Learning to Generate Word- and Phrase-Embeddings for Efficient Phrase-Based Neural Machine Translation

Chan Young Park, Yulia Tsvetkov
{chanyoun, ytsvetko}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

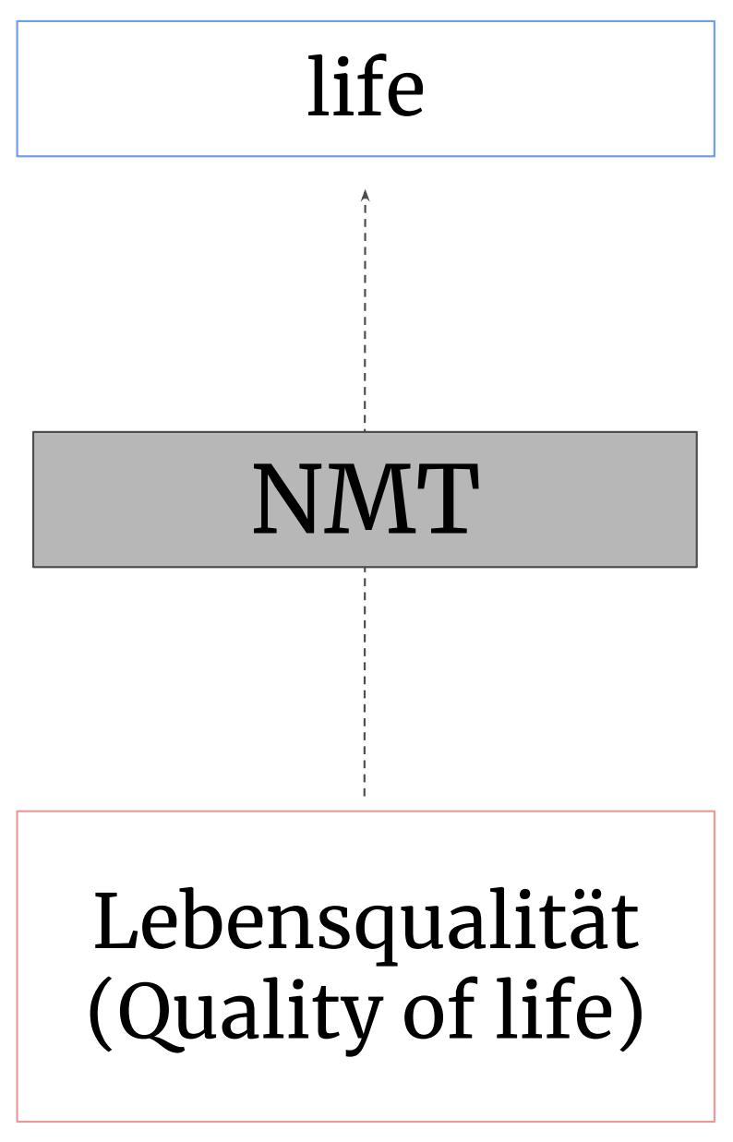
Carnegie Mellon University
Language Technologies Institute

Summary

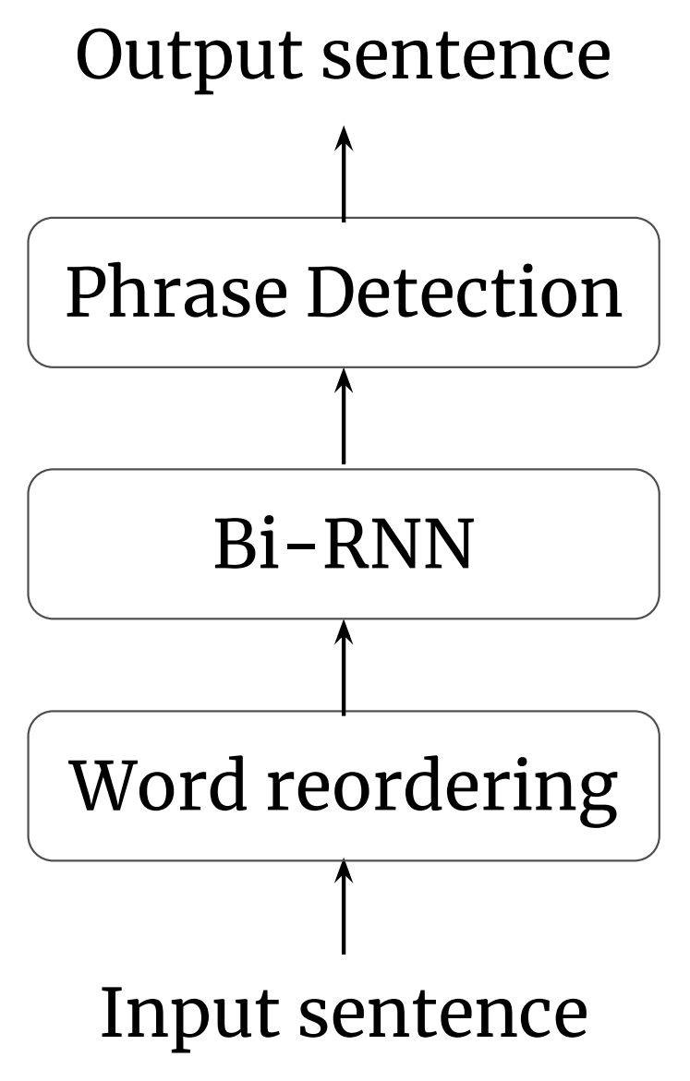
- **Question :**
  How to make the phrase-based NMT systems more efficient?
- **Method:**
  Continuous-output layer + phrase embeddings + fertility
- **Contributions:**
  1) Improve translations by enabling direct word-to-phrase trans.
  2) 112x faster than the state-of-the-art baseline
  3) Proposed to integrate fertility to guide the phrase generation

Motivation

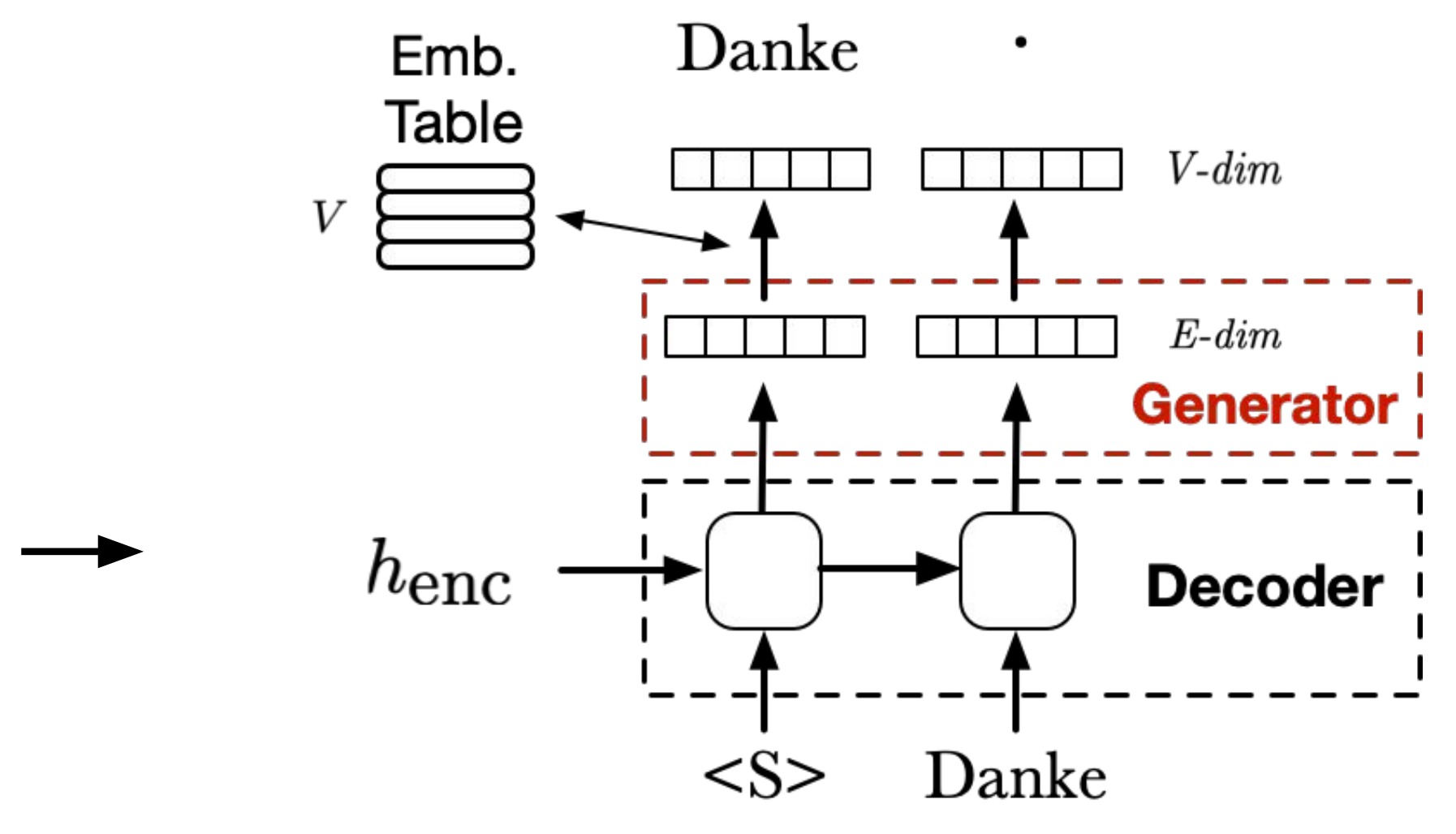
Problem1. Translation of Multi-word Expressions



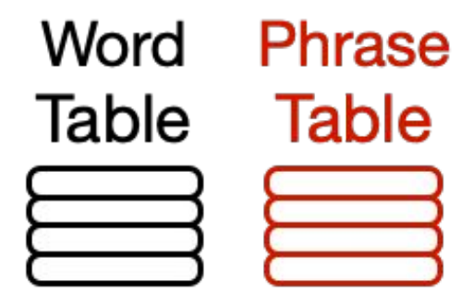
Problem2. Existing PB NMT models are expensive



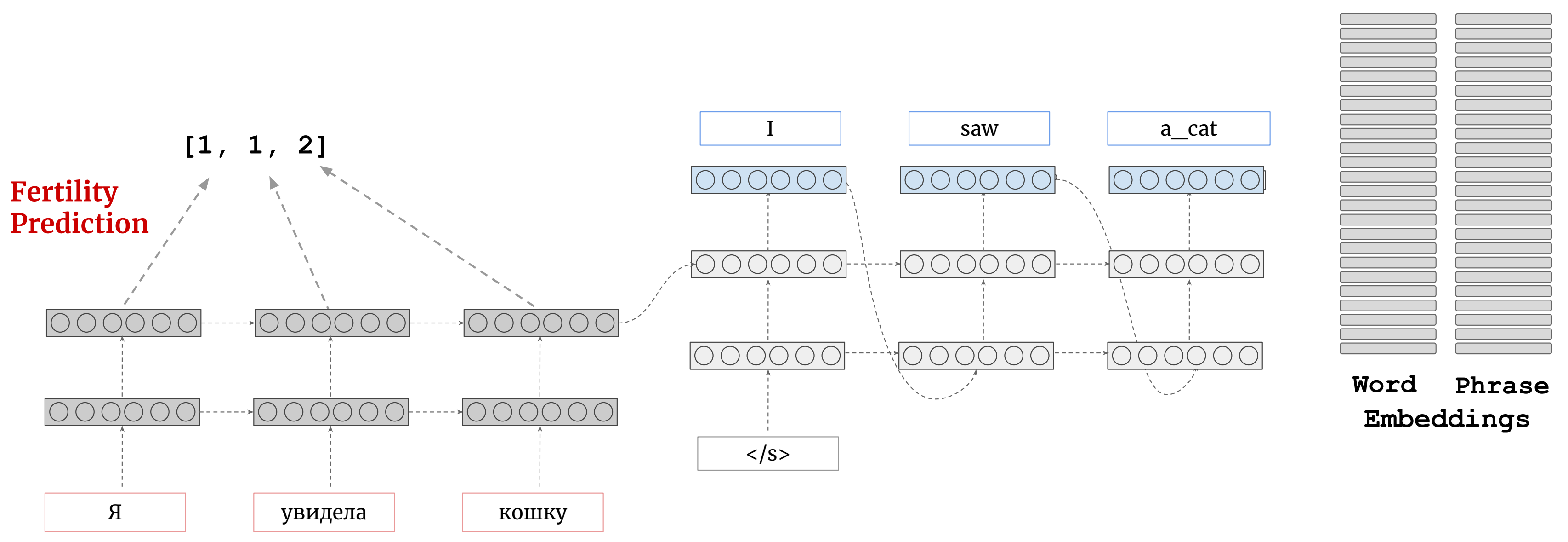
Solution 1. Continuous-output layer

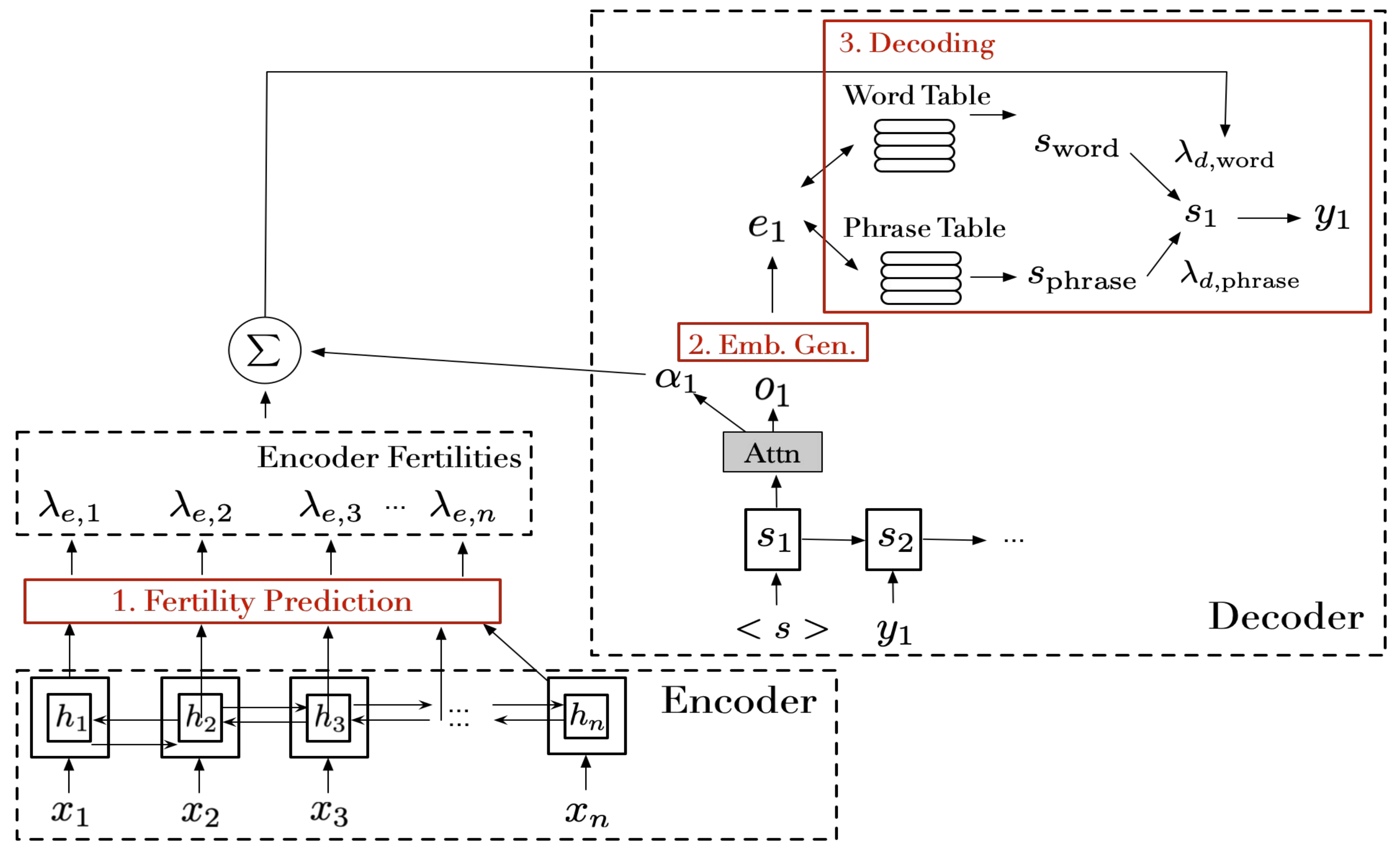


Solution2. Word/Phrase Embeddings



Proposed Model





Phrase-based Continuous-output NMT (PCoNMT)

- **Phrase list**:
  Parallel-corpus + word-alignment model
- **Word/Phrase Embeddings:**
  FastText embedding trained on the concatenated monolingual corpus
- **Fertility:**
  How many words should be generated from each source word

3 steps in decoding

1. **Fertility Prediction**
2. **Continuous-output layer**
3. **Decoding (w/ Fertility scores):**

$$\lambda_{d,\text{word}} = \begin{cases} \sum_e \mathbf{a}_{d,e}\,(\lambda_{e0} + \lambda_{e1}) & (\text{dim} = 4) \\ \sum_e \mathbf{a}_{d,e}\,[\lambda_e]_0 & (\text{dim} = 2) \end{cases}$$

$$\lambda_{d,\text{phrase}} = 1 - \lambda_{d,\text{word}}$$

Results

- **Baselines**
  - Attn (Wiseman & Rush, 2016)
  - NPMT (Huang et al., 2017): the SOTA of PBNMT
  - CoNMT (Kumar & Tsvetkov, 2019)
  - PCoNMT: our model
- **Training Efficiency**

(P)CoNMT: higher speed + faster convergence
→ **112x faster** than the baseline

	speed ↓ (samples/sec)	convergence ↑ (epochs)	total time ↑ (hours)
NPMT	15.4	40	110
CoNMT	256.0	6	1.00
PCoNMT	261.0	6	0.98

- **Translation Quality:**

De-En (full/subset-MWT): 1.4 / **3.9** BLEU ↑
Tr-En (full/subset-MWT): 1.4 / 0.9 BLEU ↑

	De-En		Tr-En	
	IWSLT	IWSLT$_{\text{MWT}}$	WMT	WMT$_{\text{MWT}}$
Attn	23.83			
NPMT	27.27	-	3.58	-
CoNMT	27.07	24.98	7.44	7.67
Our model	**28.69**	**28.89**	**8.87**	7.70
+Fertility$_4$	28.04	24.93	8.12	8.53
+Fertility$_2$	28.29	25.12	8.39	**8.61**

MWT: Subset that contains multi-word tokens

- **Fertility Prediction Eval.**

Highly imbalanced data → F1 not-so-good
Fertility$_2$ > Fertility$_4$

Class	De–En				Tr–En			
	Tot.	P	R	F-1	Tot.	P	R	F-1
$N \leq 1$	97%	0.97	0.96	0.97	97%	0.97	0.95	0.96
$N > 1$	3%	0.33	0.28	0.31	3%	0.17	0.1	0.13

Class	De–En				Tr–En			
	Total	P	R	F-1	Total	P	R	F-1
$N = 0$	10%	0.59	0.09	0.15	14%	0.56	0.30	0.39
$N = 1$	86%	0.88	0.95	0.91	83%	0.86	0.91	0.89
$N = 2$	4%	0.27	0.35	0.31	3%	0.12	0.19	0.14

Analysis of Generated Phrases

Most of the gain is coming from
Collocations and *Compound Nouns*

Category	Total #	PCoNMT	CoNMT	diff.
Compound Nouns	16%	0.63	0.25	+0.38
Verb Phrases	28%	0.5	0.57	-0.07
Collocations	56%	0.71	0.54	+0.17
Sum	100%	0.64	0.50	+0.14

Examples

German src	und Sie sollten auch an Dinge wie **Lebensqualität** denken
English ref	and you also want to think about things like **quality of life**
CoNMT	and you should think of things like **life**
PCoNMT	and you should think of things like **quality_of_life** .
German src	wer ein Gehirn hat , ist **gefährdet** .
English ref	everyone with a brain is **at risk** .
CoNMT	who has a brain is **risk** .
PCoNMT	who has a brain is **at_risk** .
German src	ich stecke **voller** Widersprüche .
English ref	I am **full of** contradictions .
CoNMT	I 'm put .
PCoNMT	I 'm **full_of** contradictions

Future Directions

- Improve the fertility module
- Phrase-to-phrase translations using SWAN
- Code-mixed output generation
- Beam search

Acknowledgement

aws
NSF