쉽게 배우는 인공지능(2편)
기계가 언어를 이해하고 번역하는 방법

2017.06.26
Company.AI
강지훈

Machine Translation

Machine Translation

A piece of text which has been written again automatically
from one language to another by a machine

번역

Machine Translation

기계 - 인공지능

번역

Machine Translation

기계 - 인공지능

Machine

...무한한 저장공간은 무한한 길이의 테이프로 나타나는데 이 테이프는 하나의 기호를 인쇄할 수 있는 크기의 정사각형들로 쪼개져있다. 언제든지 기계속에는 하나의 기호가 들어가있고 이를 "읽힌 기호"라고 한다. **이 기계는 "읽힌 기호"를 바꿀 수 있는데 그 기계의 행동은 오직 읽힌 기호만이 결정한다.** 테이프는 앞뒤로 움직일 수 있어서 모든 기호들은 적어도 한번씩은 기계에게 읽힐 것이다.

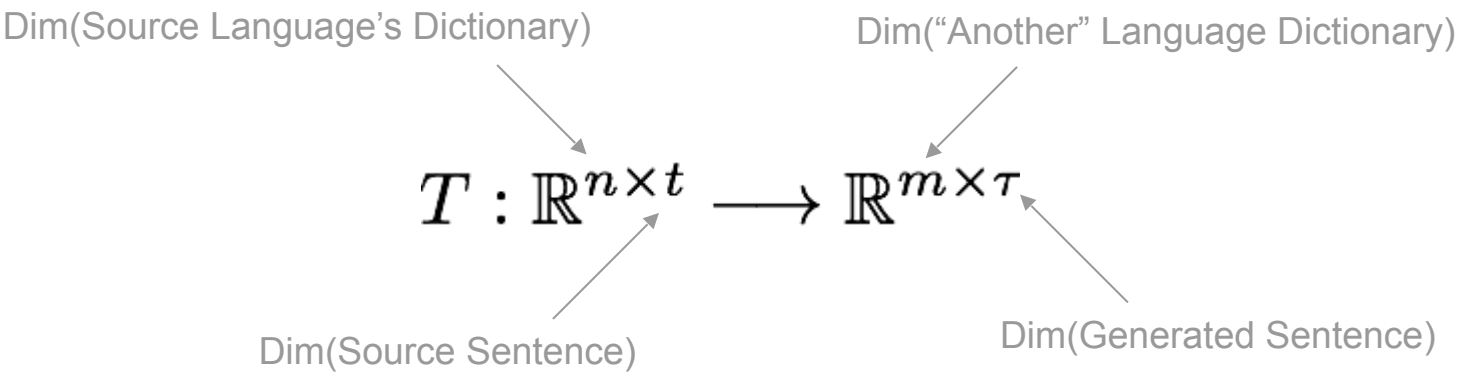
A. M. Turing

번역

Machine Translation

기계 - 인공지능

Translation



Machine Translation

Rule Based

언어마다 존재하는 내부 규칙
(=문법)을 구현하는 시스템
→ **문법에 대한 완벽한 이해?**

원문 언어와 대상 언어에 대한
단어 사전 2권이 필요함
→ **신조어의 발생?**

Statistical

기존에 발생하여 데이터로
저장된 번역으로부터 학습
→ **데이터의 확보?**

실제로 사용하는 언어로부터
결과가 도출됨
→ **품질의 일관성?**

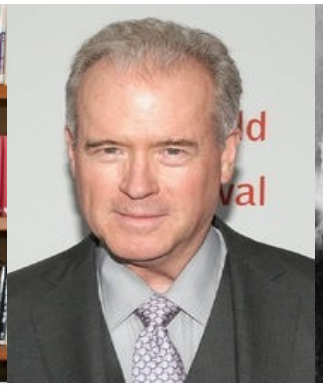
History of MT Research

Two Most Important Moments in MT Research

- 1949: Warren Weaver's Memorandum <Translation>
- 1991-1993: Statistical MT from IBM

Peter F. Brown

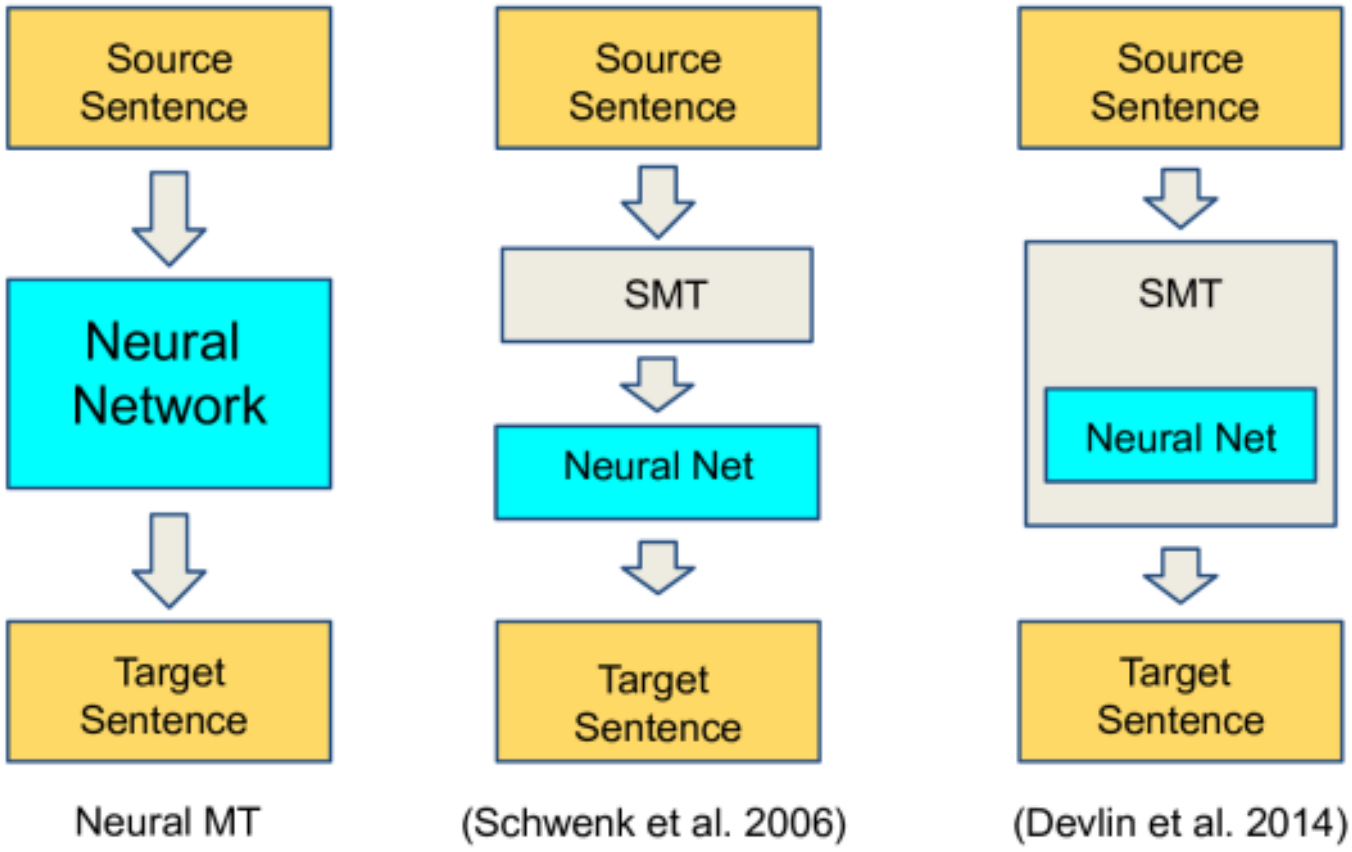
Vincent &
Stephen Della Pietra



Robert L. Mercer

Warren Weaver

Neural Machine Translation - Big Picture



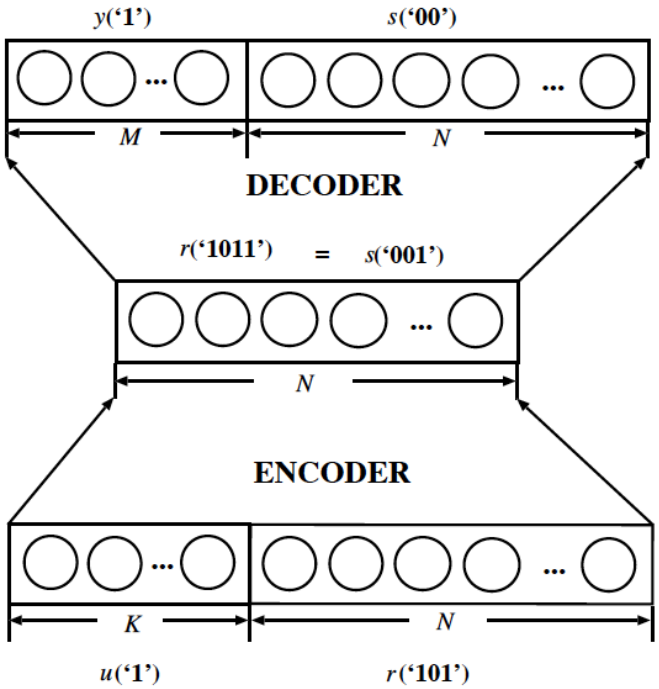
Kyunghyun Cho, Introduction to Neural Machine Translation with GPUs, NVDIA Dev Blog

Neural Machine Translation - Motivation

"우리는 동일하게 번역되는 서로 다른 문장들의
예제로부터 일반적인 번역을 학습할 수 있는
반복적 이종 결합 메모리를 제안한다"

"We propose .. **Recursive Hetero-Associative Memory** which .. may be applied **to learn general translations from examples** in which different sentences may have the same translation"

Forcada & Ñeco, 1997

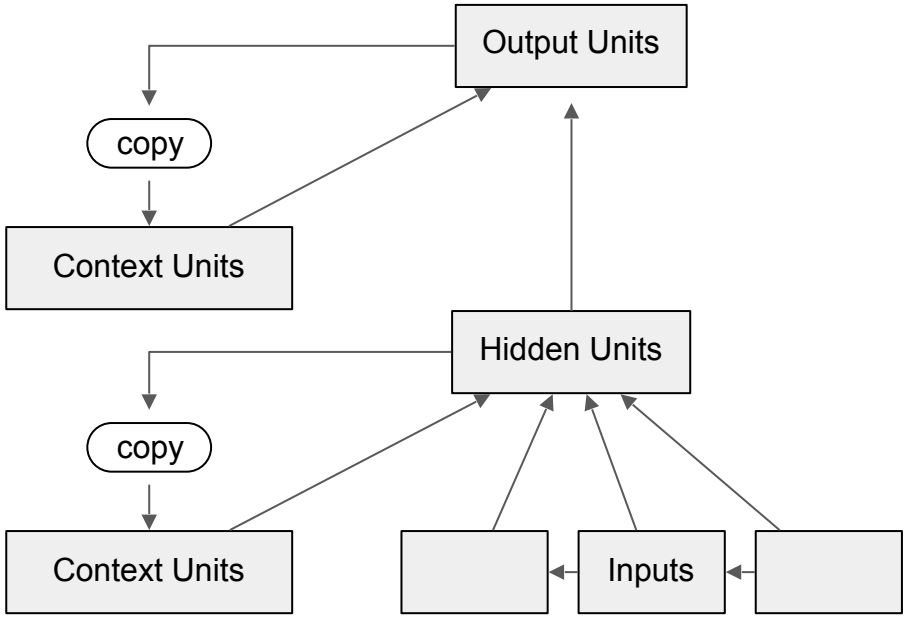


Neural Machine Translation - Motivation cont'd

"성능의 향상을 통해, 앞으로는 더욱 복잡하고 특수한 영역에서의 번역이 가능해질 것으로 보인다. **그러나 그 적용 범위에는 (그리고 학습에 걸리는 시간에도) 제약이 있을 것이다**"

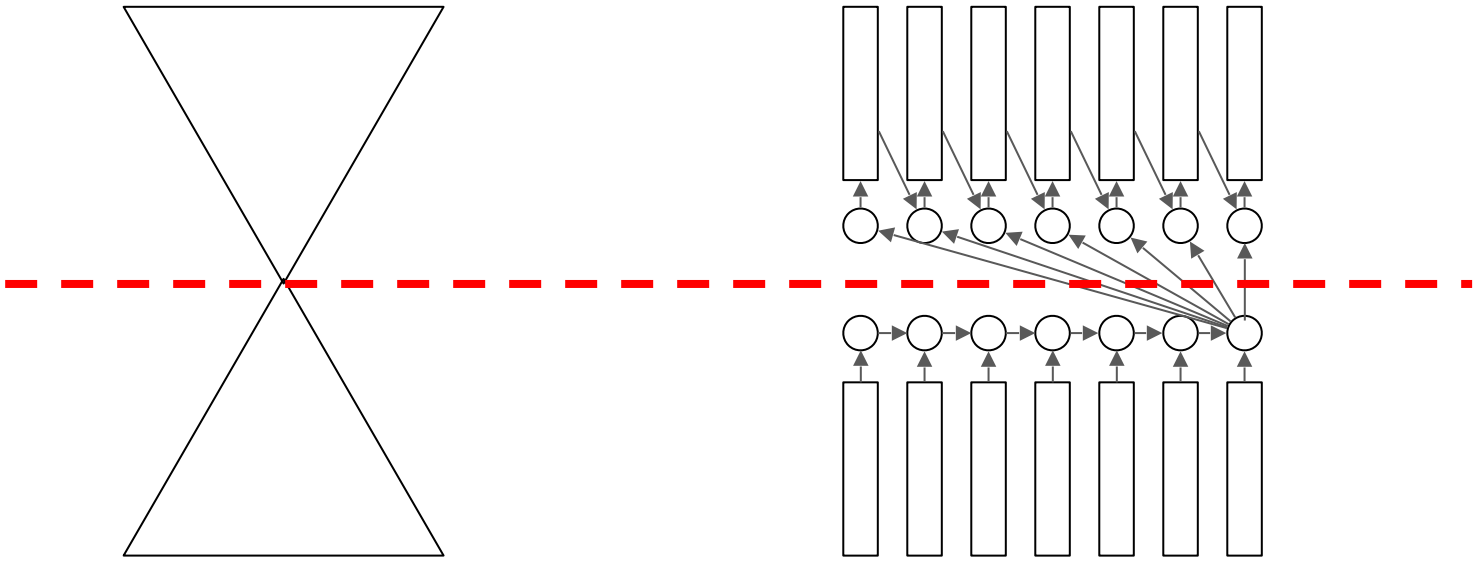
"Based on these encouraging performances, future work dealing with more complex limited-domain translations seems to be feasible. **However, the size of the applications (and consequently, the learning time) can be prohibitive**"

Castaño & Casacuberta, 1997

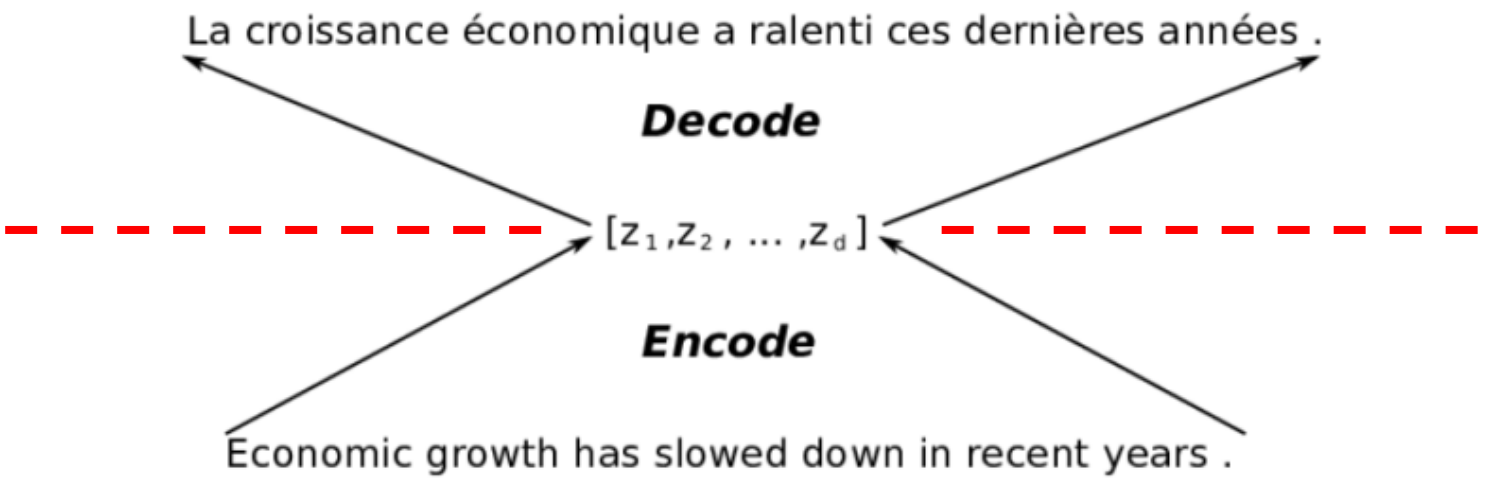


Neural Machine Translation - Basic Architecture

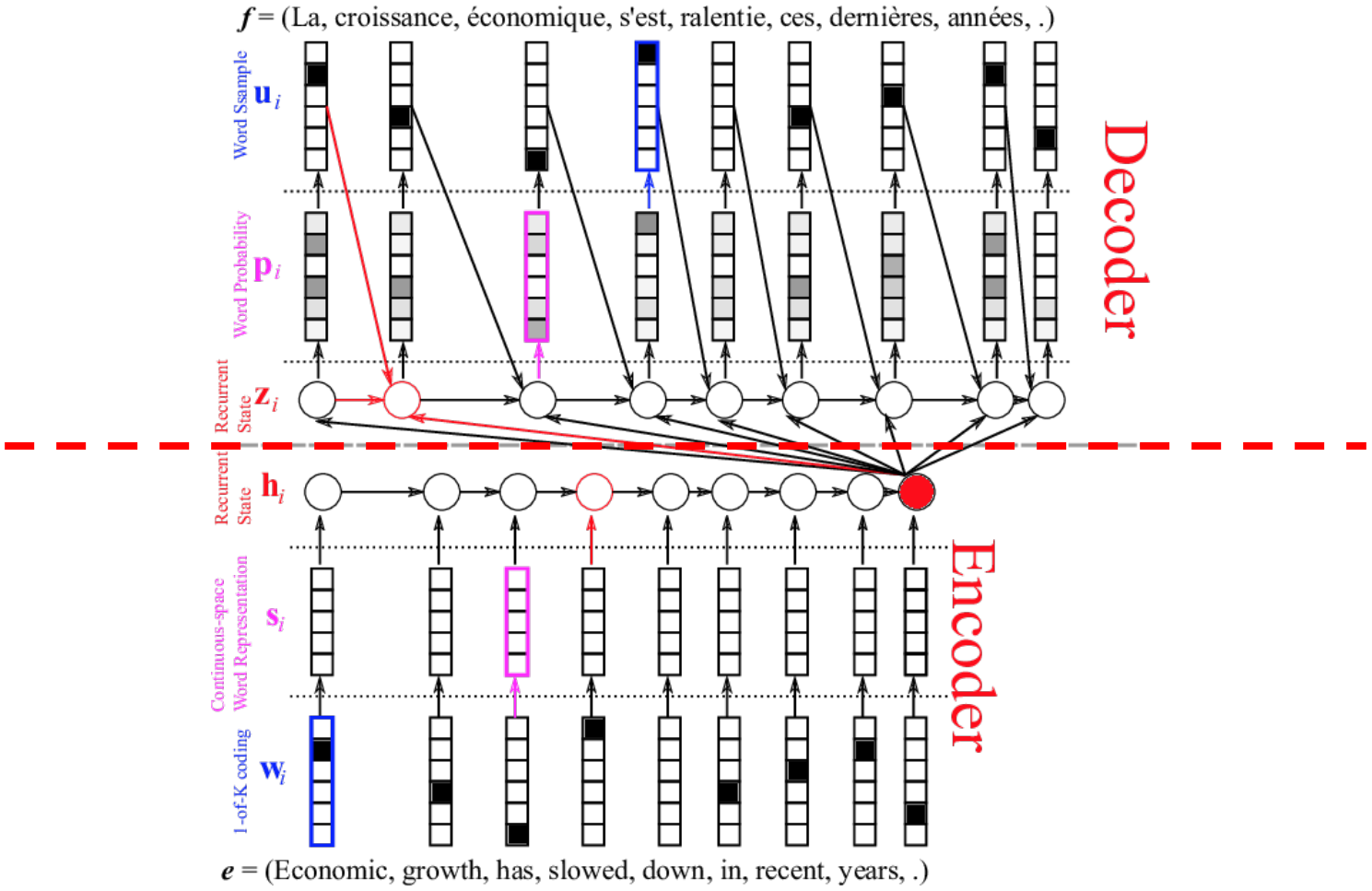
Encoder-Decoder Architecture for Machine Translation



Neural Machine Translation - Front View



Neural Machine Translation - Side View



Embedding

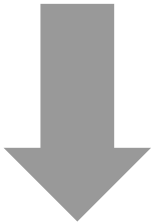
[동명사] (...에) 박아 넣음, 심음

Convert Language to Code

나라의 언어가 중국과 달라 중국어와는 서로 맞지 않아서, 백성이 말하고자 하는 바가 있어도 그 뜻을 제대로 전하지 못하는 사람이 많다.

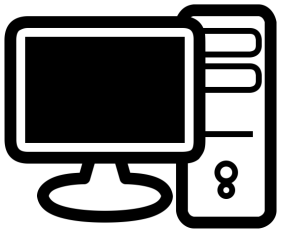
훈민정음 언해 - 한국어

Input to Machine



Kwa sababu ya lugha mbalimbali na Kichina na Kichina haziendani, hata bar Unataka watu wanasema kuna watu wengi ambao wanaweza kufikisha maana sahihi.

훈민정음 언해 - 스와힐리어



But, how to input whole language?

https://translate.google.co.kr

One Hot Encoding

Dictionary

banana 바나나
boy 소년
cat 고양이
dog 강아지
girl 소녀
king 왕
man 남자
mango 망고
woman 여자
queen 여왕

English Encoding

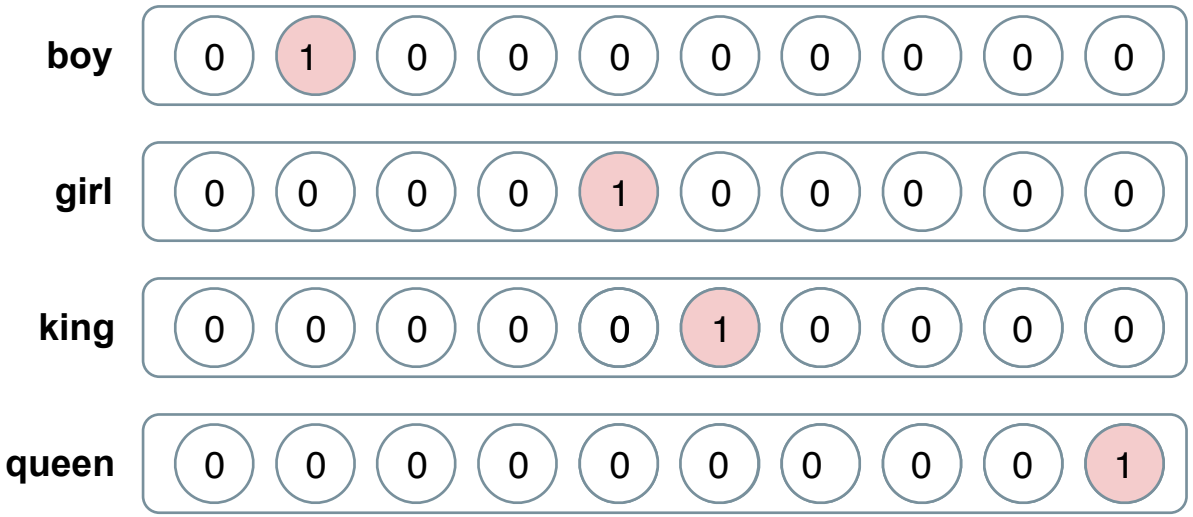
boy = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
girl = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
king = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
queen = [0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

Korean Encoding

소년 = [0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
소녀 = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
왕 = [0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
여왕 = [0, 0, 0, 0, 0, 0, 0, 0, 1, 0]

One Hot Encoding - Limitation

- Popular similarity function cosine similarity
- Words / Tokens come from some tokenization and transformation

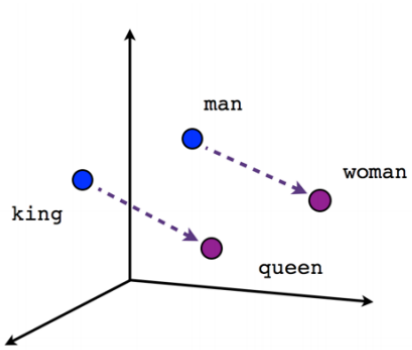


Similarity(**boy**, **girl**) = Similarity(**boy**, **king**) = Similarity(**boy**, **queen**)

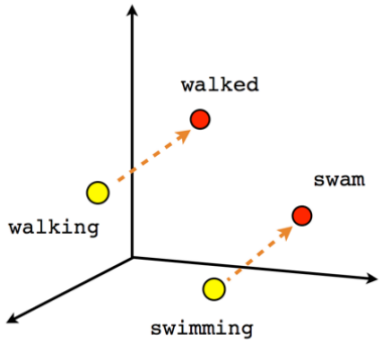
It's Meaningless Metric!

Embedding

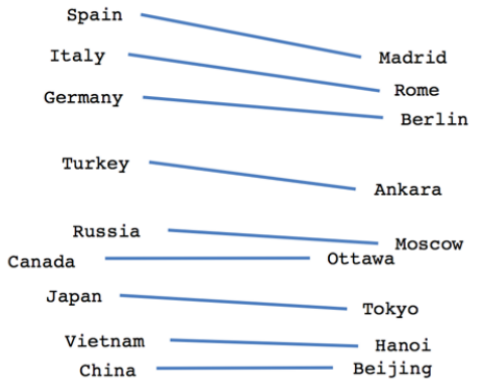
Actually, we understand our language as such...



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/versions/master/tutorials/word2vec

Embedding - Count based vs Direct prediction

Count based

LSA, HAL (Lund & Burgess),
COALS (Rohde et al),
Hellinger-PCA (Lebret & Collobert)

- 빠른 학습
- 통계적 정보를 효율적으로 이용

- 주로 단어 유사도 측정을 위해 사용

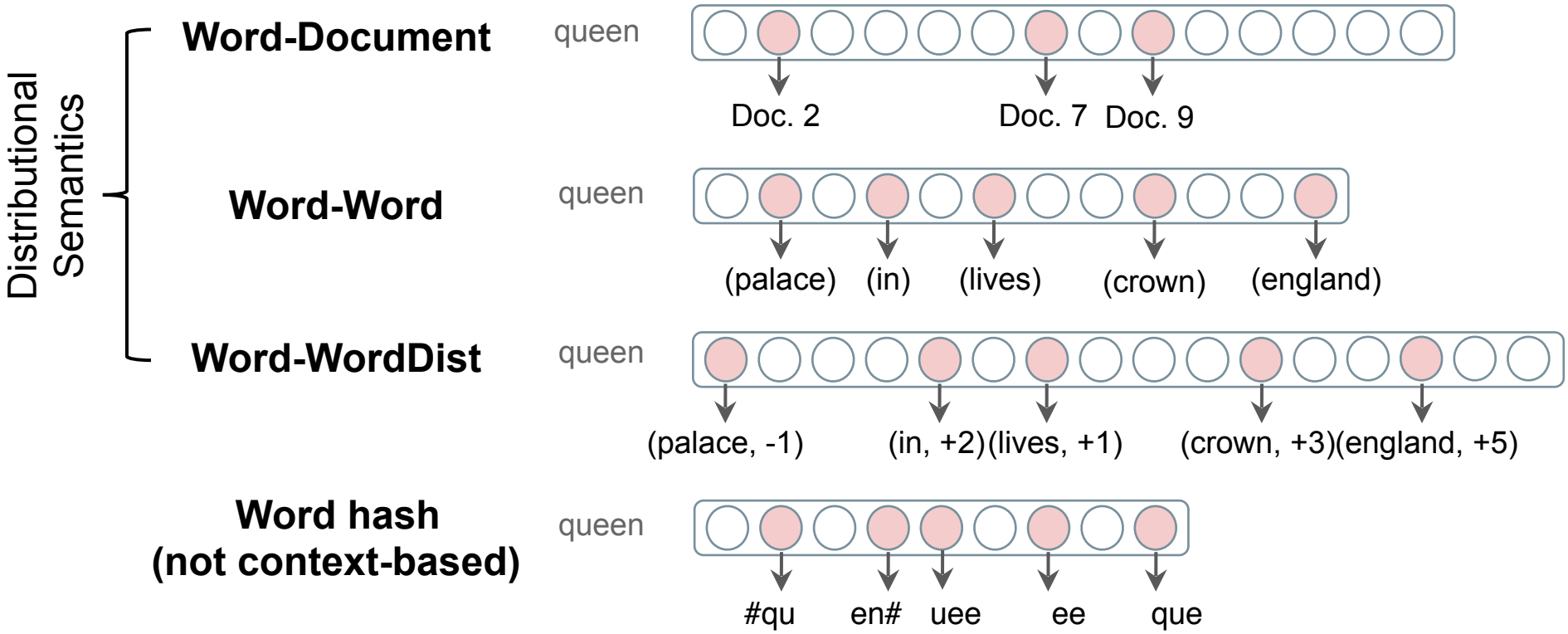
Direct Prediction

NNLM, HLBL, RNN,
word2vec Skip-gram/CBOW
(Bengio et al; Collobert & Weston;
Huang et al; Mnih & Hinton; Mikolov et al;
Mnih & Kavukcuoglu)

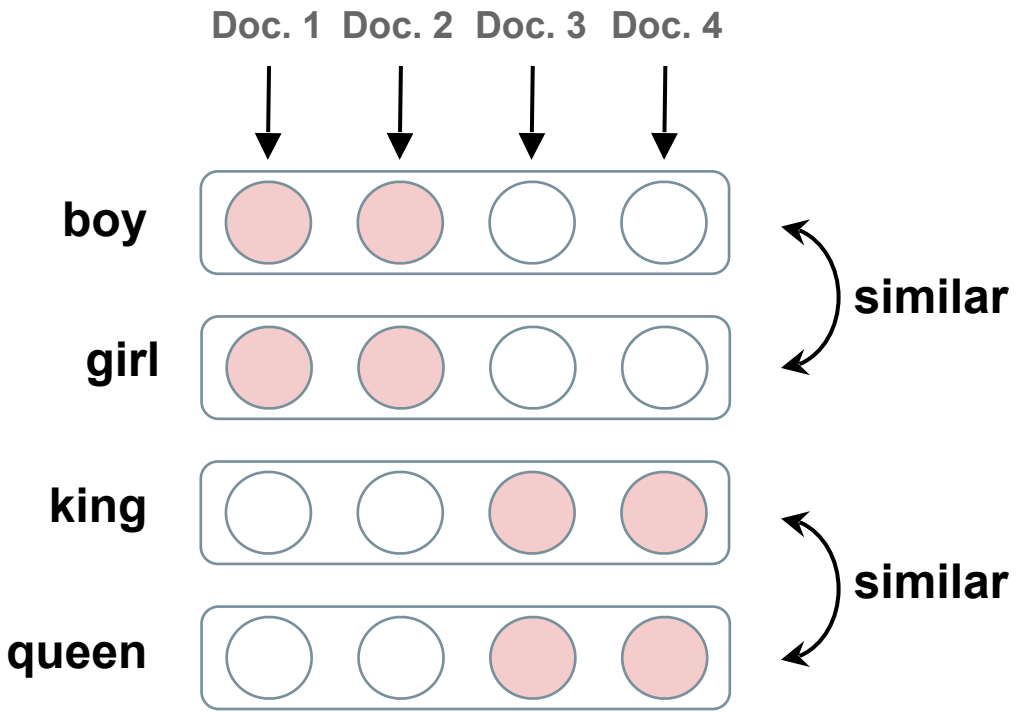
- Corpus 크기에 따라 scale 됨
- 통계 정보 이용이 어려움

- 단어 유사도 이상의 복잡한 단어의 패턴을 알아낼 수 있음

Embedding - Count Based

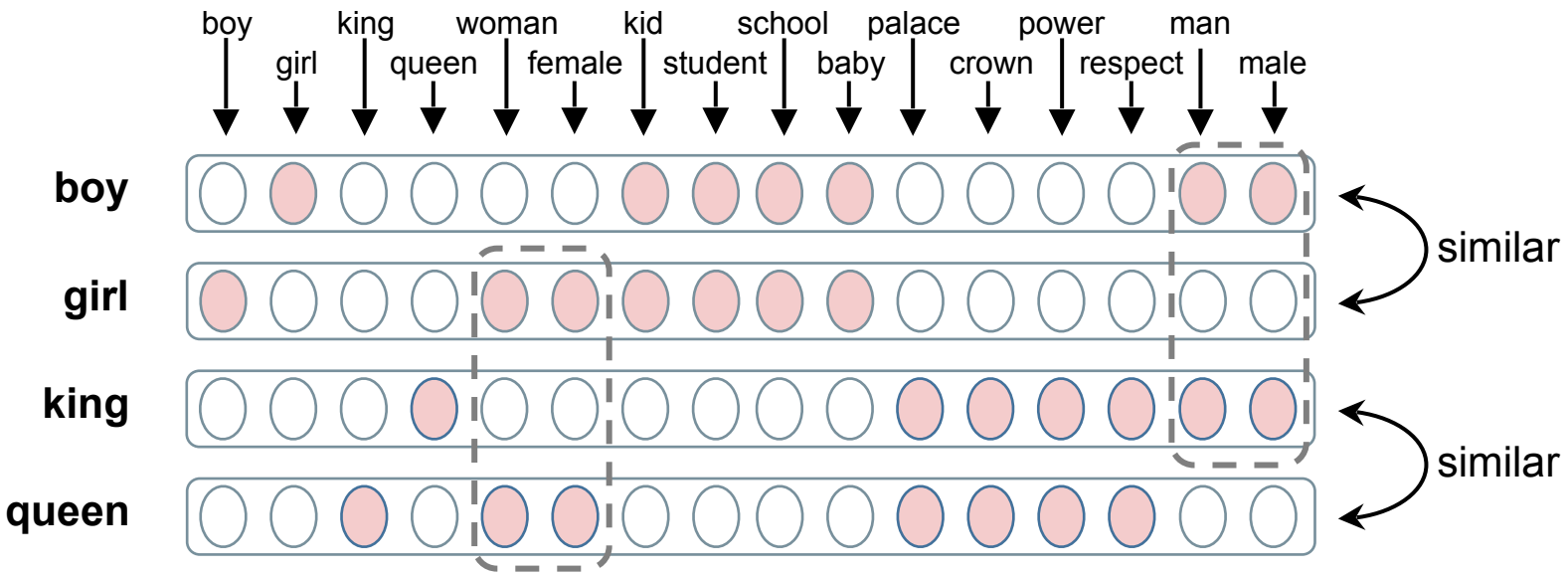


Using Word-Document context



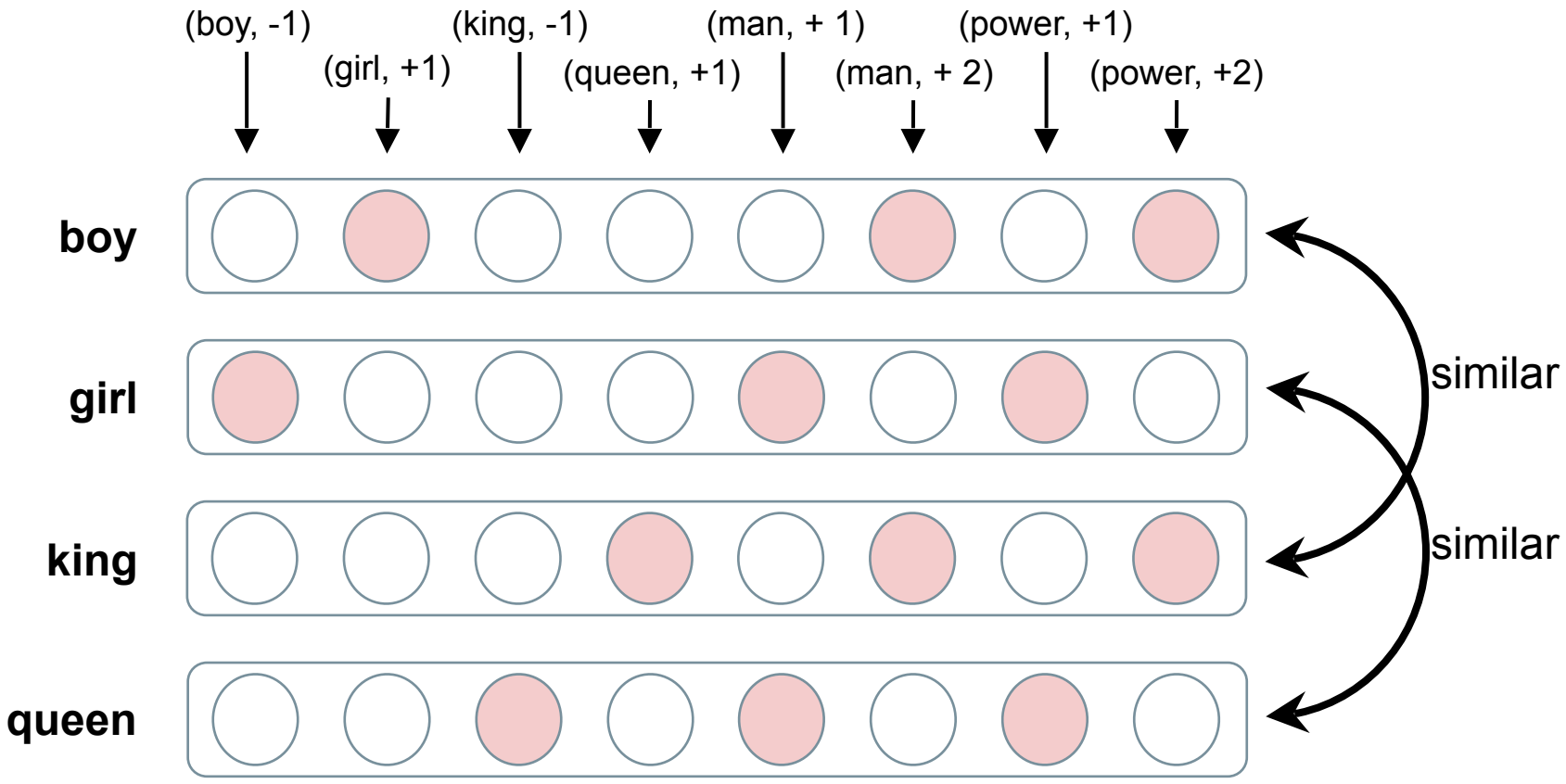
- This is *Topical* or *Syntagmatic* similarity.

Using Word-Word context



1. Word-Word is less sparse than Word-Document (Yan et al., 2013)
2. A mix of topical and typical similarity (function of window size)

Using Word-WordDist context



- This is *Typical* or *Paradigmatic* Similarity.

Vector Space Models

For a given task:

Choose matrix

choose s_{ij} weighting:

could be binary, could be raw counts

PPMI weighting for Word-Word matrix
TF-IDF weighting for Word-Document matrix

	c_0	c_1	c_2	…	c_j	…	$c_{\|C\|}$
W_0							
W_1							
W_2							
…							
W_i					S_{ij}		
…							
$w_{\|V\|}$							

[Example] Positive Pointwise Mutual Information (Word-Word Matrix)

V: vocabulary, C: set of contexts, S: sparse matrix |V| x |C|

$$S_{ij} = PPMI(w_i, c_j)$$

$$PPMI(w,c) = \begin{cases} 0 & PMI(w,c) < 0 \\ PMI(w,c) & otherwise \end{cases}$$

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} = \log \frac{freq(w,c)|corpus|}{freq(w)freq(c)}$$

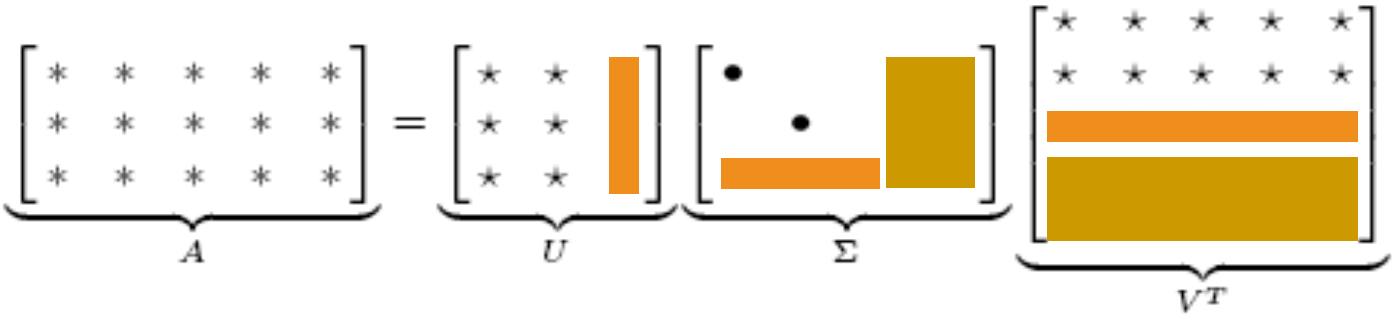
Turney and Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 2010

Embedding - LSA

LSA (Latent Semantic Analysis) : Count model

- (Weighted or Log-scaled) term-document Matrix (Deerwester et al. 1990), 또는 Word-Context Matrix (Schütze 1992)을 SVD로 분해함
- 임의의 *k*개 singular values로 generalization

ex. SVD with *k* = *2*



[Cf. Baroni: Don't count, predict! A systematic comparison of context- counting vs. context-predicting semantic vectors. ACL 2014]

Embedding - Word2Vec

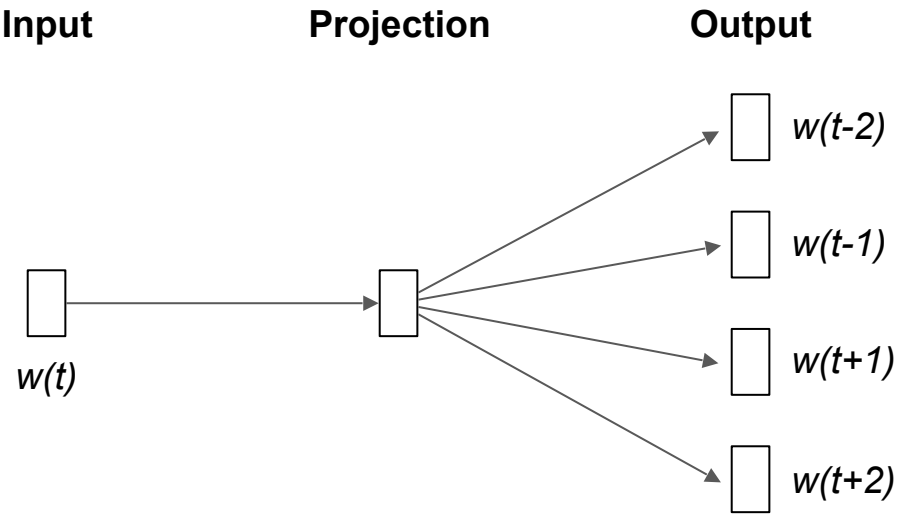
CBOW / SkipGram [Mikolov et al. 2013] **: Predict model**

- 단어 벡터의 학습을 위해

  Predict a word given its bag-of-words context (CBOW)

  Predict a context word (position independent) from the center word

- 예측이 충분히 잘 될 때까지 단어 벡터의 업데이트를 반복



Embedding - Glove Word similarities

Nearest words to ‘frog’

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

[Pennington et al., EMNLP 2014]

litoria

leptodactylidae

rana

eleutherodactylus

http://nlp.stanford.edu/projects/glove/

Named Entity Recognition Performance

F1 score of CRF trained on CoNLL 2003 English with 50 dim word vectors

Model on CoNLL	CoNLL ’03 dev	CoNLL ’03 test	ACE 2	MUC 7
Categorical CRF	91.0	85.4	77.4	73.4
SVD (log tf)	90.5	84.8	73.6	71.5
HPCA	92.6	88.7	81.7	80.7
C&W	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	93.2	88.3	82.9	82.2

Embedding - Conclusion

Glove는 단어간 동시 발생 수 (co-occurrence counts)와 단어 사이의
의미적 연관성을 단어 vector 공간에서의 선형 관계로 변환함

Glove는 발생 빈도의 Count!와 Prediction! 사이의 연관성을 보여줌
count에 대한 적절한 scaling은 예측 모델의 퍼포먼스를 향상시킴

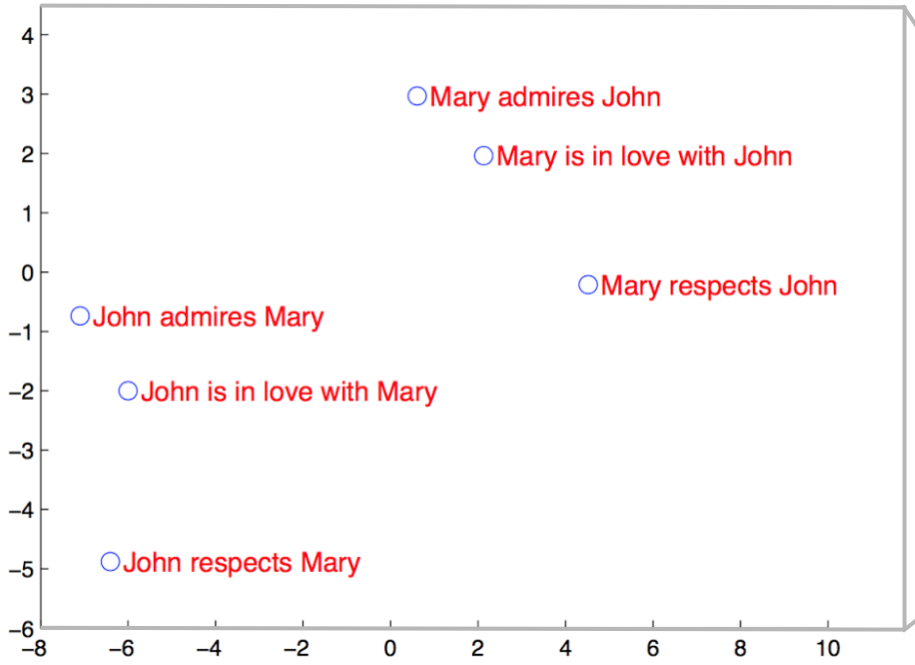
주요 관련 연구

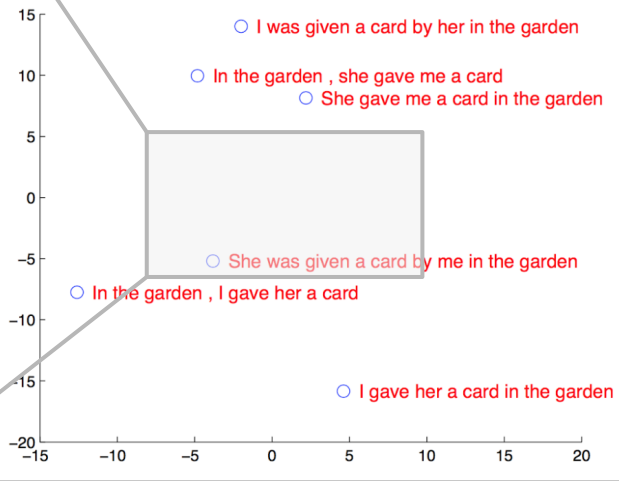
Dependency-Based Word Embeddings
[Levy & Goldberg, 2014]

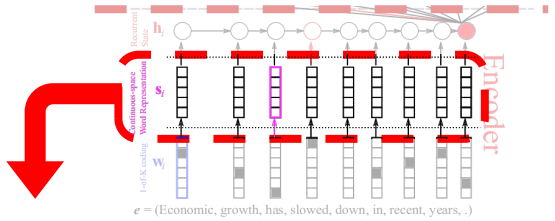
Enriching Word Vectors with Subword Information
[Joulin, Armand, et al, 2016]

[Arora, Li, Liang, Ma & Risteski, 2015]
[Hashimoto, Alvarez-Melis & Jaakkola, 2016]

Embedding Example in NMT







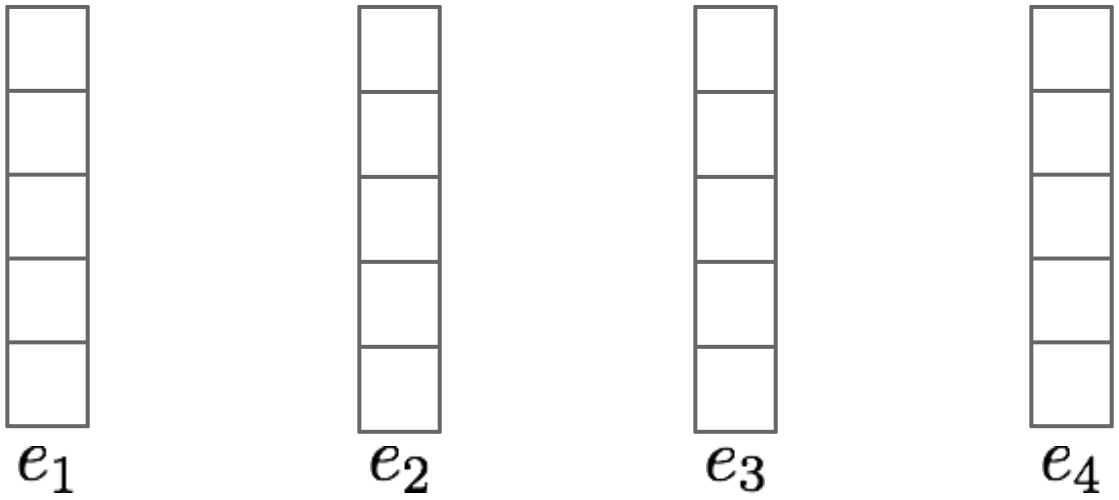
RNN

Recurrent Neural Network

Our input is...

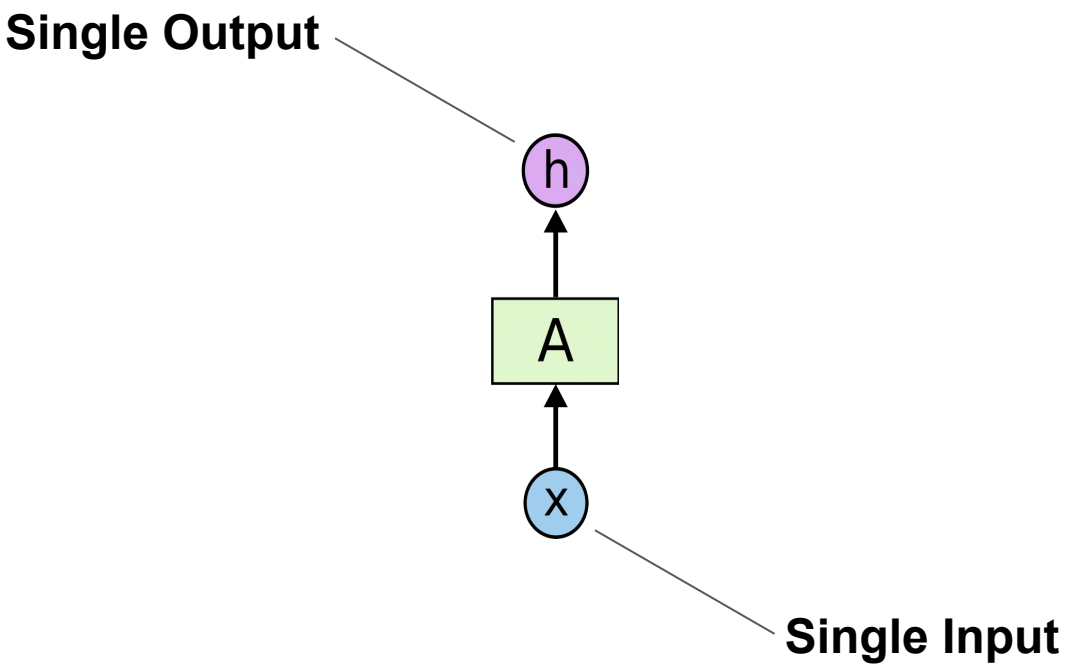
나라의 언어가 중국과 달라 중국어와는 서로 맞지 않아서, …(후략)

나라/NNG+의/JKG 언어/NNG+가/JKS 중국/NNP+과/JKB 달/VV+라/EC …(후략)



Our inputs are Sequential!

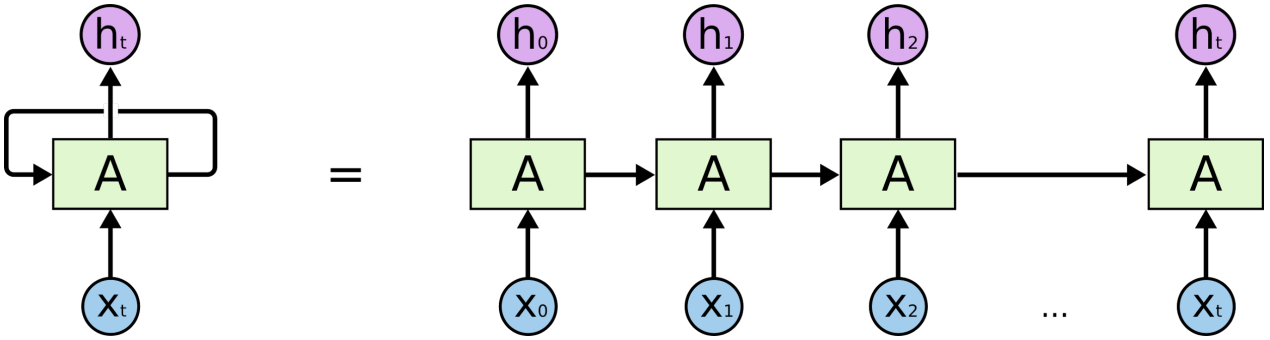
NN is not sufficient



Basic neural network can’t manage sequential inputs

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

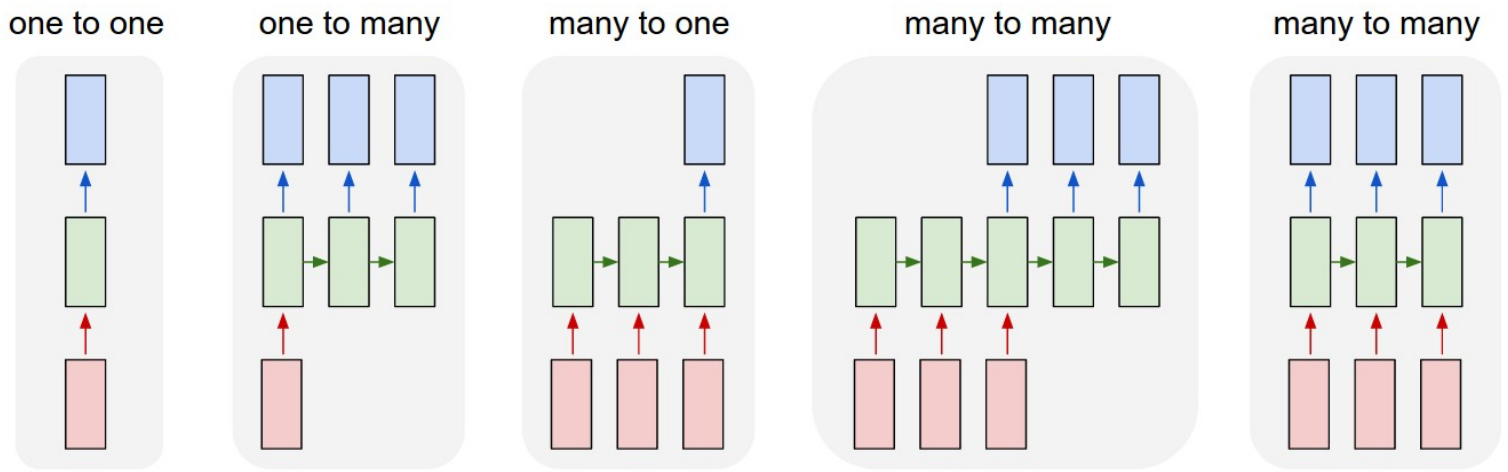
Recurrent Neural Network



RNN can handle a sequential input

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

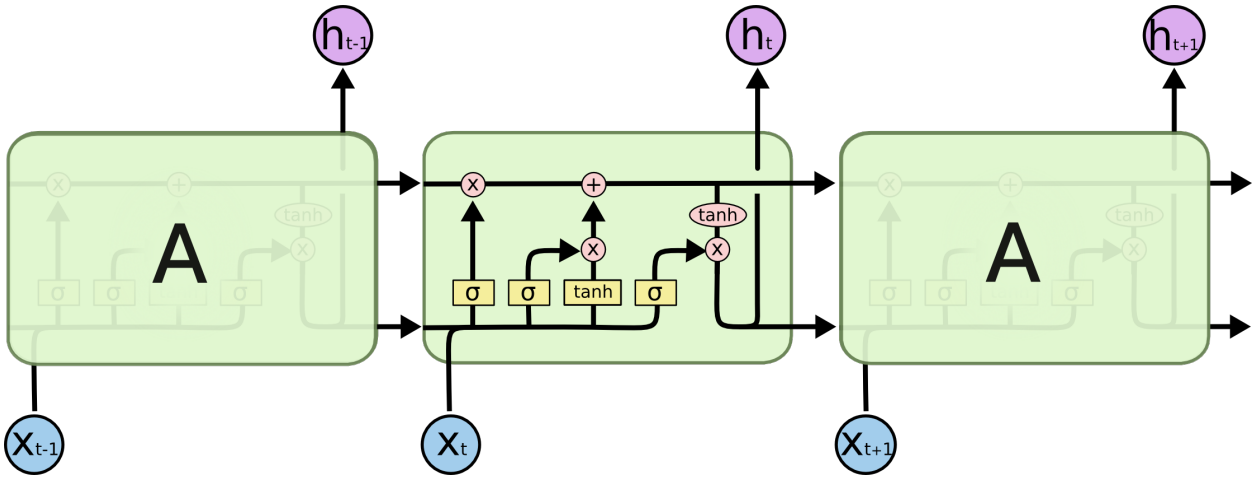
RNN- Applications



Structure	Task	IO
1 to 1	Image Classification	Image → Class
1 to N	Image Captioning	Image → Sentence
N to 1	Sentiment Analysis	Sentence → Class; positive or Negative
N to N	Machine Translation	Sentence; source lang. → Sentence; target lang.
N to N	Chat Bot	Sentence; question → Sentence; answer

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

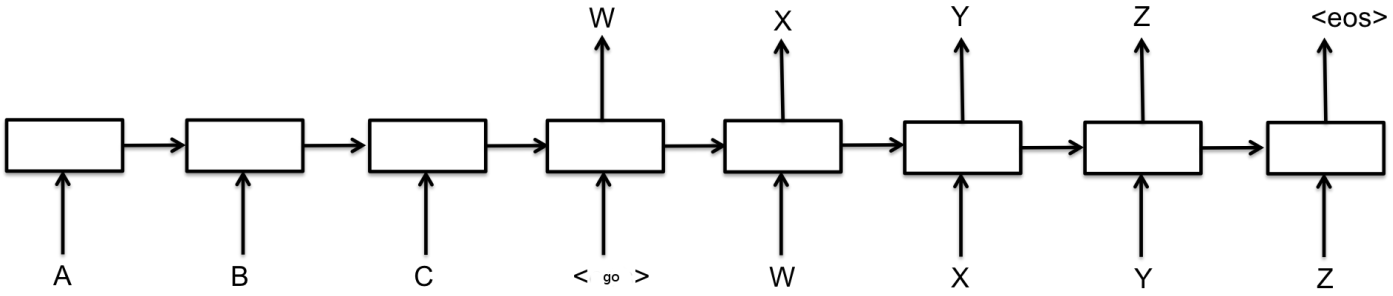
Long Short Term Memory



In short, we have an advanced RNN

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

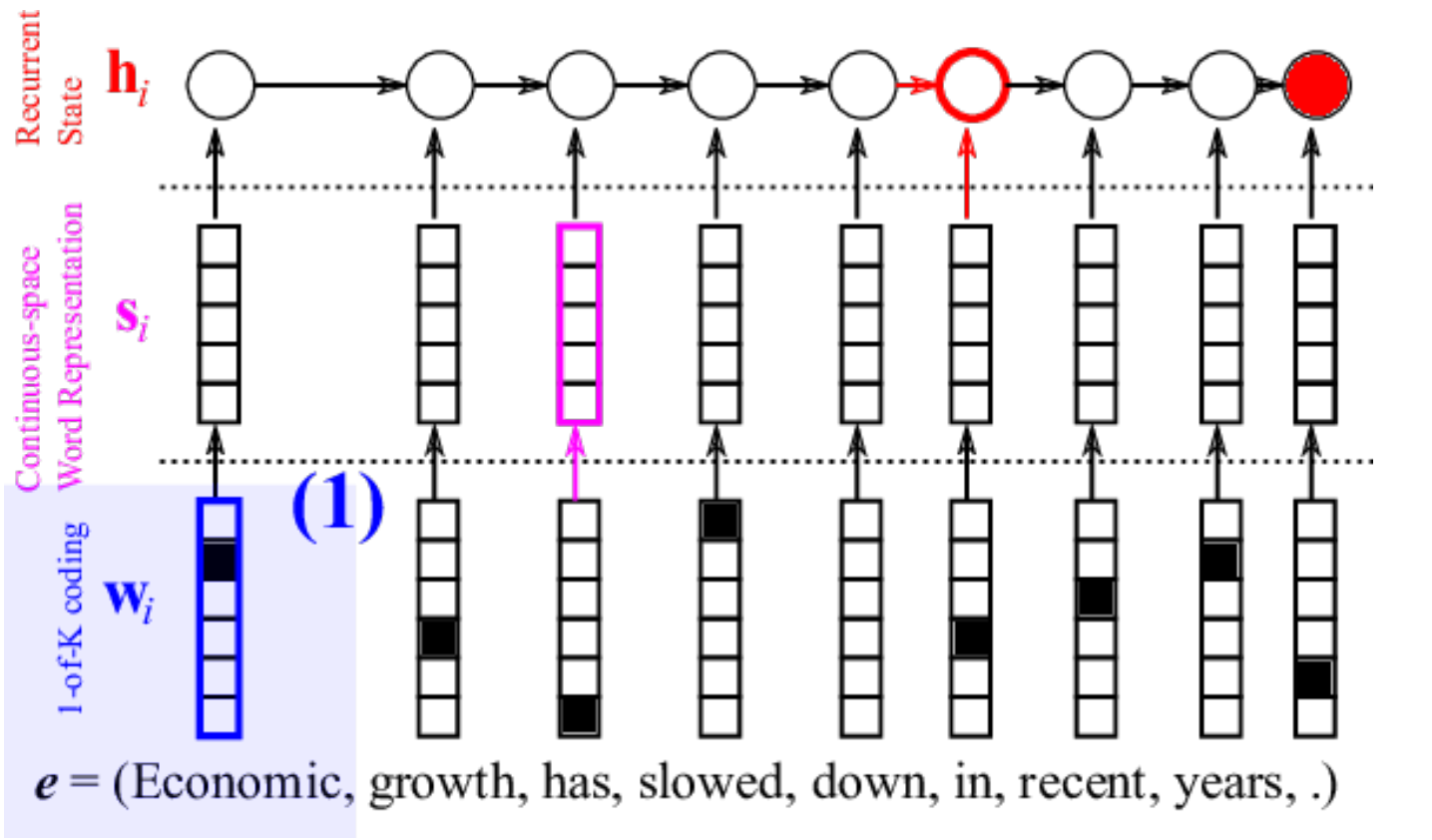
Seq-2-Seq



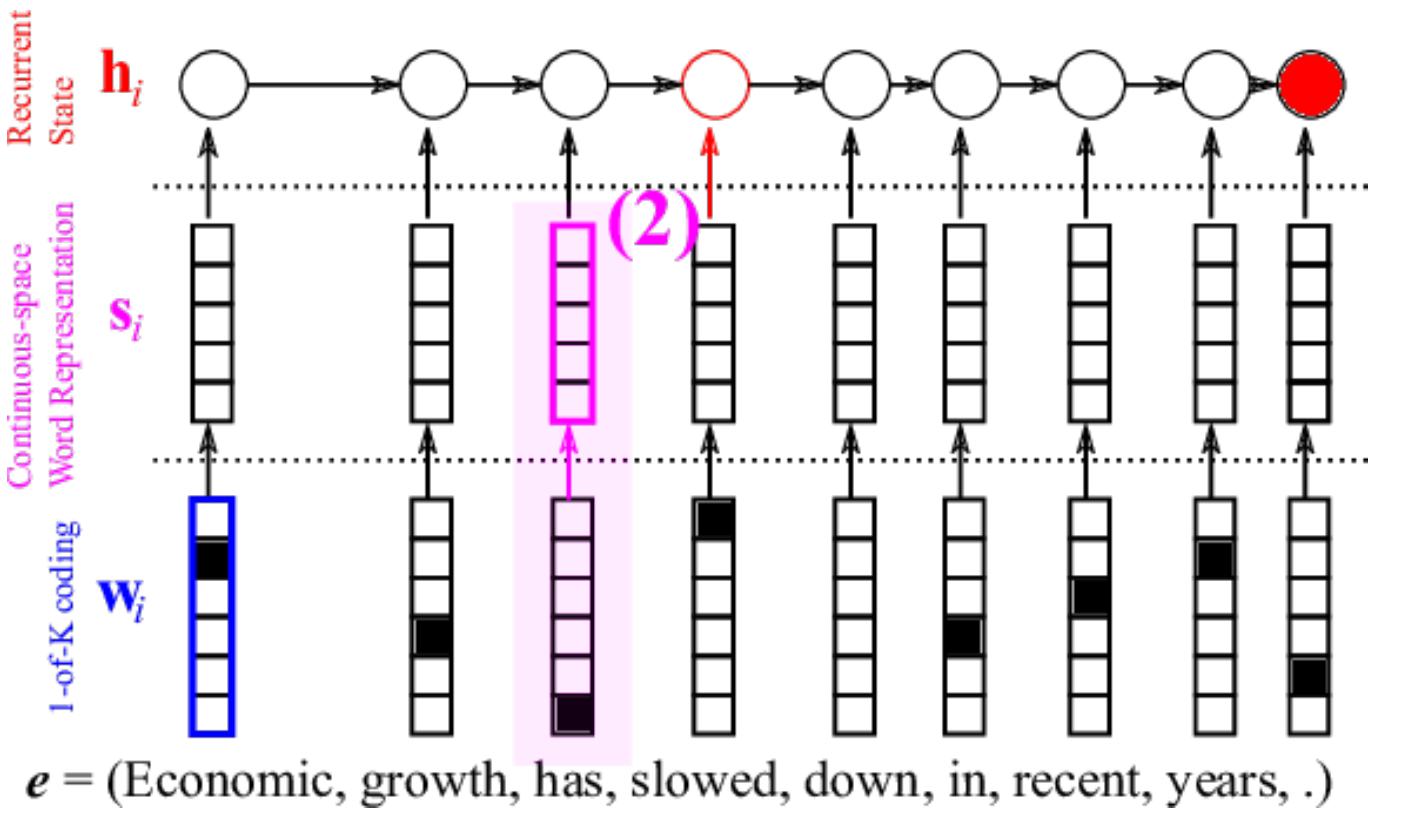
Let's see how seq-2-seq works in NMT

https://tensorflowkorea.gitbooks.io/tensorflow-kr/content/g3doc/tutorials/seq2seq/

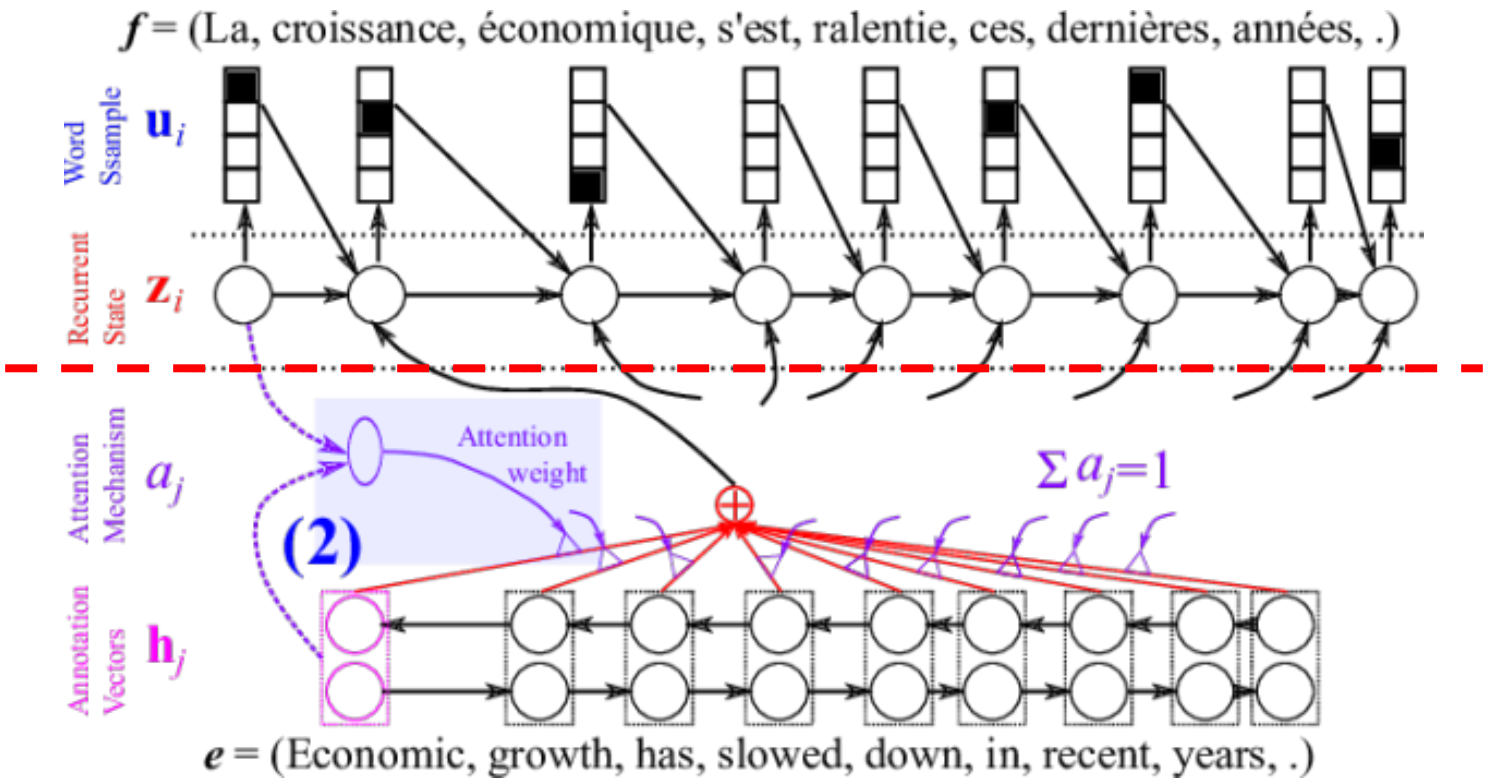
Neural Machine Translation - Encoder



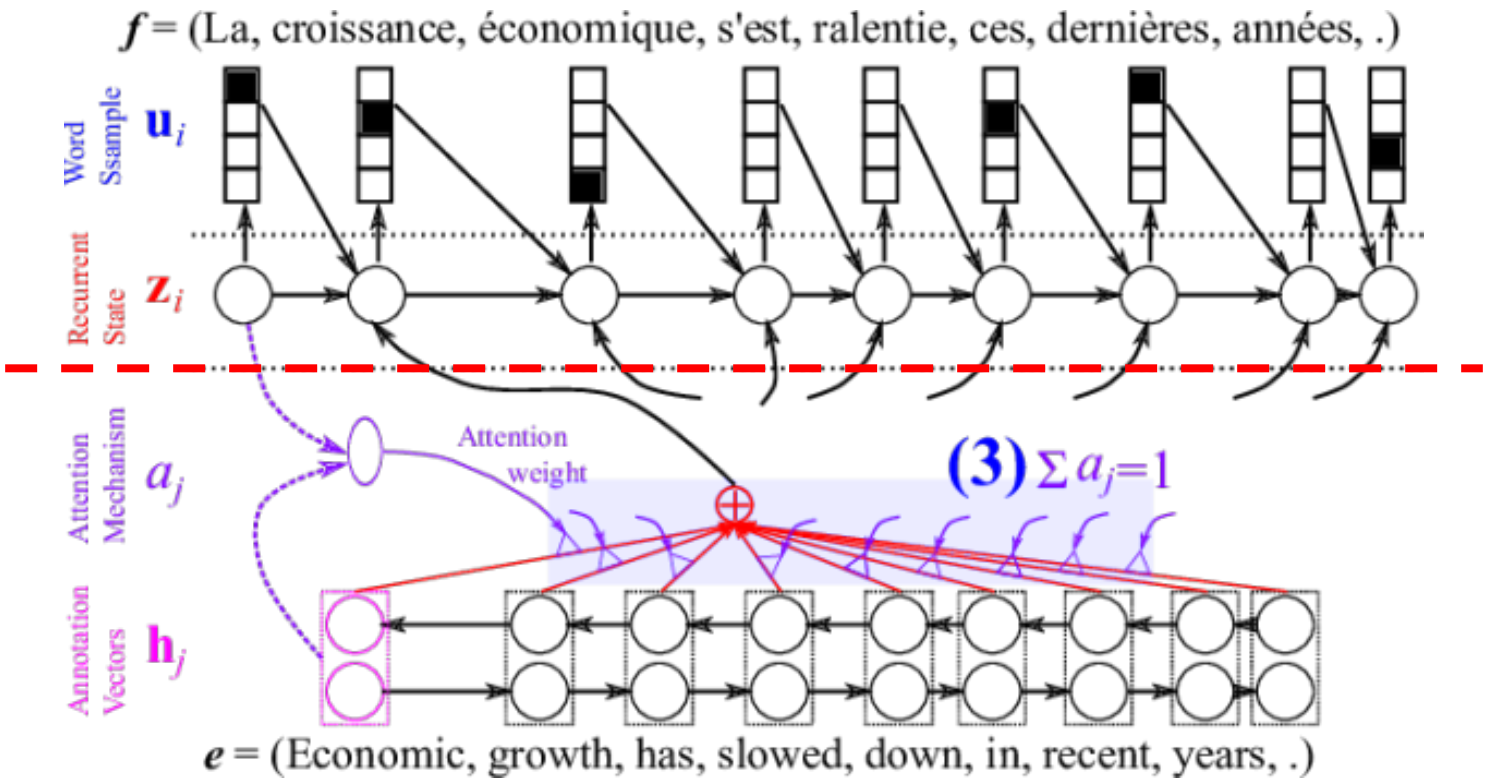
Neural Machine Translation - Encoder cont'd



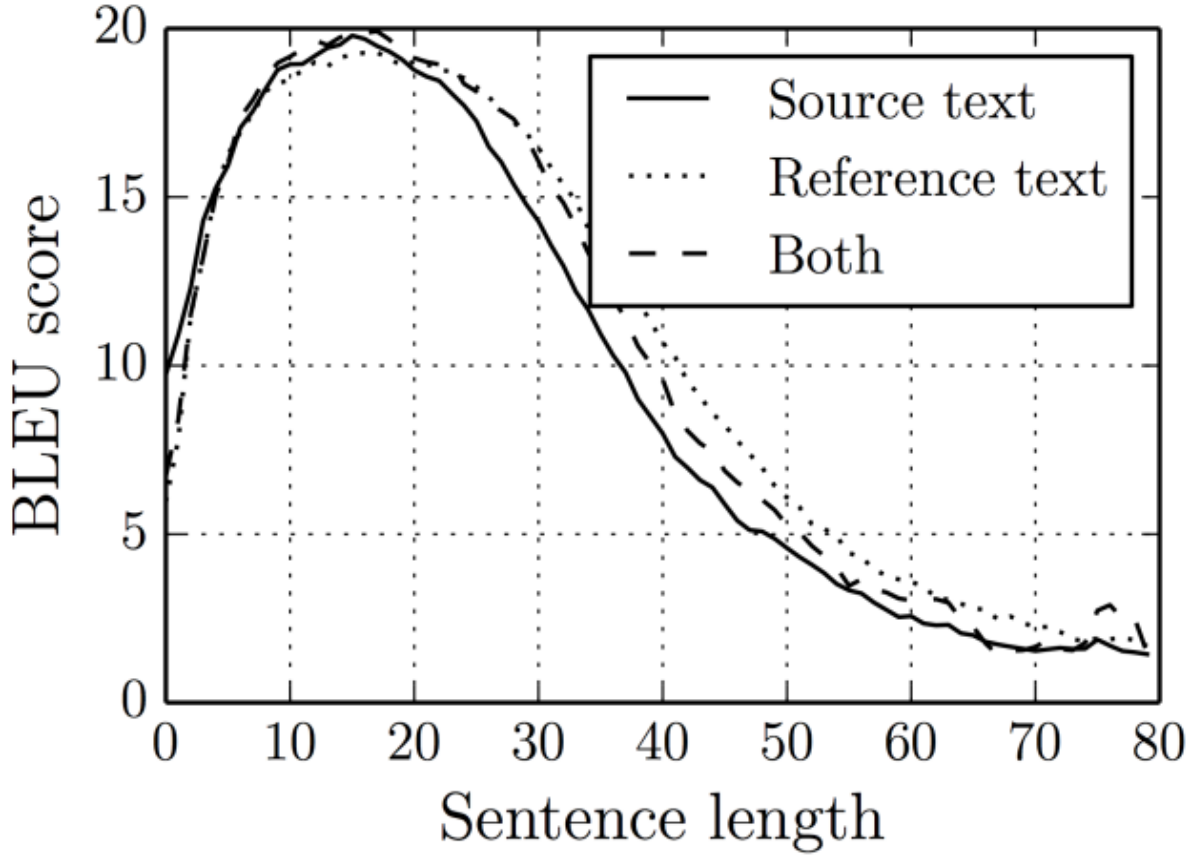
Neural Machine Translation - Decoder



Neural Machine Translation - Decoder cont'd



Neural Machine Translation - Results



Alignment

- Applying attention mechanism in the decoder for machine translation

- The attention mechanism builds attention pairs between source sentences and target sentences.

- Such pairs can be even transformed into heatmap-like matrix to demonstrate a kind of soft alignment between these two sentences.

Bahdanau et. al., ICLR 2015, Neural Machine Translation by Jointly Learning to Align and Translate

Alignment - cont’d

Benefits

- It can handles long length sentence without merging or folding their semantics into a vague and incomplete representation.
- It can conquer the order variation and discrepency between sources and targets by soft alignment, which can be learned without any external knowledge.

Alignment - cont'd

New Decoder

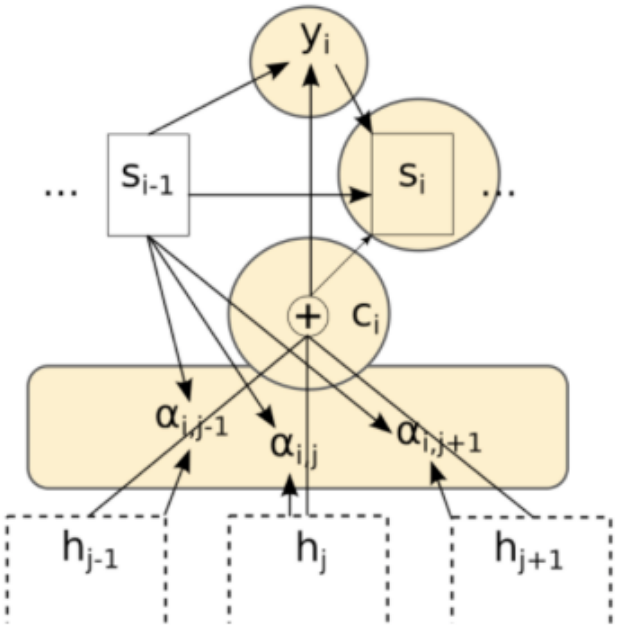
Step i:

compute alignment

compute context

generate new output

compute new decoder state



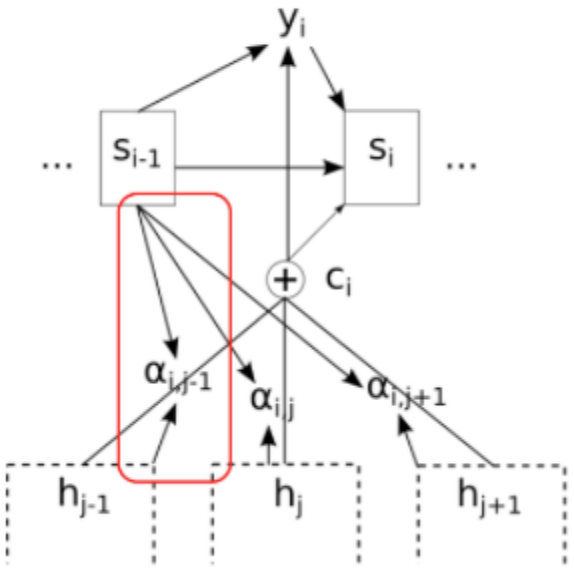
Alignment - cont'd

Alignment Model

$$e_{ij} = v^T \tanh(W s_{i-1} + V h_j) \quad (1)$$

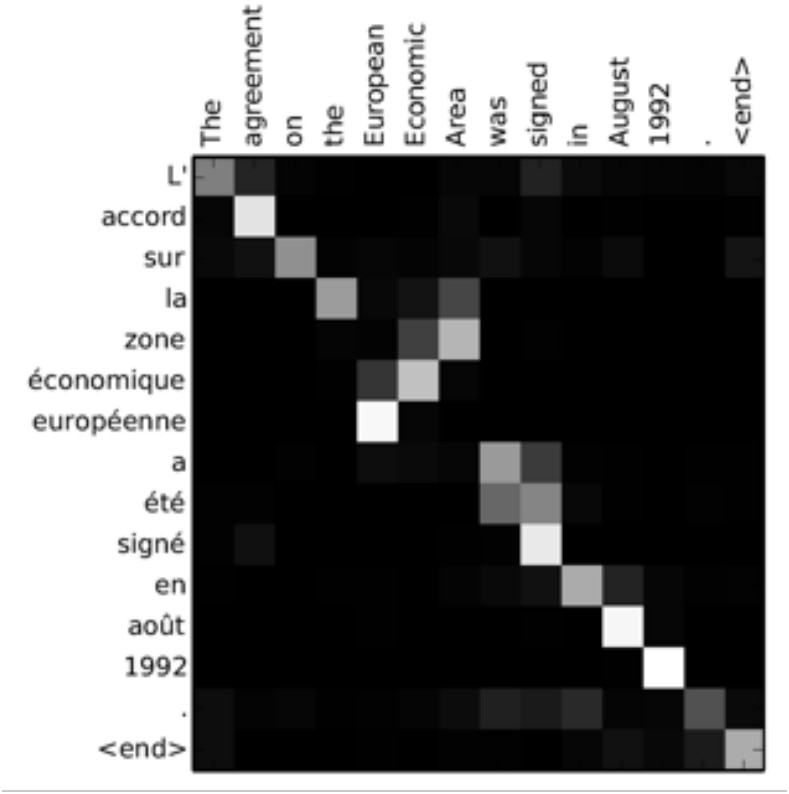
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L} \exp(e_{ik})} \quad (2)$$

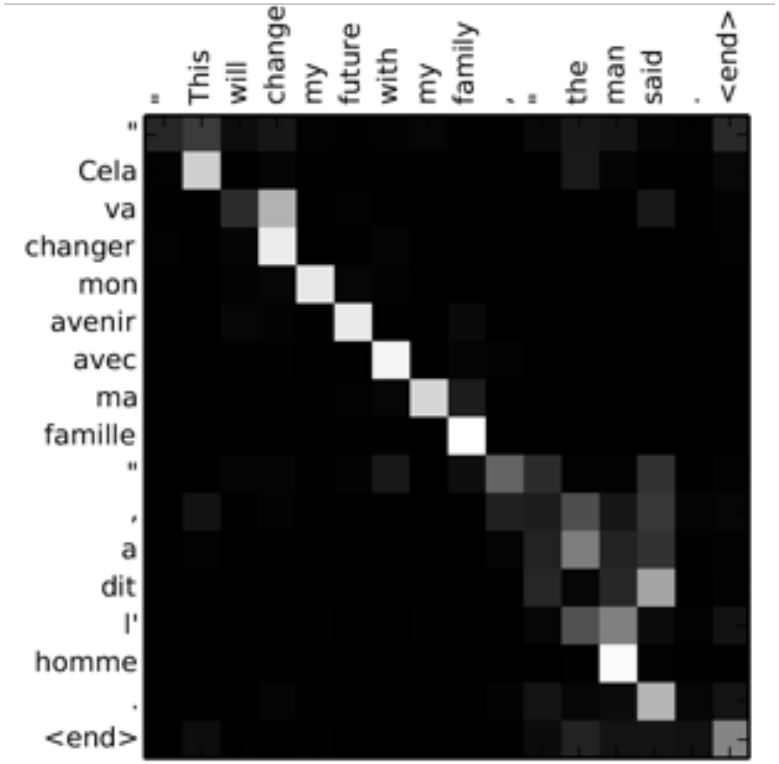
- nonlinearity (tanh) is crucial!
- simplest model possible
- Vh_j is precomputed => quadratic complexity with low constant



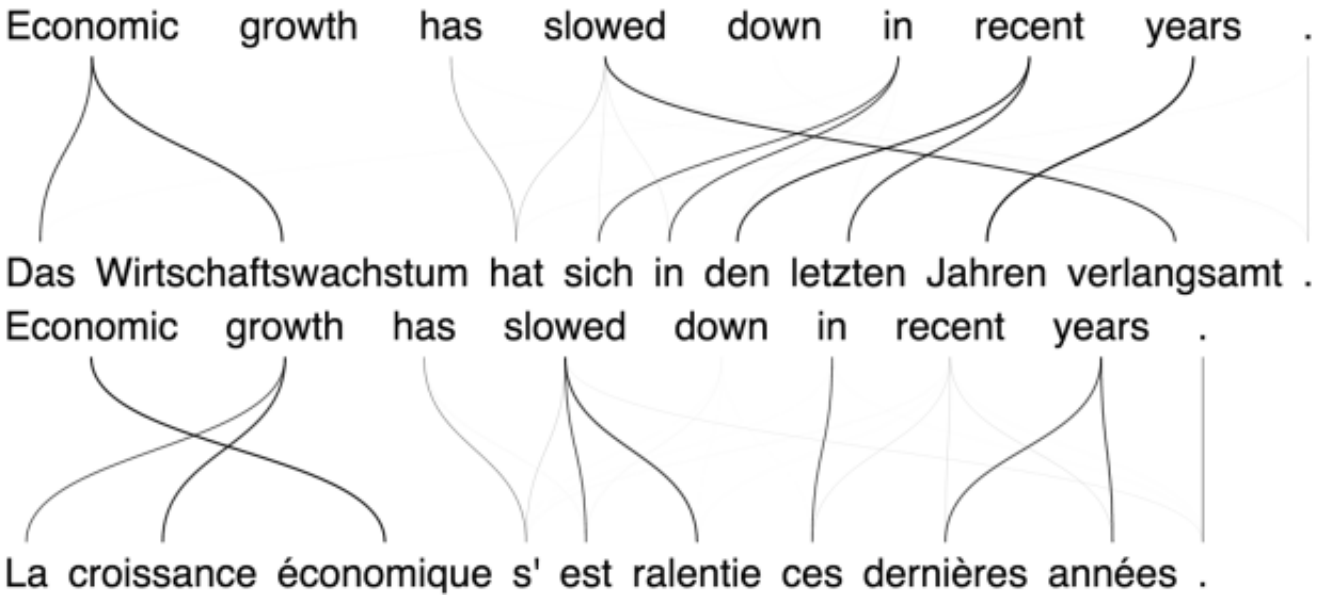
Alignment - cont’d

Learnt alignments



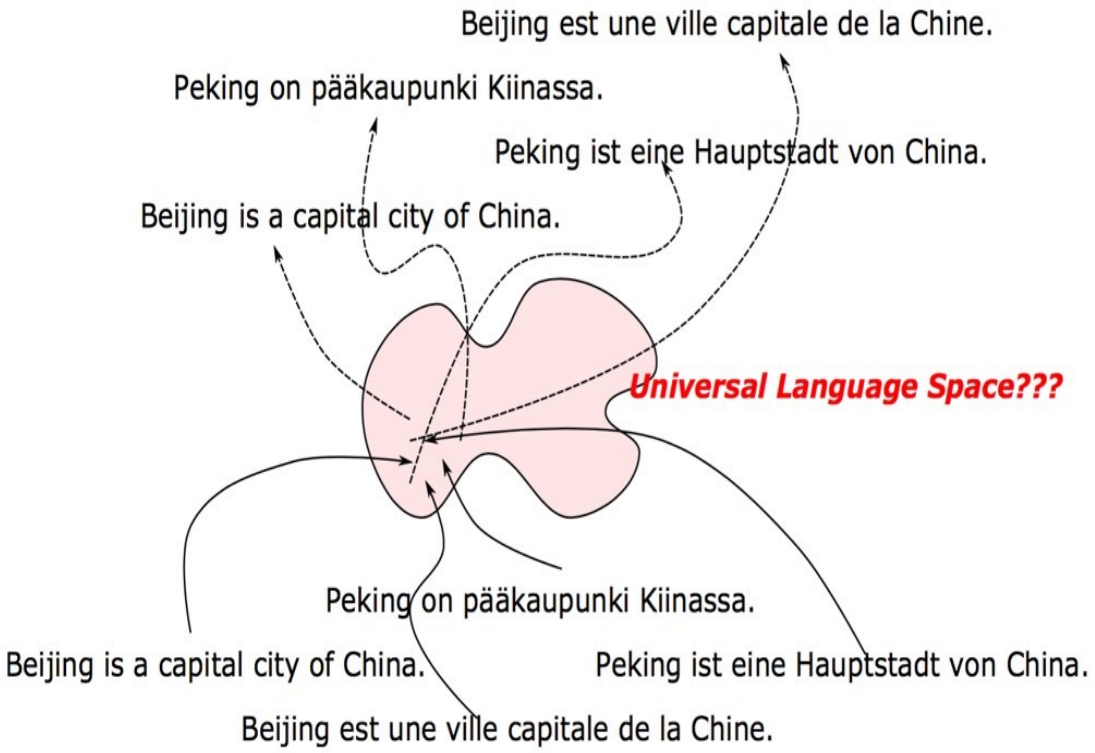


Neural net will capture underlying structures



As long as the structures are needed to achieve the goal

Multilingual Translation



Dong et al. (2015)

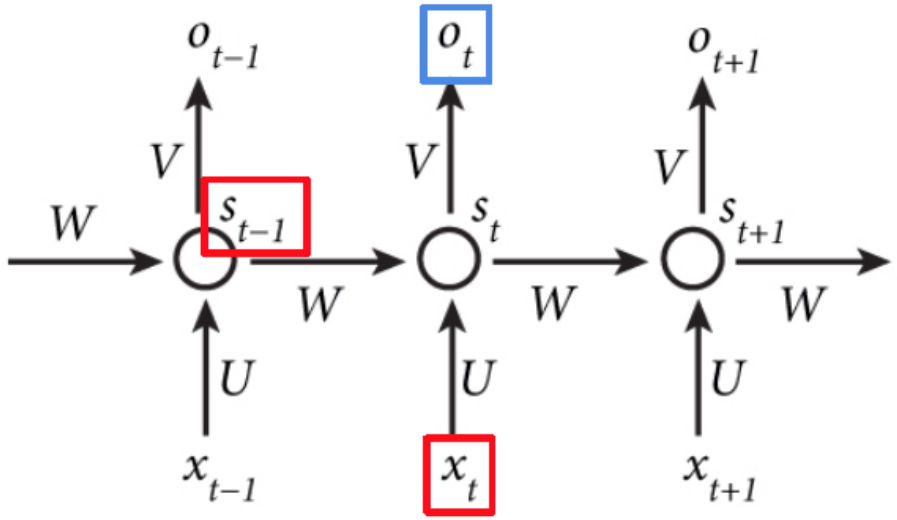
Definition of Optimization

Optimization

Minimization
Consider a objective function $J : \Theta \to \mathbb{R}$.
For any $\theta \in \Theta$, finds the θ^* such that $J(\theta^*) \leq J(\theta)$.

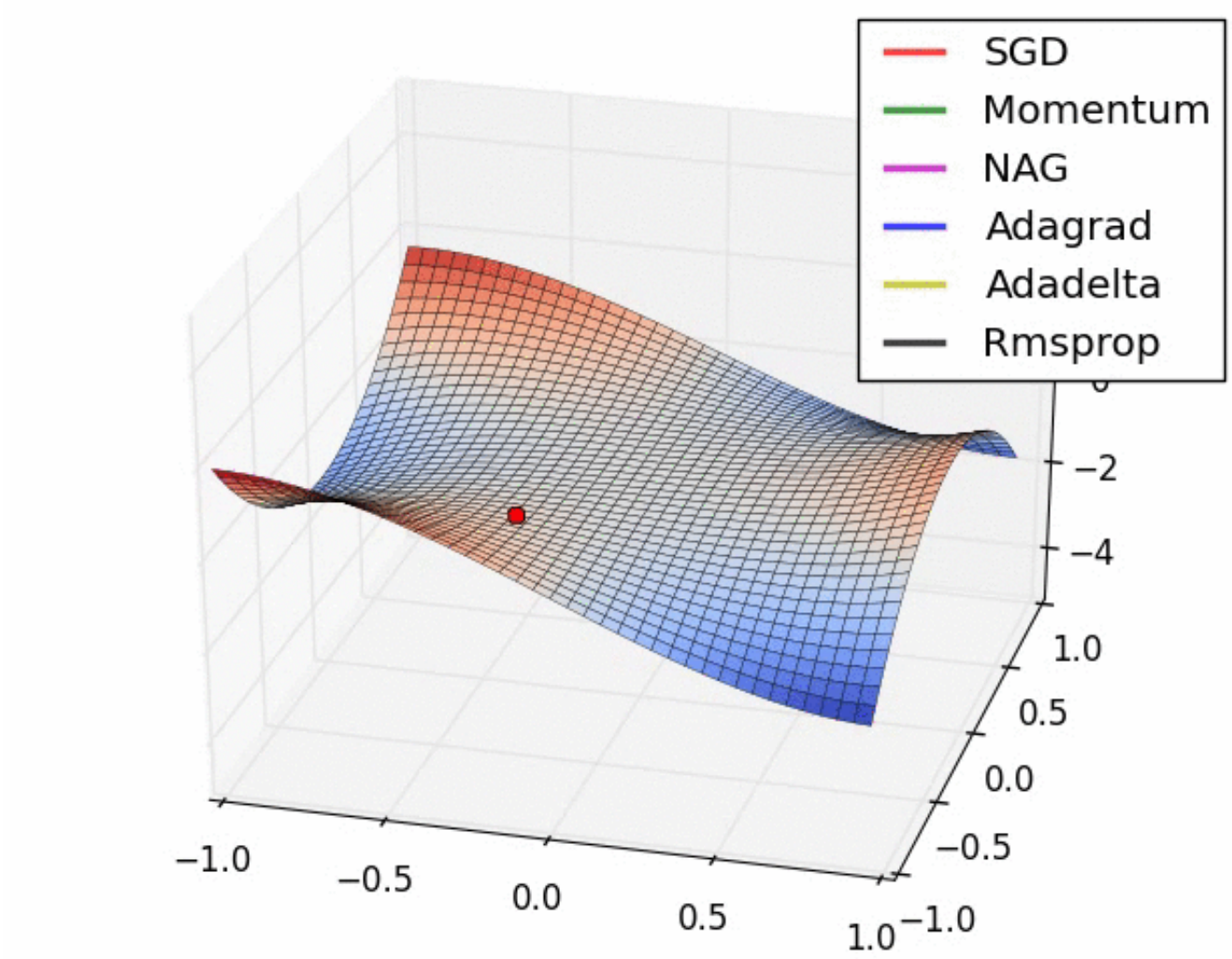
Maximization
Consider a objective function $J : \Theta \to \mathbb{R}$.
For any $\theta \in \Theta$, finds the θ^* such that $J(\theta^*) \geq J(\theta)$.

Optimization in RNN



학습을 하면서, input vector를 통해 RNN cell에서 예측한 prediction과
실제 데이터의 label과의 오차 함수를 최소화시키는 U, W, V 를 구함

Optimization Simulation



http://www.denizyuret.com/2015/03/alec-radfords-animations-for.html

Adam Optimizer

Update Equation

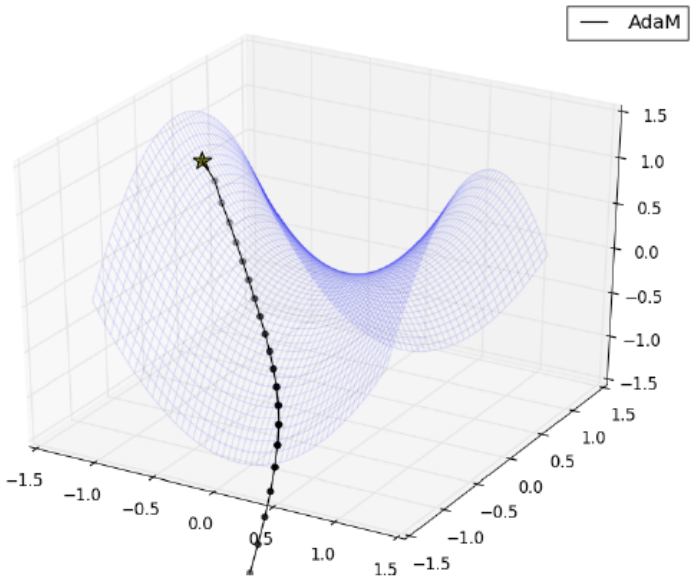
$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$$

Key Idea

- AdaDelta가 Hessian term에 momentum을 취한 것처럼, gradient와 Hessian에 모두에 momentum approach를 적용함

Adam Simulation

$\eta = 0.1,\ \beta_1 = \beta_2 = 0.9,\ \epsilon = 10^{-8},\ \text{iter} = 20$



NAG Optimizer

Update Equation

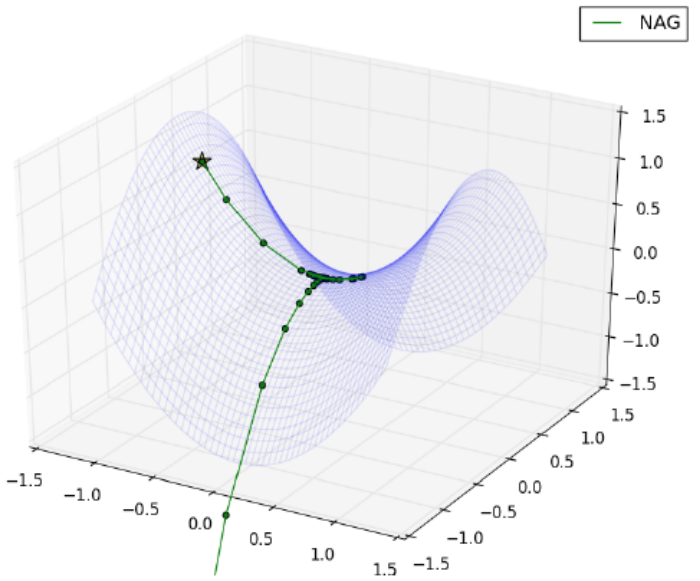
$$v_t = \gamma\ v_{t-1} + \eta \nabla_\theta J(\theta_t - \gamma\ v_{t-1})$$
$$\theta_{t+1} = \theta_t - v_t$$

Key Idea

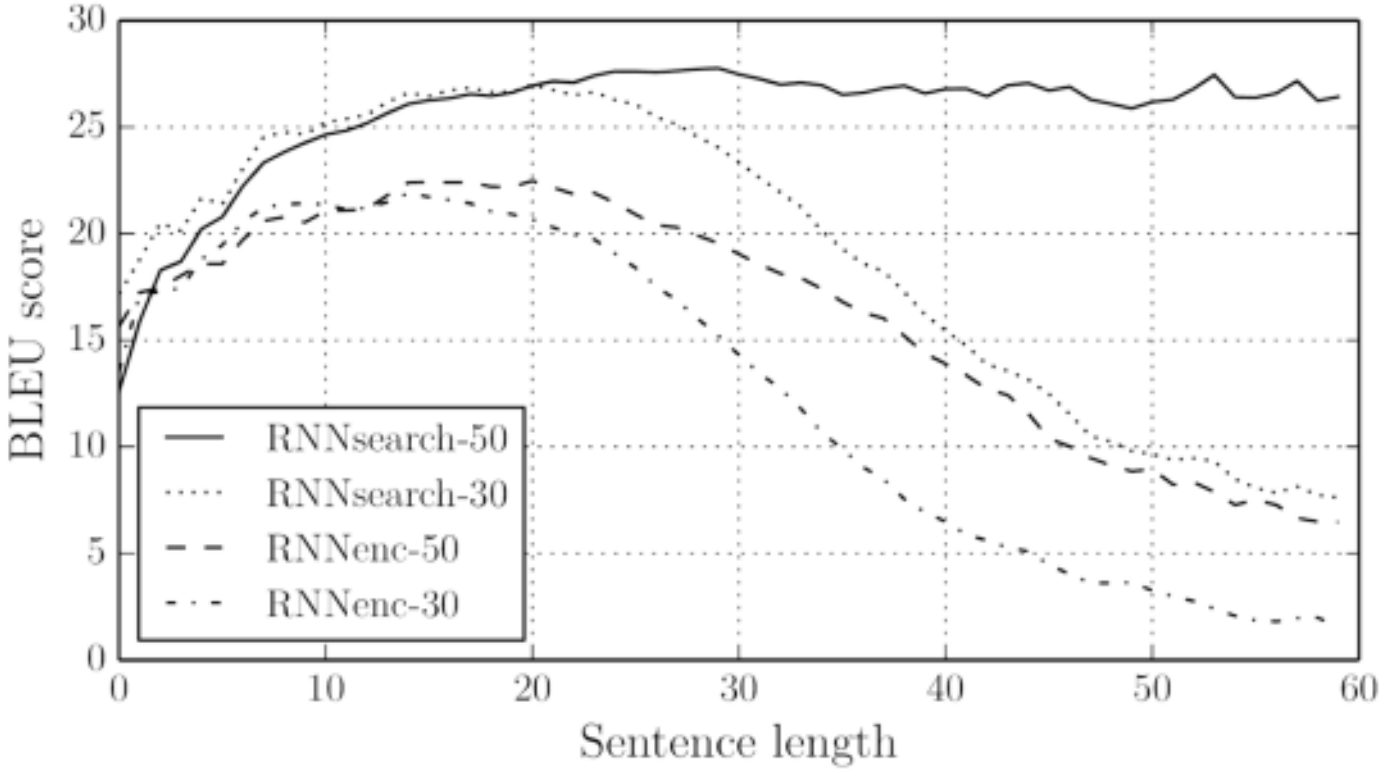
- Momentum 방식과 유사하나 gradient의 측정 지점이 다름
- 다음 iteration으로 estimation 된 지점에서의 gradient를 측정함

NAG Simulation

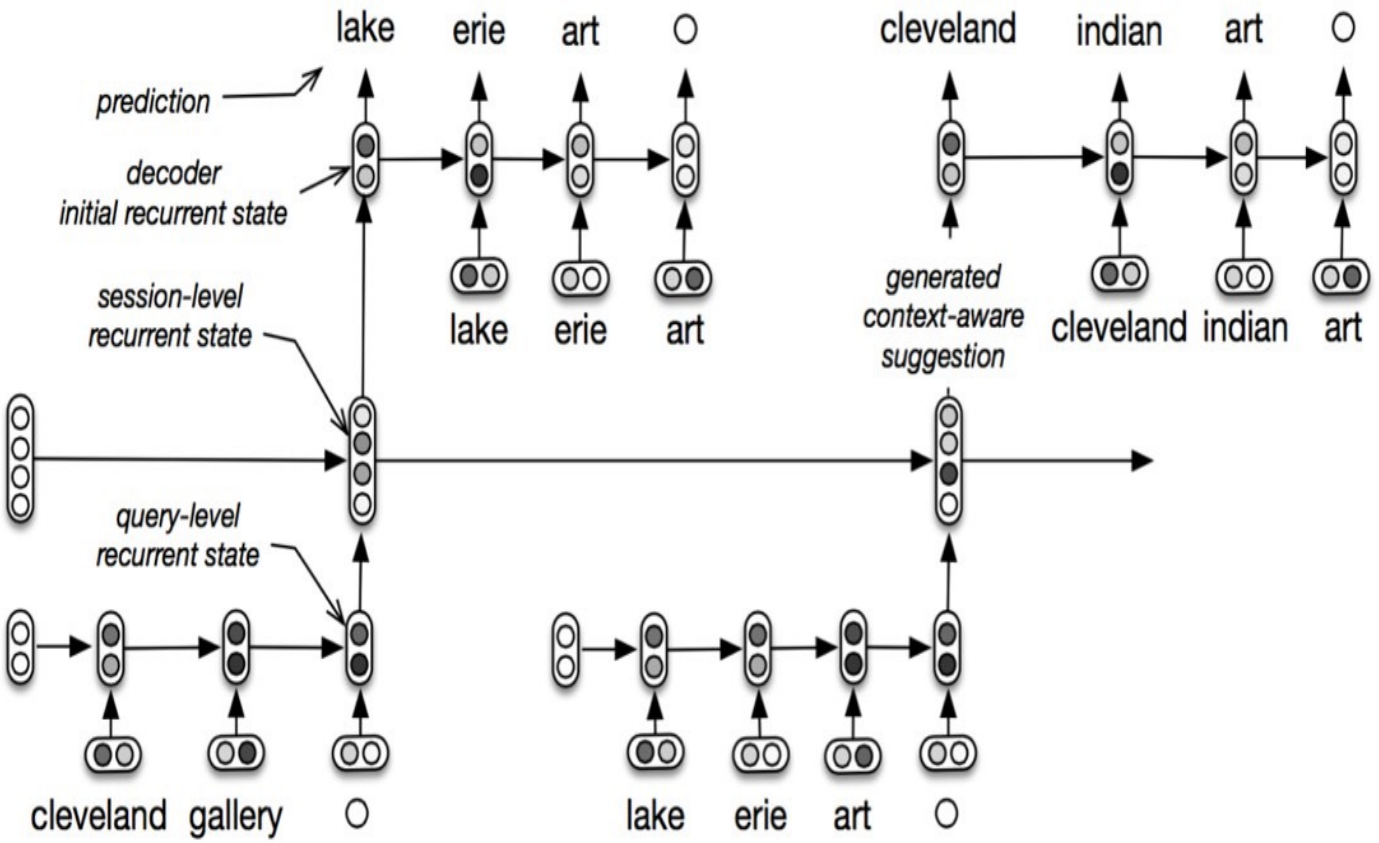
$\eta = 0.1,\ \gamma = 0.9,\ \text{iter} = 60$



Performances of RNN with different size



Toward Discourse-level MT



Hierarchical Recurrent Encoder–Decoder (HRED) by Sordoni et al. (2015)

독해

Machine Comprehension

기계 - 인공지능

Machine Comprehension

기존 기술

온톨로지 기반 기계 독해

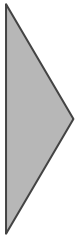
: 개체 관계 계층 및 속성을 명확하고
정형화된 **규칙**을 통하여 추론하는 방식

장점

질의가 정형화된 규칙에 맞도록 입력된
경우에 한하여 추론 결과가 정확함

단점

- 관계 및 속성의 스키마 관리 비용
- 개체의 계층별 관계 및 속성의 탐색과 추론 비용
- 비정형화된 실제 데이터와 정형화된 규칙 사이의 괴리에 인한 성능 열화
- 어휘 배열의 다양성을 고려하지 못함



개선 기술

딥러닝 기반 기계 독해

: 주어진 지문으로부터 인공지능이 **학습**을 통해 질의에 대한 답변을 추론하는 방식

기계 독해 알고리즘

RNN, BiLSTM, Paragraph2Question, Pointer Network …

맥락 파악 모델

State-based Response Model,
Multi-context Response Model …

특징

인공지능의 언어 이해력을 이용하며, 지속적으로 학습을 수행
대화 성능 검증 플랫폼과 아키텍처
딥러닝 학습데이터 구축 도구 및 서비스 프레임워크 제작

Machine Comprehension - cont'd

외국 현황

방대한 분량의 기계 독해 학습 데이터 공개

→ Microsoft (MCTest, MSMARCO), Facebook (bAbI), Stanford (SQuAD), CNN/DailyNews 등

최근 딥러닝 자연어 처리 연구의 주류 분야

→ Microsoft, Google, Facebook, IBM, Salesforce 등 세계 유수의 기업과 Stanford, Carnegie Mellon, Singapore 대학 등 학계에서 다수의 연구 실적

국내 현황

한국어의 기계 독해 공개 학습 데이터 전무

한국어 기반 기계 독해 분야 상용 서비스 사례 전무

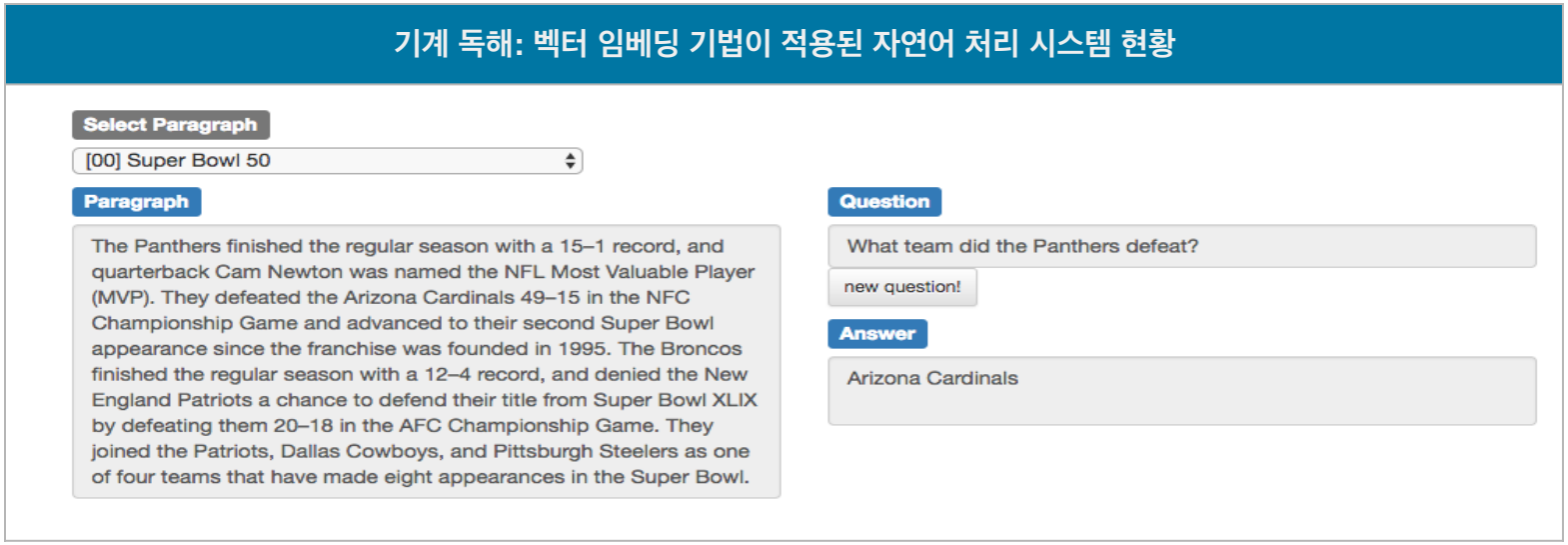
기계 독해 한국어 연구 성과 거의 없음

Deep Learning based Machine Comprehension

- Machine Comprehension 문제 해결 방법
  - 주어진 지문에 대한 이해
  - 입력된 질문에 대한 이해
  - 지문과 질문 사이의 관계 파악
  - 지문 내에서 질문에 대한 답변을 찾아 제공

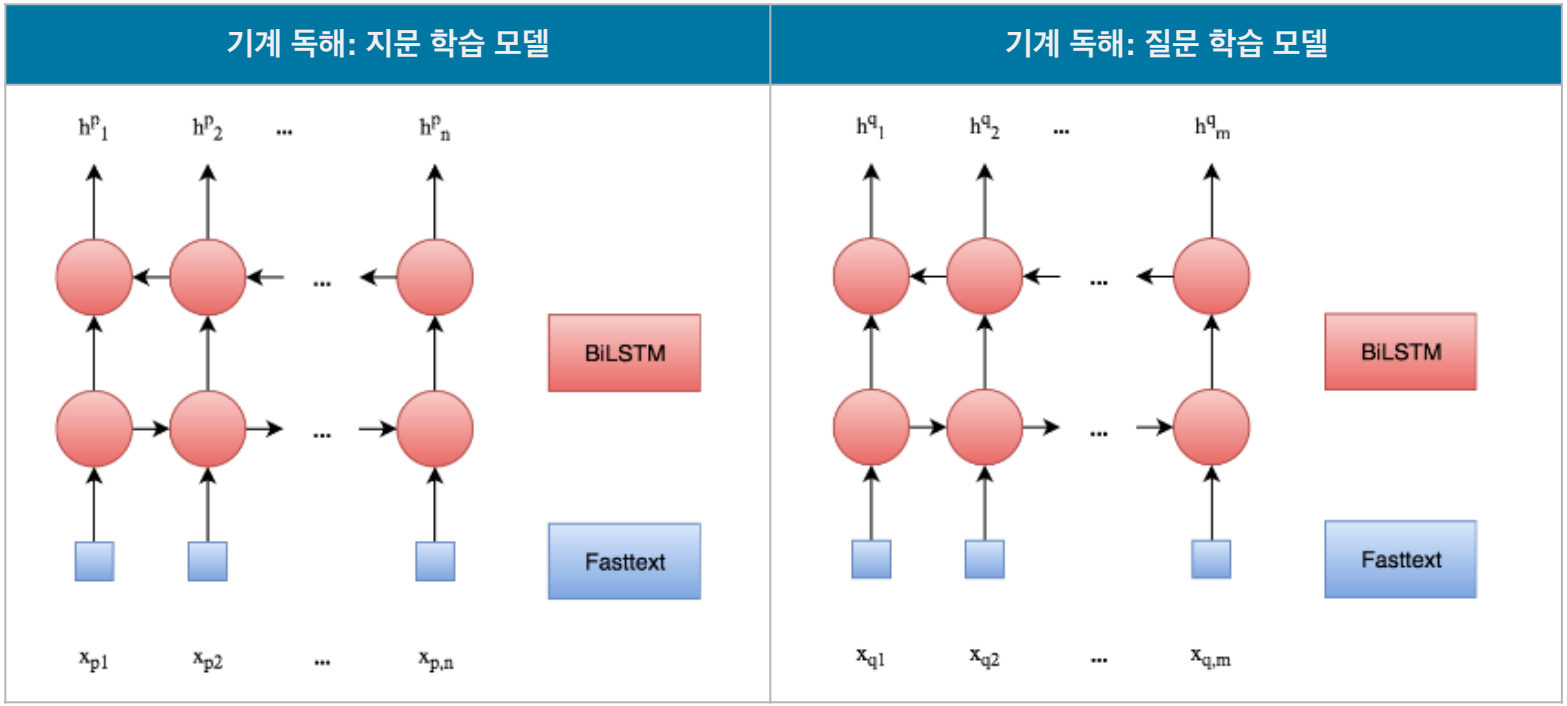
Deep Learning based Machine Comprehension - cont'd

- 딥러닝 모델을 기본으로 한국어 특성에 맞게 최적화하여 활용 시도
- 비정형 데이터에 최적인 딥러닝 기반 벡터 임베딩 기법 적용
  → 지문 및 질문의 문맥 파악 정확도 향상
- 방대한 온톨로지 지식의 구축 및 추론이 불필요
  → 시스템 관리 및 리소스 비용 절감



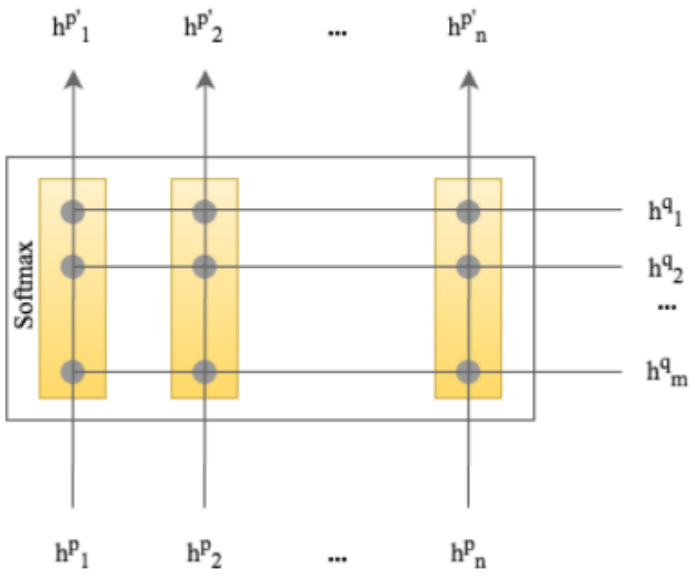
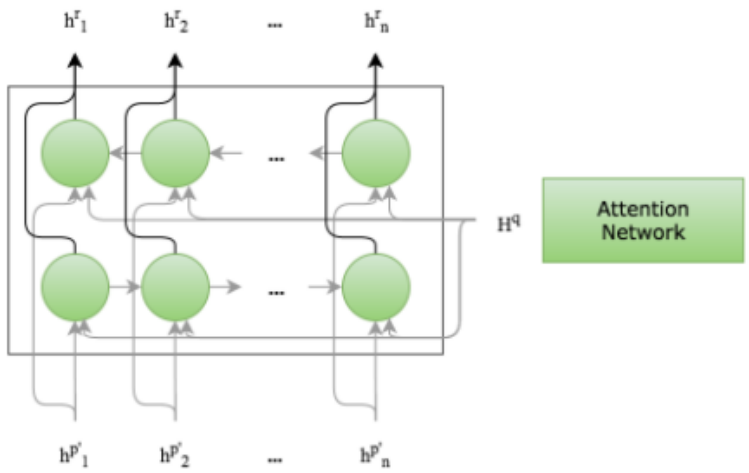
Paragraph/Question Understanding

- 주어진 지문에 대한 이해 & 입력된 질문에 대한 이해
- 지문 및 질문은 단어들의 시퀀스 형태로 변환
- 각 단어들은 GloVe, fasttext 등의 기법을 적용하여 임베딩 벡터 형태로 변환
- 딥러닝 기반의 BiLSTM 모델을 적용하여 지문과 질문의 내용의 임베딩 계산



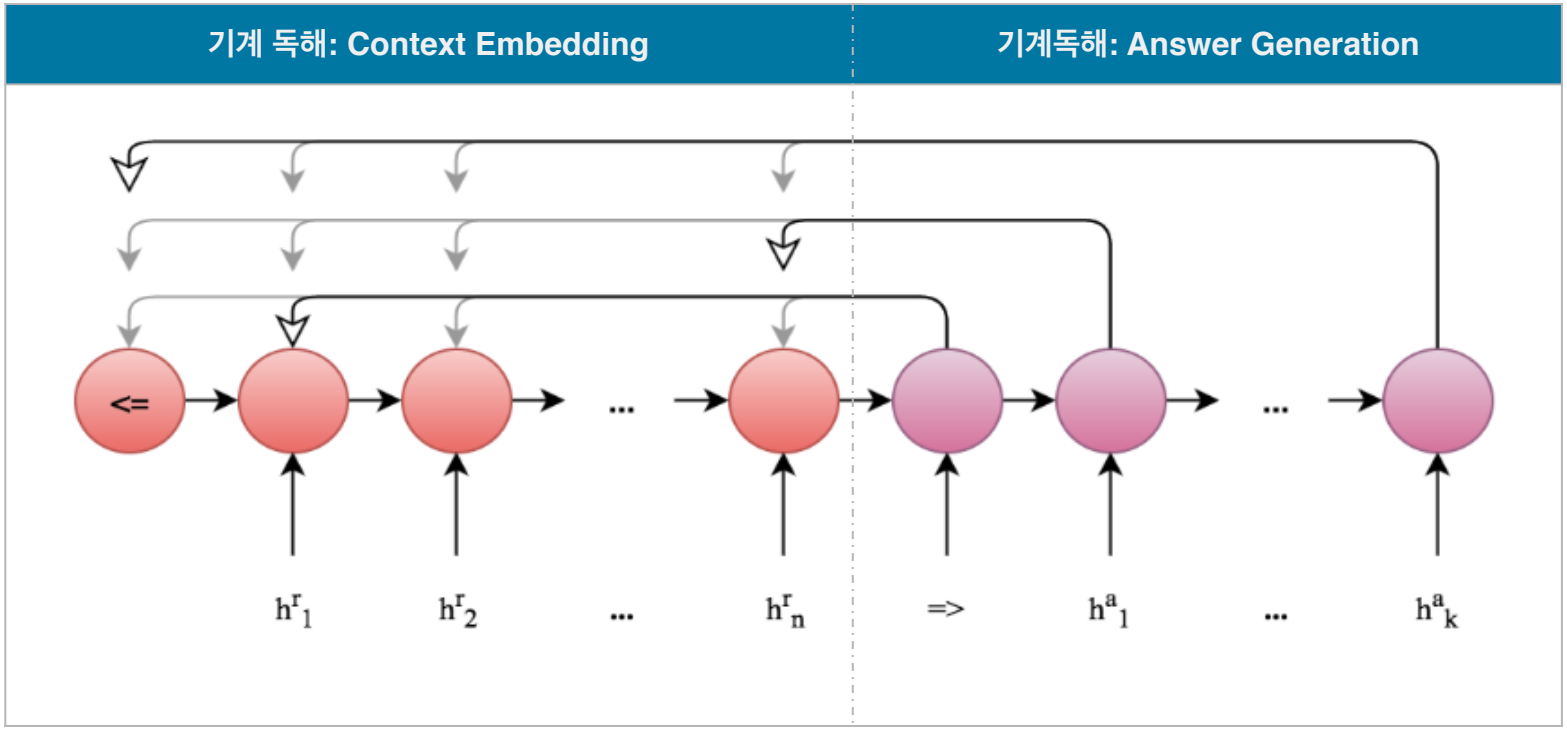
Paragraph and Question Coupling

- 지문과 질문 사이의 관계 파악
- Attention Network 기법을 활용하여 질문의 내용과 지문의 관련도를 측정
- 지문의 관점에서 질문의 각 단어 사이의 연관성 파악
- 질문의 관점에서 지문의 각 단어 사이의 연관성 파악

기계 독해: Paragraph2Question 모델	기계독해: Question2Paragraph 모델
	

Answer Finding

- 지문 내에서 질문의 답변에 해당하는 내용을 찾아서 제공
- 기계 독해 문제에서 답변은 지문 안에 있는 표현을 찾아서 보여주는 방식
- 이전 단계에서 파악한 질문과 지문의 관계 정보 및 Pointer Network 기법을 적용하여 지문의 단어 중 어느 단어들을 조합하여 답변을 생성하는지 계산

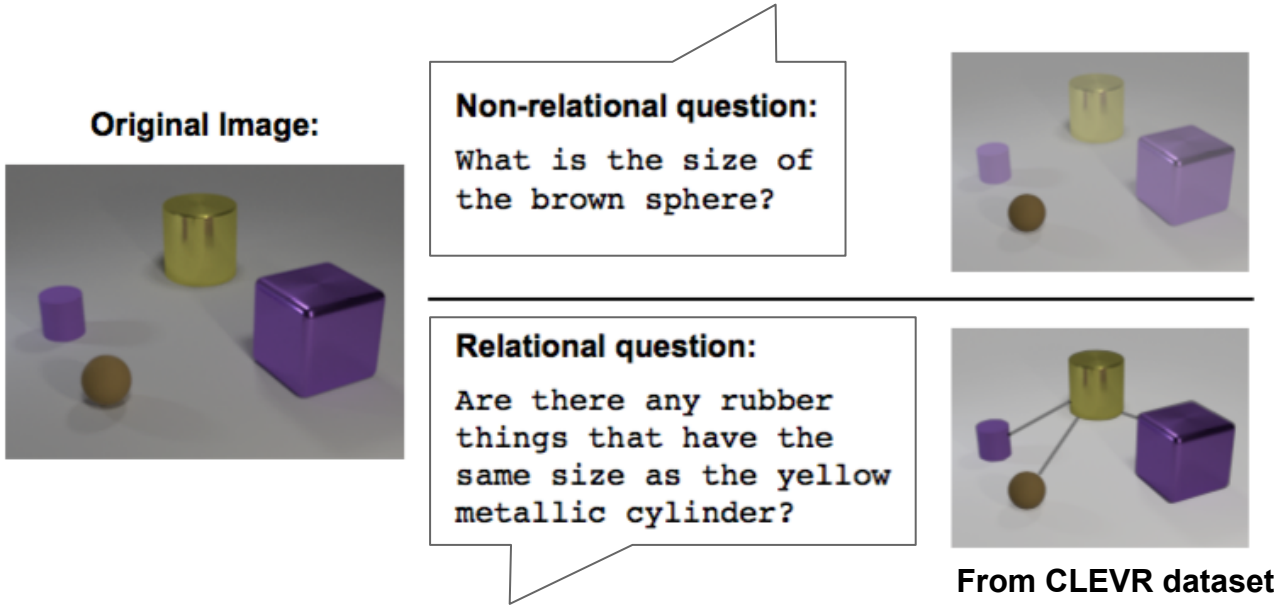


Relational Reasoning

객체들 및 그들의 속성들 사이의 관계를 파악하는 것

비관계형 질문

갈색공 하나에 대한 정보만 알고 있으면 답변 가능



From CLEVR dataset

관계형 질문

노란색 원기둥 모양의 물체에 대한 정보 이외 그림 내
다른 사물들과의 관계 정보를 파악해야만 답변 가능

Relation Networks

A simple neural network module for relational reasoning DeepMind (2017)

Basic Model

Neural network to compute
(potential) relations between all object pairs

$$RN(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right)$$

Neural networks to infer pairwise relations

For Visual QA

$$RN(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j, q) \right)$$

Relation Networks Performances

CLEVR 데이터를 사용한 visual question answering 문제에 적용

현재까지 알려진 모델 대비 월등한 성능 향상

사람보다도 더 뛰어난 결과를 보임

Model	**Overall**	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	**95.5**	**90.1**	**97.8**	**93.6**	**97.9**	**97.1**

Generative Models for Natural Languages

Ideas

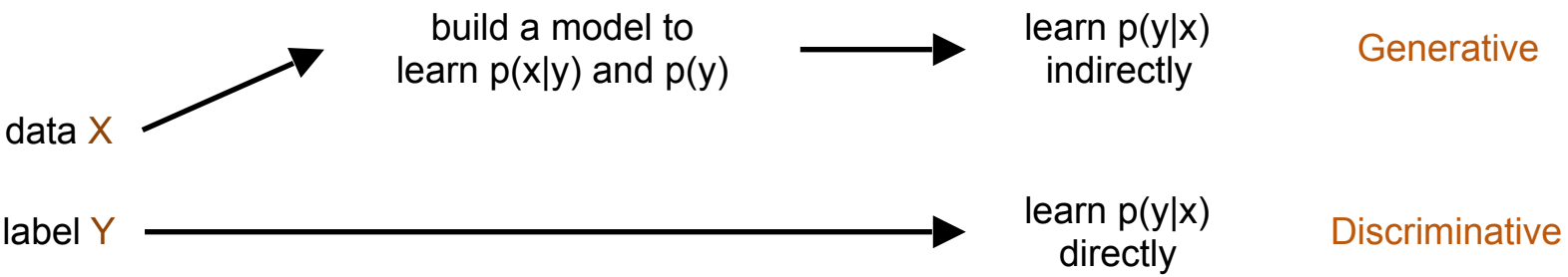
A generative model describes **how data is generated**

"What I cannot create, I do not understand."

Richard Feynman[1)]

A quick sketch

comparison between discriminative models



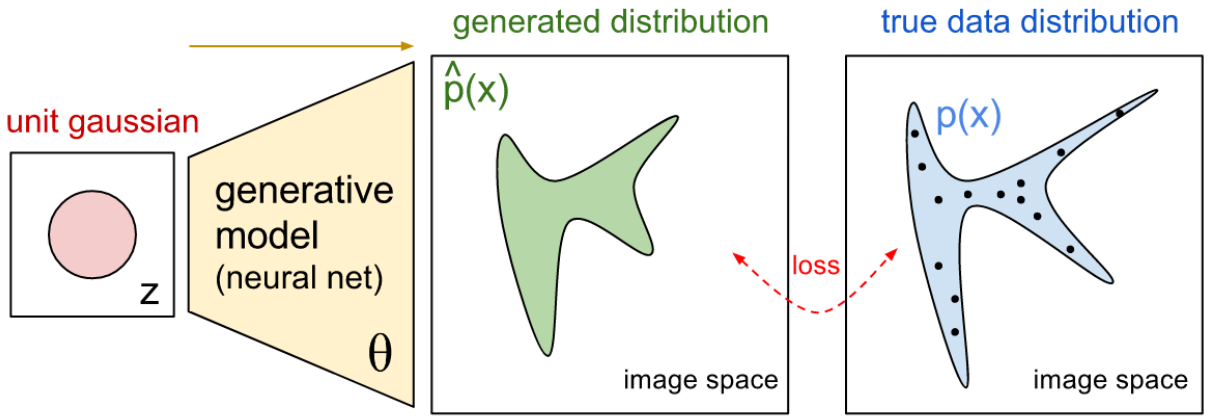
1) OpenAI blog: Generative Models (https://blog.openai.com/generative-models/)

Ideas - cont'd

Intuition

A model for randomly generating observable data values,

typically given some hidden parameters



The Generative Models in Natural Language

Generative models in Language that we are familiar with :

Naive Bayes, GMM, LDA, HMM, Boltzman machines, Seq-to-Seq,…

Three most popular approaches that utilize neural networks :

- Autoregressive models: pixel-RNN (char-RNN)
- Generative Adversarial Networks (GANs)
- Variational Autoencoders(VAEs)

VAE as a generative model

- Generative models are known to have following drawbacks:
    1. They might require strong assumptions about the structure in the data
    2. They might make severe approximations, leading to suboptimal models
    3. They might rely on computationally expensive inference procedures like MCMC

- VAE handle these problems by
    1. Building the model on top of standard function approximators — neural networks
    2. Training with stochastic gradient descent, making the training faster

- Why so popular: weak assumptions, fast training, relatively small error in approximation

Latent Variable models with Neural Network

Recall
The generative model

$$p(x, z) = p(z)\, p(x|z)$$

- We assume $p(x, z)$ is parameterized by neural networks with parameter θ

$$p_\theta(x, z)$$

- Train to find θ that maximize some objective function

Variational Autoencoders

- Typically in VAE, we define by

$$p_\theta(z) = \mathcal{N}(0, I)$$

Prior

$$p_\theta(z|x) = \mathcal{N}\left(x|\mu_\theta(z), \sigma_\theta^2(z)\right)$$

Generation

$$p_\theta(x) = \int p_\theta(x|z) p_\theta(z)\, dz$$

$$= \int \mathcal{N}\left(\mu_\theta(z), \sigma_\theta^2(z)\right) p_\theta(z)\, dz$$

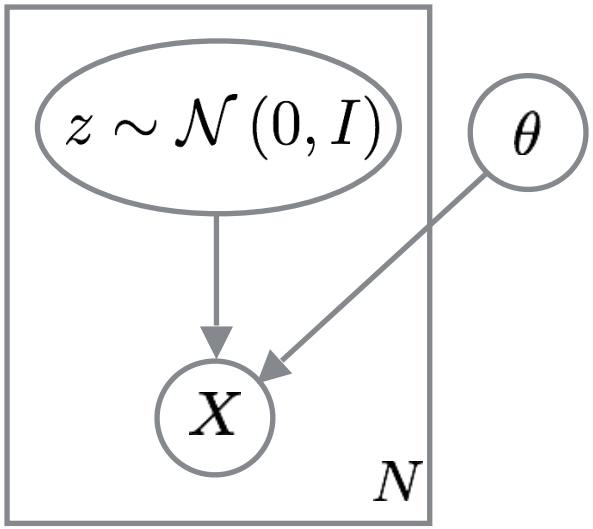
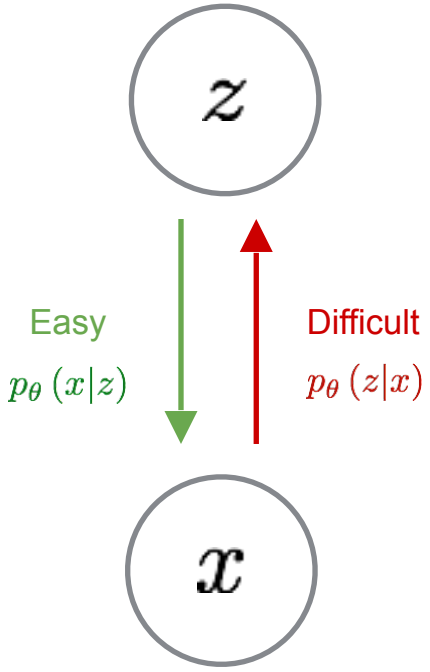


plate notation
we can sample N times from z and X, while θ fixed

Difficulties

- Posterior $p_\theta(z|x)$ is intractable

$$
\begin{aligned}
p_\theta(z|x) &= \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} \\
&= \frac{p_\theta(x|z)p_\theta(z)}{\int p_\theta(x,z')dz'} \\
&= \frac{p_\theta(x|z)p_\theta(z)}{\int p_\theta(x|z')p_\theta(z')dz'}
\end{aligned}
$$



- Note that it is **not able to calculate** the integral analytically
- when z' is in high dimension, the integral is difficult to estimate

Variational Inference

Instead of $p_\theta(z|x)$,

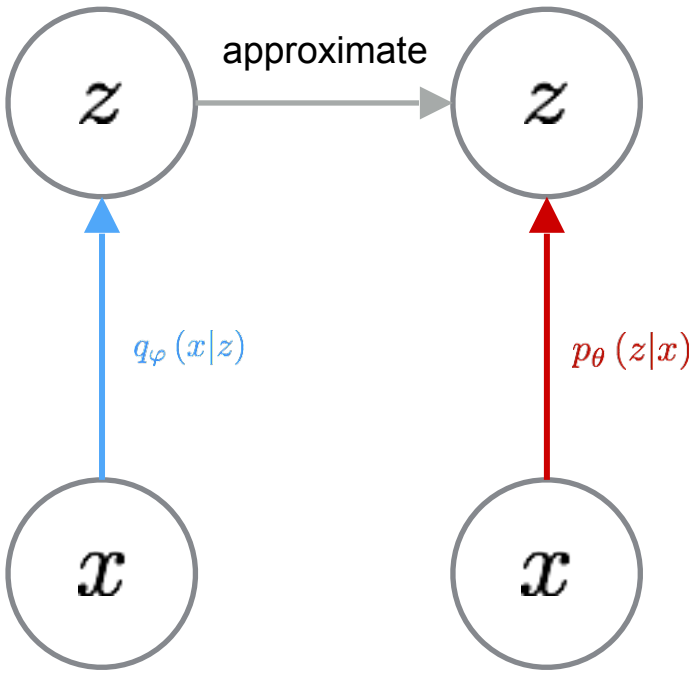
we approximate it with $q_\varphi(x|z)$

In addition to θ, we find φ

that approximates $p_\theta(z|x)$ well

Choice of $q_\varphi(x|z)$ should be easy

to be calculated or to be sampled from



Evidence Lower Bound (ELBO)

- The objective : maximize ELBO instead of the original

$$\log p_\theta(x) \longrightarrow \mathcal{L}(x;\theta)$$

$$= \log \int p_\theta(x,z)dz \qquad \uparrow$$

$$= \log \int \frac{q_\varphi(z|x)p_\theta(x,z)}{q_\varphi(z|x)}dz \qquad \mathcal{D}_{KL}\left(q\left(z|x\right) \parallel p\left(z|x\right)\right) \qquad \downarrow$$

$$\geq \int \frac{q_\varphi(z|x)\log p_\theta(x,z)}{q_\varphi(z|x)}dz \longrightarrow \tilde{L}\left(x;\theta,\varphi\right)$$

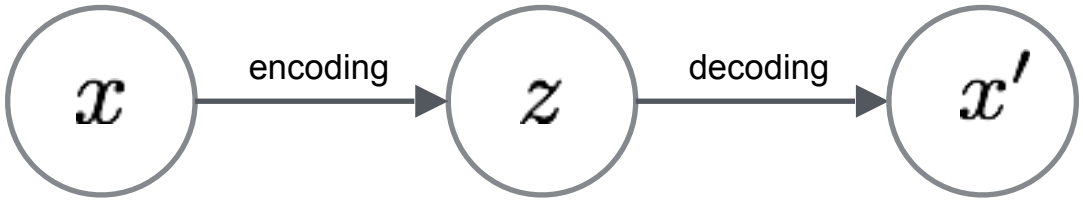
Calculating the gradients

- Reinforce
- Control variate
- Reparameterization trick (Stochastic Gradient Variational Bayes, SGVB)

  Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes

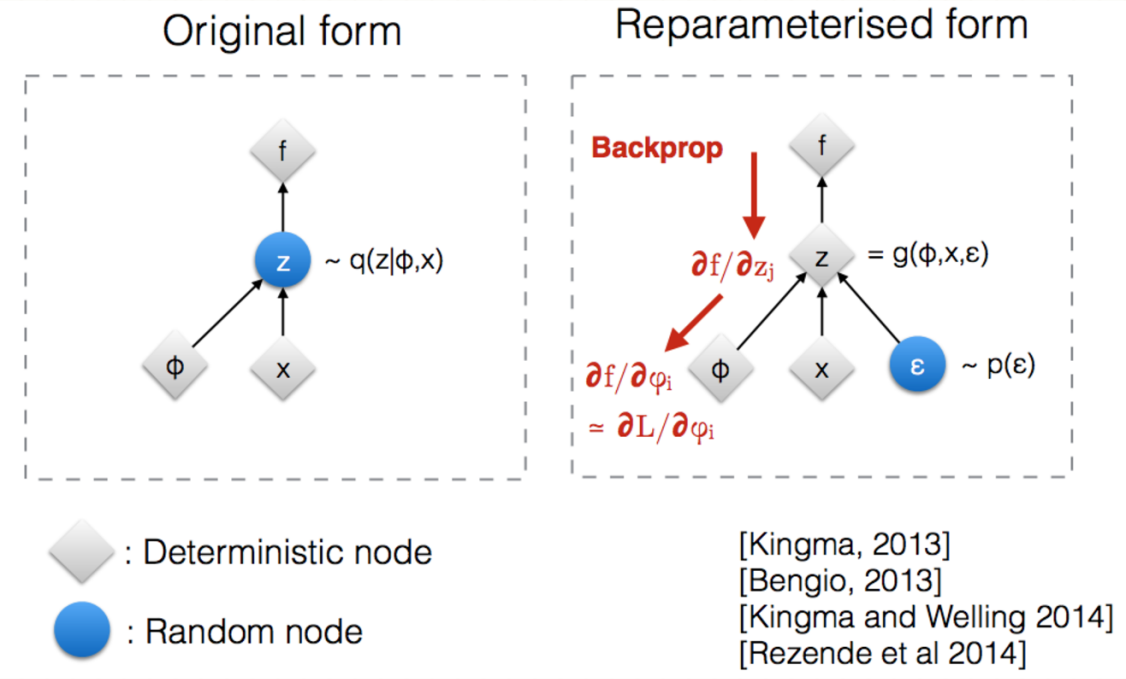
  - move the sampling to an input layer

    Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models.



Why reparametrization trick

- Backpropagation cannot flow through a random node
- Introducing a new parameter ϵ allows to reparameterize z in so that backprop to flow through the deterministic nodes

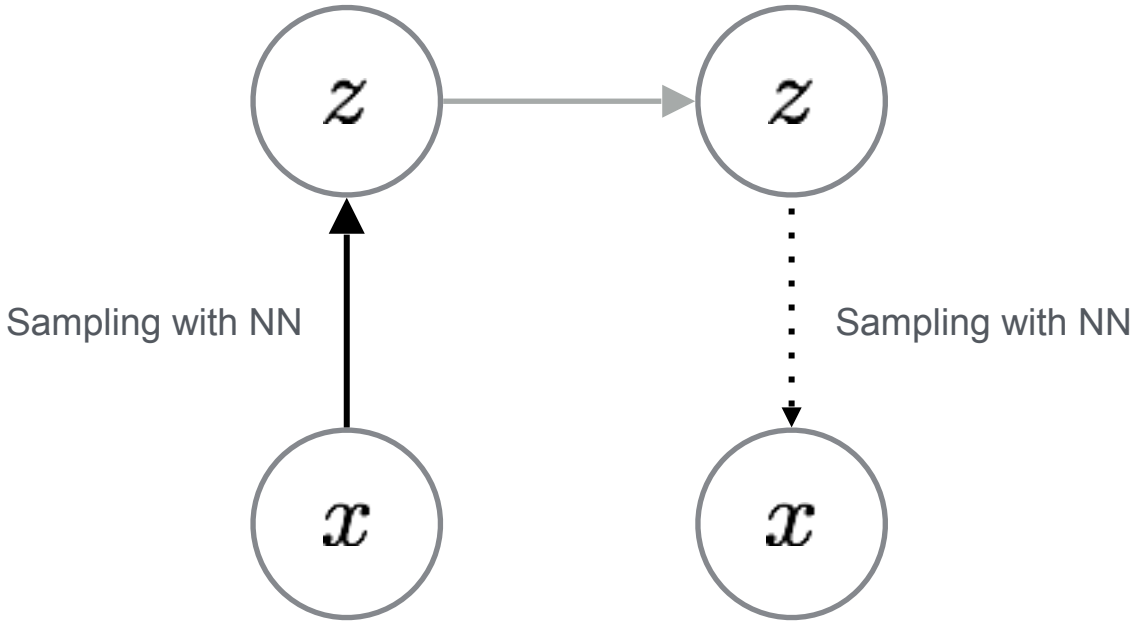


Training Procedure

The ELBO can be decomposed into 2 parts

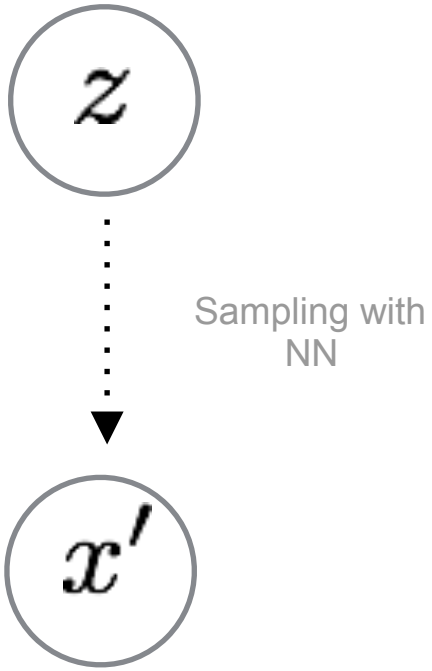
$$E_{z\sim Q}[\log p(x|z)] - D_{KL}[q(z|x)||p(x|z)]$$

Reconstruction loss Regularization loss



Generation

We can generate data points with trained generative models



VAE applied to sentence

Neural Variational Inference for Text Processing Yishu Miao, Lei Yu, Phil Blunsom (2015)

Modeling

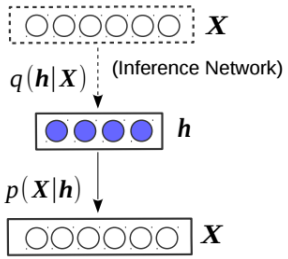


Figure 1. NVDM for document modelling.

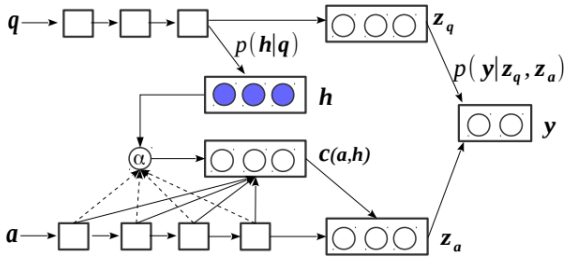


Figure 2. NASM for question answer selection.

http://proceedings.mlr.press/v48/miao16.pdf

VAE applied to sentence

Empirical Results

Q1	how old was sue lyon when she made lolita
A_{NASM}	the actress who played lolita , sue lyon , was fourteen at the time of filming .
A_{LSTM}	the actress who played lolita , sue lyon , was fourteen at the time of filming .
Q2	how much is centavos in mexico
A_{NASM}	the peso is subdivided into 100 centavos , represented by " _UNK_ "
A_{LSTM}	the peso is subdivided into 100 centavos , represented by " _UNK_ "
Q3	what does a liquid oxygen plant look like
A_{NASM}	the blue color of liquid oxygen in a dewar flask
A_{LSTM}	the blue color of liquid oxygen in a dewar flask

http://proceedings.mlr.press/v48/miao16.pdf

VAE applied to conversation

Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders

Tiancheng Zhao, Ran Zhao and Maxine Eskenazi (2017)

Intuition

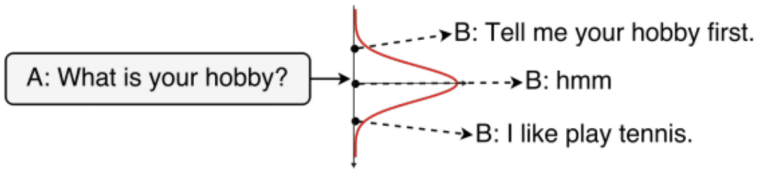
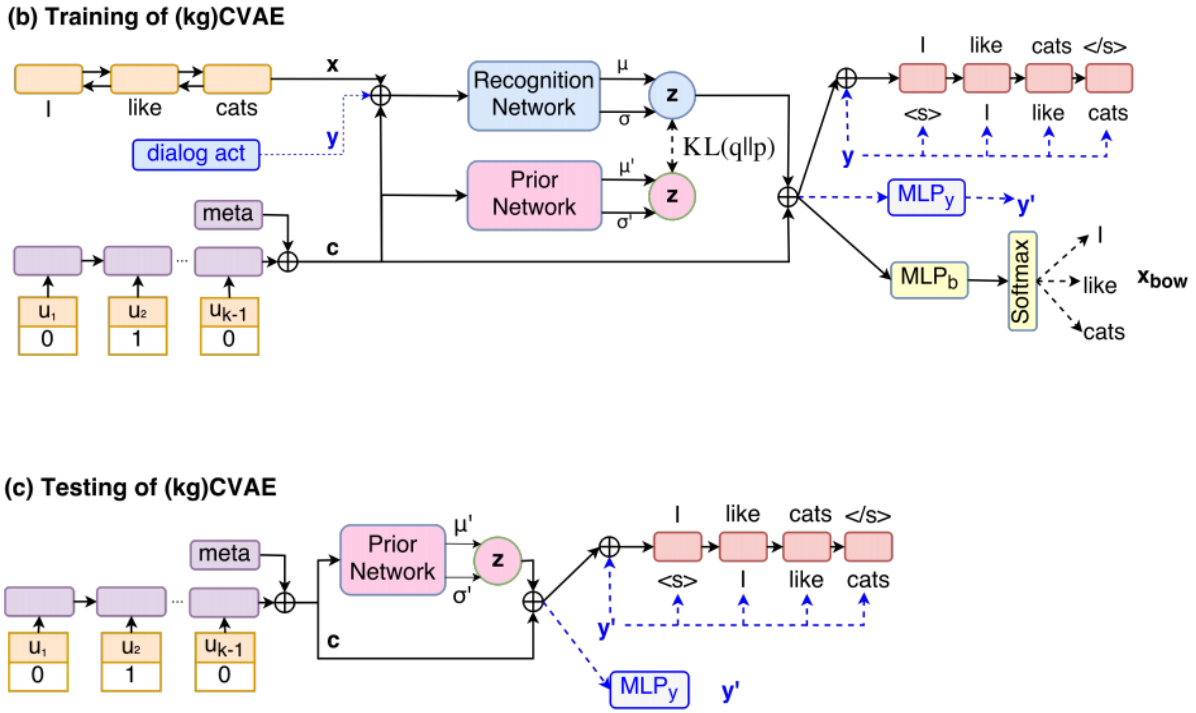


Figure 1: Given A's question, there exists many valid responses from B for different assumptions of the latent variables, e.g., B's hobby.

https://arxiv.org/pdf/1703.10960.pdf

VAE applied to conversation - cont'd

Training & Testing



https://arxiv.org/pdf/1703.10960.pdf

VAE applied to conversation - cont'd

Empirical Results

Example 1-Topic: Recycling **Context**: A: are they doing a lot of recycling out in Georgia? **Target-B** (statement): well at my workplace we have palaces for aluminium cans and we have a separate trash can for recyclable paper	
Baseline+Sampling	**kgCVAE+Greedy**
1. well I'm a graduate student and have two kids	1. (non-understand) pardon
2. well I was in last year and so we've had lots of recycling	2. (statement) oh you're not going to have a curbside pick up here
3. I'm not sure	3. (statement) okay I am sure about a recycling center
4. well I don't know I just moved here in new york	4. (yes-answer) yeah so

Example 2-Topic: Child Care **Context**: A: you know a private home to take their children to when they're young until they hit the preschool age and they **Target-B** (backchannel): uh-huh	
Baseline+Sampling	**kgCVAE+Greedy**
1. um - hum	1. (backchannel) uh-huh
2. yeah	2. (turn-exit) um-hum
3. um - hum	3. (backchannel) yeah
4. uh-huh	4. (statement) oh yeah I think that's part of the problem

https://arxiv.org/pdf/1703.10960.pdf

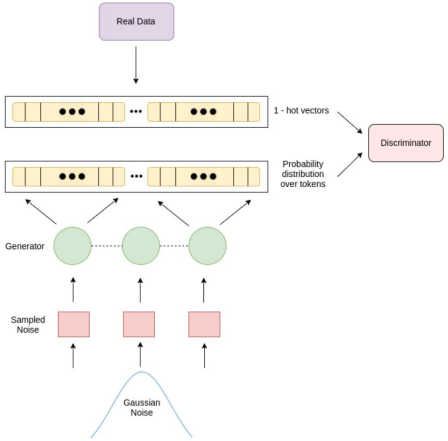
Deep Learning and NLP : A Controversy

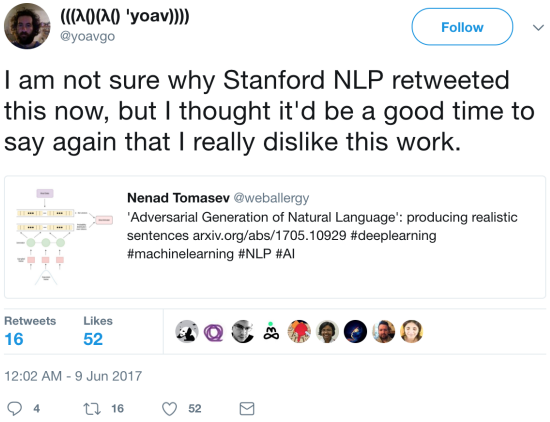
The original Paper: Adversarial Generation of Natural Language (https://t.co/HGEijW2KkA)

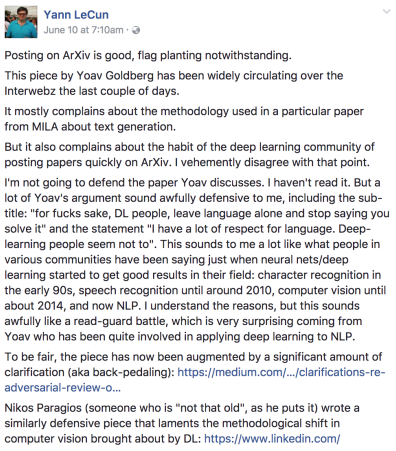
An Adversarial Review of “Adversarial Generation of Natural Language (https://t.co/bkj2EbOIcP)

Yann’s Response (https://www.facebook.com/yann.lecun/posts/10154498539442143)

A Response to Yann LeCun’s Response. (https://t.co/elwEJXpXvq)







End of Document

Company.AI

facebook.com/companyai
twitter.com/companyai
http://www.company.ai
all@company.ai