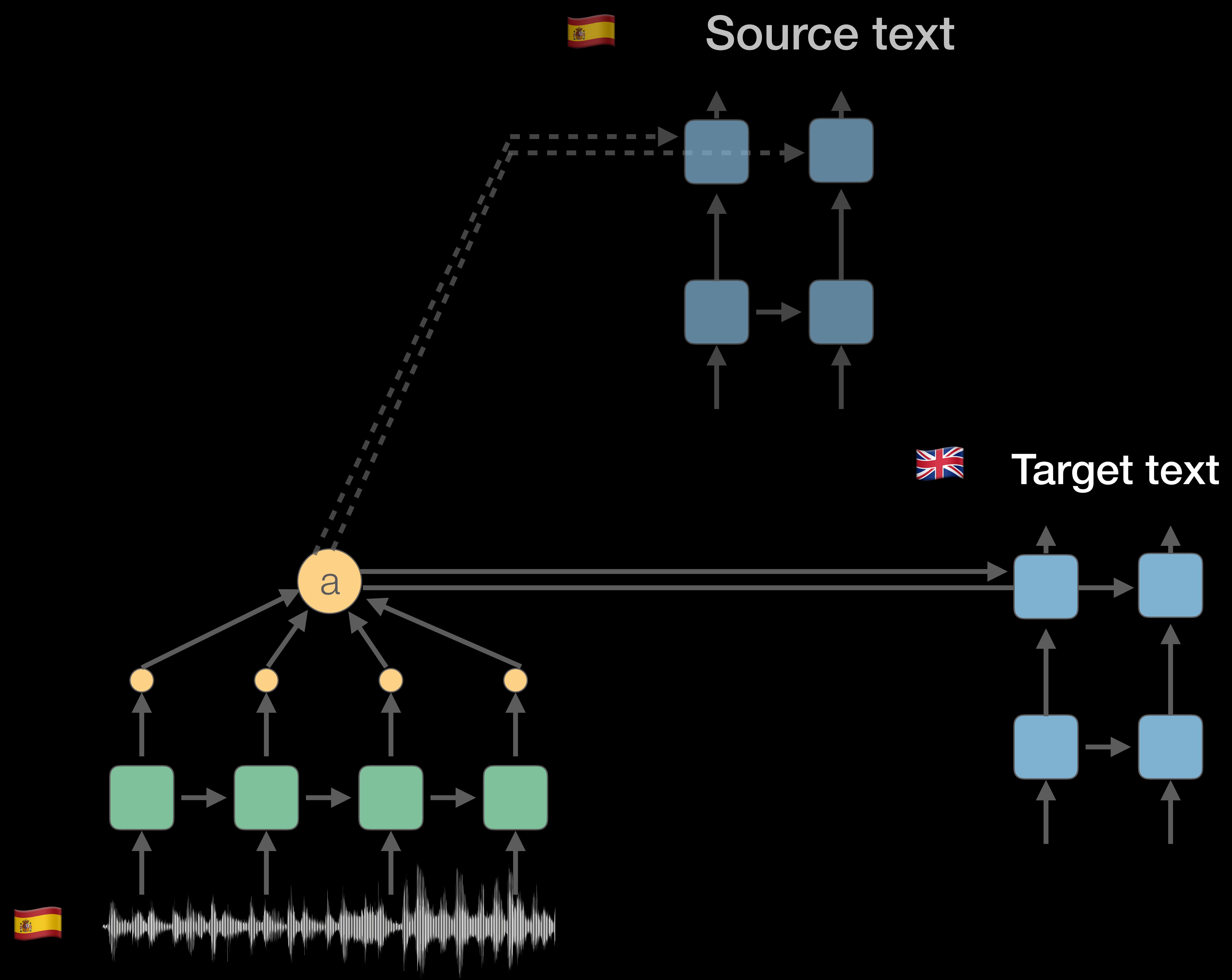
[Weiss+2017; Berard+2018]

- ASR & MT tasks



Pretraining
[Bansal+2019]

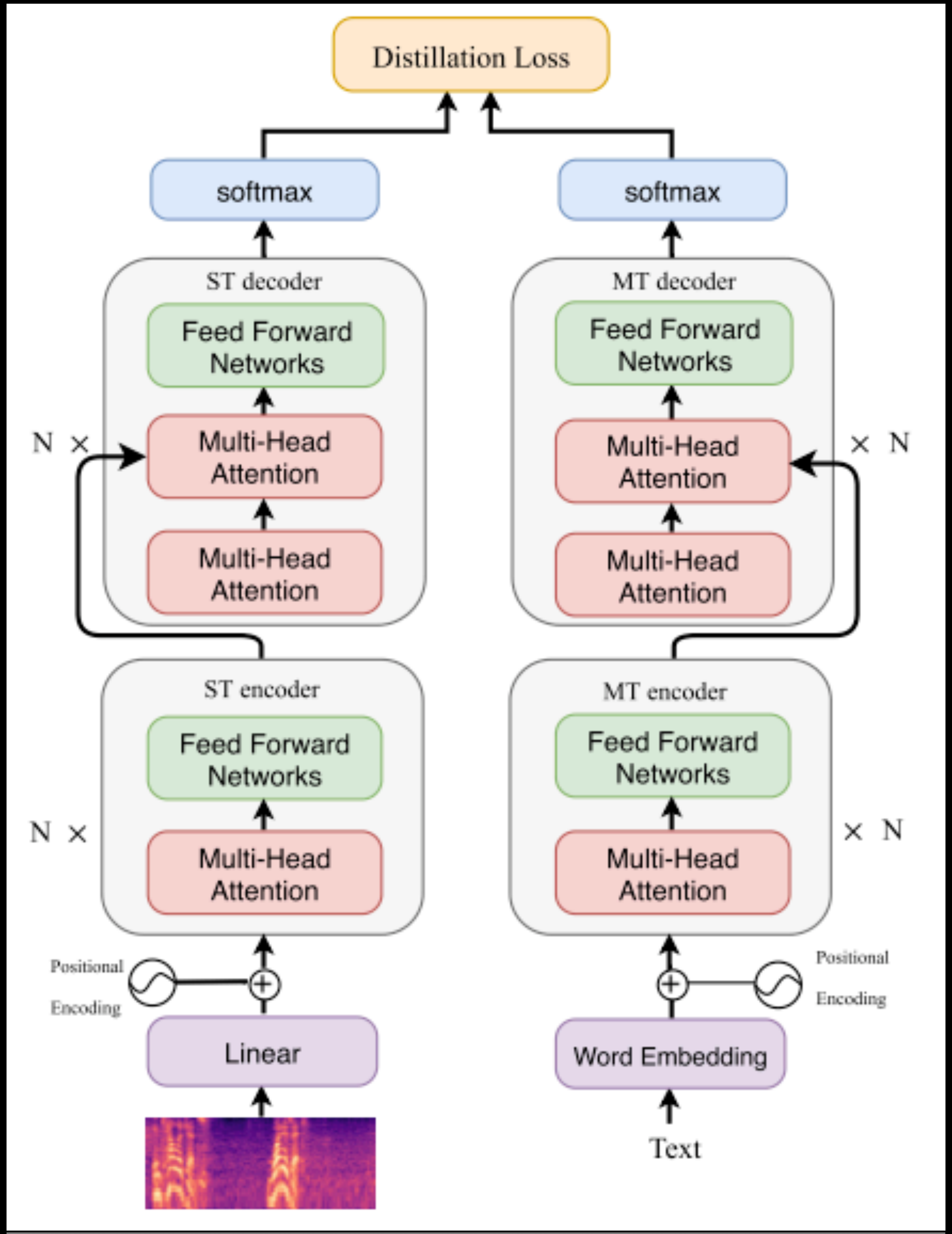
- Pretrain on ASR task
- Finetune on ST task
- Pretraining:
  - Possibly using larger ASR data
  - Helps even for unrelated ASR language!



Knowledge Distillation

[Liu+2019]

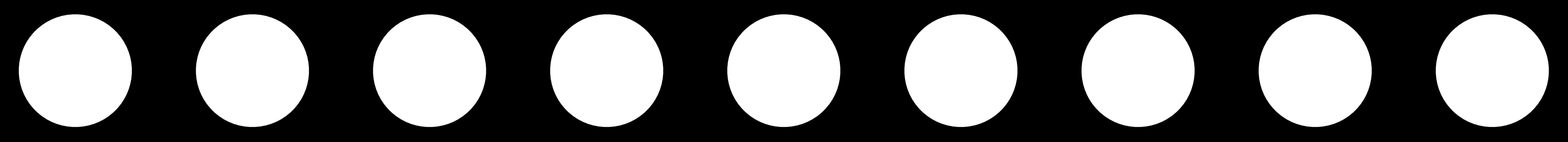
- Teacher: text translation model
- Student: speech translation model
  - Trained on teacher's softmax probabilities to imitate how teacher generalizes



Phoneme-level representations

[Salesky+2019]

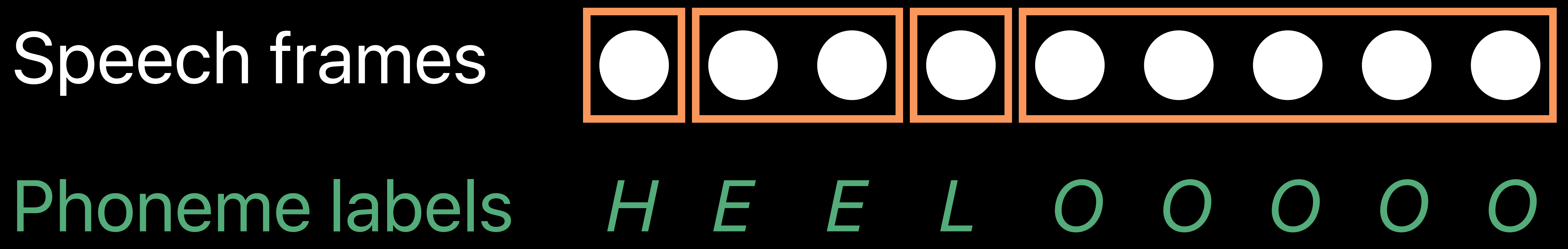
Speech frames



Phoneme-level representations
[Salesky+2019]

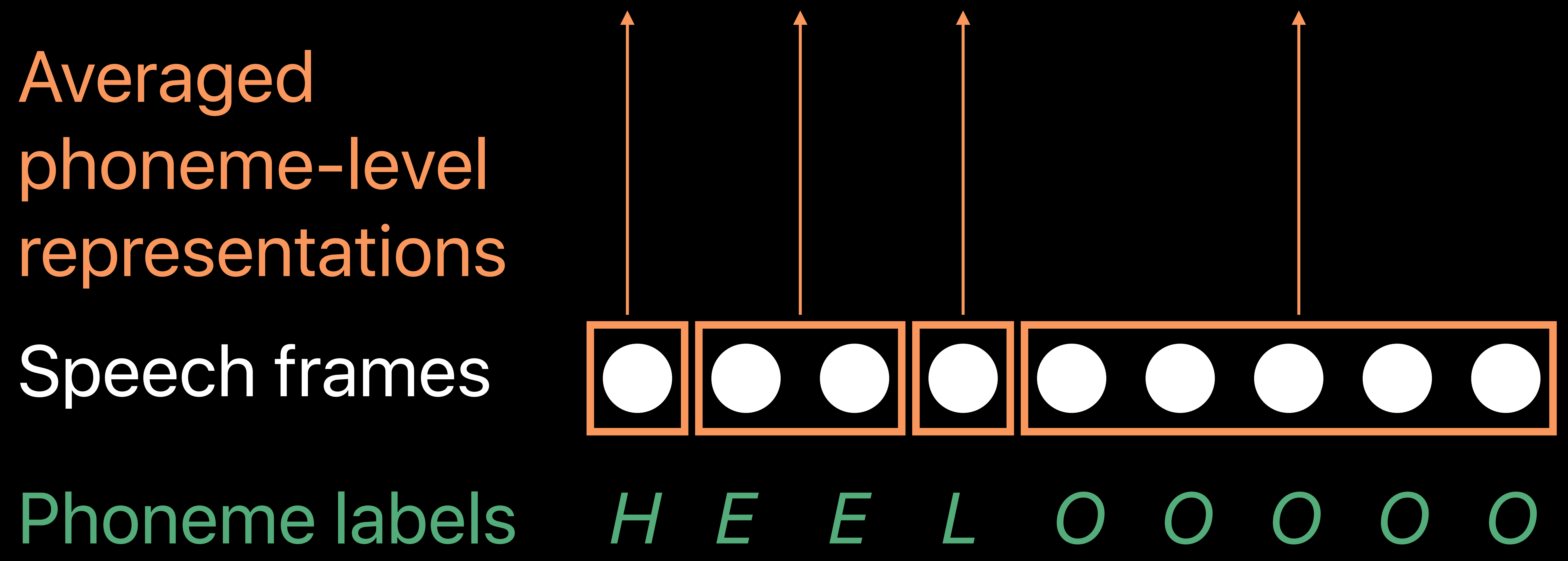
Phoneme-level representations

[Salesky+2019]



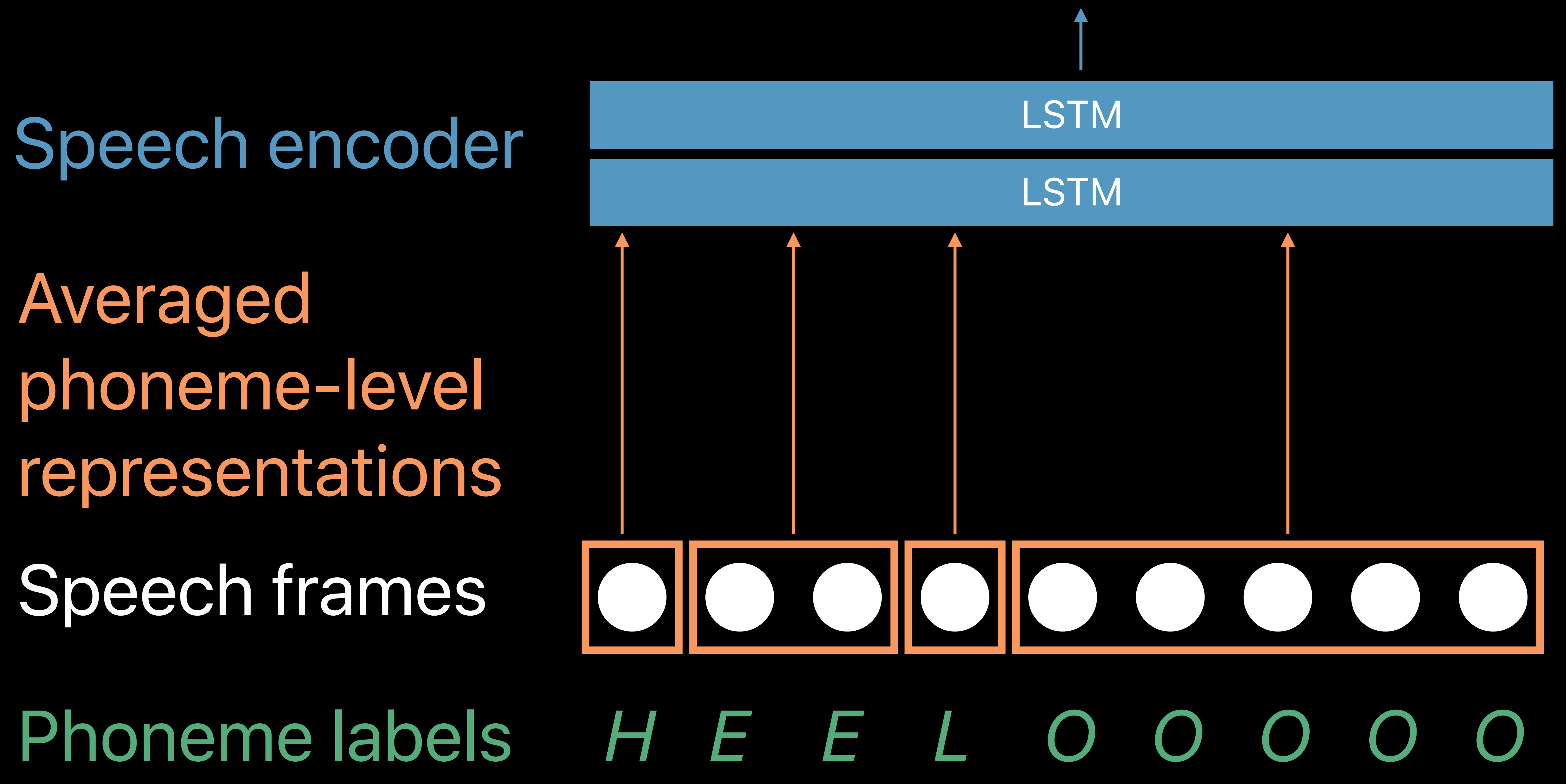
Phoneme-level representations

[Salesky+2019]



Phoneme-level representations

[Salesky+2019]



Phoneme-level representations

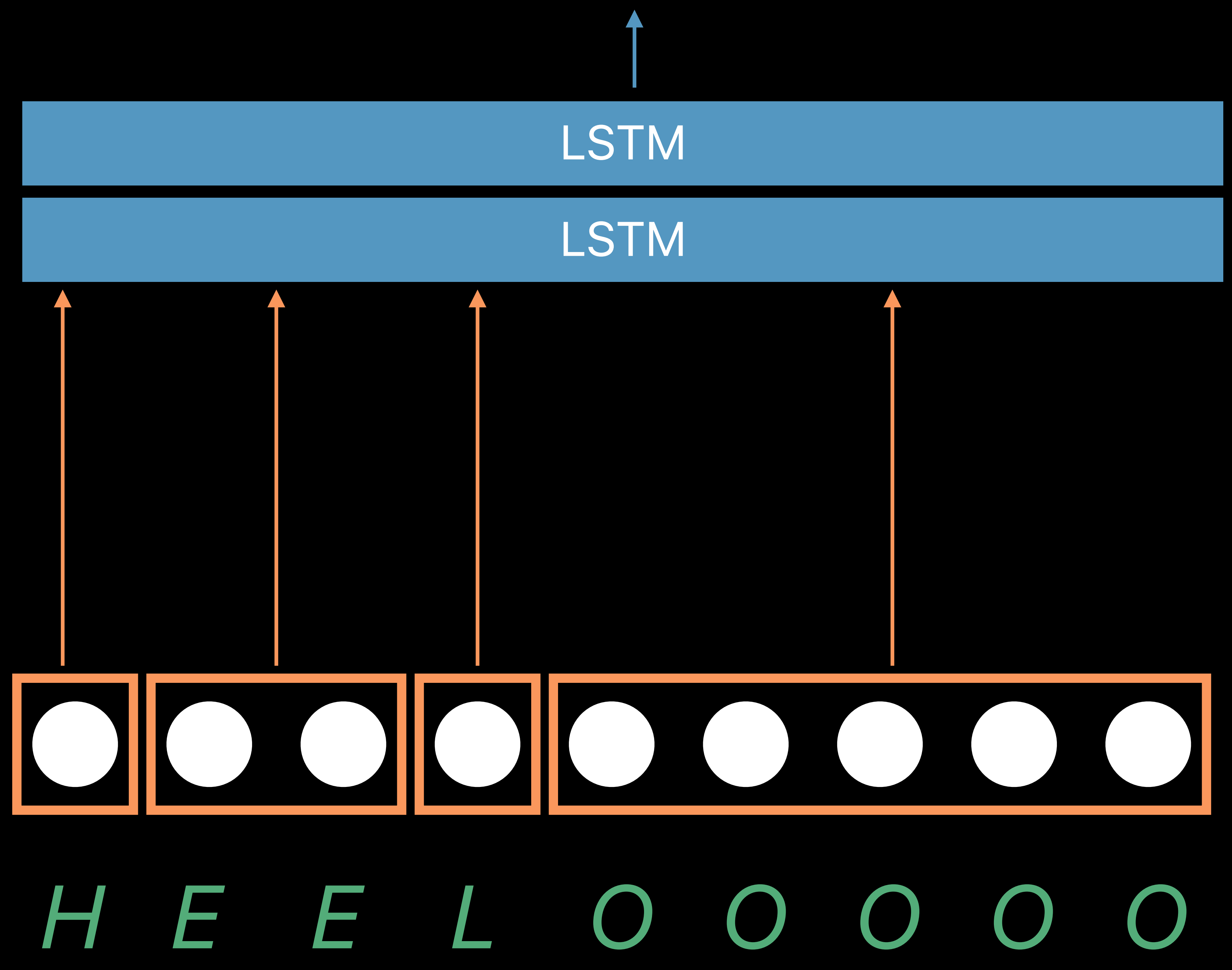
[Salesky+2019]

Speech encoder

Averaged phoneme-level representations

Speech frames

Phoneme labels



Data	Frames dev	Frames test	Phonemes dev	Phonemes test	BLEU Δ	Time Δ
Full	32.4	33.7	37.6	38.8	+5.2	–67%
40hr	19.5	17.4	21.0	19.8	+2.0	–52%
20hr	9.8	8.9	11.1	10.0	+1.2	–65%

Cascade vs. direct model

[Sperber+2019]

Cascade vs. direct model

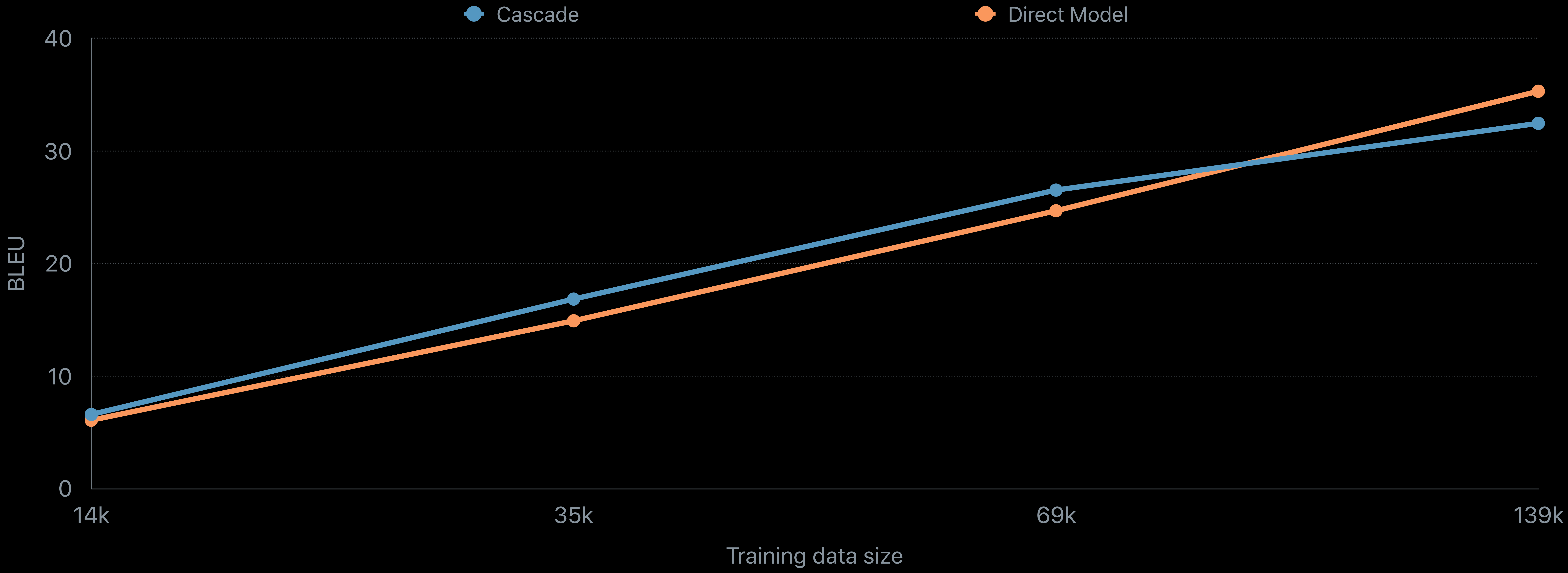
[Sperber+2019]

- Direct model works better if we have enough data

Cascade vs. direct model

[Sperber+2019]

- Direct model works better if we have enough data



Data efficiency

Analysis

[Sperber+2019]

Data efficiency

Analysis

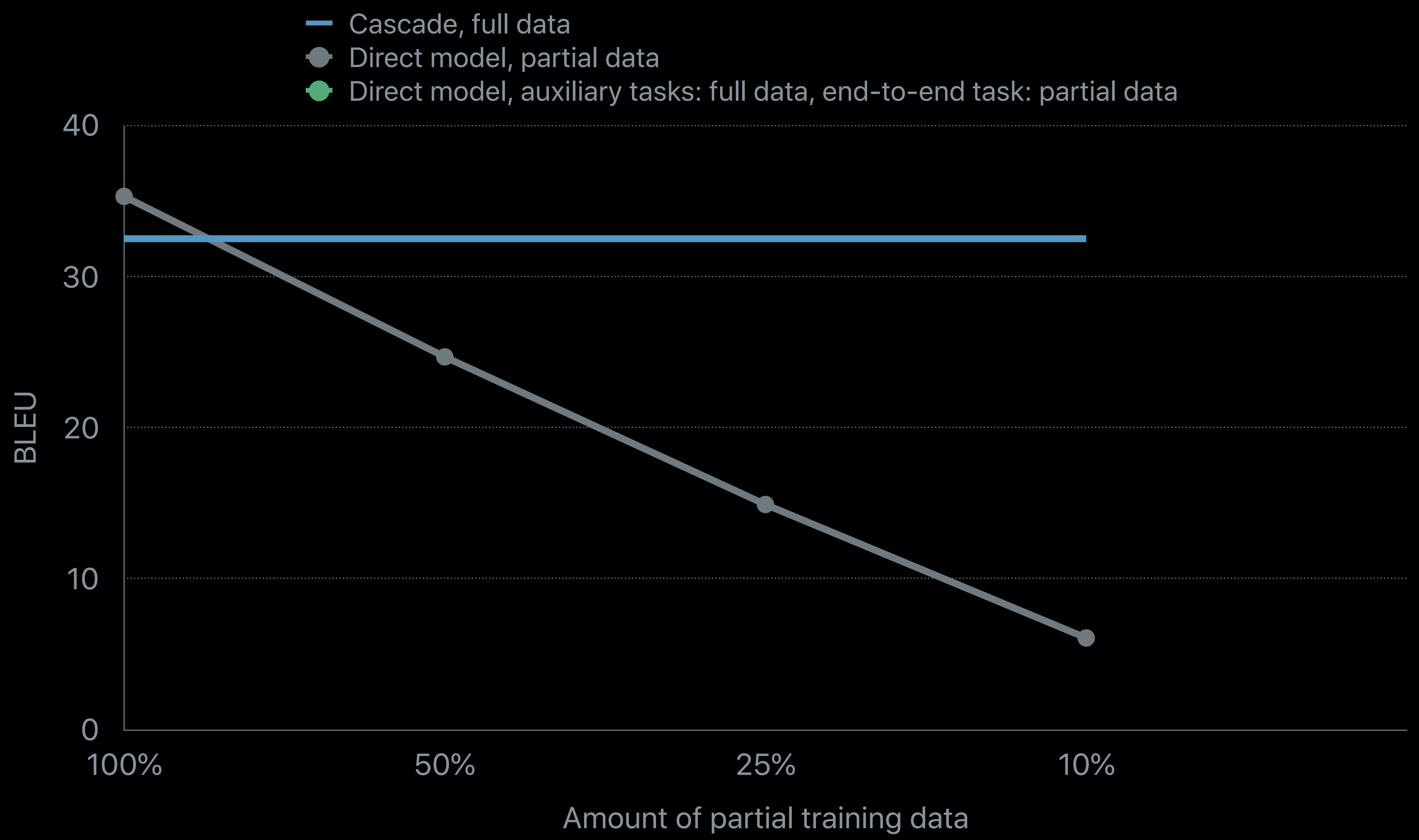
[Sperber+2019]

Data efficiency

Analysis

[Sperber+2019]

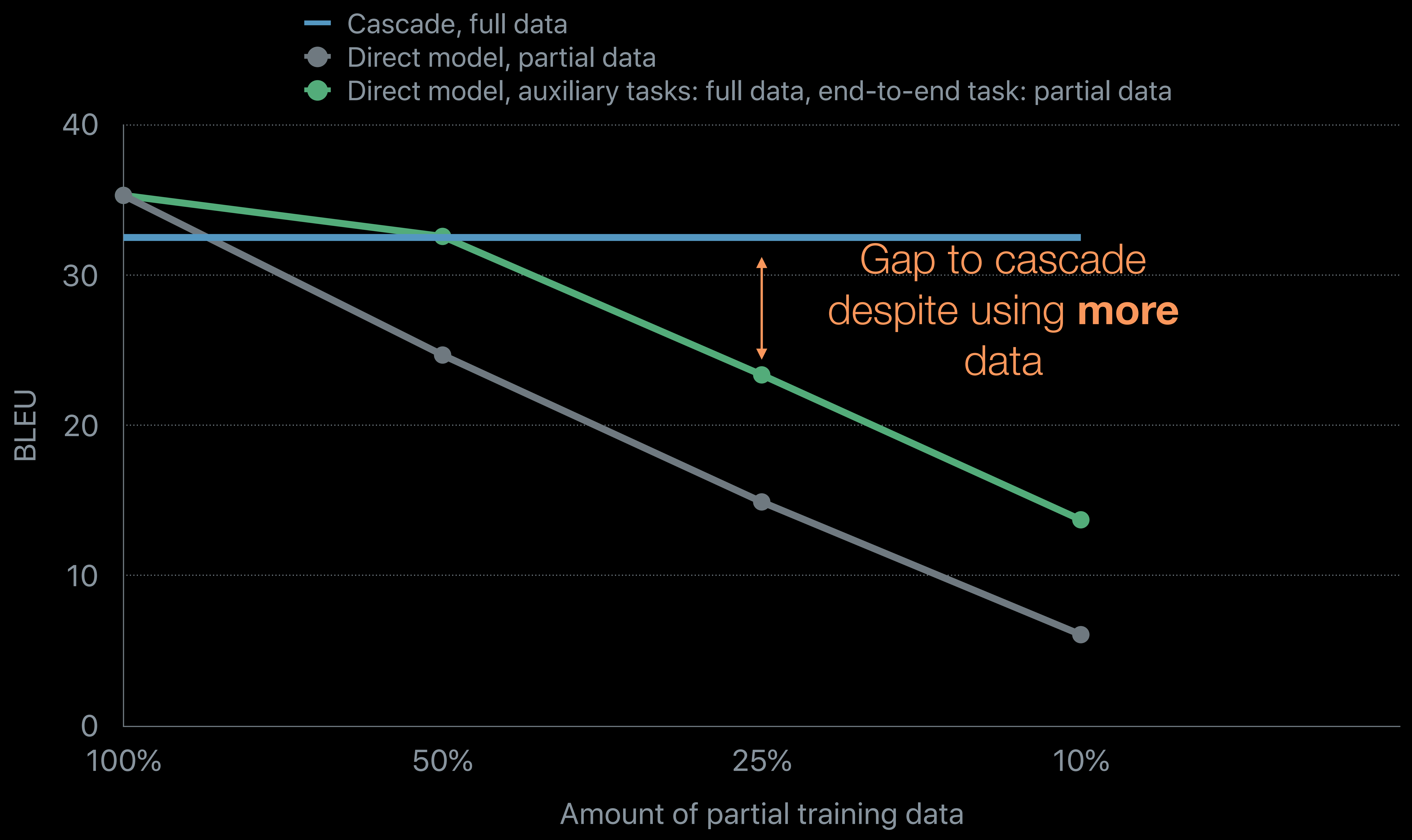
Data efficiency
Analysis
[Sperber+2019]



Data efficiency

Analysis

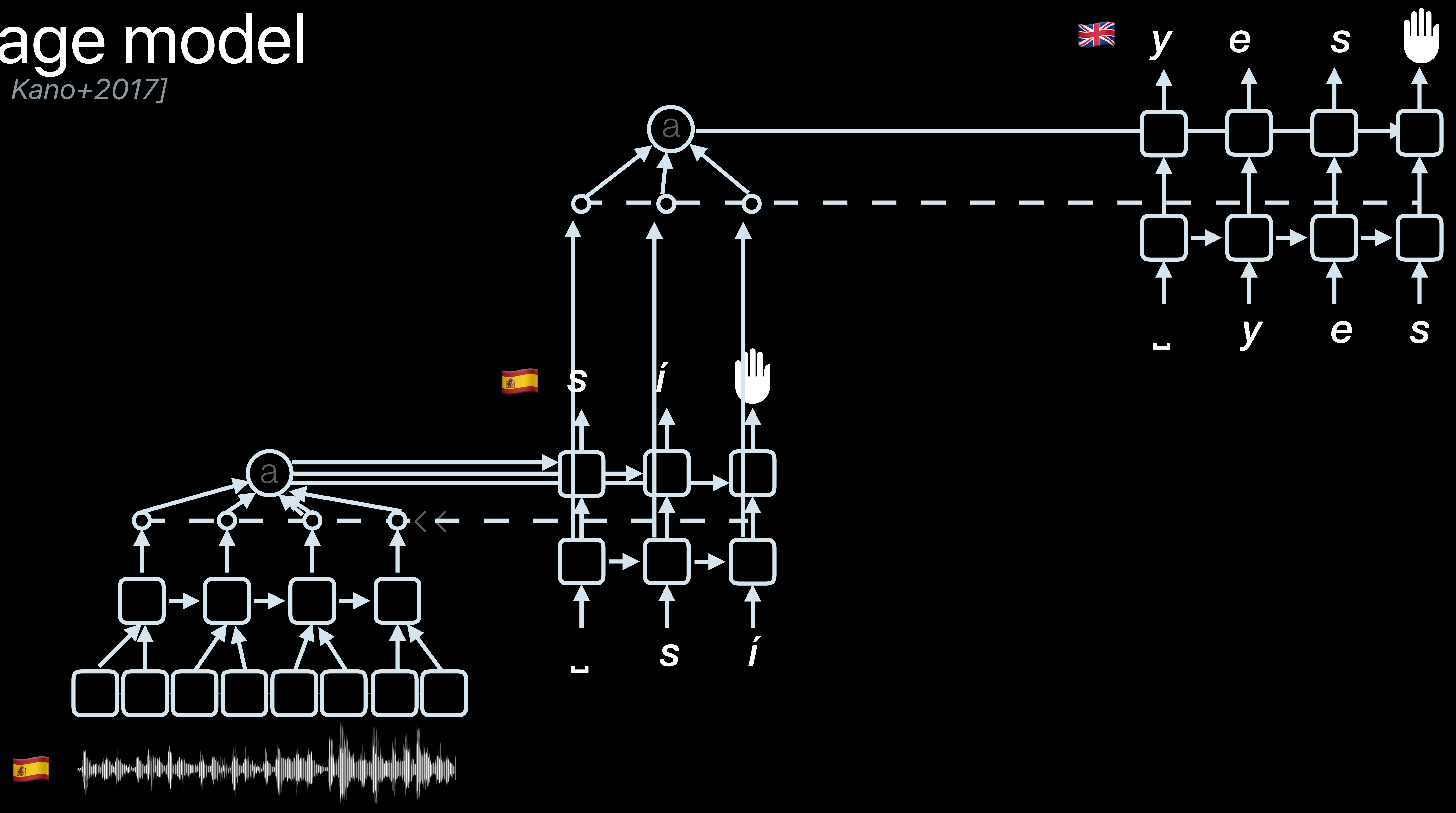
[Sperber+2019]



Improving data efficiency

2-stage model

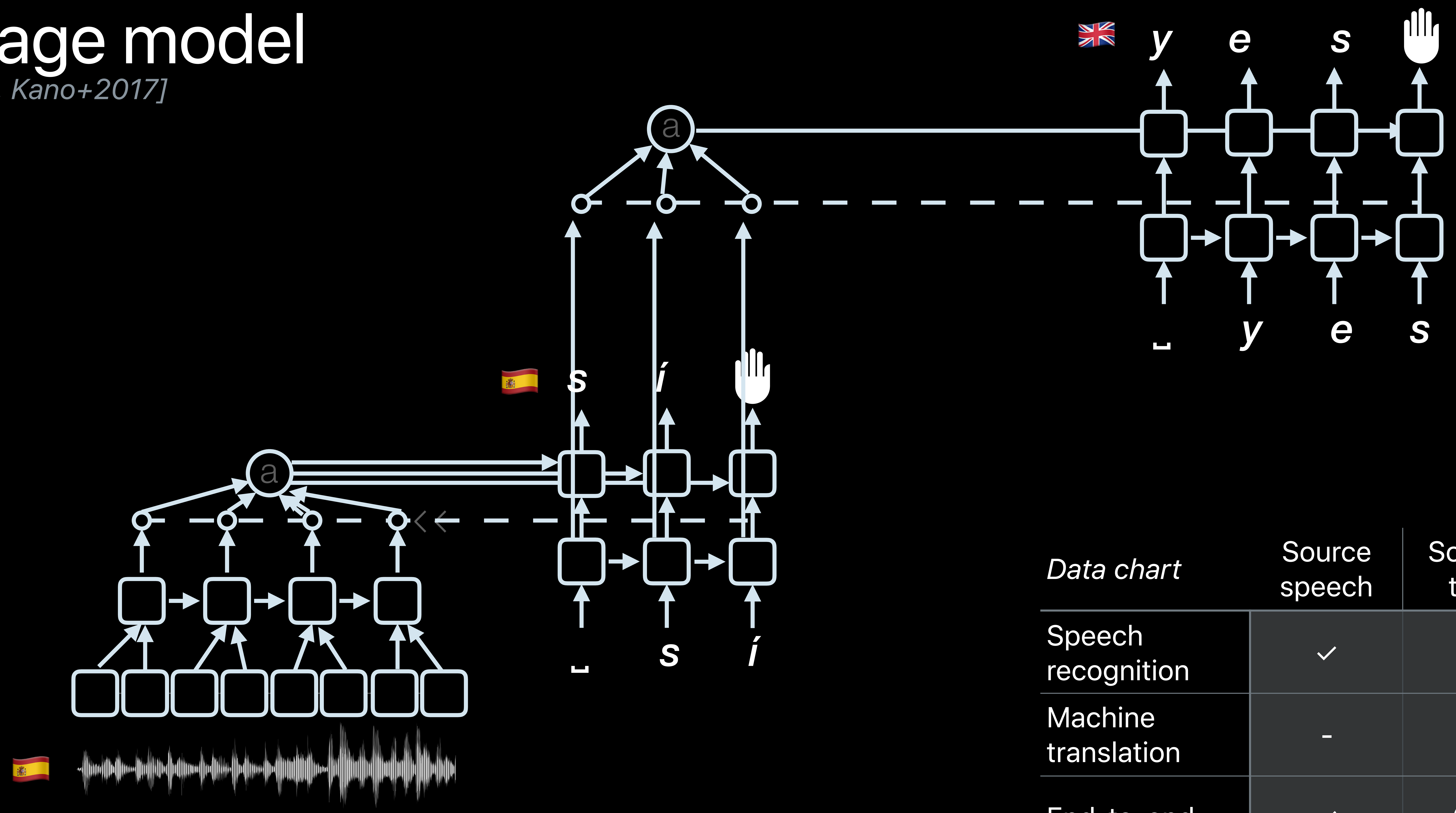
[Tu+2016, Kano+2017]



Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

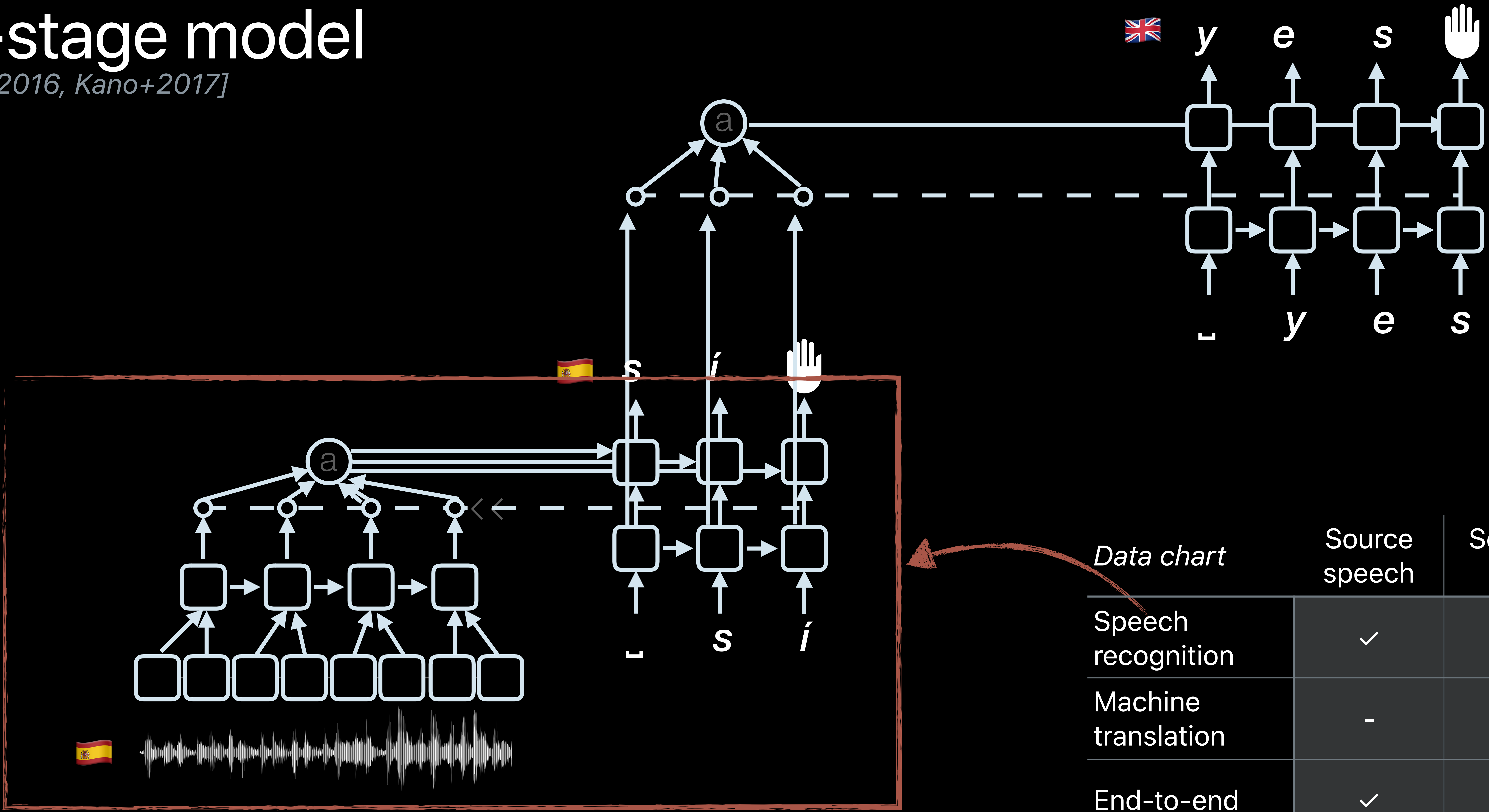


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	–
Machine translation	–	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

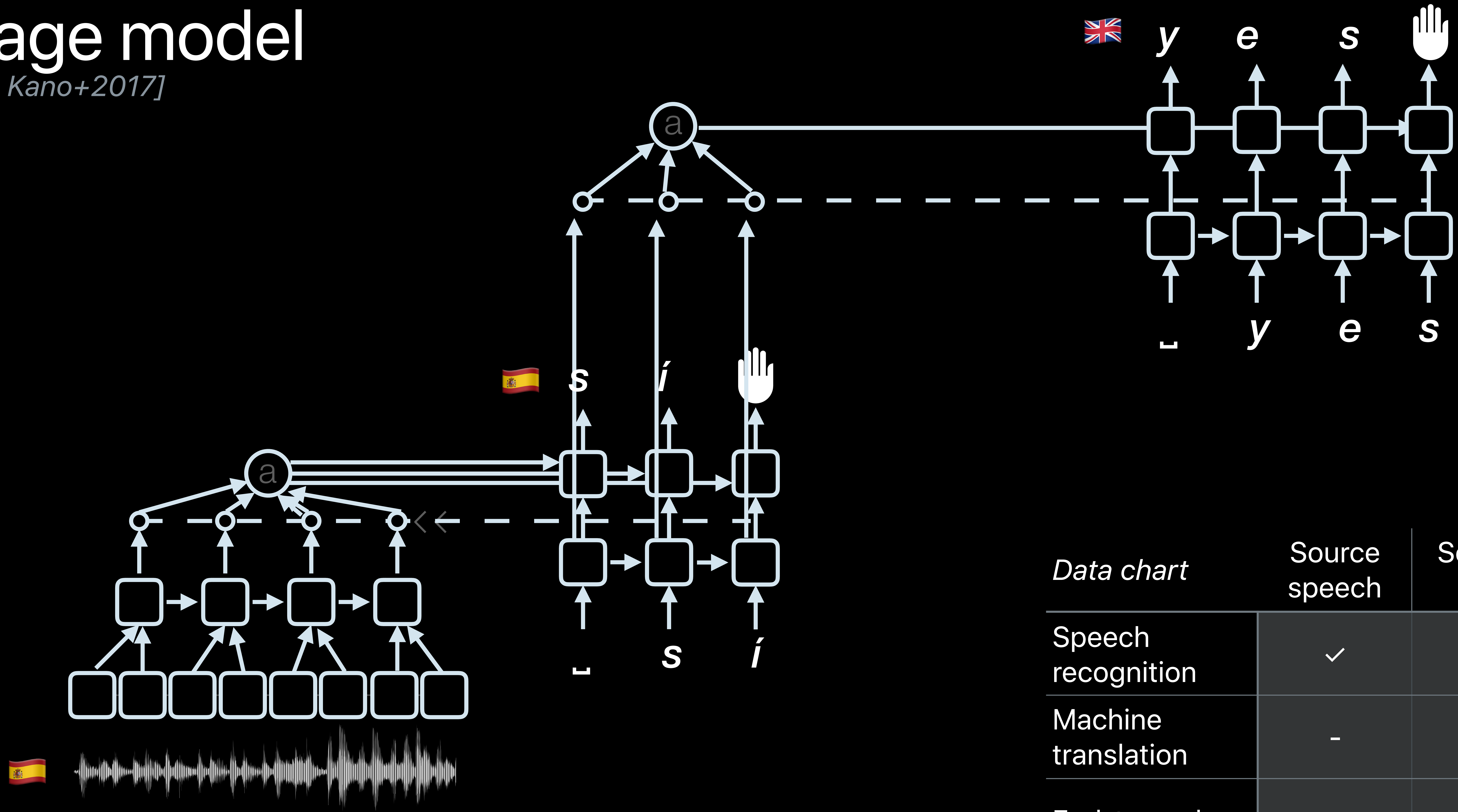


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	-
Machine translation	-	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

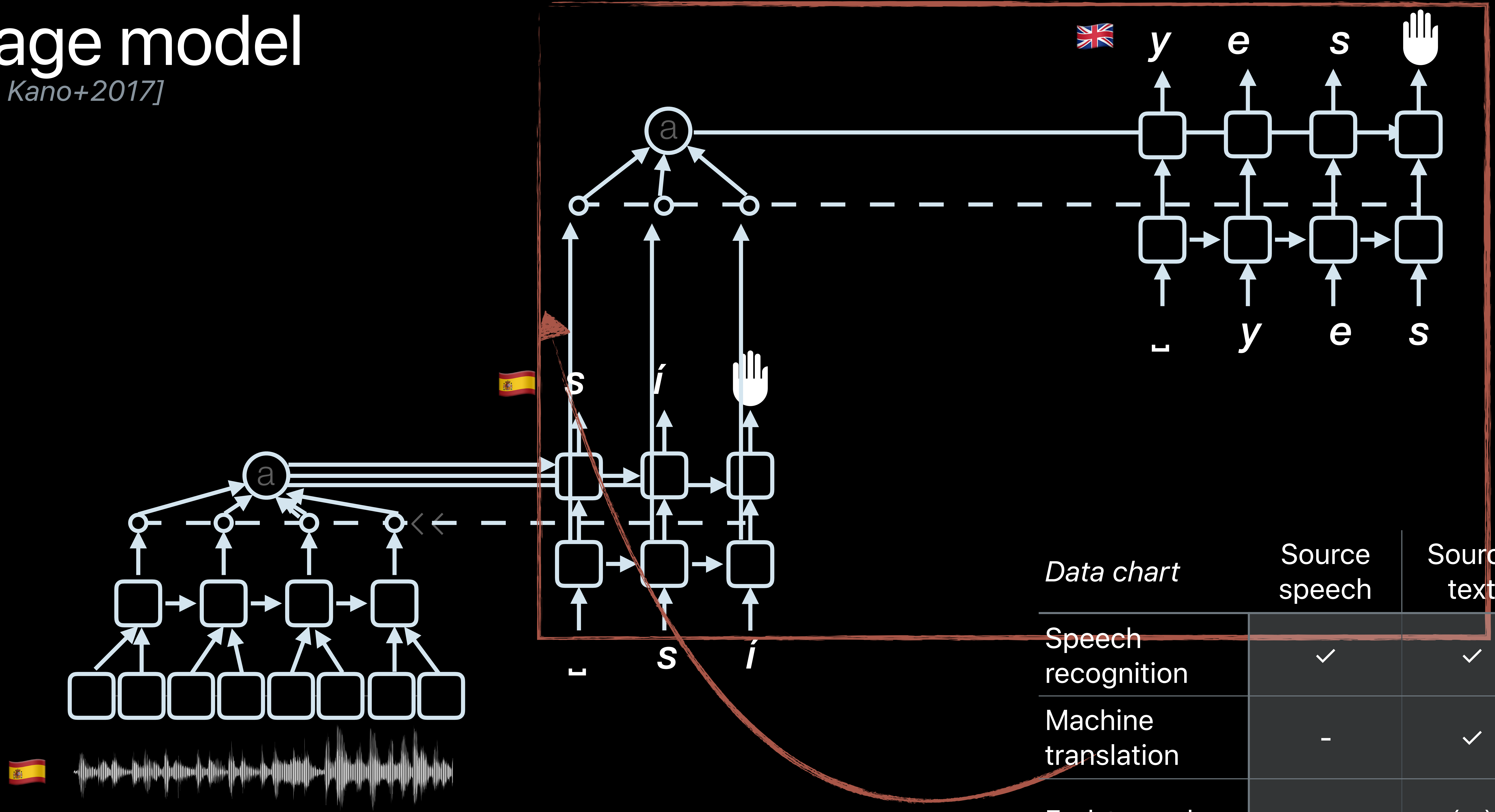


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	-
Machine translation	-	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

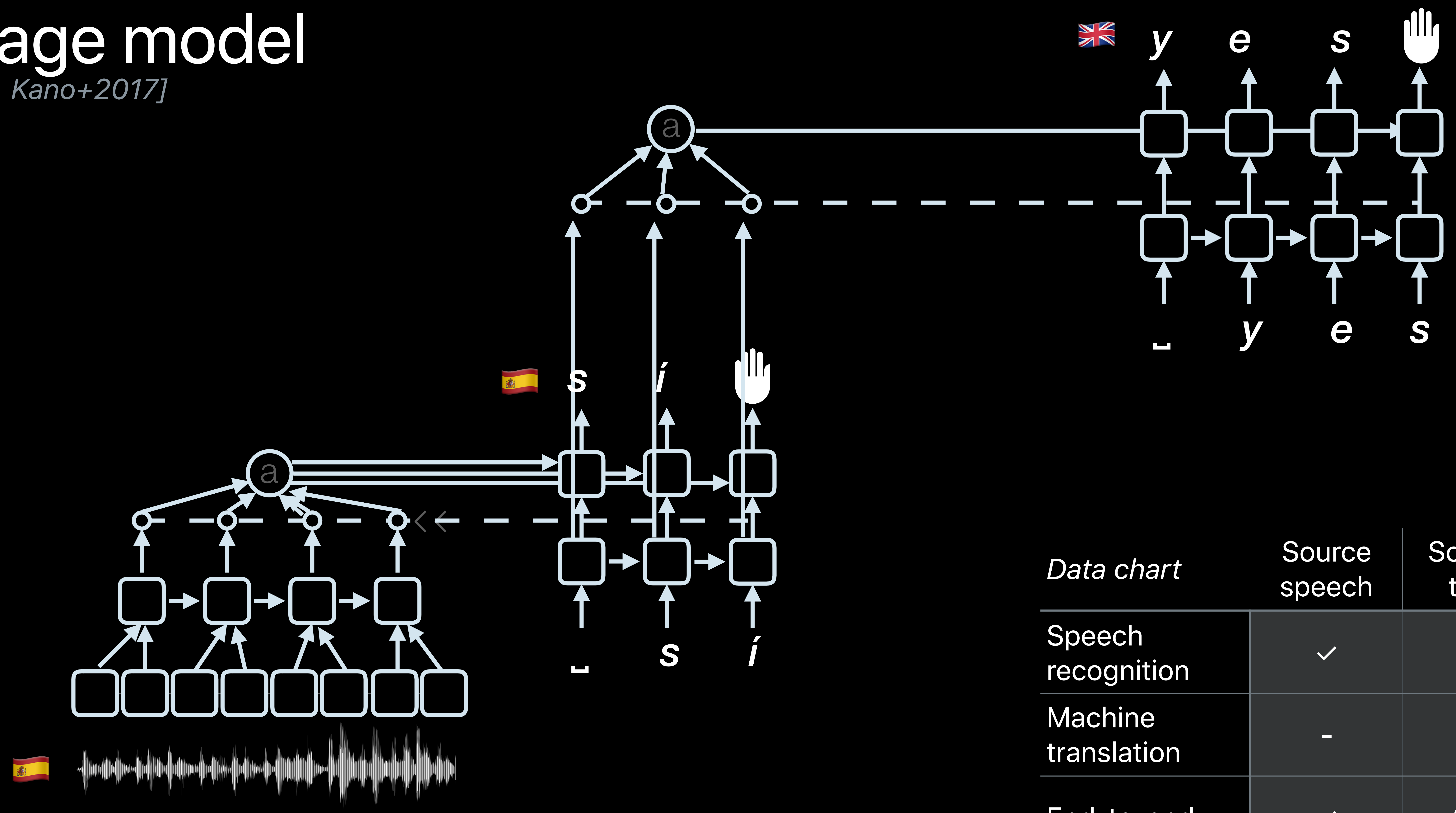


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	-
Machine translation	-	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

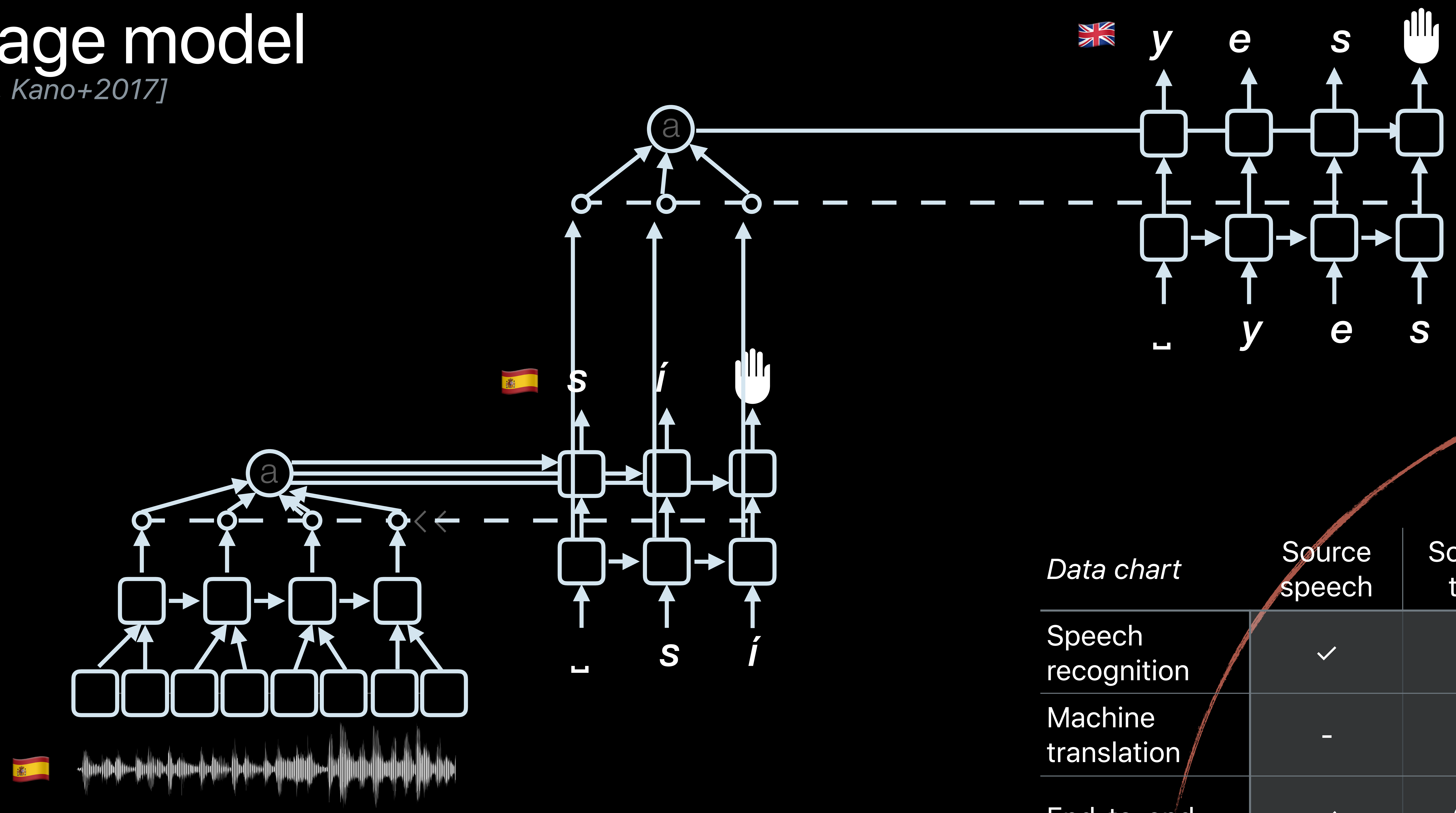


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	–
Machine translation	–	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

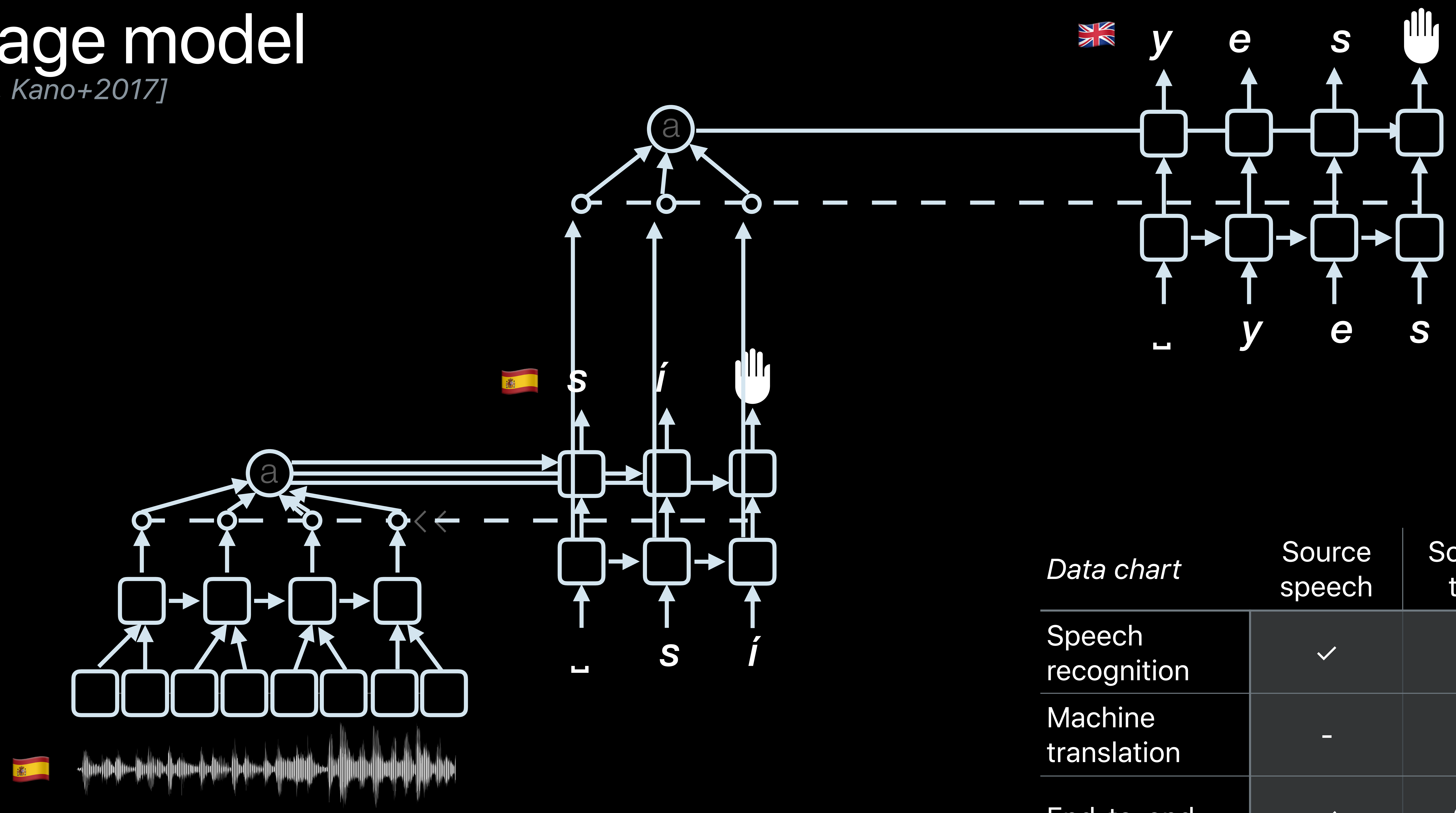


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	–
Machine translation	–	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

2-stage model

[Tu+2016, Kano+2017]

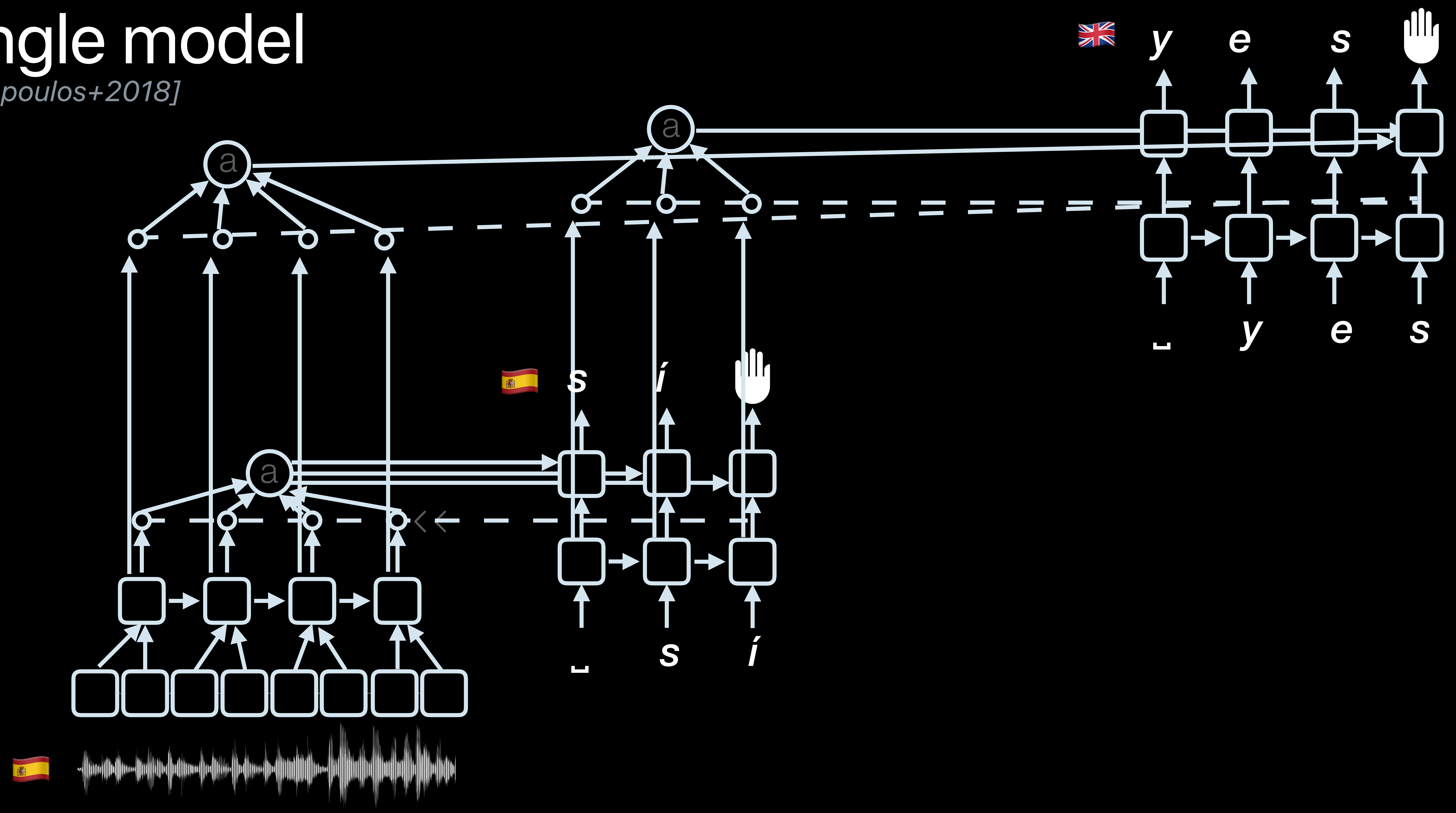


Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	–
Machine translation	–	✓	✓
End-to-end	✓	(✓)	✓

Improving data efficiency

Triangle model

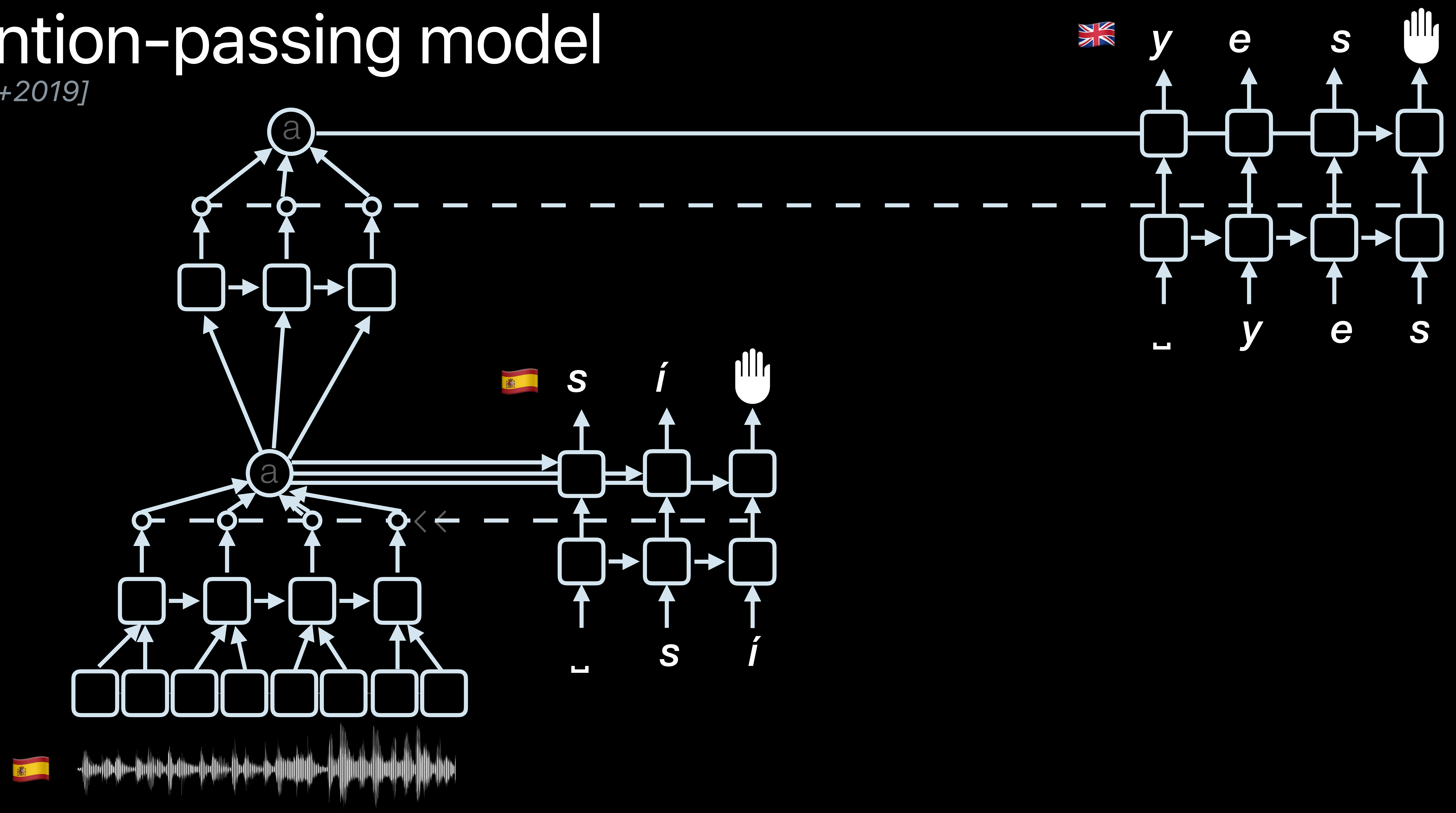
[Anastasopoulos+2018]



Improving data efficiency

Attention-passing model

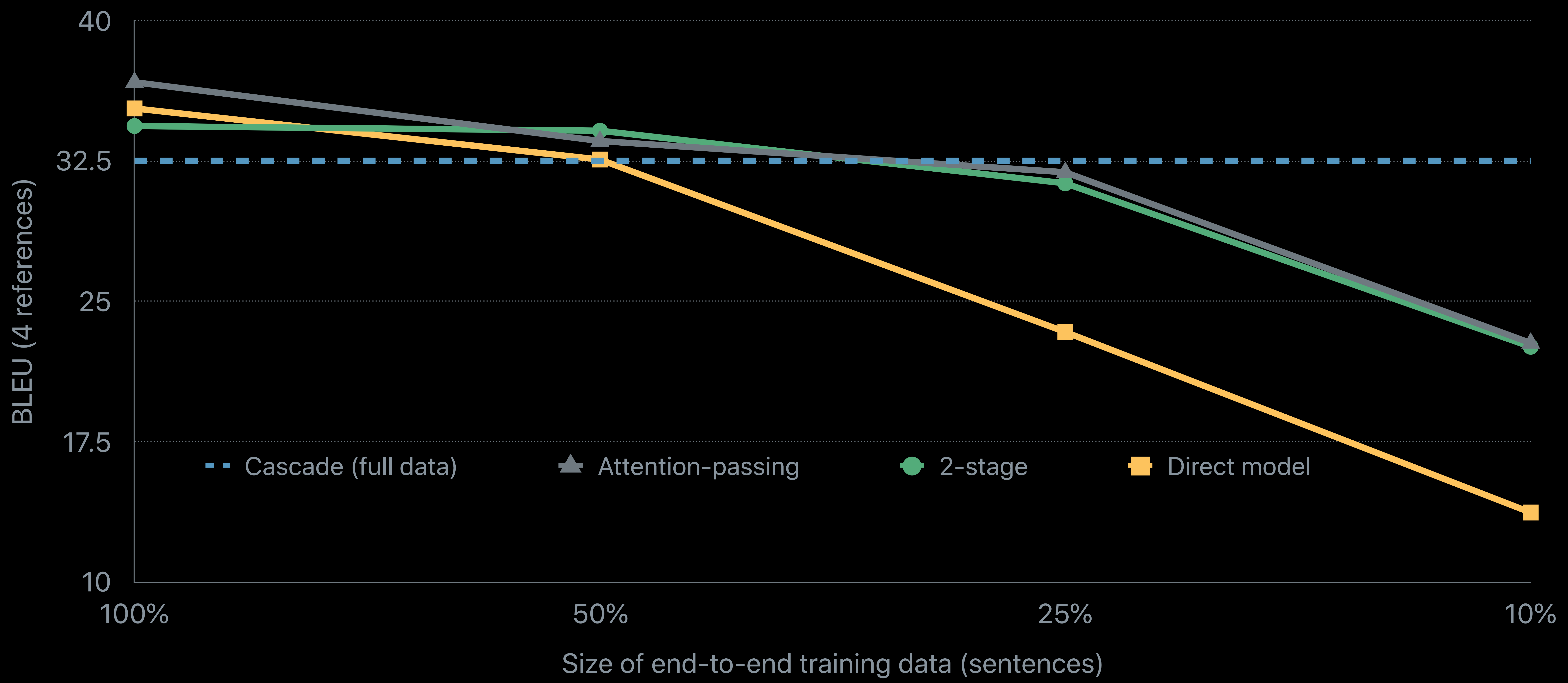
[Sperber+2019]



Data efficiency

Analysis

[Sperber+2019]

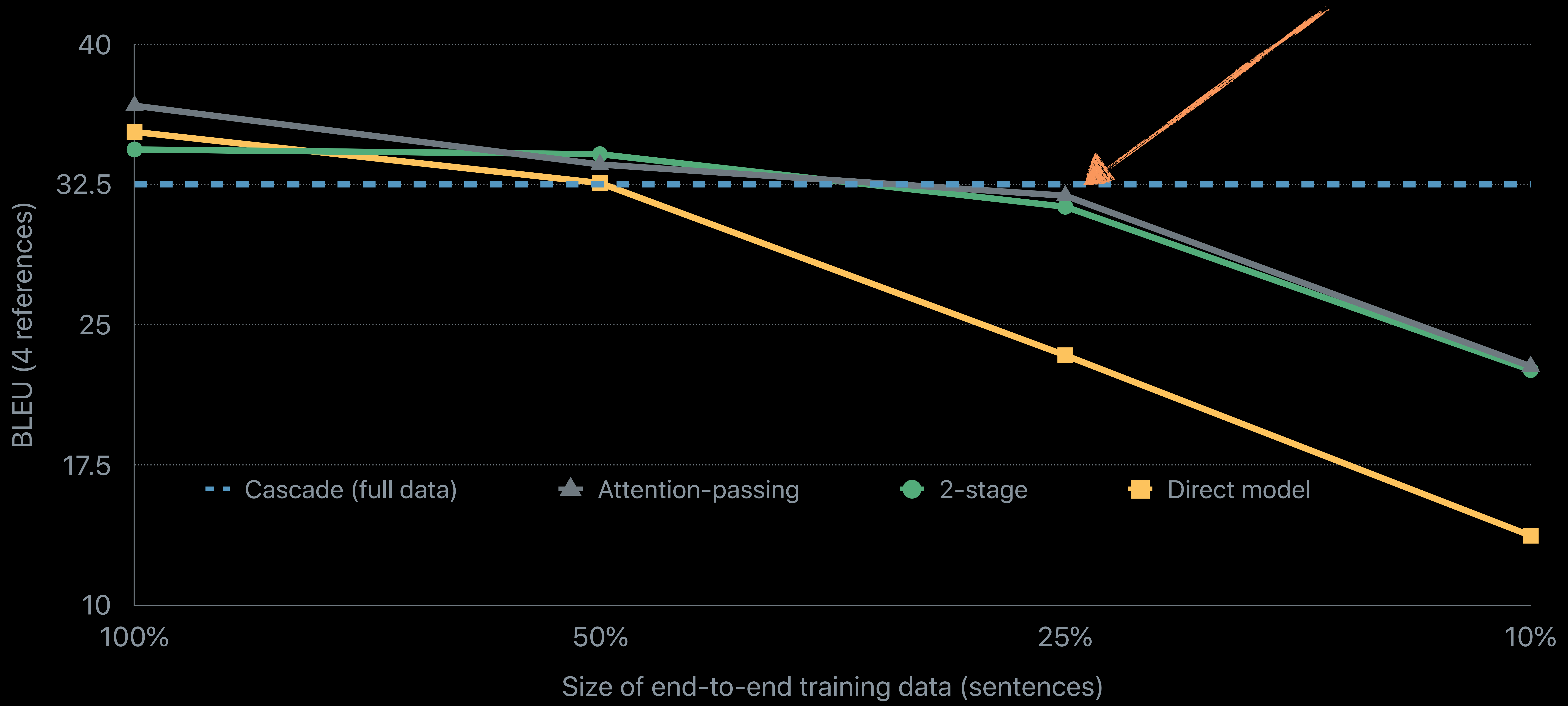


Data efficiency

Analysis

[Sperber+2019]

Attention-passing & 2-stage models work with much less e2e data!



Data efficiency

Synthesizing missing data points

[Jia+2019]

Data chart	Source speech	Source text	Target text
Speech recognition	✓	✓	synthesize (MT)
Machine translation	synthesize (TTS)	✓	✓
End-to-end	✓	(✓)	✓

Fine-tuning set	In-domain	Out-of-domain
Real	55.9	19.5
Real + TTS synthetic	59.5	22.7
Real + MT synthetic	57.9	26.2
Real + both synthetic	**59.5**	**26.7**
Only TTS synthetic	53.9	20.8
Only MT synthetic	42.7	26.9
Only both synthetic	**55.6**	**27.0**

Are the cascade’s problems solved?

- Problem 1: Error propagation
- Problem 2: Domain mismatch
- Problem 3: Information loss

Are the cascade's problems solved?

→ Yes (direct model; but: not enough data)

- Problem 1: Error propagation
- Problem 2: Domain mismatch
- Problem 3: Information loss

Are the cascade's problems solved?

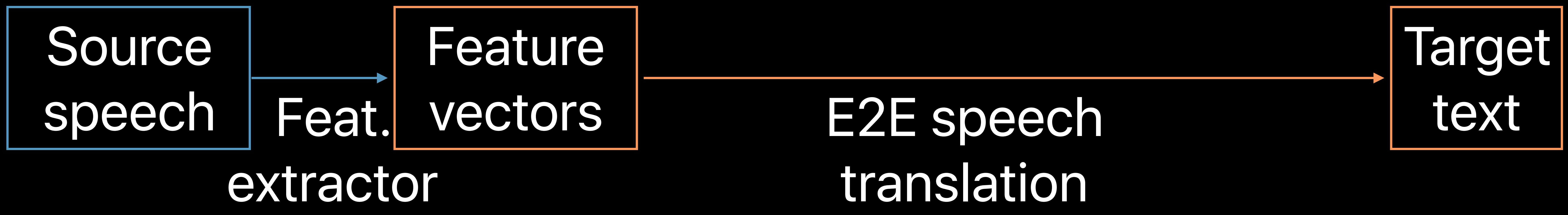
→ Yes (direct model; but: not enough data)

→ Only partly (2-stage, multi-tasking, synthesized data, etc.)

- Problem 1: Error propagation
- Problem 2: Domain mismatch
- Problem 3: Information loss

Can we do even more “end-to-end”?

Can we do even more “end-to-end”?



Can we do even more "end-to-end"?

Can we do even more “end-to-end”?

Can we do even more "end-to-end"?

Can we do even more “end-to-end”?

Translate + remove disfluency

[Salesky+2019]

- Input: source speech
- Output: target text with disfluencies already removed

Segment comparison: **Deletion** **Insertion** **Shift**

Disfluent:	**and that** you see it **well** but you are **not** sure that **you're** there
Fluent:	you **don't** see it but you are sure that **they are** there
Disfluent:	**and well that even if they** don't see
Fluent:	**although you** don't see
Disfluent:	yes **yes**
Fluent:	yes

★ Better n-gram match

★ Similar semantic match

Model	Metric	dev 1Ref	dev 2Ref	test 1Ref	test 2Ref
Disfluent	BLEU	13.0	16.2	13.5	17.0
Fluent	BLEU	14.6	18.1	14.6	18.1
Disfluent	METEOR	22.2	23.9	23.1	24.8
Fluent	METEOR	22.3	24.0	23.1	24.9

Speech-to-speech

[Jia+2019]

- Based on the "Tacotron" end-to-end text-to-speech model

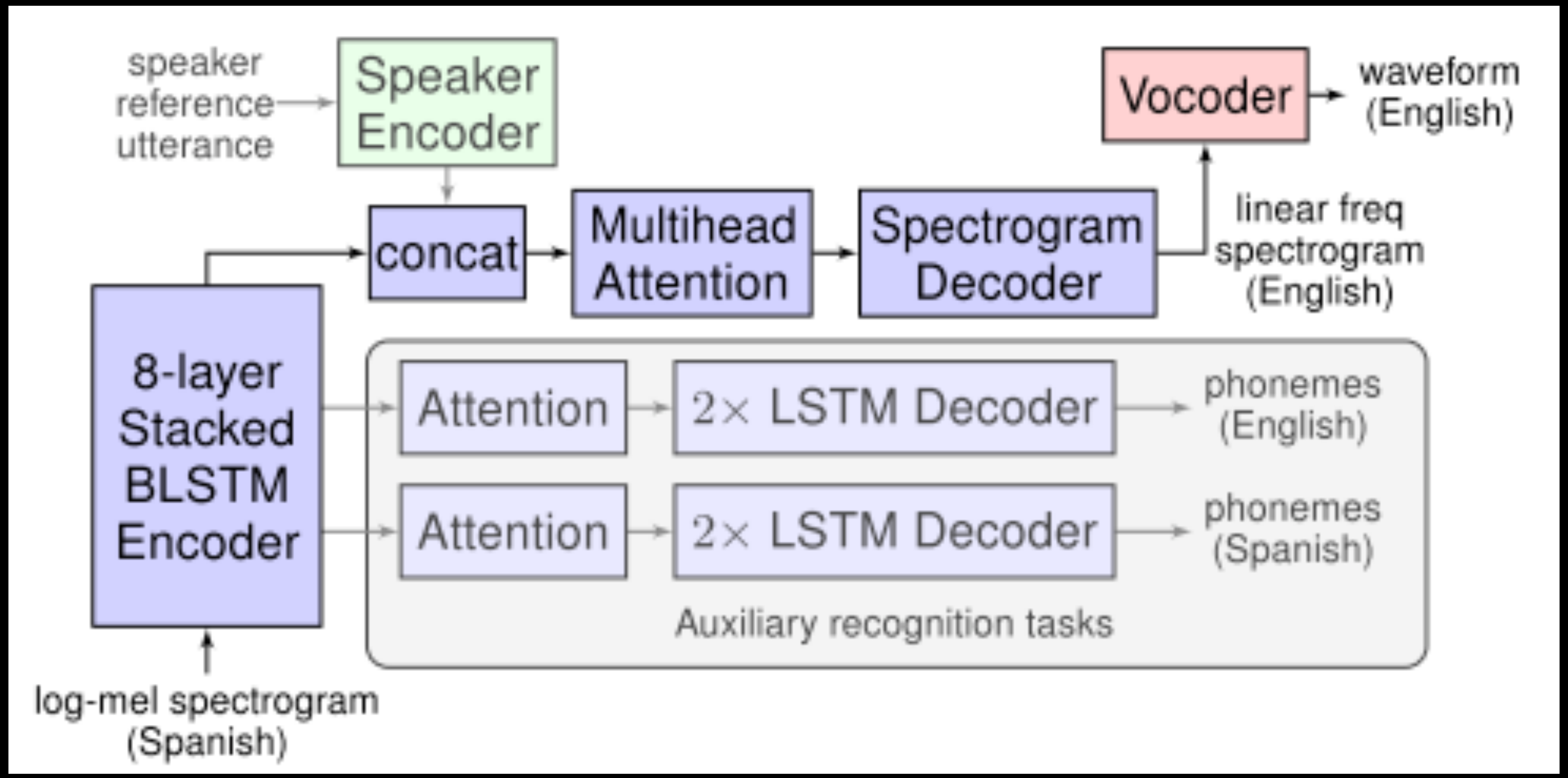


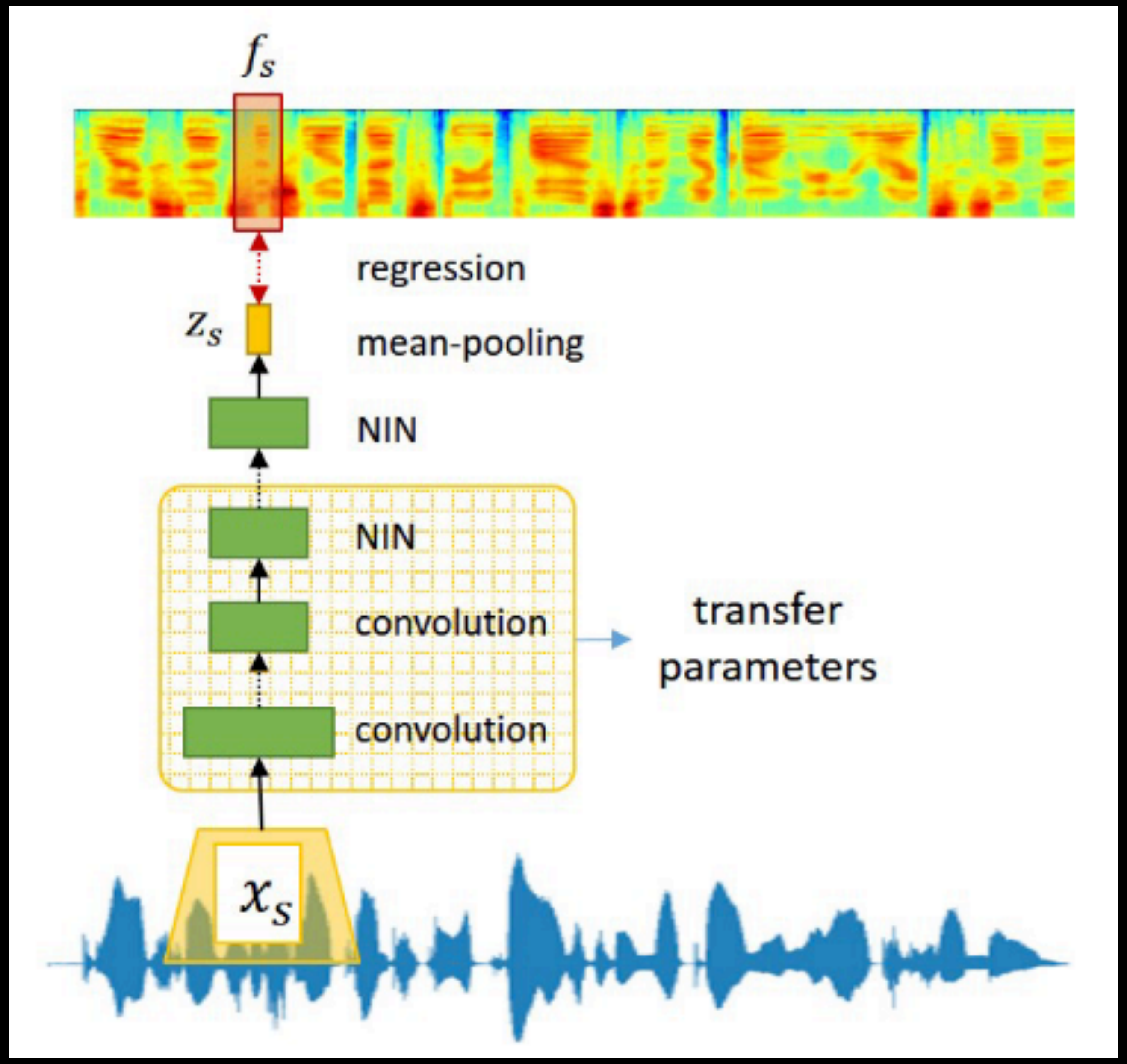
Table 2: *Conversational test set performance. Single reference BLEU and Phoneme Error Rate (PER) of aux decoder outputs.*

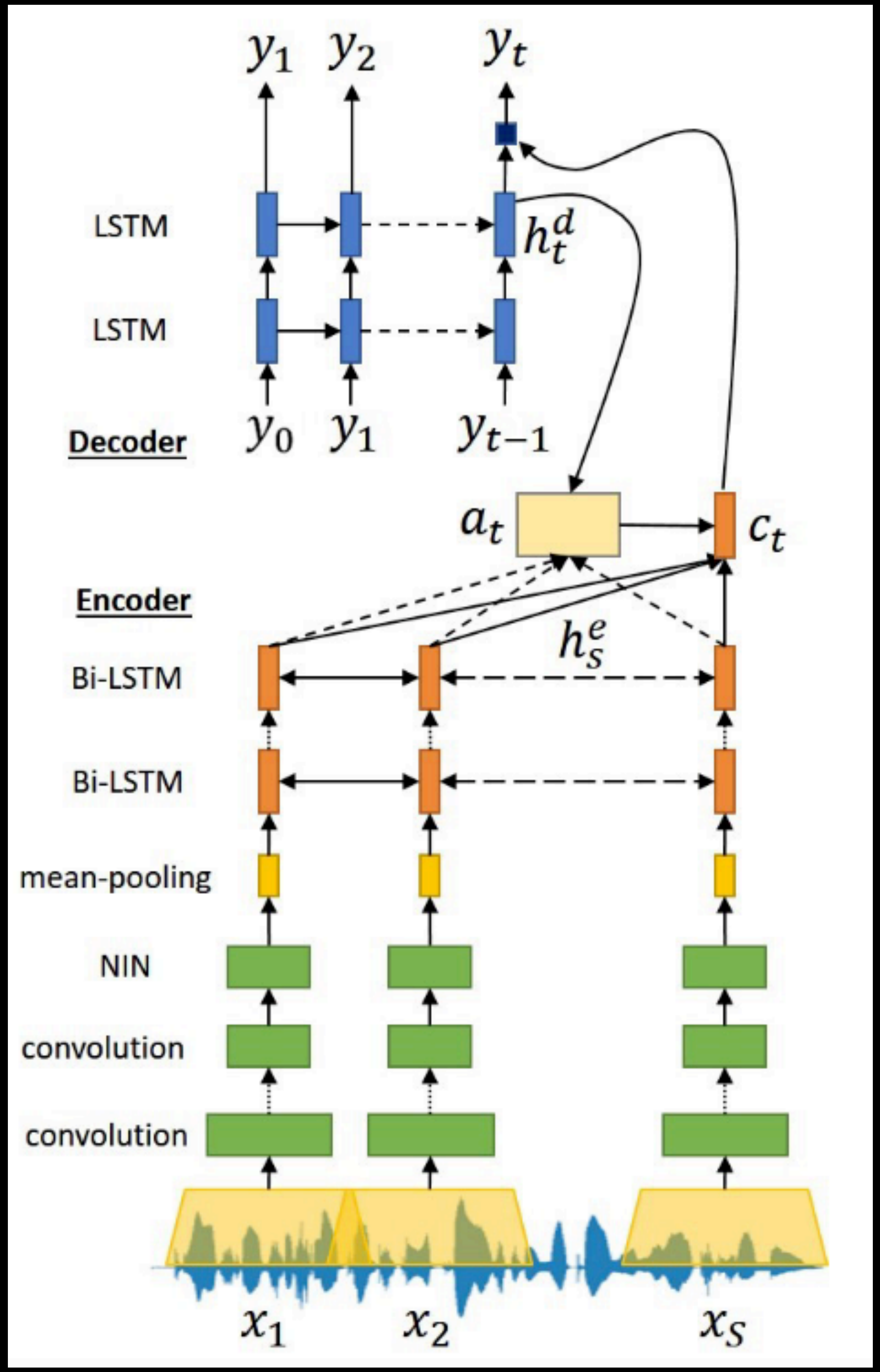
Auxiliary loss	BLEU	Source PER	Target PER
None	0.4	-	-
Source	42.2	5.0	-
Target	42.6	-	20.9
Source + Target	42.7	5.1	20.8
ST [21] $\rightarrow$ TTS cascade	48.7	-	-
Ground truth	74.7	-	-

Raw speech inputs

[Tjandra+2017]

- Can we skip the feature preprocessing step?





Summary

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What’s next?

★Simplification through E2E ASR

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What’s next?

★Simplification through E2E ASR

★Model prosody

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What's next?

★Simplification through E2E ASR

★Model prosody

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What's next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What’s next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

★imitate interpreting strategies (simplification, …)

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What's next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

★imitate interpreting strategies (simplification, …)

★Create more data

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What's next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

★imitate interpreting strategies (simplification, ...)

★Create more data

★Transfer techniques from multilingual & low-resource NMT

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

What's next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

★imitate interpreting strategies (simplification, ...)

★Create more data

★Transfer techniques from multilingual & low-resource NMT

★...

Summary

- Cascaded Models
  - Error propagation (lattices, robustness)
  - Domain mismatch (segmentation, adaptation, disfluencies)
  - Information loss (alignment, markup)
- Simultaneous Translation
  - Segment-based
  - Streaming models
  - Translate & revise
- End-to-end
  - Transcript-free (language preservation)
  - Including ASR/MT corpora, data efficiency

Thanks for your attention

What’s next?

★Simplification through E2E ASR

★Model prosody

★go end-to-end?

★imitate interpreting strategies (simplification, ...)

★Create more data

★Transfer techniques from multilingual & low-resource NMT

★...