Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

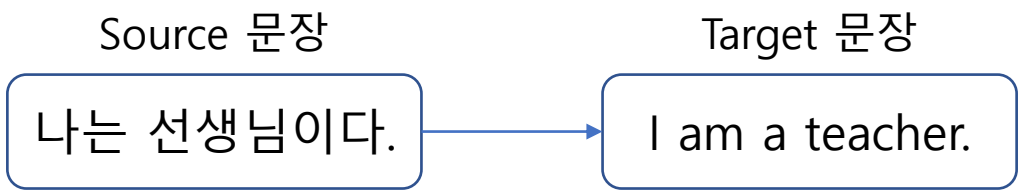
윤예준

https://arxiv.org/abs/1406.1078

Introduction

Translation?

- Source 문장을 Target문장으로 변환하는 것.



Introduction

Statistical Machine Translation?

- 통계적 방법론을 활용하여 번역을 하는 방법.

Statistical Machine Translation 예시 (unigrams, bigrams)

나는 먹는 것을 좋아합니다.

나는 / 먹는 / 것을 / 좋아

I / eat / to / like

$f_1 = p(i|\text{나는})p(eat|\text{먹는})p(to|\text{것을})p(like|\text{좋아})$

나는 먹는 것을 좋아합니다.

나는 먹는 / 것을 좋아

I eat / to like

$f_2 = p(i\ eat|\text{나는 먹는})p(to\ like|\text{것을 좋아})$

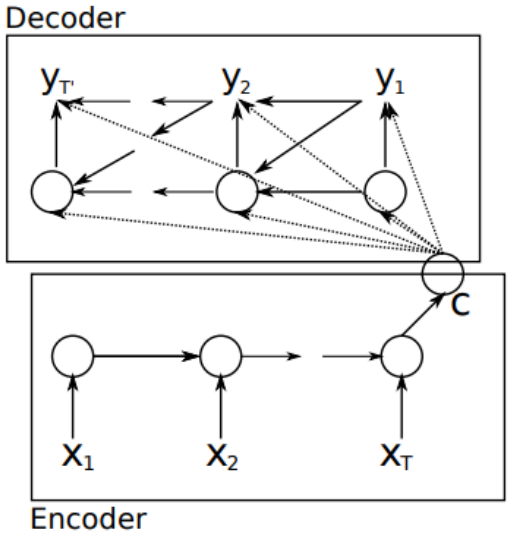
Introduction

Gated Recurrent Unit(GRU)

- 순차적으로 입력(가변적 길이)을 받아서 고정된 크기의 벡터 형태로 나타냄.
- LSTM과 비슷하게 작동하지만 보다 간단한 구조이고 계산량을 줄임.

Sequence-to-Sequence

- 가변적 길이의 Source 문장의 문법적, 의미적 특징을 고정된 크기의 벡터로 나타낼 수 있음.
- 고정된 크기의 벡터로부터 문법적, 의미적 특징을 고려한 가변적인 길이의 Target 문장을 생성할 수 있음.

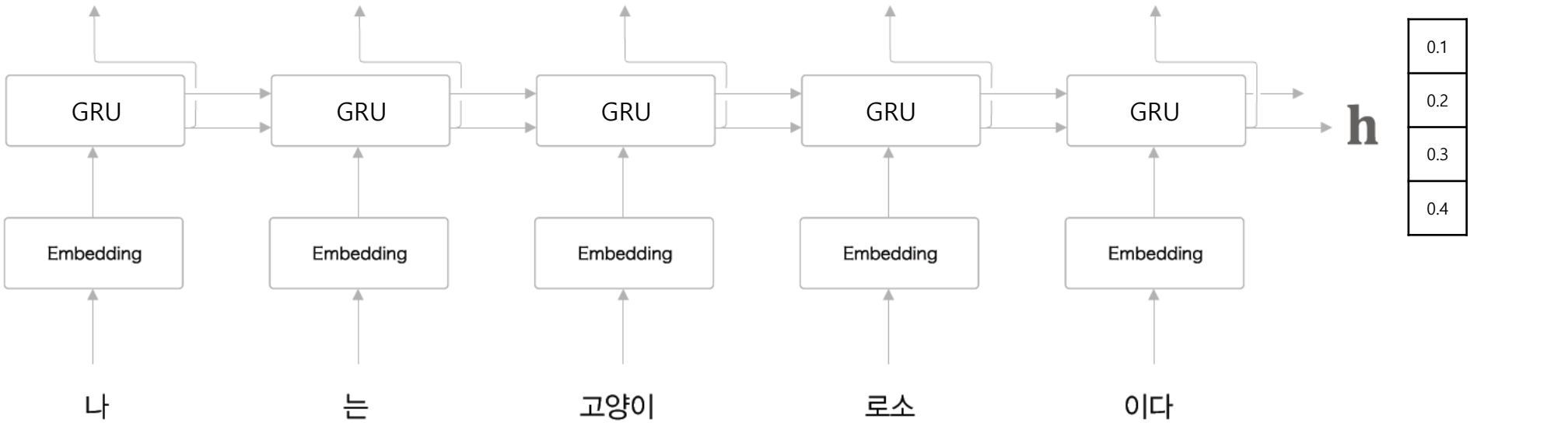


Model

Gated Recurrent Unit(GRU)

- 순차적으로 입력을 받아서 고정된 크기의 벡터 형태로 압축
- 출력은 항상 동일한 크기의 벡터

그림 7-6 Encoder를 구성하는 계층

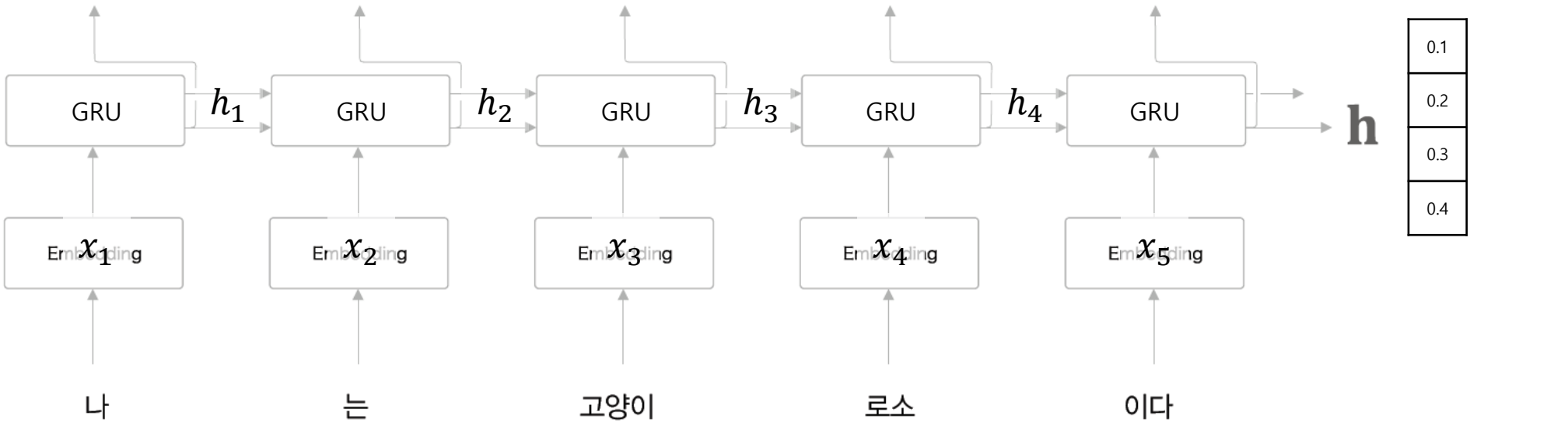


Model

Gated Recurrent Unit(GRU)

- 순차적으로 입력을 받아서 고정된 크기의 벡터 형태로 압축
- 출력은 항상 동일한 크기의 벡터

그림 7-6 Encoder를 구성하는 계층



Model

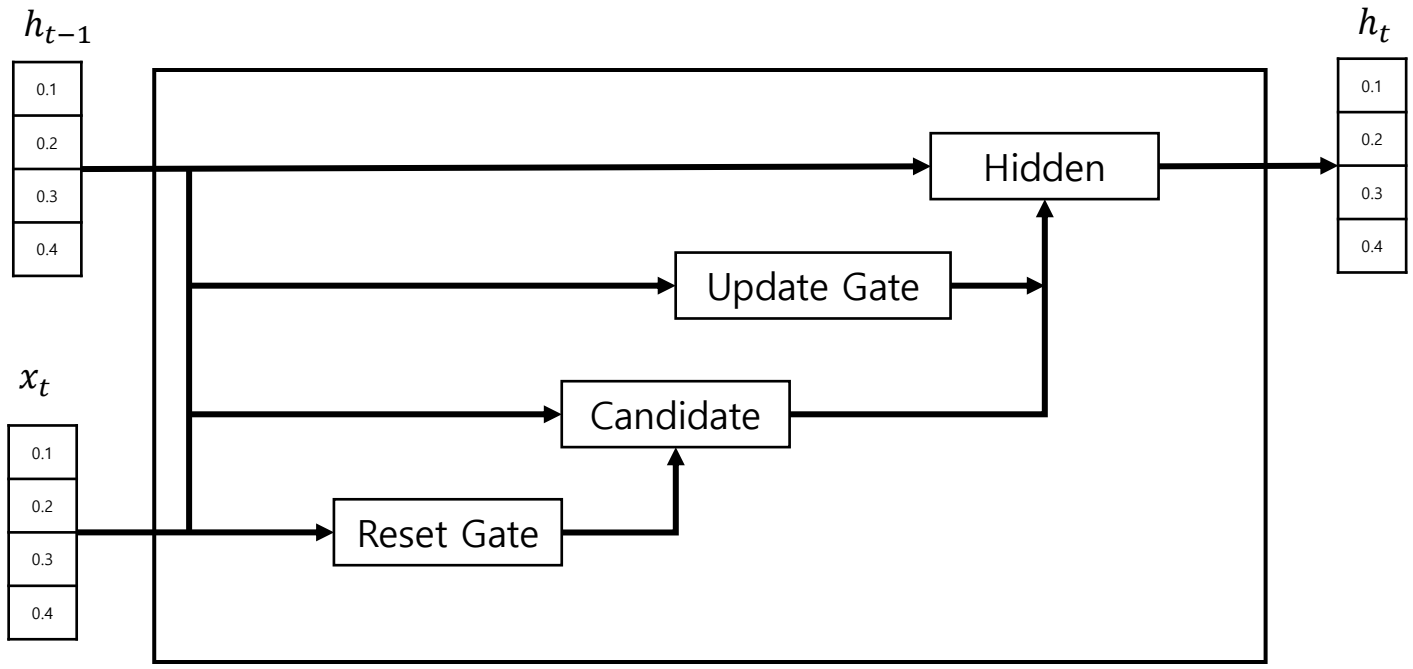
Gated Recurrent Unit(GRU) 아키텍처

$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

$$\tilde{h}_j^{\langle t\rangle} = \phi\left(\left[\mathbf{W}\mathbf{x}\right]_j + \left[\mathbf{U}\left(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle}\right)\right]_j\right)$$

$$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1-z_j)\tilde{h}_j^{\langle t\rangle}$$



Model

Gated Recurrent Unit(GRU) 아키텍처

- Reset Gate : Candidate 계산 과정에서 과거의 정보를 어느정도 제거할지에 대한 값을 도출하는 역할 (0~1 사이의 값으로 이루어진 벡터)

$$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

- Candidate : 현시점의 정보와 Reset Gate를 통해 줄어든 과거 정보를 취합하여 정보 후보군을 계산하는 단계

$$\tilde{h}_j^{\langle t\rangle} = \phi\left(\left[\mathbf{W}\mathbf{x}\right]_j + \left[\mathbf{U}\left(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle}\right)\right]_j\right)$$

Model

Gated Recurrent Unit(GRU) 아키텍처

- Update Gate : Hidden을 계산하는 과정에서 과거의 정보와 현재 정보 결합 비율에 대한 값을 도출하는 역할 (0~1 사이의 값으로 이루어진 벡터)

$$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

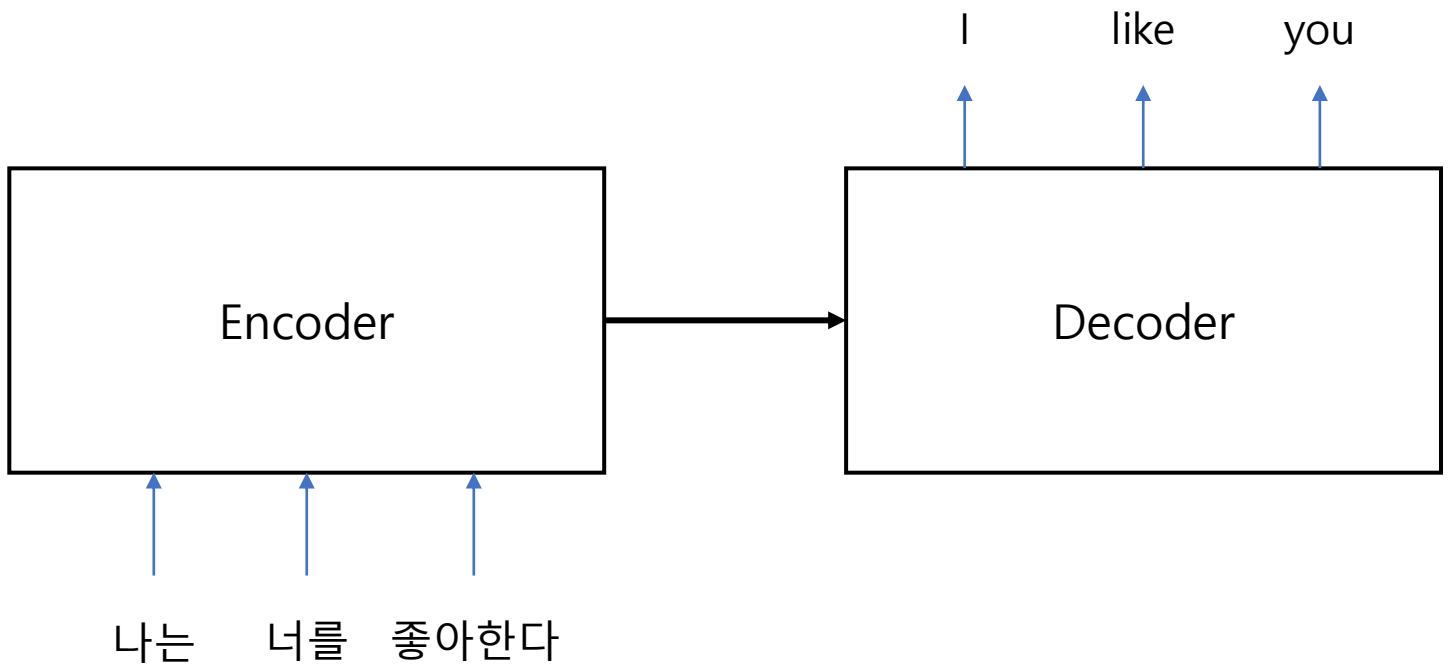
- Hidden : Update Gate를 통해 현재 정보와 과거 정보를 조합하여 GRU의 최정 결과인 hidden vector를 계산하는 과정

$$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1 - z_j)\tilde{h}_j^{\langle t\rangle}$$

Model

Sequence-to-Sequence 아키텍처

- 가변적인 길이의 Source 문장의 문법적, 의미적 특징을 고정된 크기의 벡터로 압축할 수 있음
- 고정된 크기의 벡터로부터 문법적, 의미적 특징을 고려한 가변적인 길이의 Target 문장을 생성할 수 있음.
- Encoder와 Decoder로 구성되어 있음

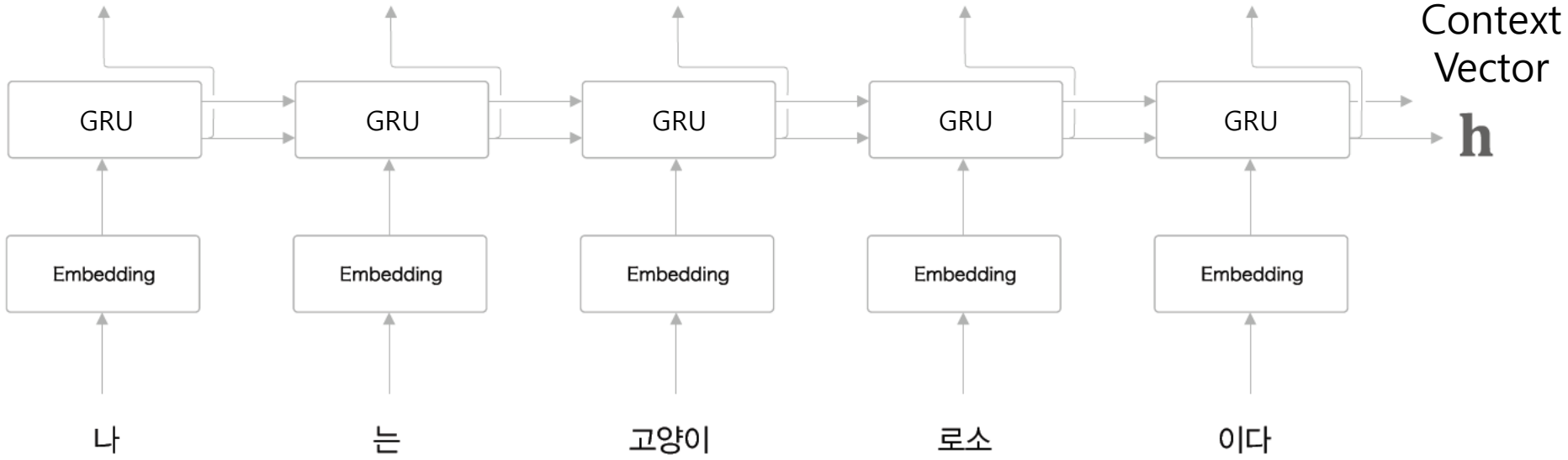


Model

Sequence-to-Sequence 아키텍처

- Encoder
  1) 가변적인 길이의 Source 문장의 문법적, 의미적 특징을 고정된 크기의 벡터로 압축
  2) GRU 아키텍처 활용

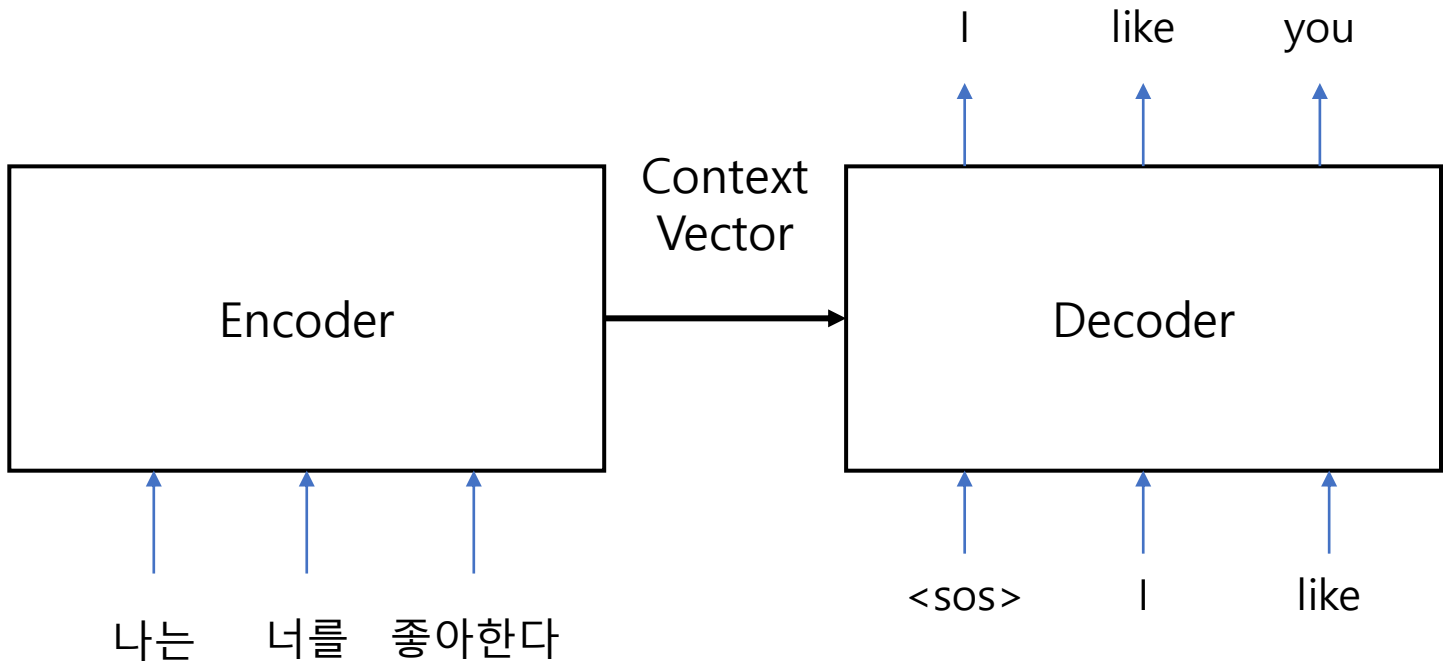
그림 7-6 Encoder를 구성하는 계층



Model

Sequence-to-Sequence 아키텍처

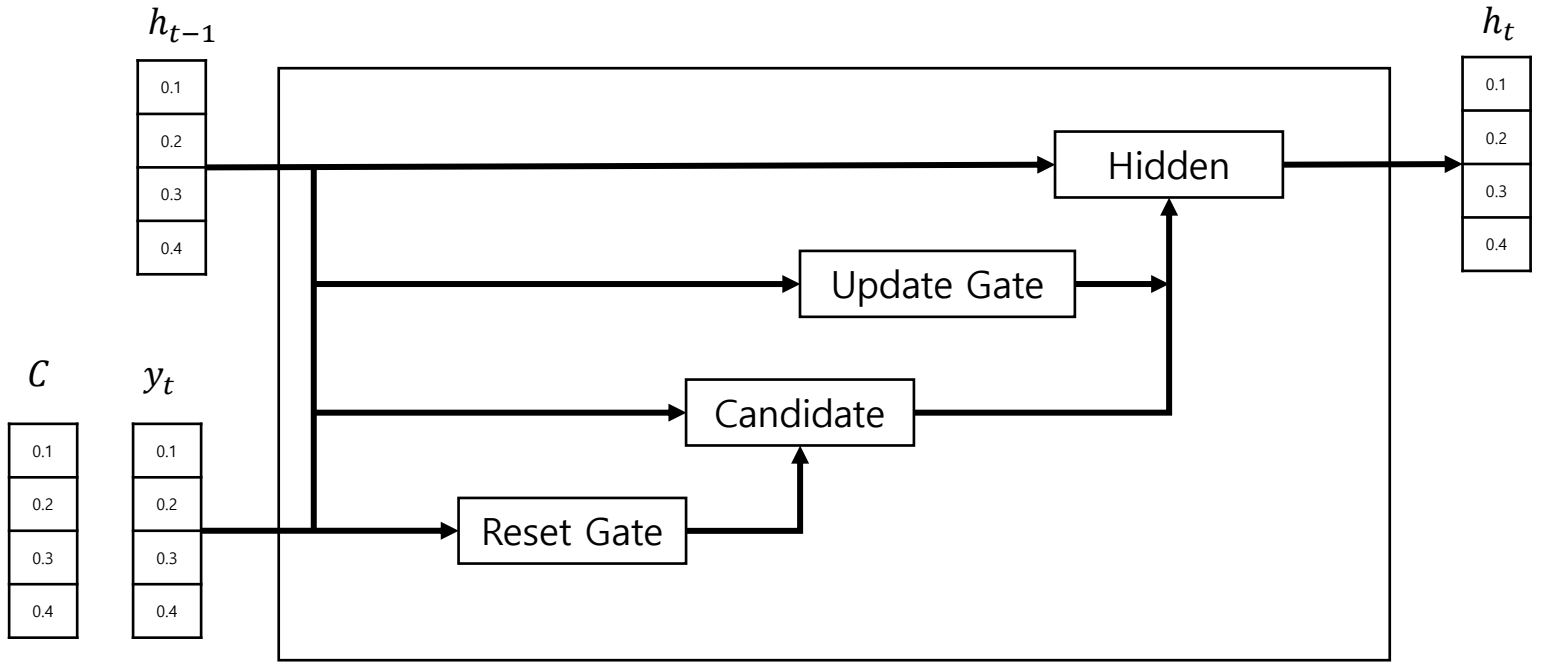
- Decoder
  1) Context Vector를 추가로 입력받는 GRU 아키텍처 활용
  2) Context Vector를 활용하여 Target 문장을 생성하는 역할



Model

Sequence-to-Sequence 아키텍처

- Dncoder GRU 아키텍쳐
  1) 아키텍처는 동일하나 입력으로 Context Vector 추가됨.



$$z = \sigma(Wx + Uh + Cc)$$

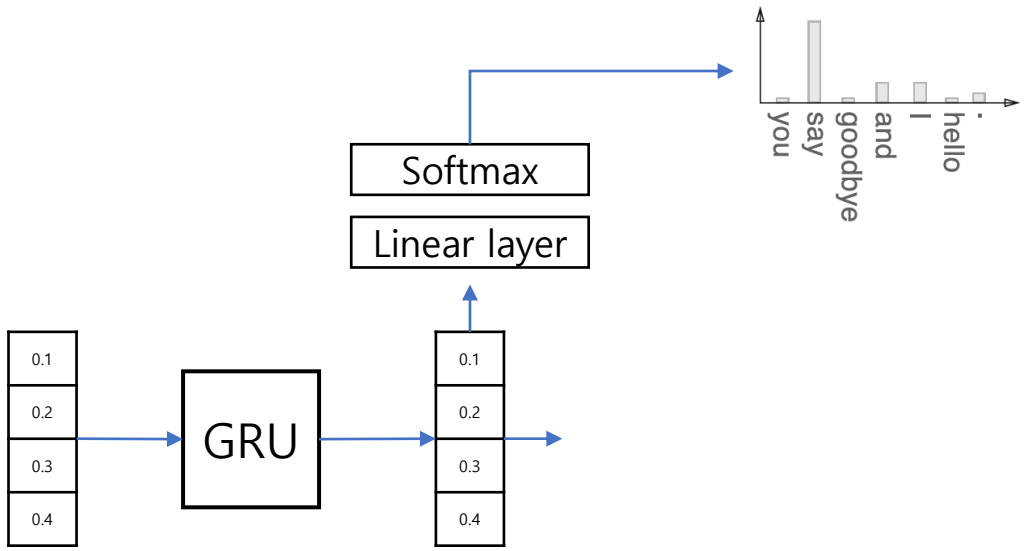
$$\overline{h} = \phi(Wx + r \odot (Uh + Cc))$$

$$r = \sigma(Wx + Uh + Cc)$$

Model

Sequence-to-Sequence 아키텍처

- Decoder로부터 문장 생성 방법
  1) Linear layer와, Softmax 함수를 취해 특정 단어가 나올 확률을 의미하는 Vector 생성



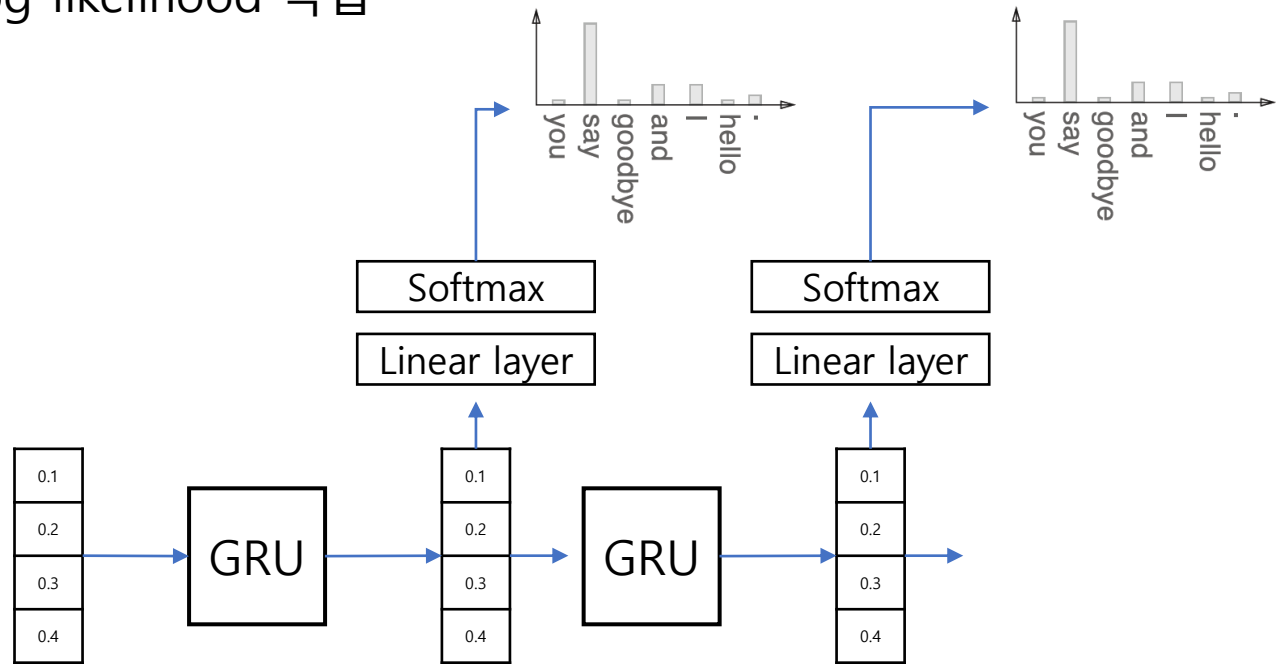
Model

Sequence-to-Sequence 아키텍처

- Sequence-to-Sequence 학습하는 방법
  1) Target 문장을 Label로 활용하여 Decoder로부터 나온 확률을 기반으로 학습.
  2) maximize the conditional log-likelihood 학습

학습 목표

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{y}_n \mid \mathbf{x}_n),$$

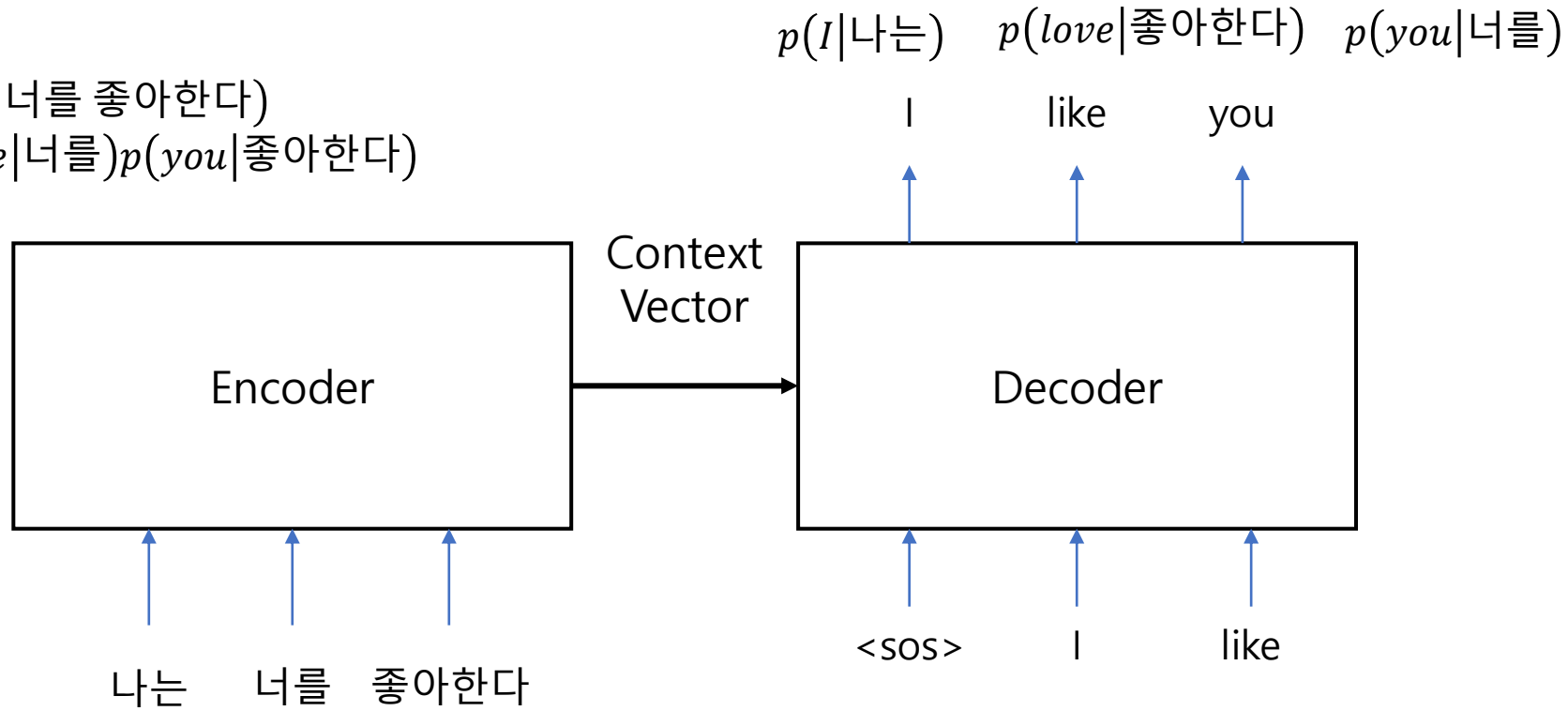


Model

Sequence-to-Sequence 아키텍처

- 학습한 모델 활용 방법
  학습된 모델로부터 Source 문장에 대한 Target 문장의 확률을 도출

$p(I\ like\ you|\text{나는 너를 좋아한다})$
$= p(I|\text{나는})p(like|\text{너를})p(you|\text{좋아한다})$



Experiments

English/French translation

- Task
  WMT'14 Dataset을 활용하여 영어를 프랑스어로 번역

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

Experiments

English/French translation

- Word representation

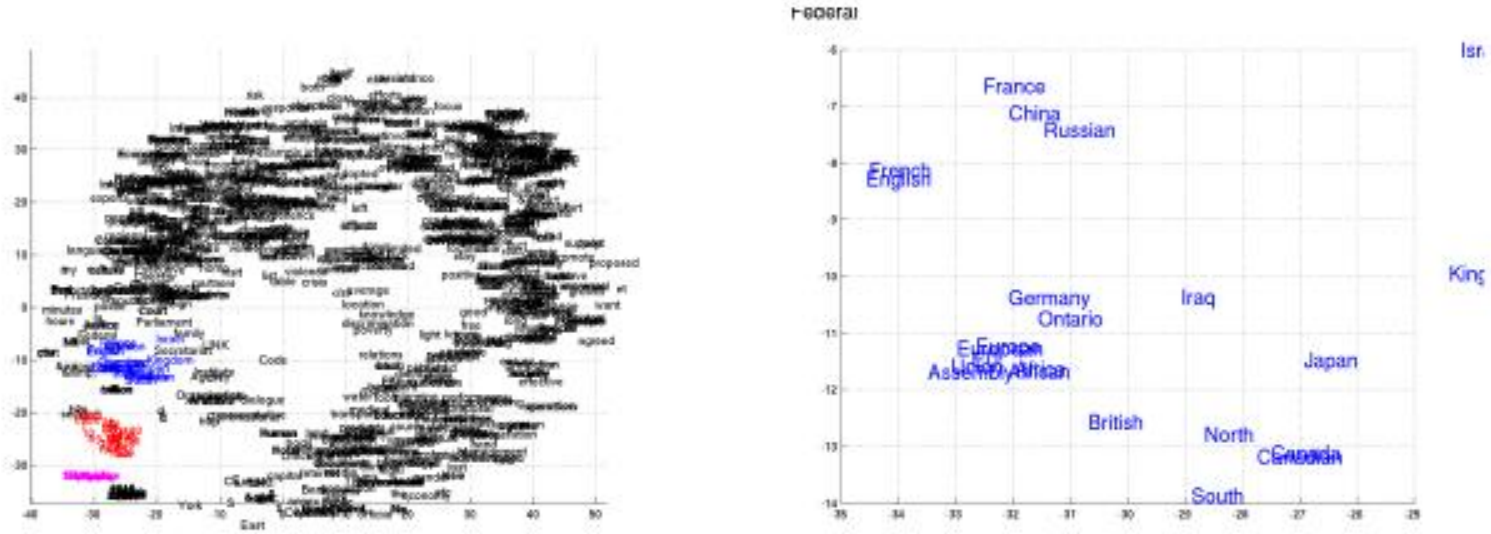
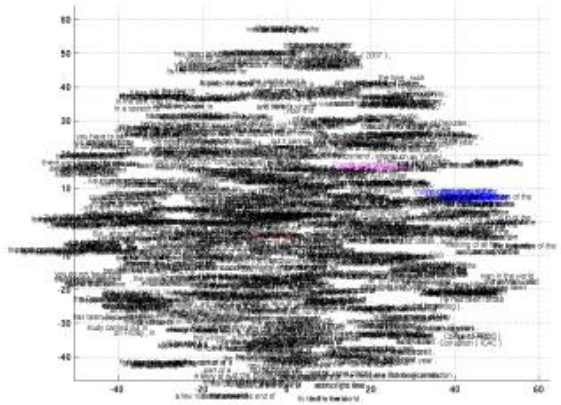


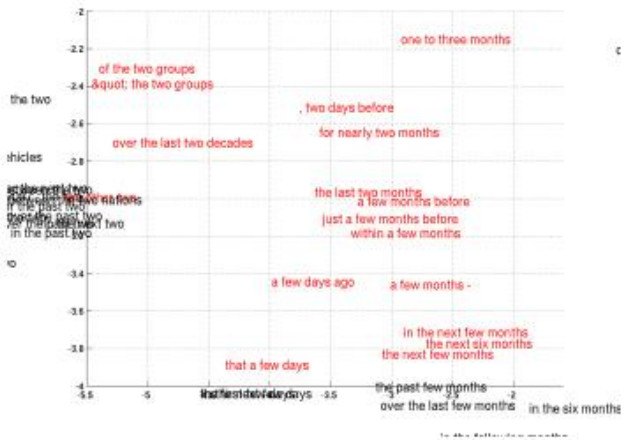
Figure 4: 2–D embedding of the learned word representation. The left one shows the full embedding space, while the right one shows a zoomed-in view of one region (color–coded). For more plots, see the supplementary material.

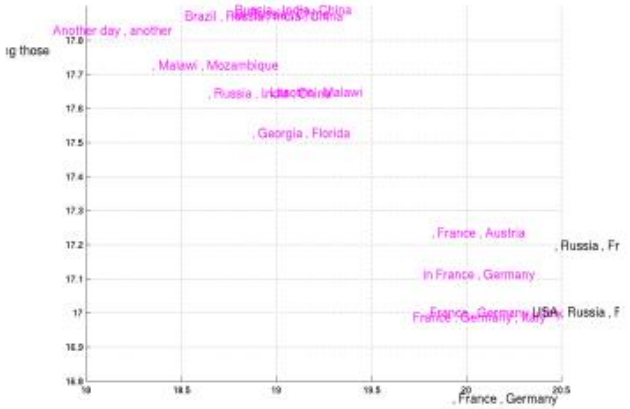
Experiments

English/French translation

- phrase representation







Conclusion

Summary

- GRU architecture
- Seq-to-seq architecture