NLP 南京大学
自然语言处理实验室
Natural Language Processing Group

Neural Machine Translation with Word Predictions

Rongxiang Weng, Shujian Huang, Zaixiang Zheng, Xinyu Dai and **Jiajun Chen**
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China
`{wengrx, huangsj, zhengzx, daixy, chenjj}@nlp.nju.edu.cn`

Outline

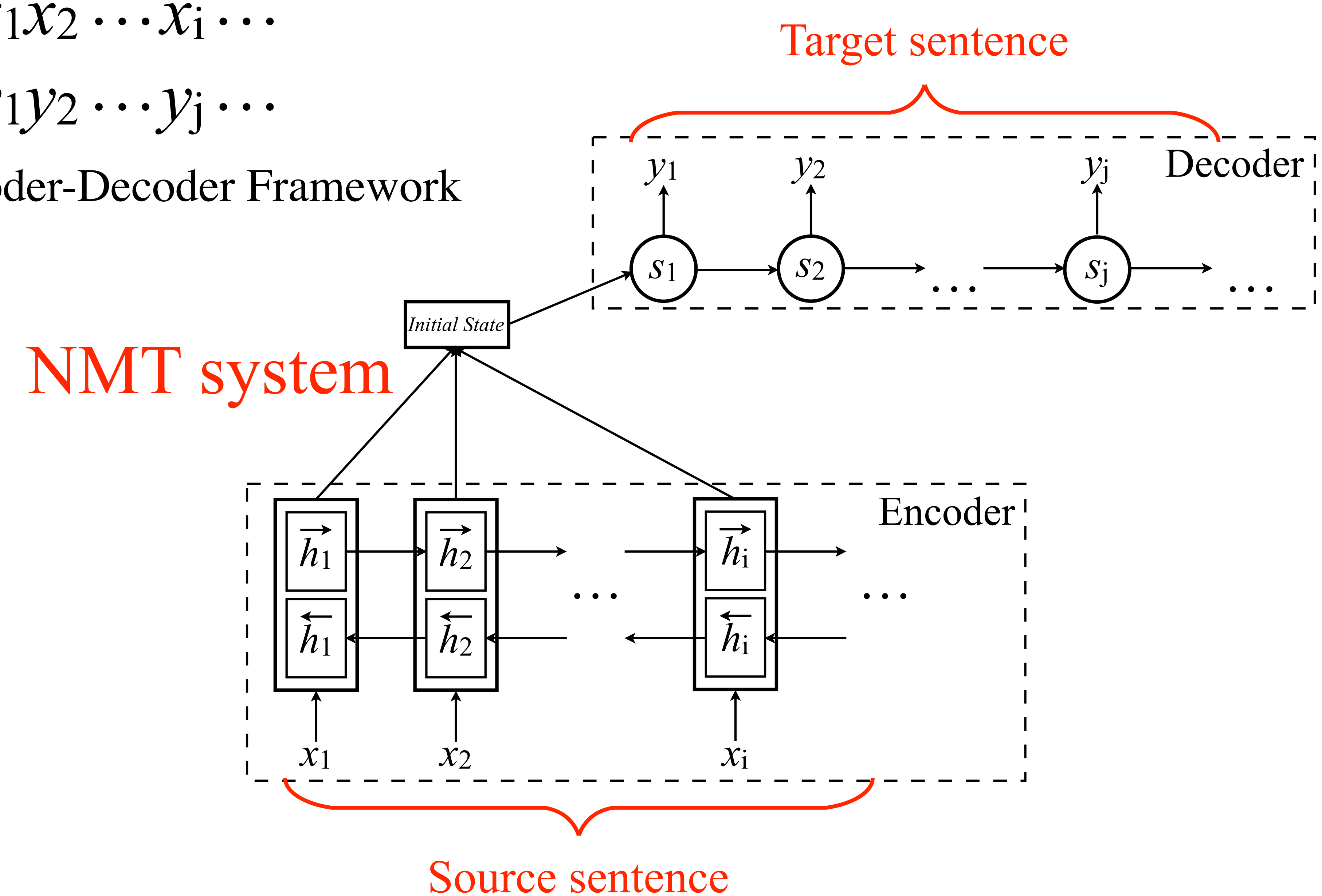
- Background
- Motivation
- Approach
- Experiment
- Conclusion

Outline

- Background
- Motivation
- Approach
- Experiment
- Conclusion

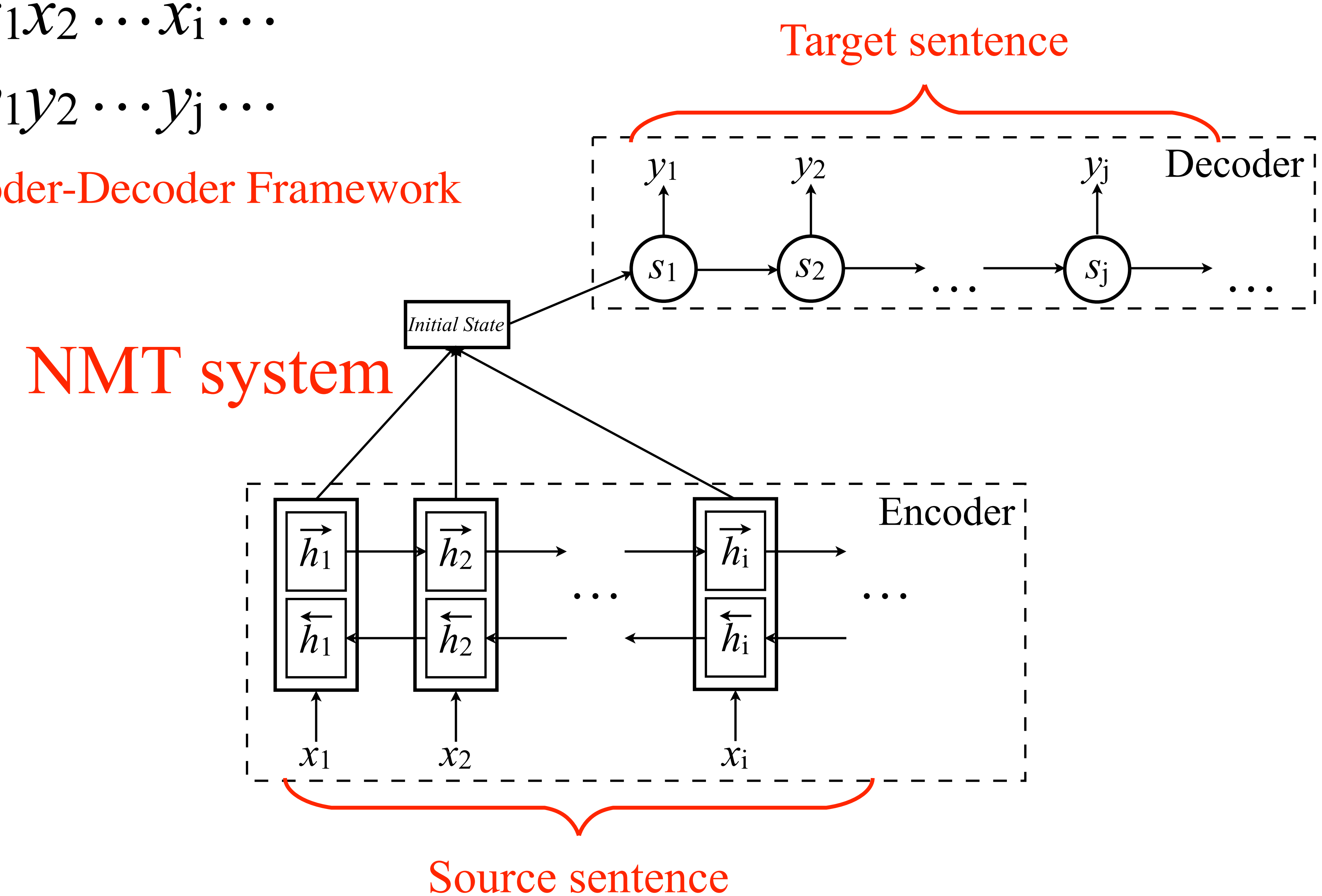
Neural Machine Translation

- Source sentence: $x_1 x_2 \cdots x_i \cdots$
- Target sentence: $y_1 y_2 \cdots y_j \cdots$
- NMT system: Encoder-Decoder Framework



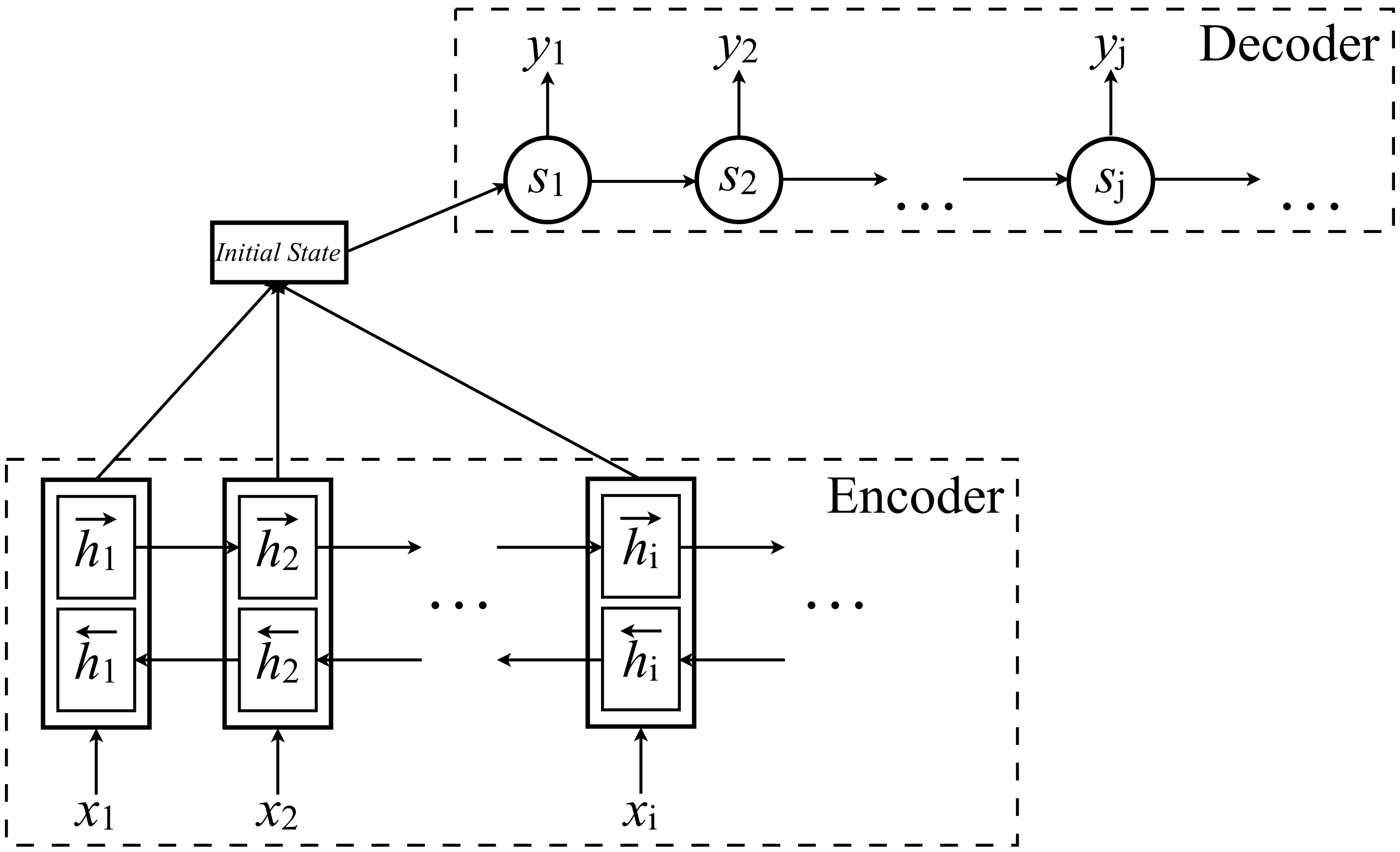
Neural Machine Translation

- Source sentence: $x_1 x_2 \cdots x_i \cdots$
- Target sentence: $y_1 y_2 \cdots y_j \cdots$
- NMT system: Encoder-Decoder Framework



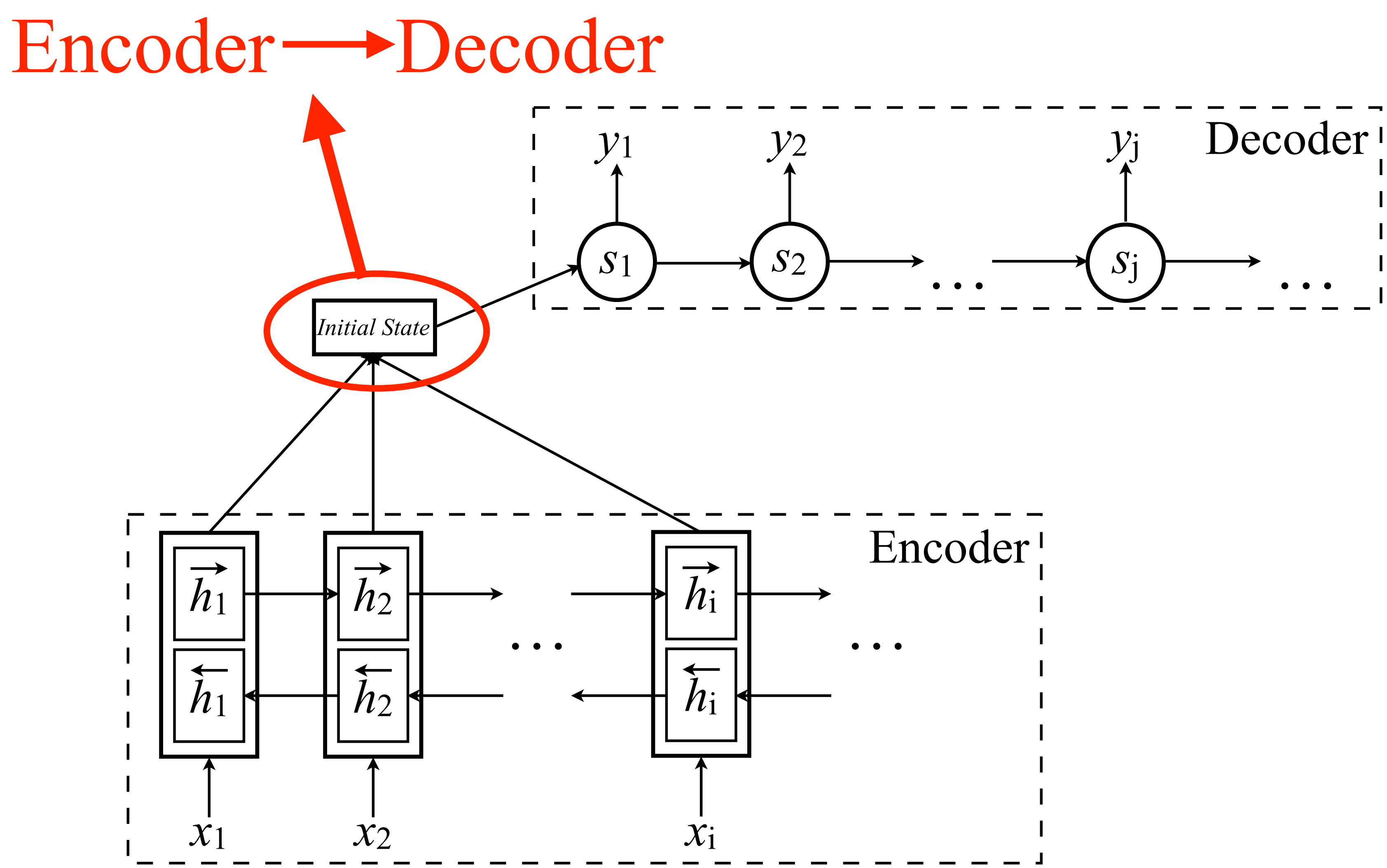
Encoder-Decoder Framework

- Encoder: encode all information of source sentence and generate the Initial State
- Decoder: decode target sentence start from Initial State



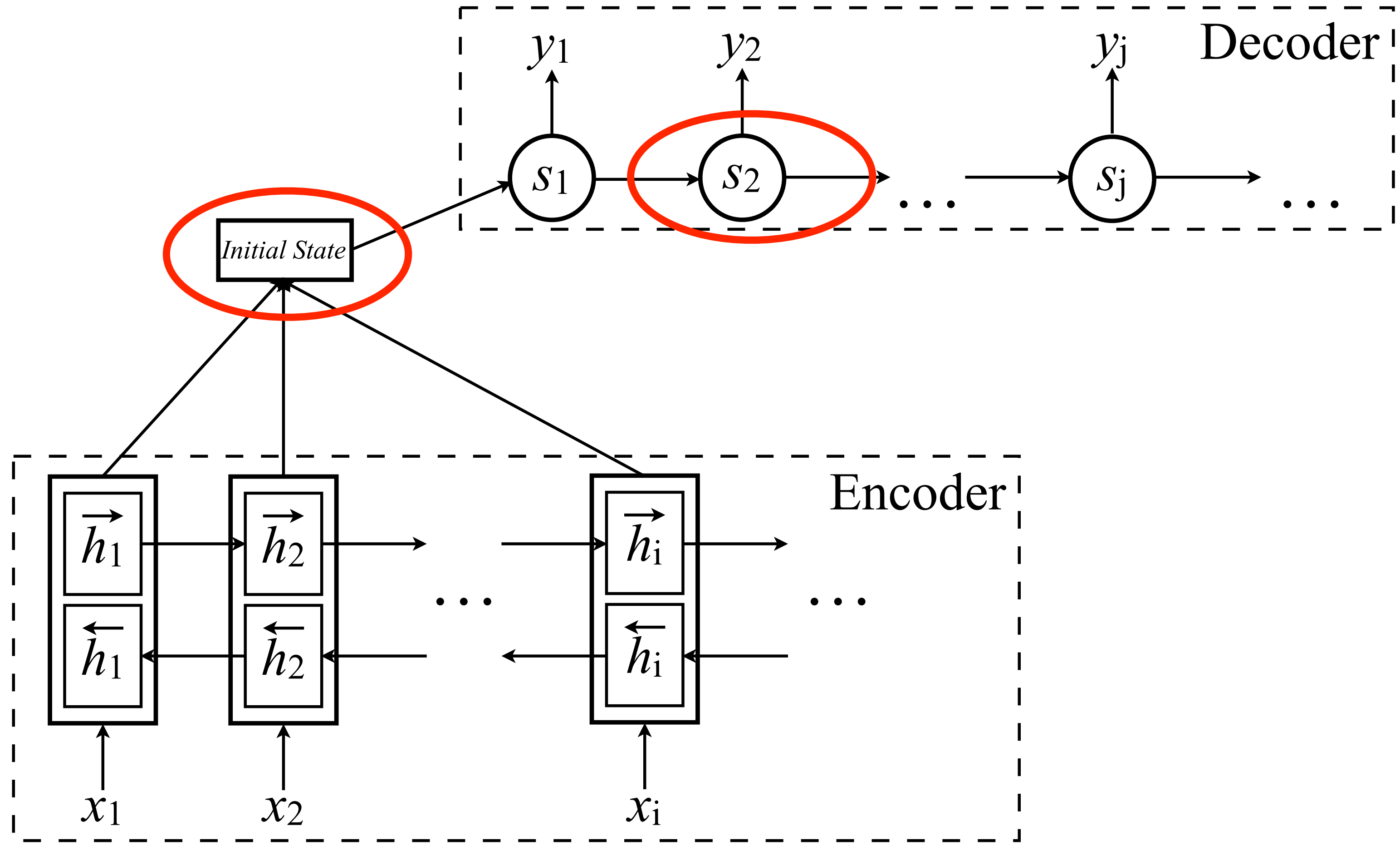
Encoder-Decoder Framework

- Encoder and Decoder are connected by Initial State



Encoder-Decoder Framework

- Initial State has all target information
- Hidden States of Decoder have target information which have not been generated

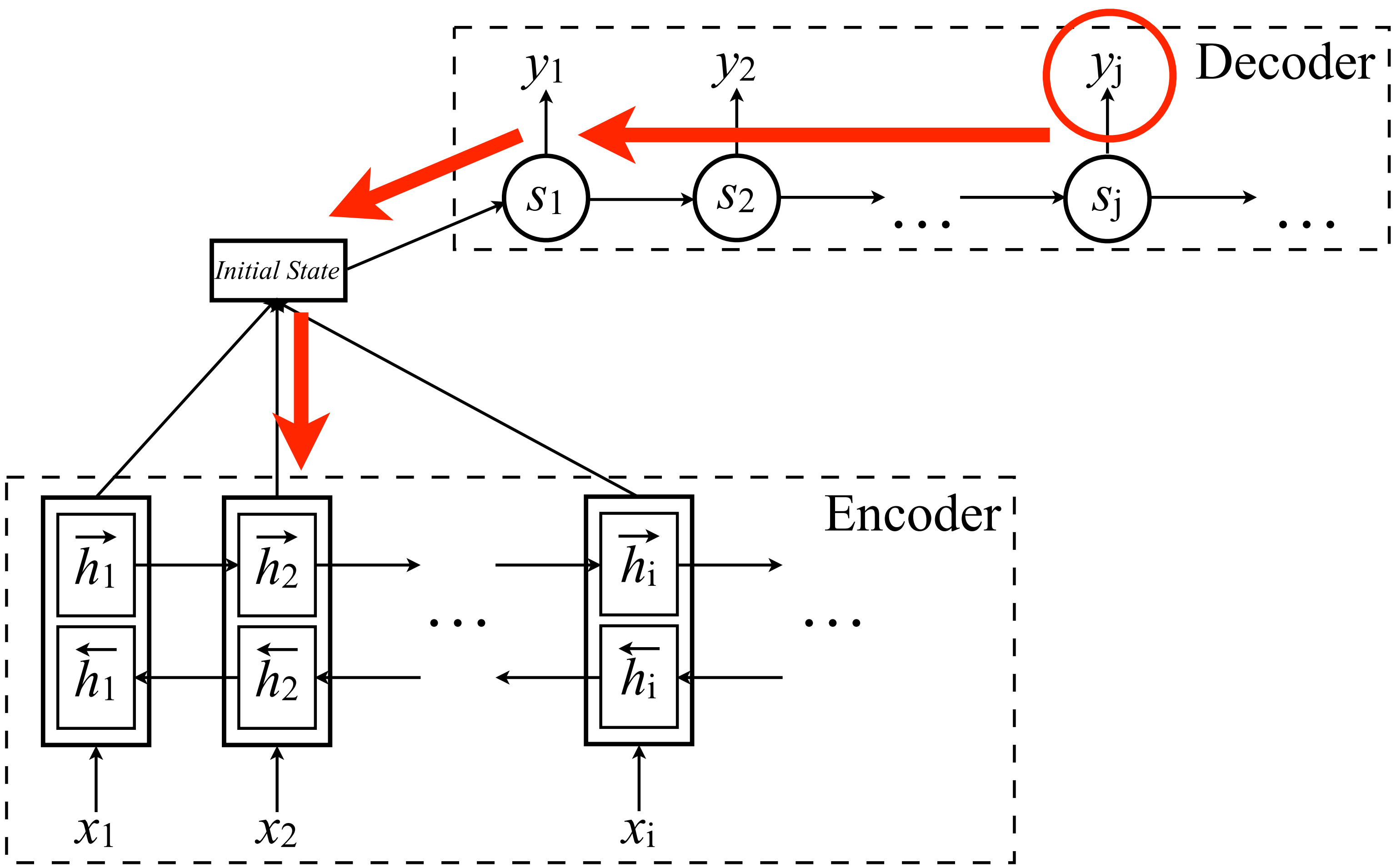


Outline

- Background
- Motivation
- Approach
- Experiment
- Conclusion

Shortage

- Initial State does not have a direct control
- Hidden States of Decoder are just supervised by current word



Motivation

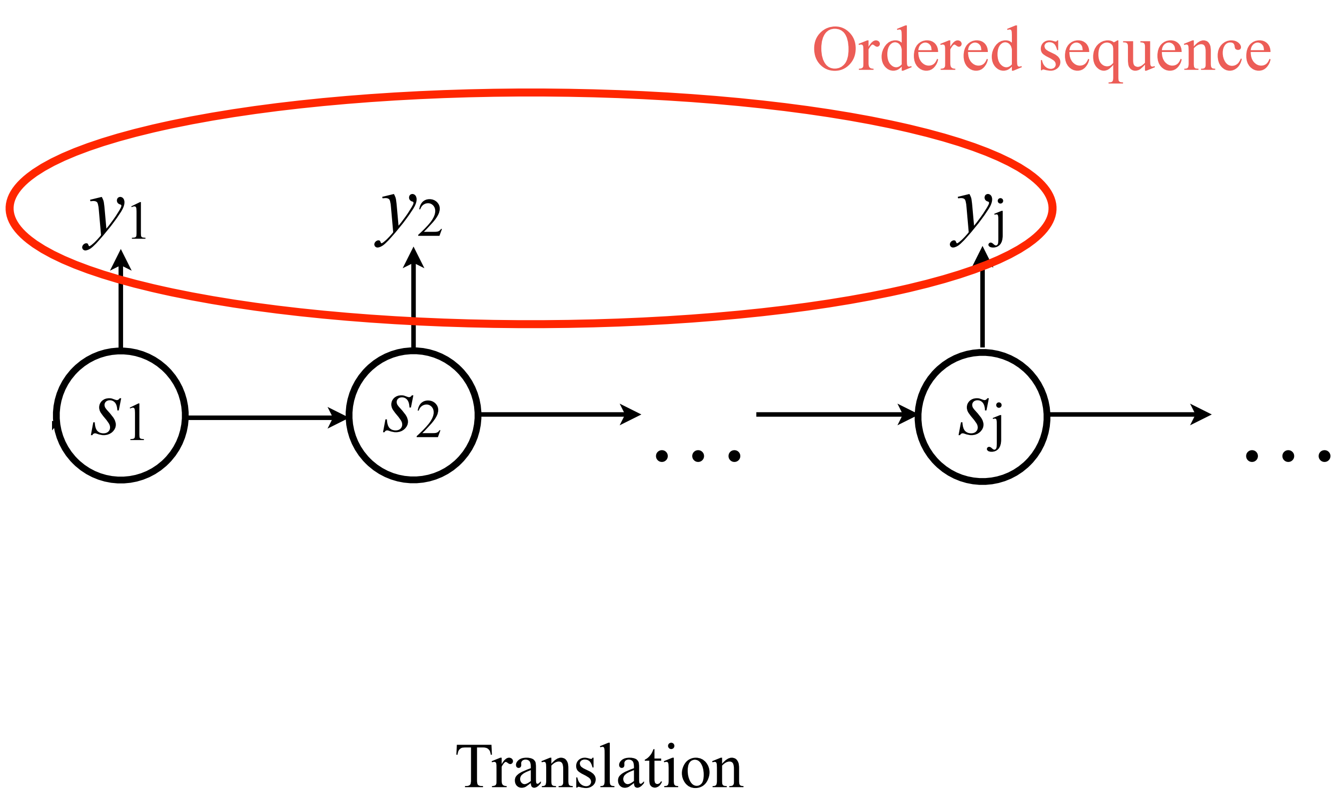
- The initial state and hidden state plays an important role of translation, but it does not have a good control in the currently research
- Propagating translation errors through the end-to-end recurrent structures is not enough of control the hidden states

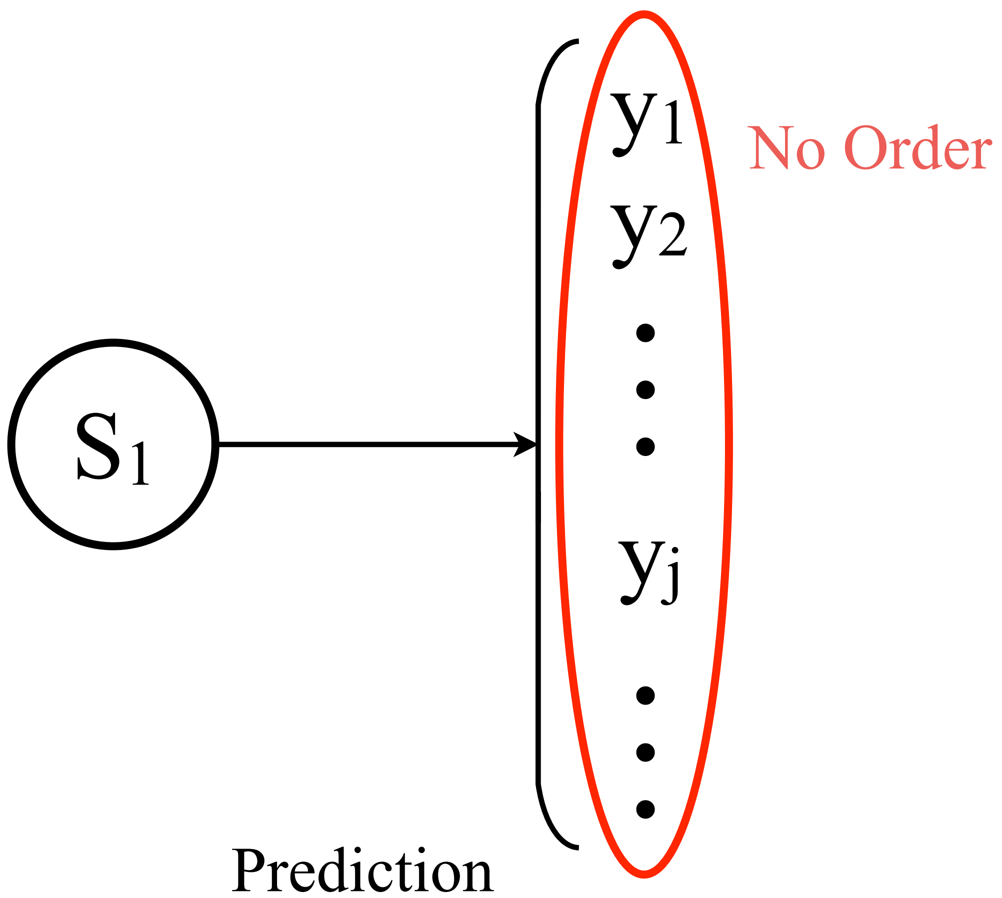
Outline

- Background
- Motivation
- Approach
- Experiment
- Conclusion

Word Prediction

- Translation task to generate an ordered sequence
- The goal of word prediction is to generate several words which is no order





Word Prediction

- Words in the target sentence could be viewed as a natural annotation
- Initial State and Hidden States should contain information about words in target sentence

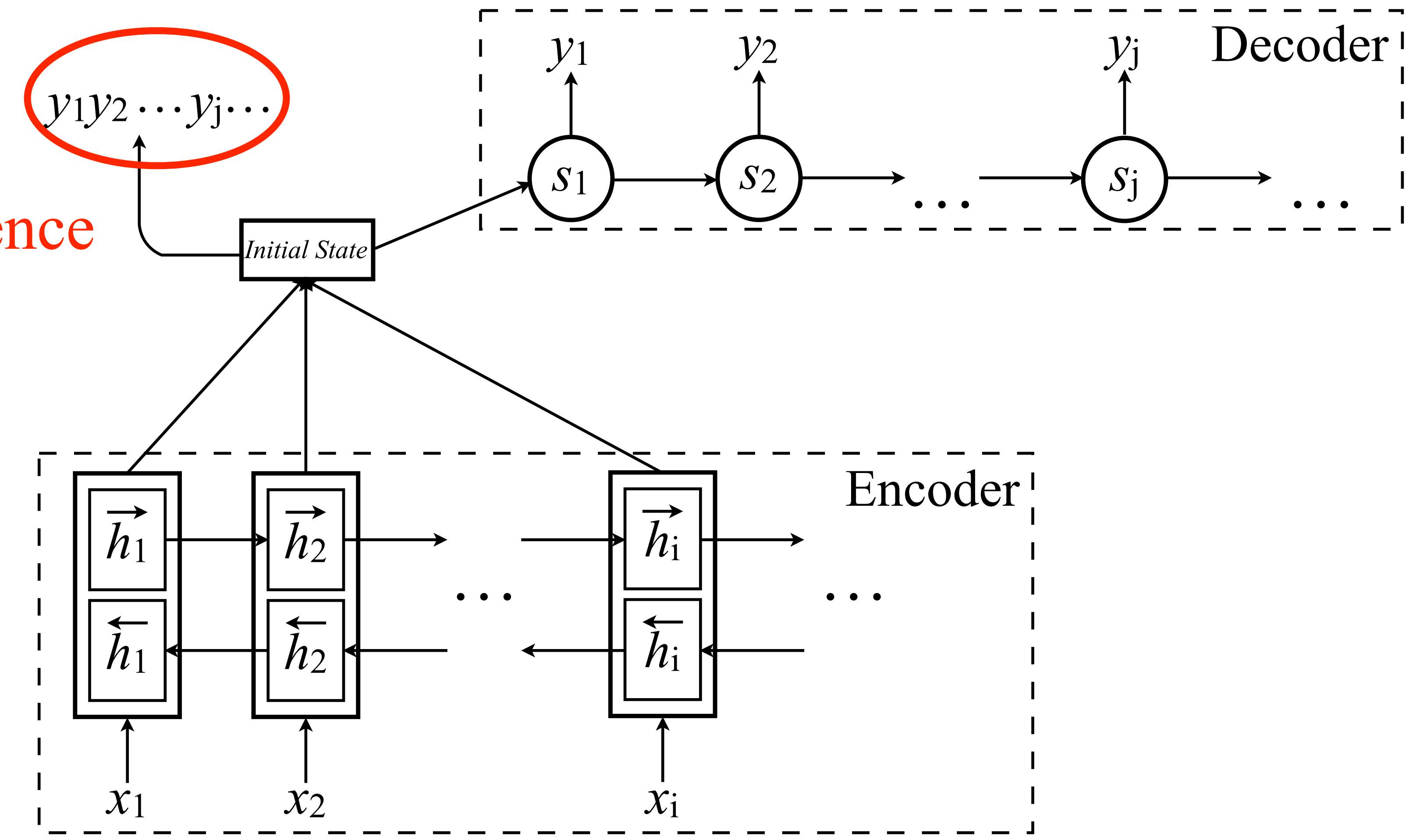
Word Prediction

- For the Initial State (WP_E)
- For Decoder's Hidden States (WP_D)

WP for the Initial State

- Initial State is responsible for the translation of whole target sentence, it should contain information of each word in the target sentence

Predict all words in target sentence

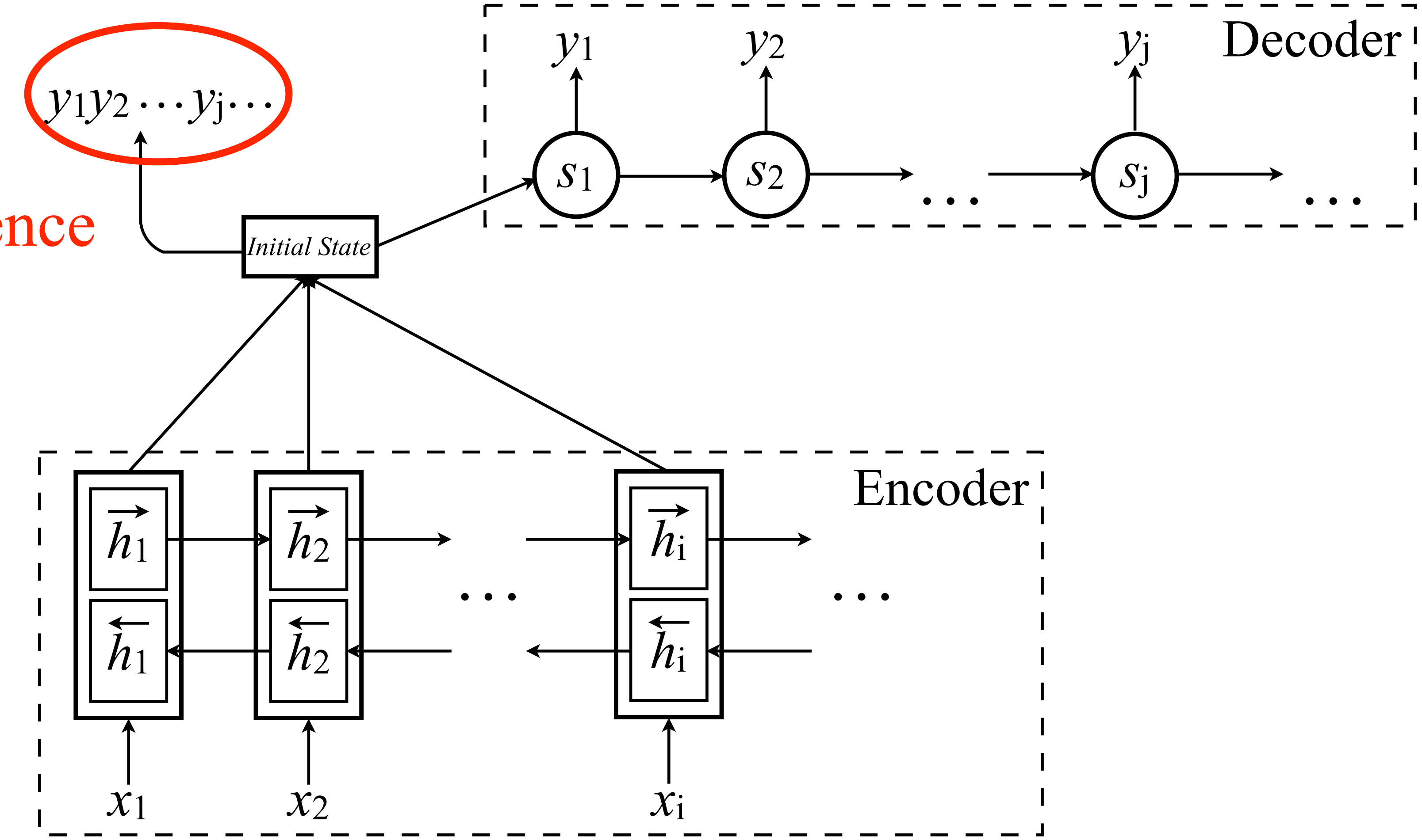


WP for the Initial State

- Initial State is responsible for the translation of whole target sentence, it should contain information of each word in the target sentence

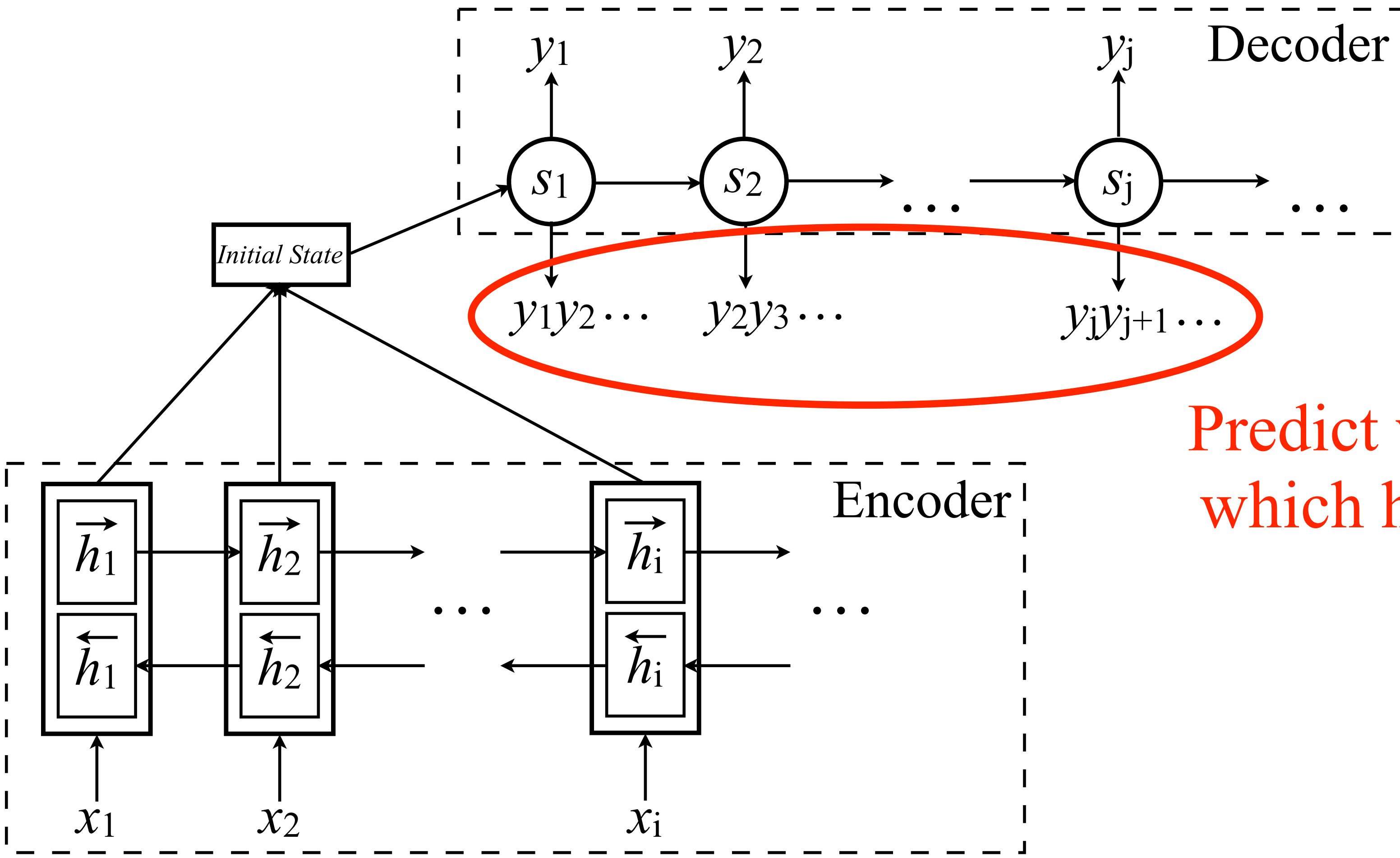
Predict all words in target sentence

$$P_{\mathrm{WP_E}}(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{|\mathbf{y}|} P_{\mathrm{WP_E}}(y_j|\mathbf{x})$$



WP for Decoder's Hidden States

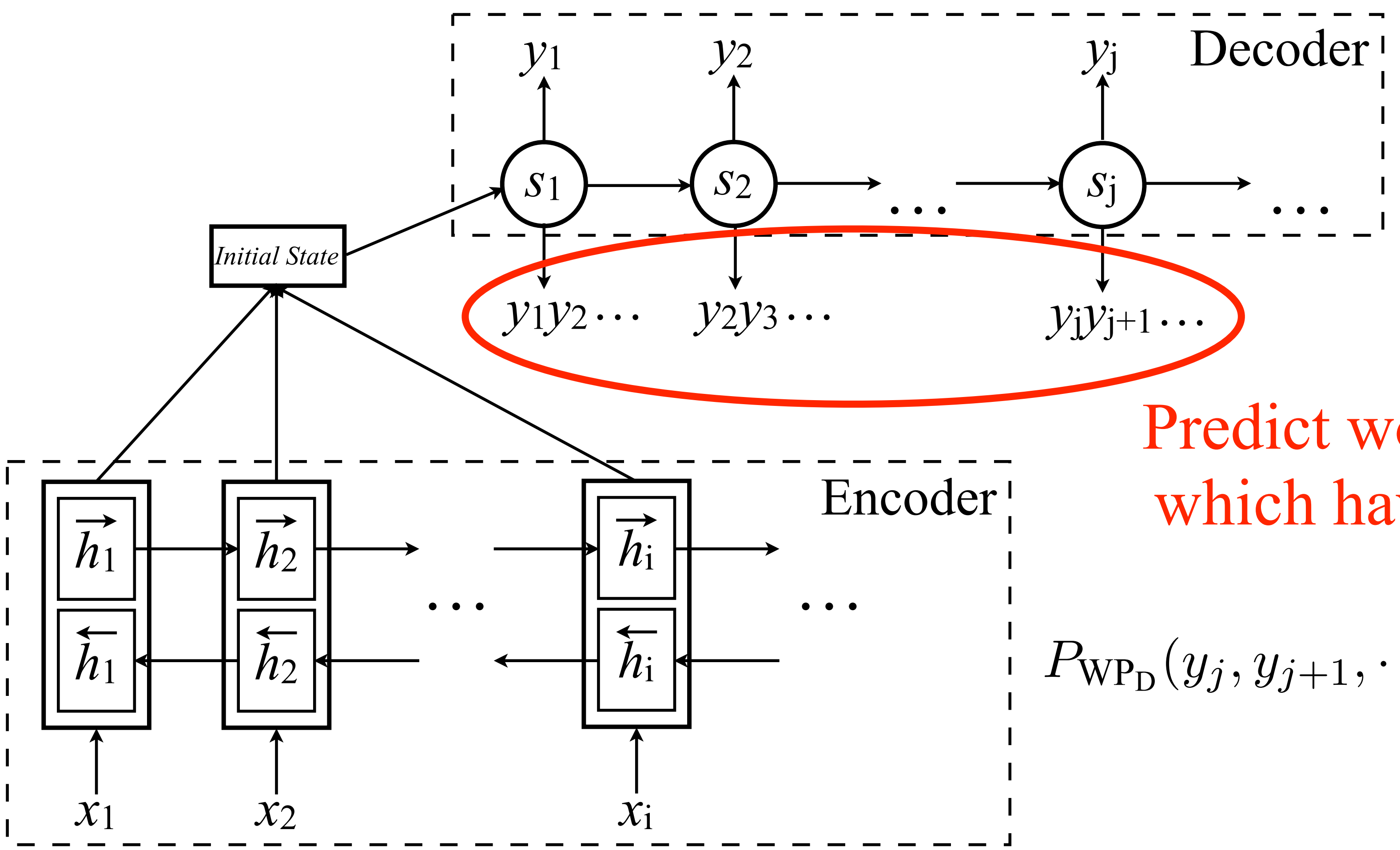
- The hidden states of Decoder are responsible for the translation of target words, and they should contain information of each word which have not been translated



Predict words in target sentence which have not been translated

WP for Decoder's Hidden States

- The hidden states of Decoder are responsible for the translation of target words, and they should contain information of each word which have not been translated



Predict words in target sentence which have not been translated

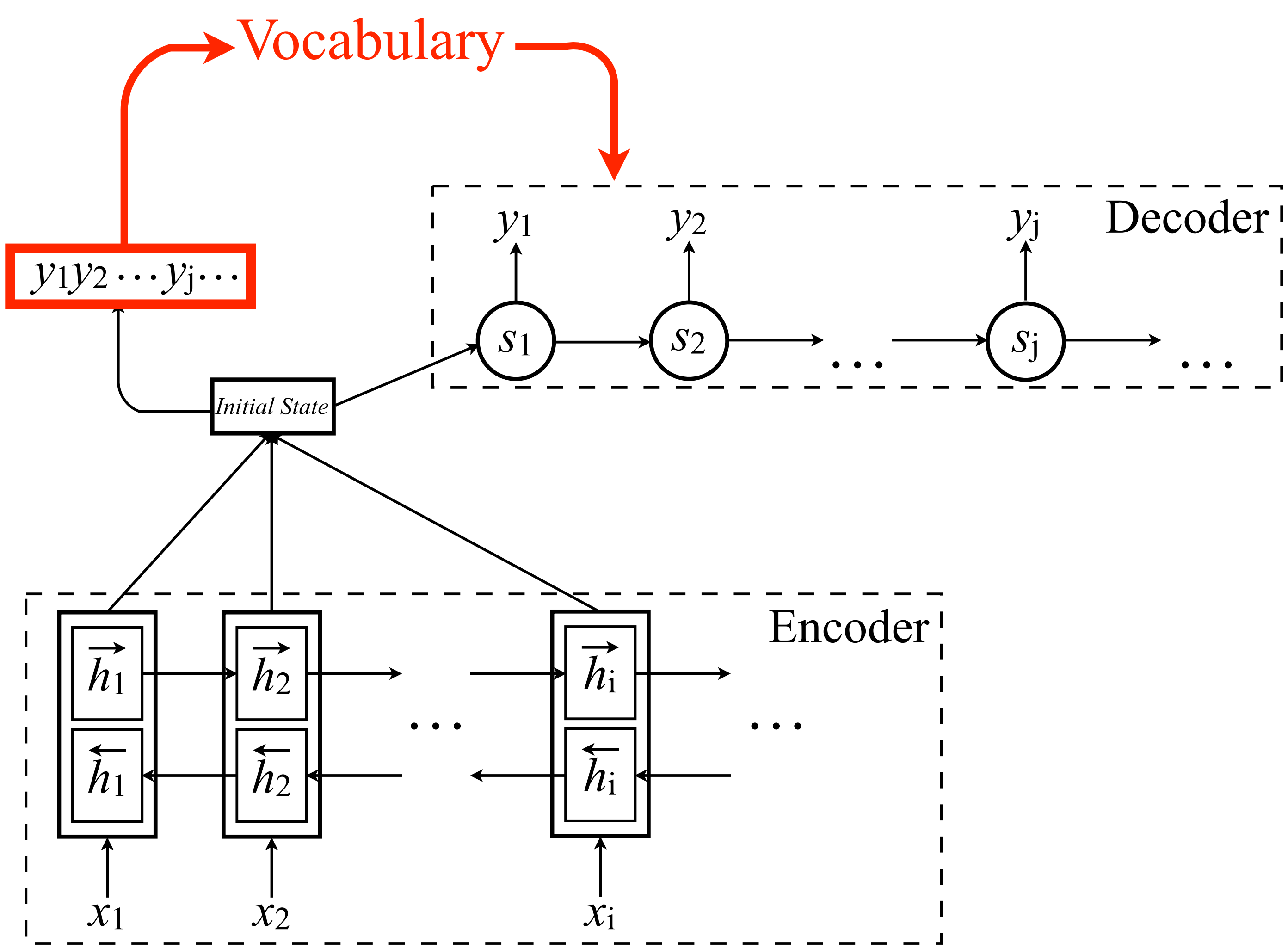
$$P_{\mathrm{WP_D}}(y_j, y_{j+1}, \cdots, y_{|\mathbf{y}|} | y_{<j}, \mathbf{x}) = \prod_{k=j}^{|\mathbf{y}|} P_{\mathrm{WP_D}}(y_k | y_{<j}, \mathbf{x})$$

Make use of word predictor

- Using large vocabulary will reduce decoding efficiency
- Exact small vocabulary will produce better translation effects
- In the testing stage, word prediction mechanism can predict a small vocabulary to decode

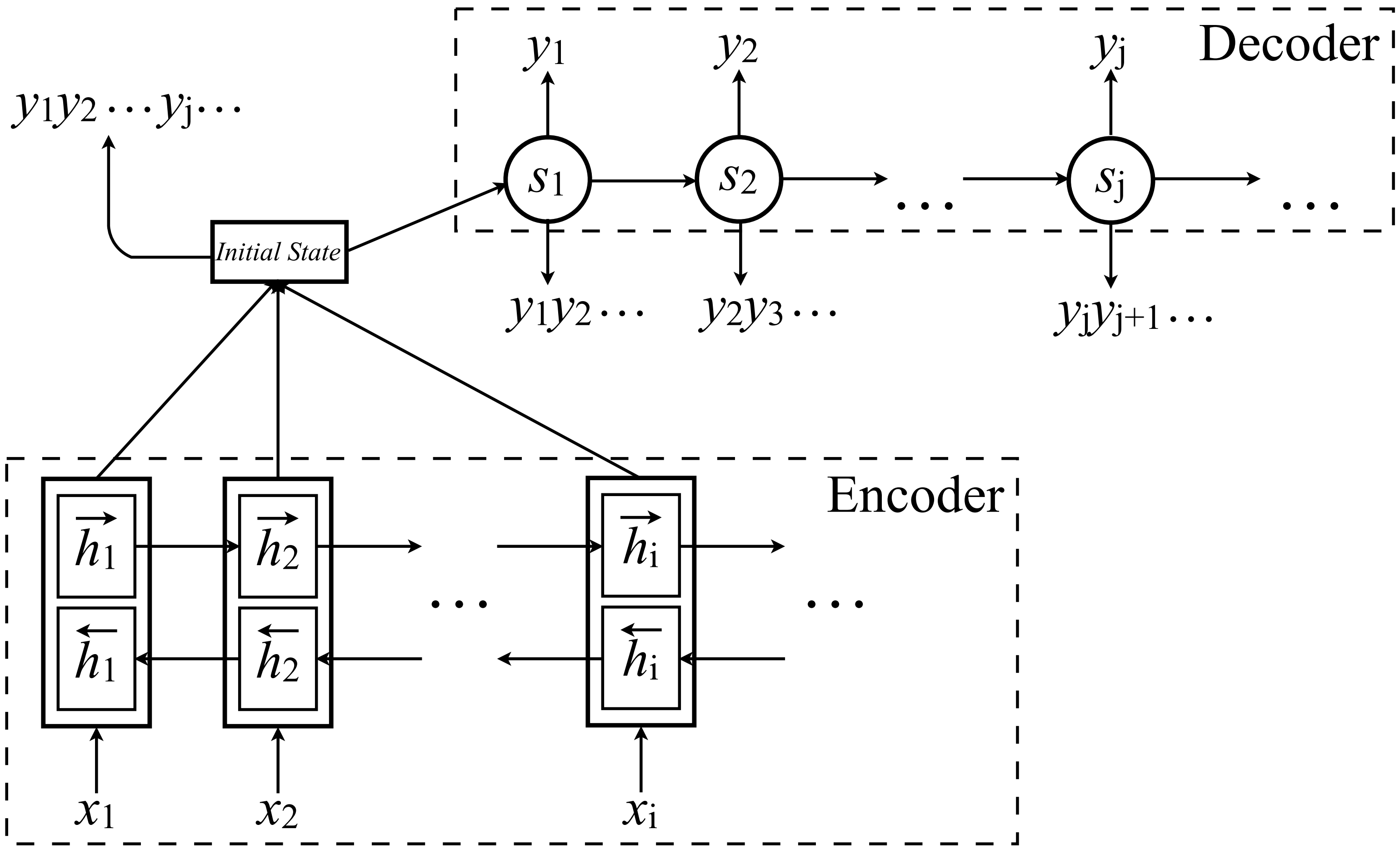
Make use of word predictor

- Predicting top-k words as new vocabulary
- Using the new vocabulary to decode



WP_E+WP_D (WP_{ED})

- Training stage
  - WP_E mechanism
  - WP_D mechanism
- Testing stage
  - WP_E as word predictor



Outline

- Background
- Motivation
- Approach
- Experiment
- Conclusion

Data and Setting

- Chinese-English (CH-EN)
  - 8M LDC data set as training set
  - MT02 as validation set
  - MT03, MT04 and MT05 as test sets
  - Both validation set and test sets have 4 references
- German-English (DE-EN)
  - WMT14 as training set
  - Newstest 2012 as validation set
  - Newstest 2013 as test set
  - Both validation set and test set have 1 reference

Data and Setting

- The source and target vocabularies are limited to the most frequent 30K words
- The out-of-vocabulary words mapped to a special token *UNK*.
- Using *EOS* as the end-of-sentence symbol
- Sentences were encoded using byte-pair encoding (BPE) on DE-EN experiments

Translation quality

- Using WP_{ED} technique improves the baseline by **4.53** BLEU on the CH-EN experiment and **1.3** BLEU on the DE-EN experiment

Models	MT02(dev)	MT03	MT04	MT05	Test Ave.	IMP
baseNMT	34.04	34.92	36.08	33.88	34.96	–
WP_E	39.36	37.17	39.11	36.20	37.49	**+2.53**
WP_D	40.28	38.45	40.99	37.90	39.11	**+4.15**
WP_{ED}	40.25	39.50	40.91	38.05	39.49	$\boxed{\textbf{+4.53}}$

Chinese-English

Models	NST13(dev)	NST14	IMP
baseNMT	23.56	20.68	–
WP_E	24.44	21.09	**+0.41**
WP_D	25.31	21.54	**+0.86**
WP_{ED}	25.97	21.98	**+1.3**

German-English

Translation quality

- Using WP_{ED} technique improves the baseline by **4.53** BLEU on the CH-EN experiment and **1.3** BLEU on the DE-EN experiment

Models	MT02(dev)	MT03	MT04	MT05	Test Ave.	IMP
baseNMT	34.04	34.92	36.08	33.88	34.96	–
WP_E	39.36	37.17	39.11	36.20	37.49	**+2.53**
WP_D	40.28	38.45	40.99	37.90	39.11	**+4.15**
WP_{ED}	40.25	39.50	40.91	38.05	39.49	**+4.53**

Chinese-English

Models	NST13(dev)	NST14	IMP
baseNMT	23.56	20.68	–
WP_E	24.44	21.09	**+0.41**
WP_D	25.31	21.54	**+0.86**
WP_{ED}	25.97	21.98	**+1.3**

German-English

Compare with other techniques

- Along with ensemble method, the improvement could be up to **5.79** BLEU on the CH-EN and **1.79** BLEU on the DE-EN

Models	Test	IMP
baseNMT	34.86	–
$\mathrm{WP}_{\mathrm{ED}}$	39.49	+4.53
baseNMT-dropout	37.02	+2.06
$\mathrm{WP}_{\mathrm{ED}}$-dropout	39.25	+4.29
baseNMT-ensemble(4)	37.71	+2.75
$\mathrm{WP}_{\mathrm{ED}}$-ensemble(4)	40.75	+5.79

Chinese-English

Models	Test	IMP
baseNMT	20.68	–
$\mathrm{WP}_{\mathrm{ED}}$	21.98	+1.3
baseNMT-dropout	21.62	+0.94
$\mathrm{WP}_{\mathrm{ED}}$-dropout	21.71	+1.03
baseNMT-ensemble(4)	21.58	+0.9
$\mathrm{WP}_{\mathrm{ED}}$-ensemble(4)	22.47	+1.79

German-English

Compare with other techniques

- Along with ensemble method, the improvement could be up to **5.79** BLEU on the CH-EN and **1.79** BLEU on the DE-EN

Models	Test	IMP
baseNMT	34.86	–
WP_{ED}	39.49	+4.53
baseNMT-dropout	37.02	+2.06
WP_{ED}-dropout	39.25	+4.29
baseNMT-ensemble(4)	37.71	+2.75
WP_{ED}-ensemble(4)	40.75	$\boxed{+5.79}$

Chinese-English

Models	Test	IMP
baseNMT	20.68	–
WP_{ED}	21.98	+1.3
baseNMT-dropout	21.62	+0.94
WP_{ED}-dropout	21.71	+1.03
baseNMT-ensemble(4)	21.58	+0.9
WP_{ED}-ensemble(4)	22.47	+1.79

German-English

Compare with other techniques

- Along with ensemble method, the improvement could be up to **5.79** BLEU on the CH-EN and **1.79** BLEU on the DE-EN

Models	Test	IMP
baseNMT	34.86	–
WP_{ED}	39.49	+4.53
baseNMT-dropout	37.02	+2.06
WP_{ED}-dropout	39.25	+4.29
baseNMT-ensemble(4)	37.71	+2.75
WP_{ED}-ensemble(4)	40.75	+5.79

Chinese-English

Models	Test	IMP
baseNMT	20.68	–
WP_{ED}	21.98	+1.3
baseNMT-dropout	21.62	+0.94
WP_{ED}-dropout	21.71	+1.03
baseNMT-ensemble(4)	21.58	+0.9
WP_{ED}-ensemble(4)	22.47	$\boxed{+1.79}$

German-English

Precision and Recall

- The initial state in WP_E contains more specific information about target words

top-n	baseNMT		WP_E	
	Prec.	Recall	Prec.	Recall
top-10	45%	17%	73%	30%
top-20	33%	21%	63%	43%
top-50	21%	30%	41%	55%
top-100	14%	39%	28%	68%
top-1k	2%	67%	4%	89%
top-5k	0.7%	84%	0.9%	95%
top-10k	0.4%	90%	0.5%	97%

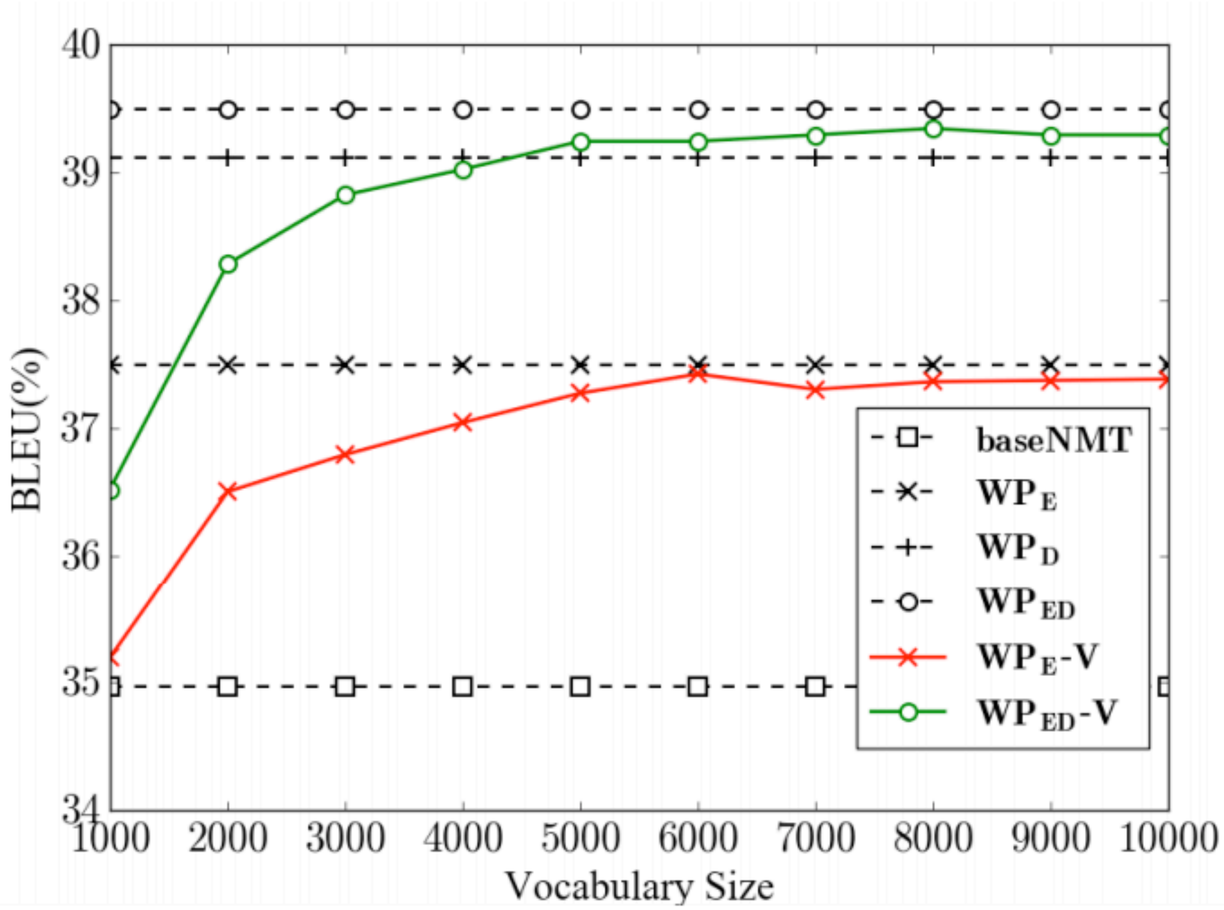
Precision and Recall

- The initial state in WP_E contains more specific information about target words

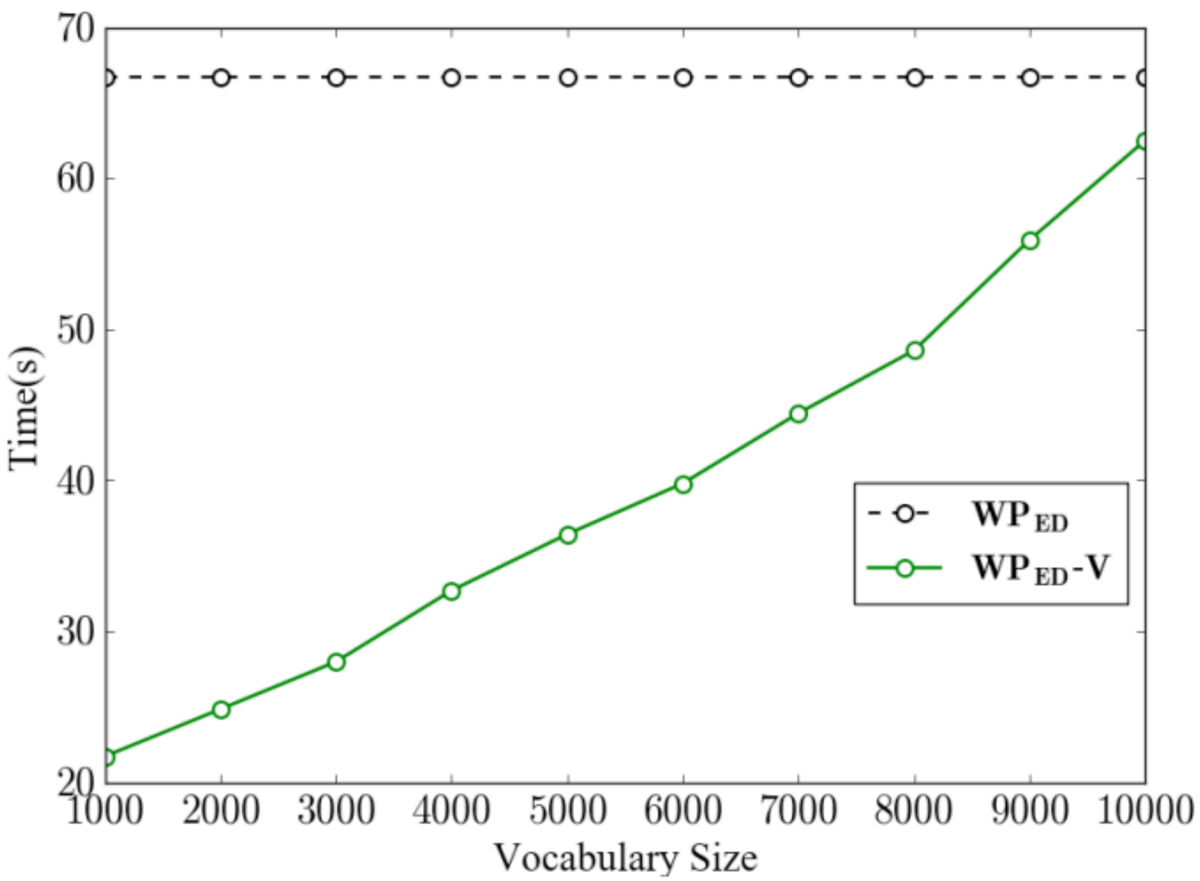
top-n	baseNMT		WP_E	
	Prec.	Recall	Prec.	Recall
top-10	45%	17%	73%	30%
top-20	33%	21%	63%	43%
top-50	21%	30%	41%	55%
top-100	14%	39%	28%	68%
top-1k	2%	67%	4%	89%
top-5k	0.7%	84%	0.9%	95%
top-10k	0.4%	90%	0.5%	97%

Decoding Efficiency

- With a 6k predicted vocabulary, the cost is about 60% of a full-vocabulary; the performance is comparable fixed-vocabulary system



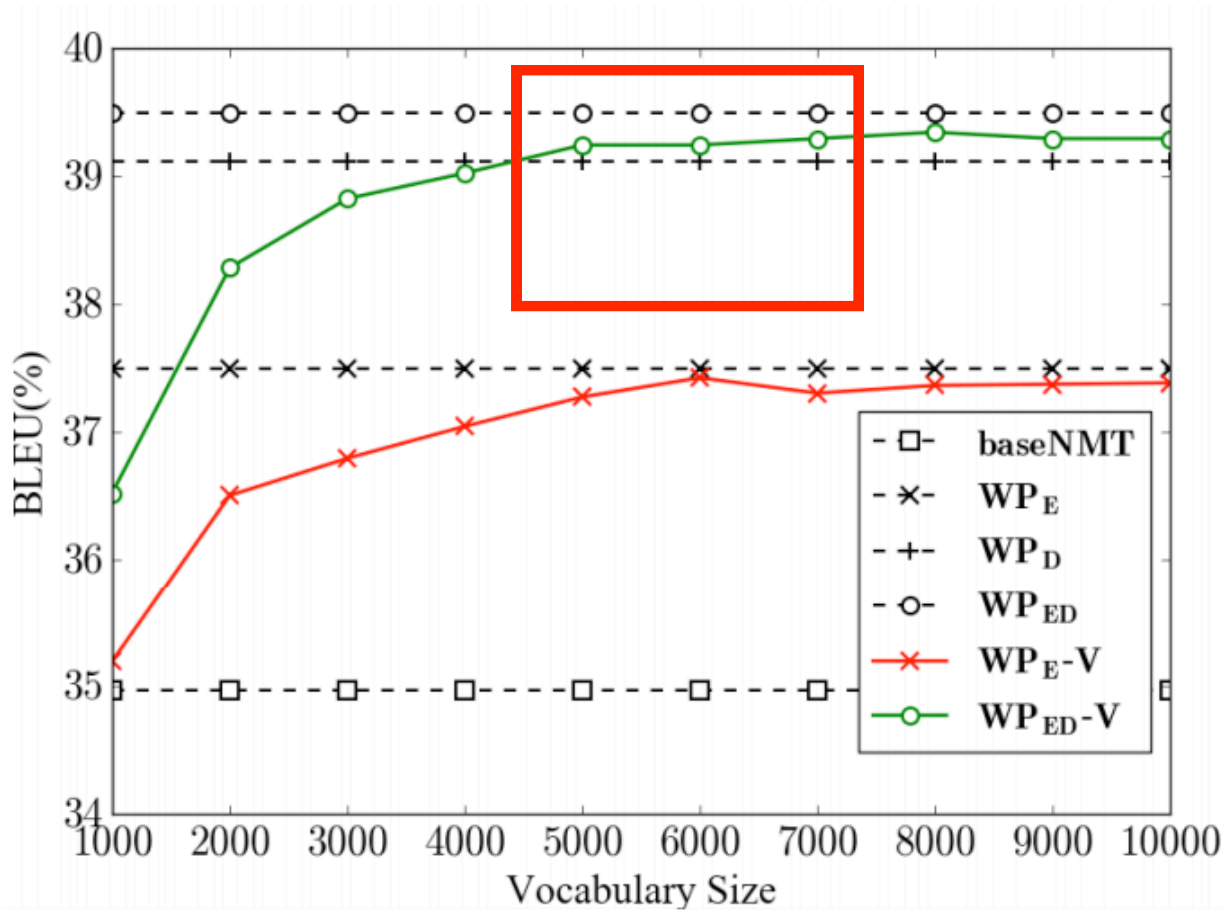
BLEU scores with different vocabulary sizes for each sentence.



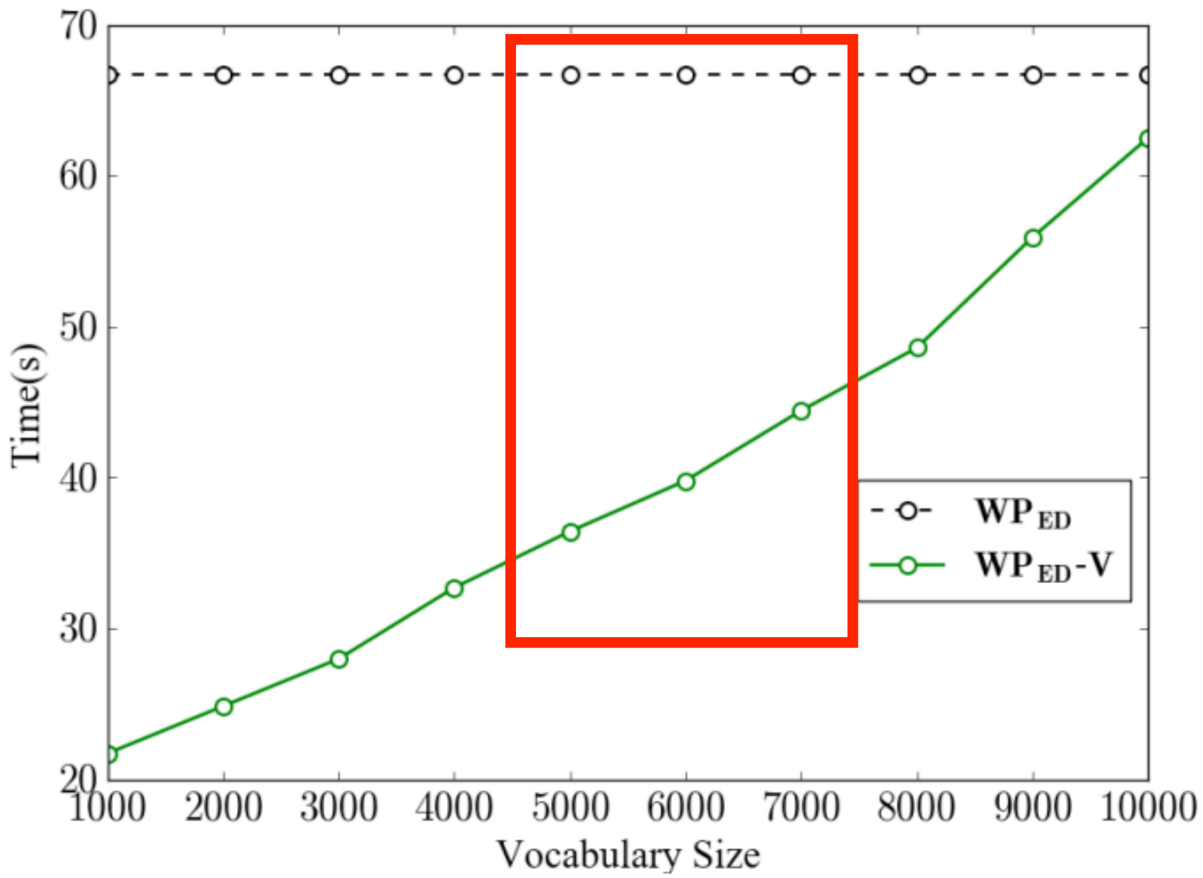
Decoding time with different vocabulary sizes for each sentence.

Decoding Efficiency

- With a 6k predicted vocabulary, the cost is about 60% of a full-vocabulary; the performance is comparable fixed-vocabulary system



BLEU scores with different vocabulary sizes for each sentence.



Decoding time with different vocabulary sizes for each sentence.

Translation Example

- WP_{ED} carries the exact information during translation, most of errors no longer exist

source	时代华纳公司的网络公司美国线上说, 它预期二○○二年的广告与商业销售将由二○○一年的二十七亿美元减少到十五亿美元。
reference	america online , the internet arm of time warner conglomerate , said it expects advertising and commerce revenue to decline from us $ 2.7 billion in 2001 to us $ 1.5 in 2002 .
baseNMT	in the *us line* , *the internet company 's internet company said on the internet* that it expected that the business sales in 2002 would fall from $ UNK billion to $ UNK billion in 2001 .
WP_{ED}	**the internet company** *of* **time warner inc.** , the *us* online , said that it expects that **the advertising and commercial** sales in 2002 will decrease from $ UNK billion in 2001 to us $ **1.5** billion .

Translation Example

- WP_{ED} carries the exact information during translation, most of errors no longer exist

source	时代华纳公司的网络公司美国线上说, 它预期二○○二年的广告与商业销售将由二○○一年的二十七亿美元减少到十五亿美元。
reference	america online , the internet arm of time warner conglomerate , said it expects advertising and commerce revenue to decline from us $ 2.7 billion in 2001 to us $ 1.5 in 2002 .
baseNMT	in the *us line* , *the internet company 's internet company said on the internet* that it expected that the business sales in 2002 would fall from $ UNK billion to $ UNK billion in 2001 .
WP_{ED}	**the internet company** *of* **time warner inc.** , the *us* online , said that it expects that **the advertising and commercial** sales in 2002 will decrease from $ UNK billion in 2001 to us $ **1.5** billion .

Outline

- Background
- Motivation
- Approach
- Experiment
- Conclusion

Conclusion

- The backpropagation provides no direct control of the information carried by the hidden states.

- Word prediction mechanism can enhance the initial state and hidden states of decoder as well.

Thanks