텐서플로와 머신러닝으로 시작하는 자연어처리

Lecture 5
2019. 06. 29

Text Generation

Text Generation

- What is Text Generation
- What is Neural Machine Translation
- Papers
    - Sequence to Sequence
- How do we evaluate?
- Neural Machine Translation Dataset

What is Text Generation

What is Text Generation

What is Text Generation

주어진 목적에 맞는 Text를 생성

What is Text Generation

주어진 목적에 맞는 Text를 생성

시 생성

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

시 생성

Generated Poem Example

```
Input: 맨발로 거친 산길을 오르는 나의 발바닥은 돌멩이에 찢겨 나뭇가지에 찢겨

먹다버린 깨진 콜라병과 눈총과 온갖 쓰레기에 치여
검붉은 피로 멍들어
불혹의 동반자로 떠도는 편지를
그 사이로 일은
곱게 물째 은행잎을 바라보지
```

https://github.com/reniew/Seq2seq_poem_generation

What is Text Generation

주어진 목적에 맞는 Text를 생성

소설 생성

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

소설 생성

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

소설 생성

인간의 이해는 거부한다!
본격 4 차 로맨스 혁명 포스트-포스트모더니즘 아방-아방가르드 소설! 그와 그녀의 피와 사랑 이야기

당신은 밖을 내다볼까 말까 몇 번이고 고민한다. 귀가 먹먹해지는 함성에 가슴이 뛴다.
용기를 내어 살짝 출구 쪽으로 얼굴을 내밀어 본다. 당신은 뛰는 가슴을 가라앉히려고 어깨를 펴고 크게 공기를 들이마신다. 향긋한 흙내음이 밀려온다.
따뜻하고 포근한 감촉. 나는 눈을 감고 있었다. 정신이 아득해지는 것을 느꼈다.
숨이 쉬어지지 않는 것이 느껴졌다. 정신을 잃었다.

" 안 돼!!"

나는 달리고 있었다. 달리는 속도가 점점 더 빨라졌다.

"이대로는 안 돼!"

나는 달리는 속도를 멈추지 않았다. 정신이 없는 것 같았다. 정신이 없는 것처럼.

https://speakerdeck.com/dreamgonfly/soseol-sseuneun-dib-reoning-pycon-korea-2018

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

Image Captioning

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

Image Captioning

What is Text Generation

주어진 목적에 맞는 Text를 생성

↓

Image Captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

What is Text Generation

주어진 목적에 맞는 Text를 생성

➔ **번역 / 대화**

What is Text Generation

주어진 목적에 맞는 Text를 생성

➔ **번역 / 대화**

➔ **Neural Machine Translation / Dialogue**

What is Text Generation

주어진 목적에 맞는 Text를 생성

➔ **번역 / 대화**

➔ **Neural Machine Translation / Dialogue**

What is Neural Machine Translation

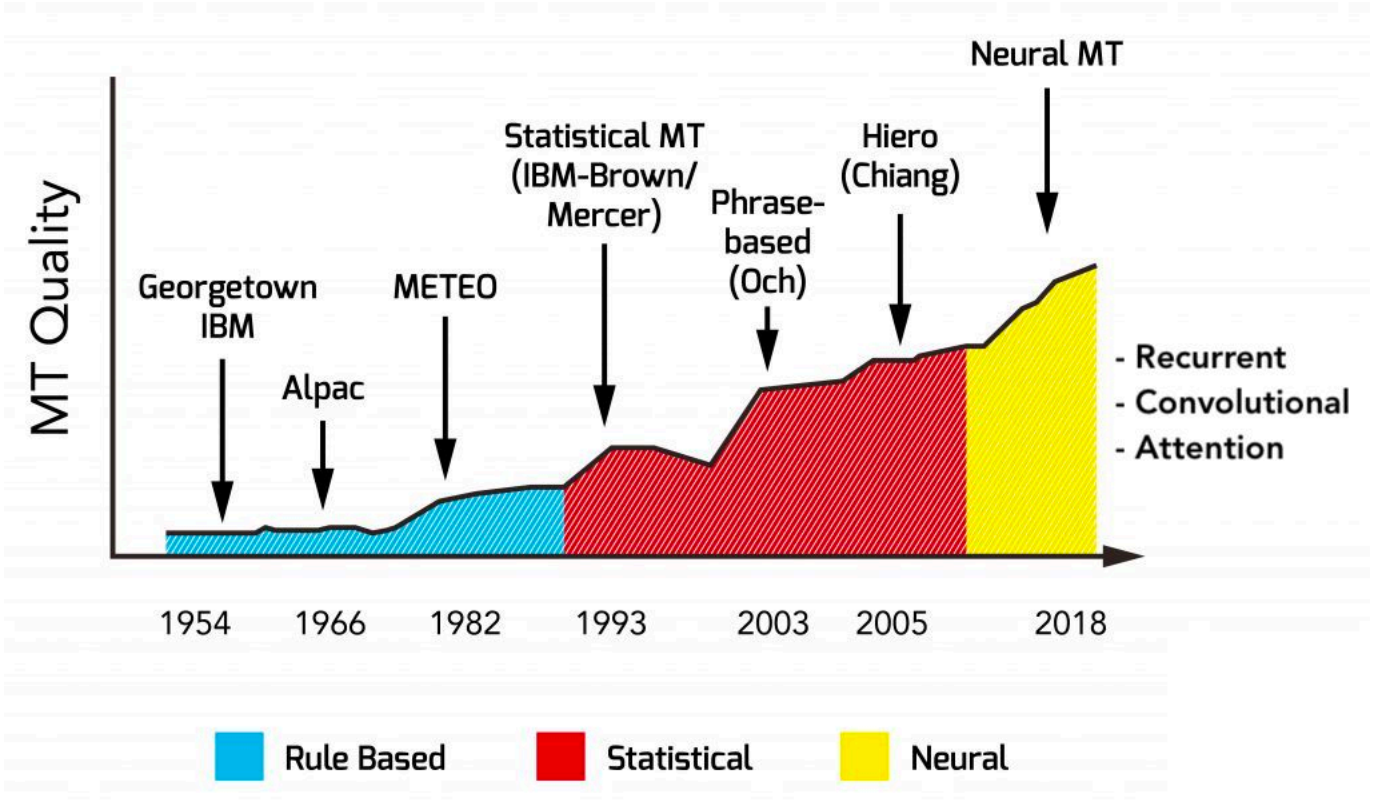
신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

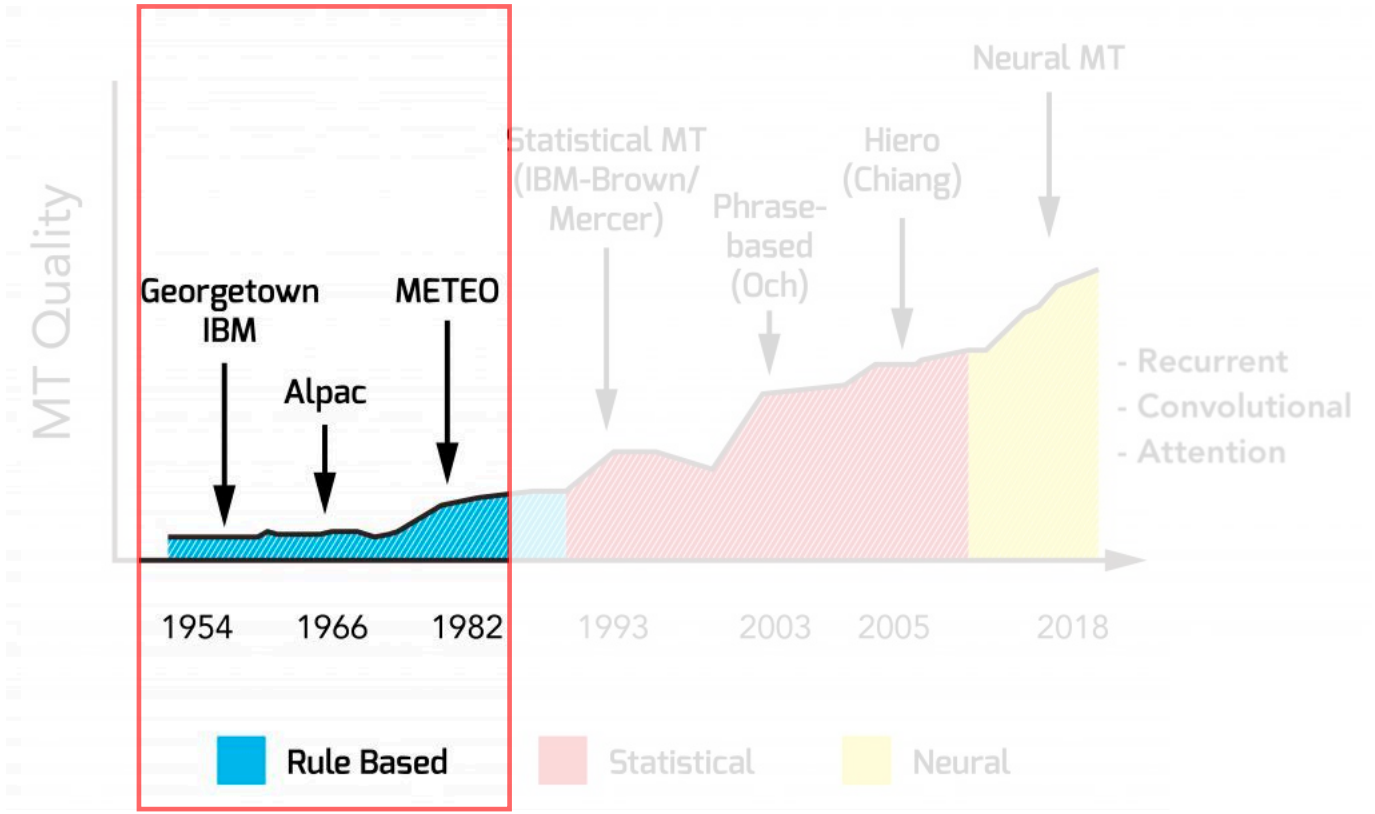
What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

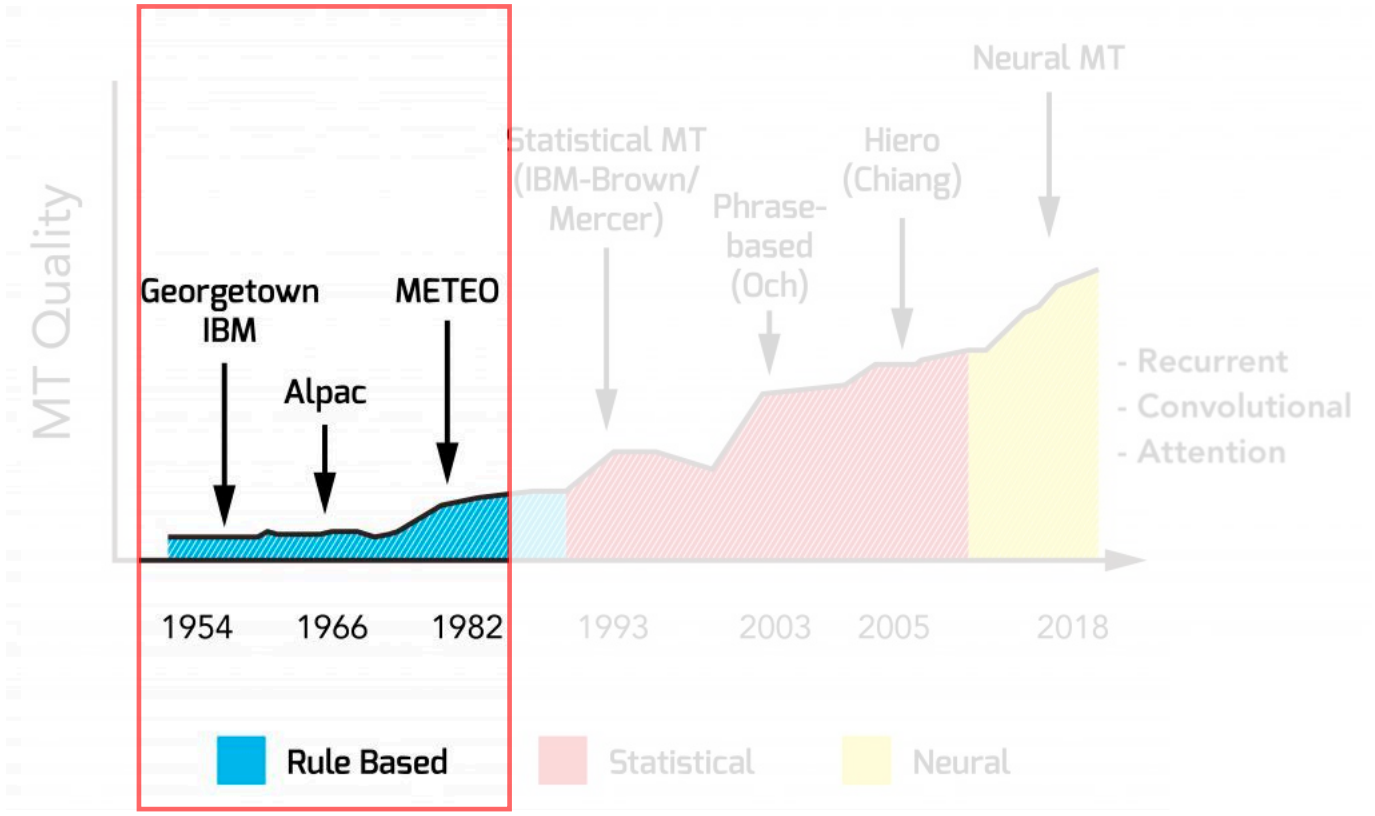


Rule Based Machine Translation (RBMT)

- 규칙 기반의 기계 번역
- 언어적인 정보를 바탕으로 규칙 수립

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



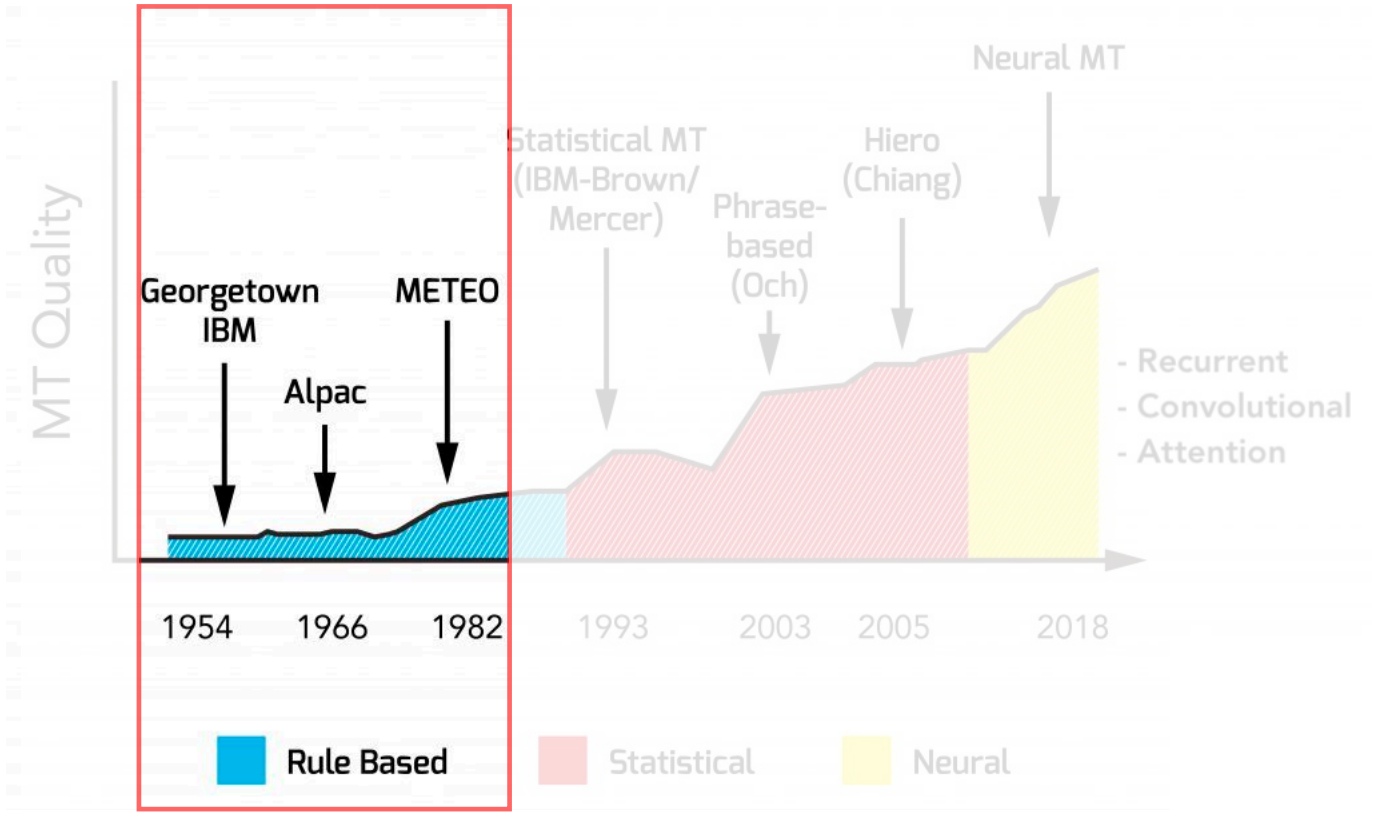
Rule Based Machine Translation (RBMT)

Source:

- **나**는 학교에 가는 중이다.
  대명사

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

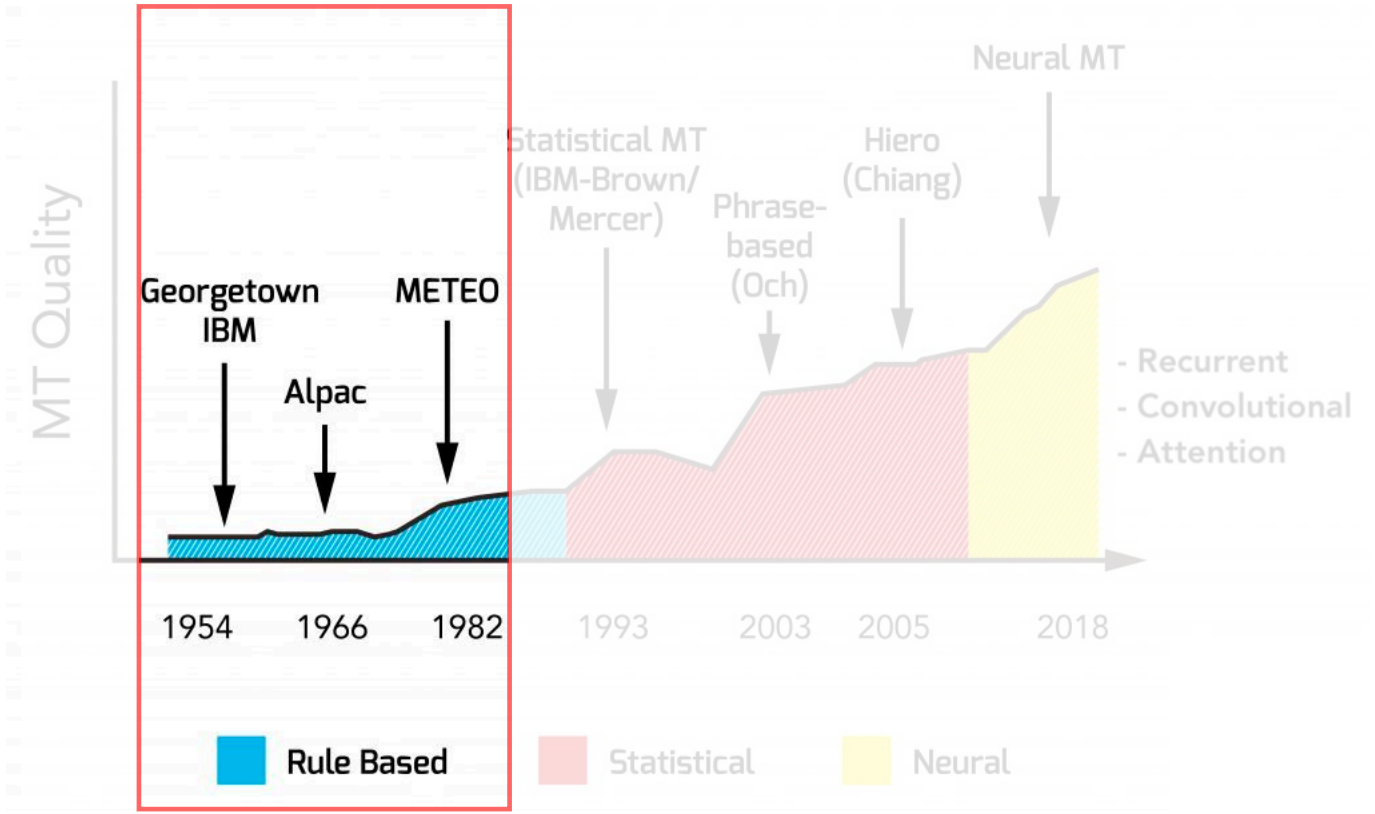
- **나**는 학교에 가는 중이다.
  대명사

  ↓

  I (NP)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



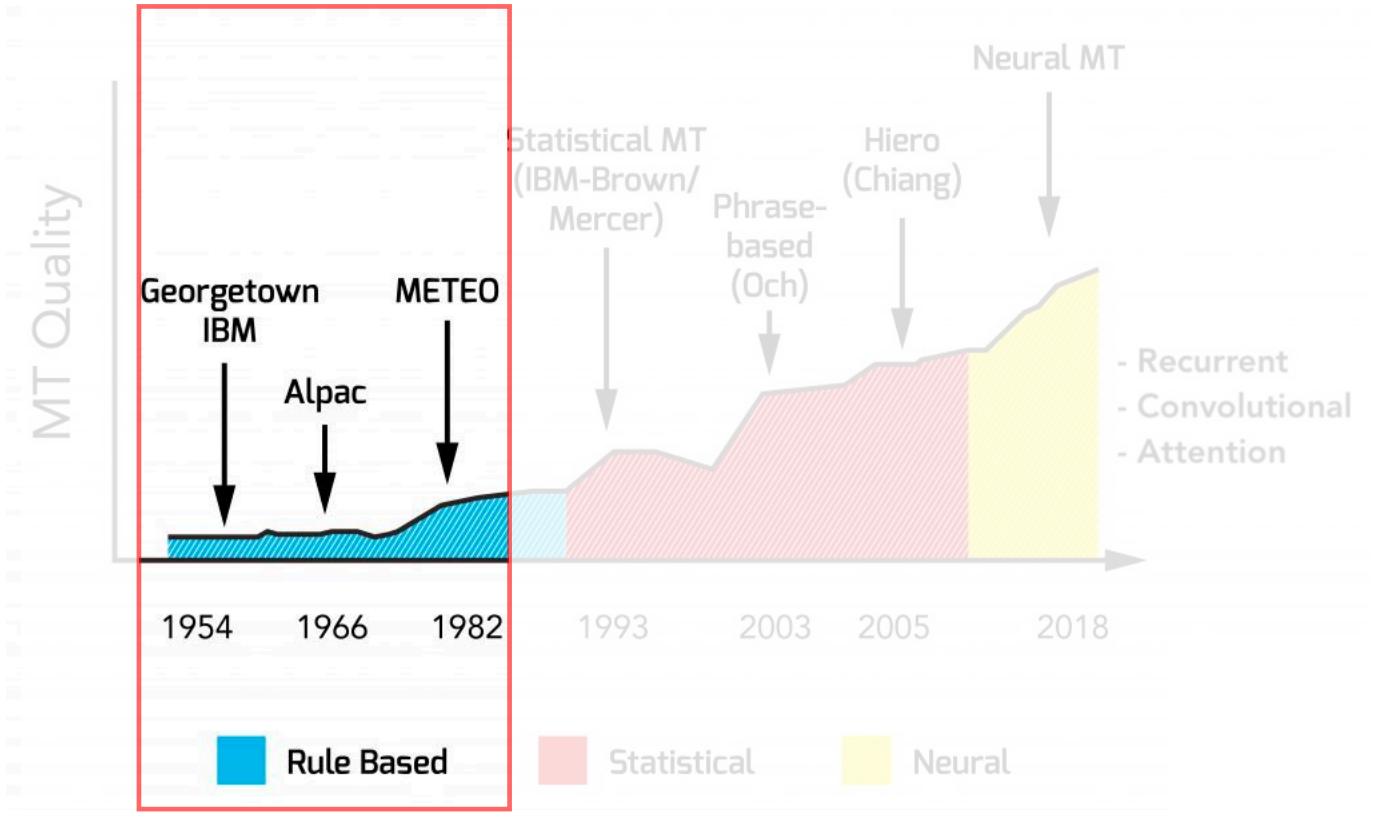
Rule Based Machine Translation (RBMT)

Source:

- 나**는** 학교에 가는 중이다.
  조사

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

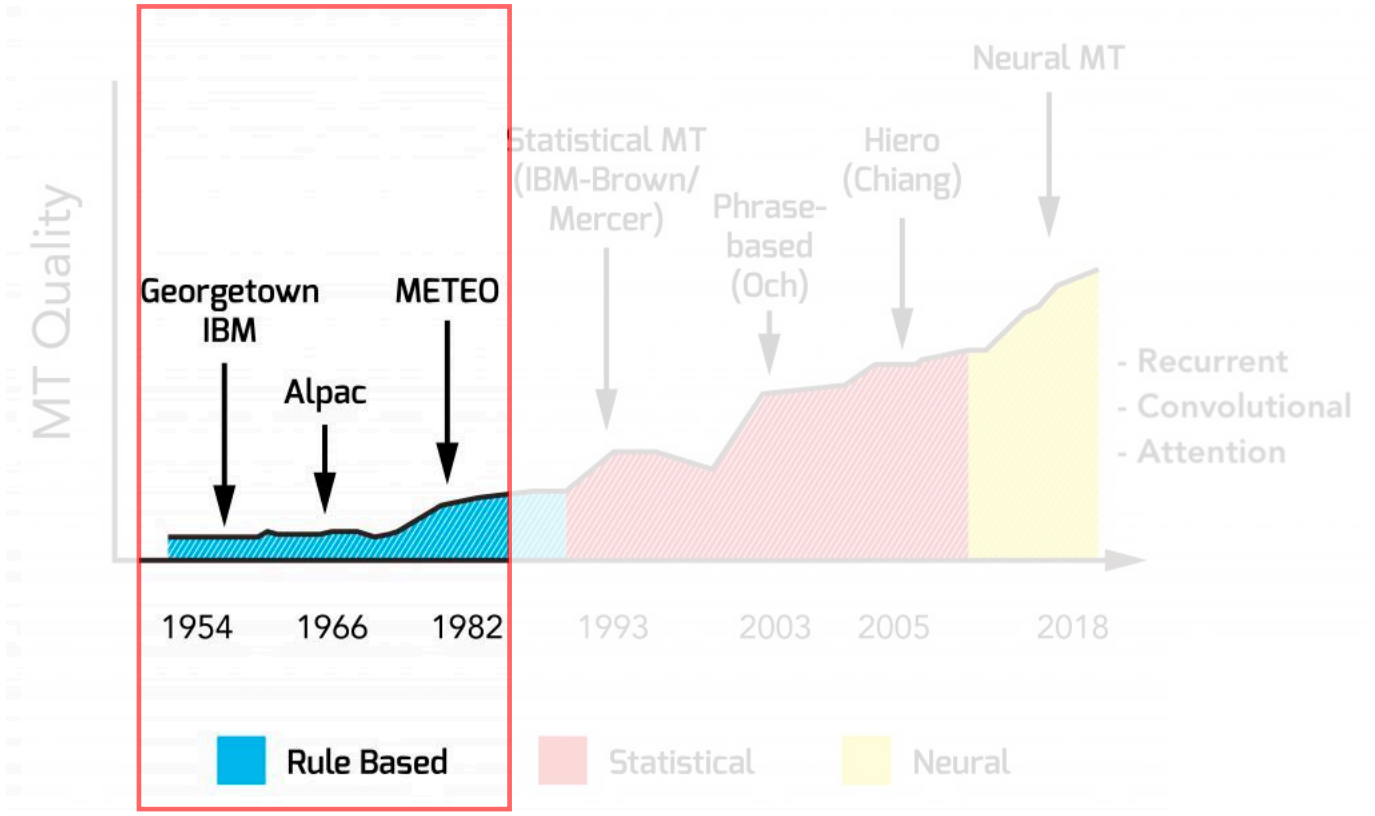
Source:

- 나**는** 학교에 가는 중이다.
  조사

am (suffix)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

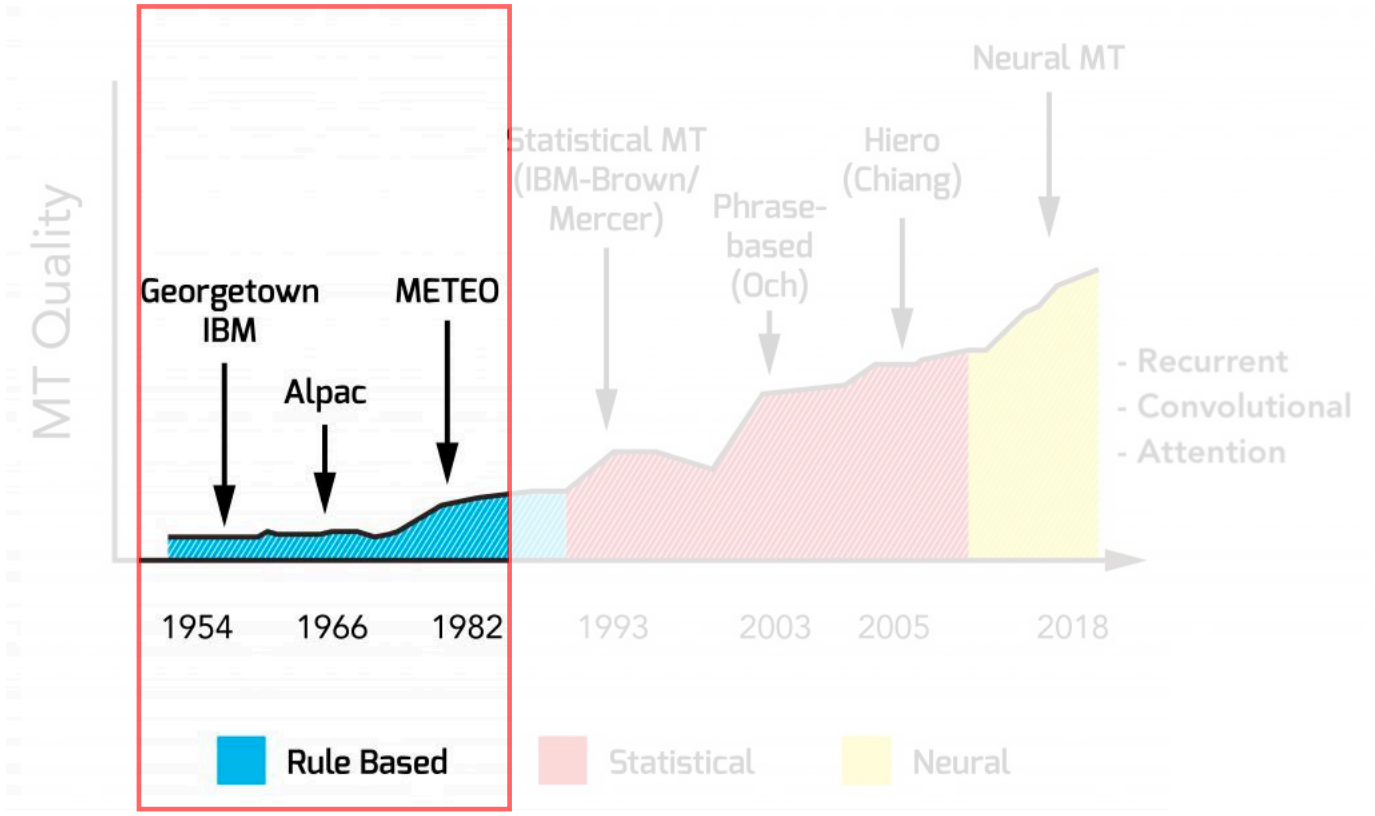
Source:

- 나는 **학교**에 가는 중이다.
  명사

school (NP)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

- 나는 학교**에** 가는 중이다.

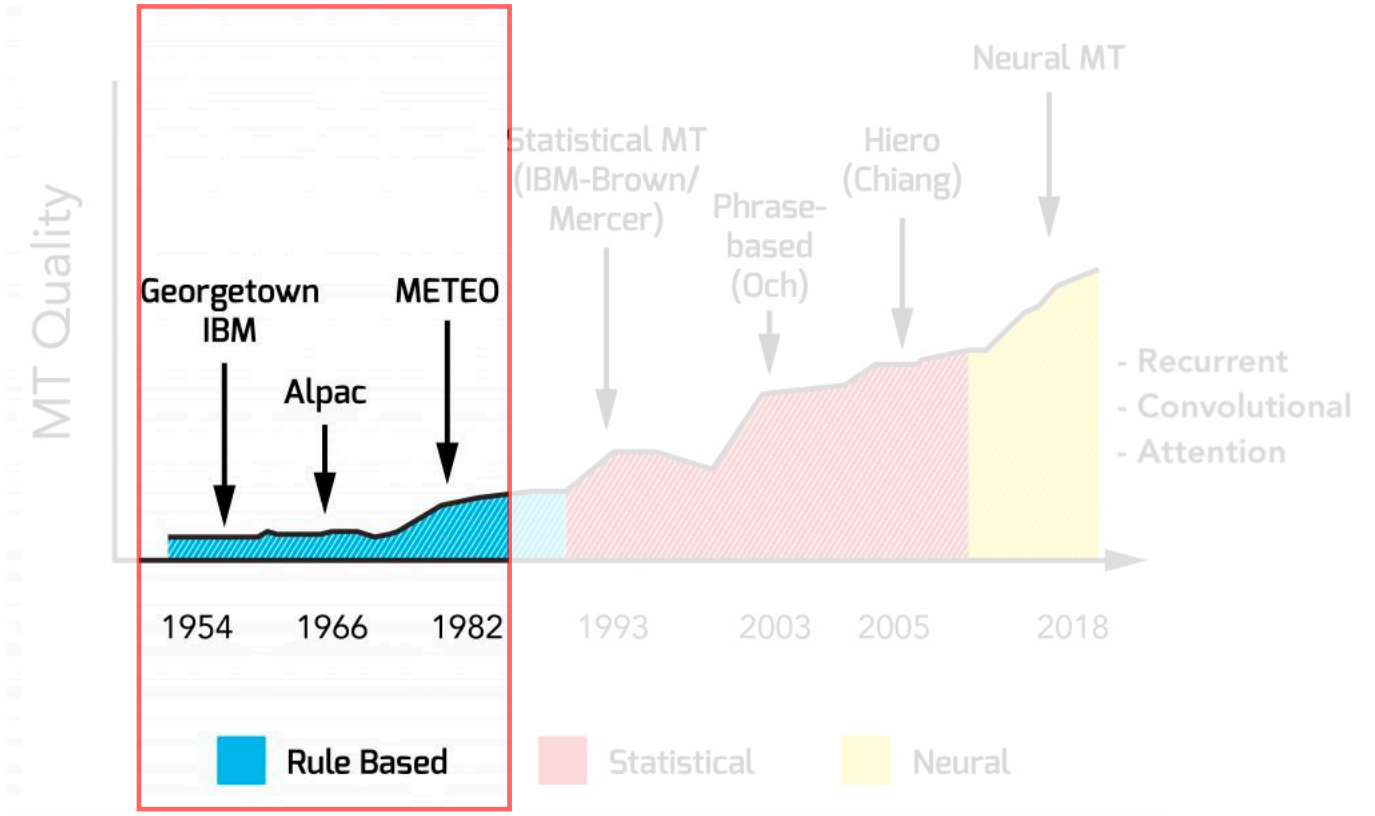
…

↓

… (…)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

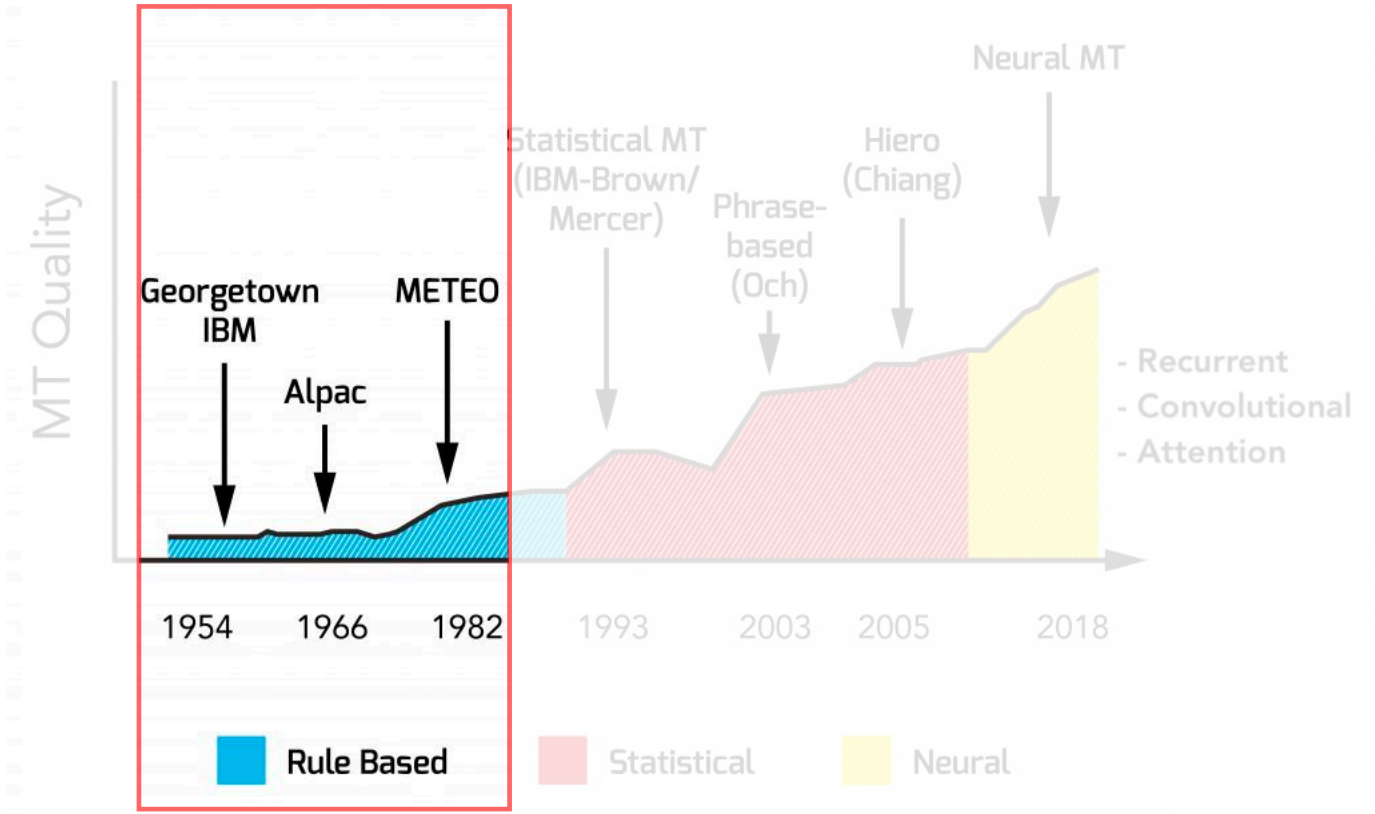
- 나는 학교에 **가는** 중이다.

…

… (…)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

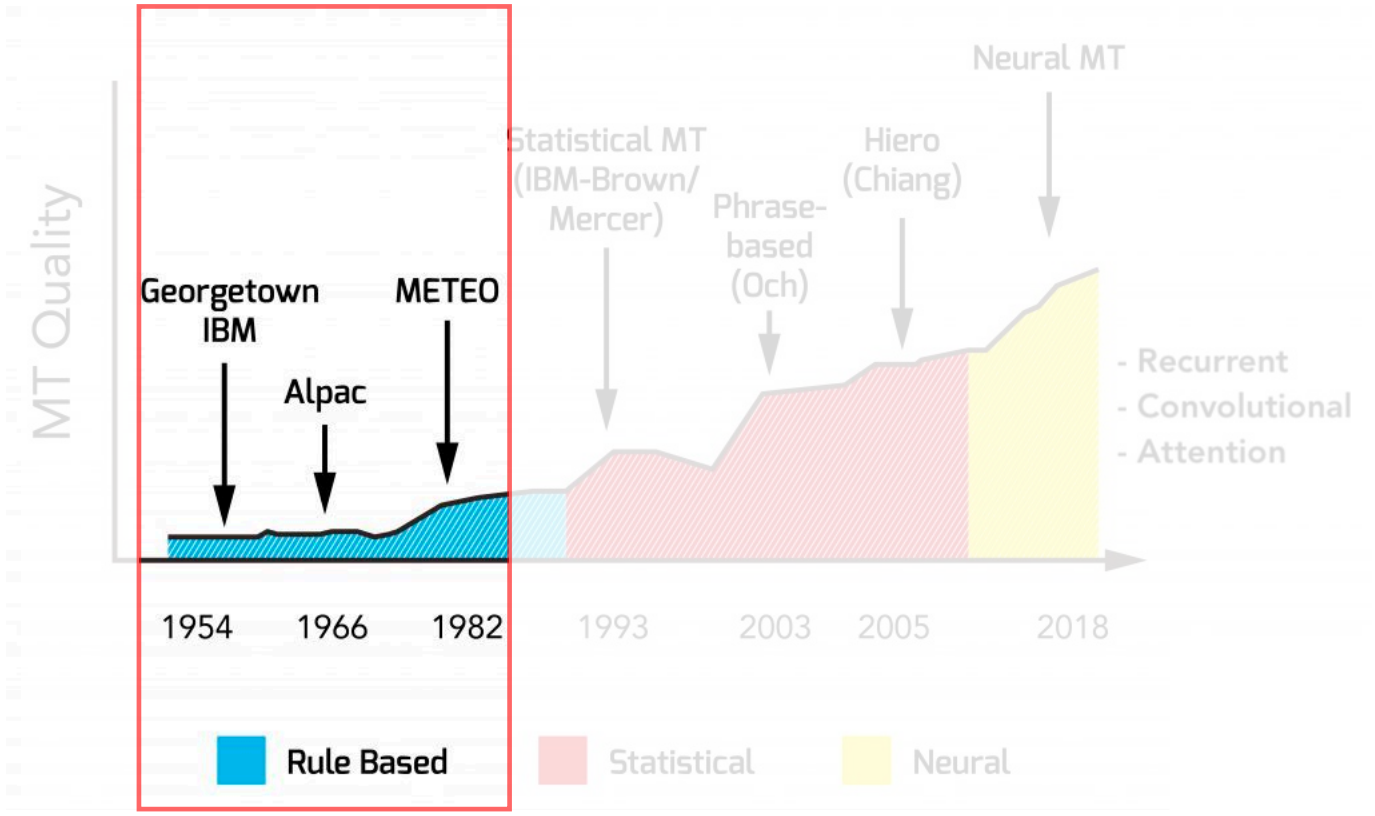
- 나는 학교에 가는 **중이다.**

…

… (…)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

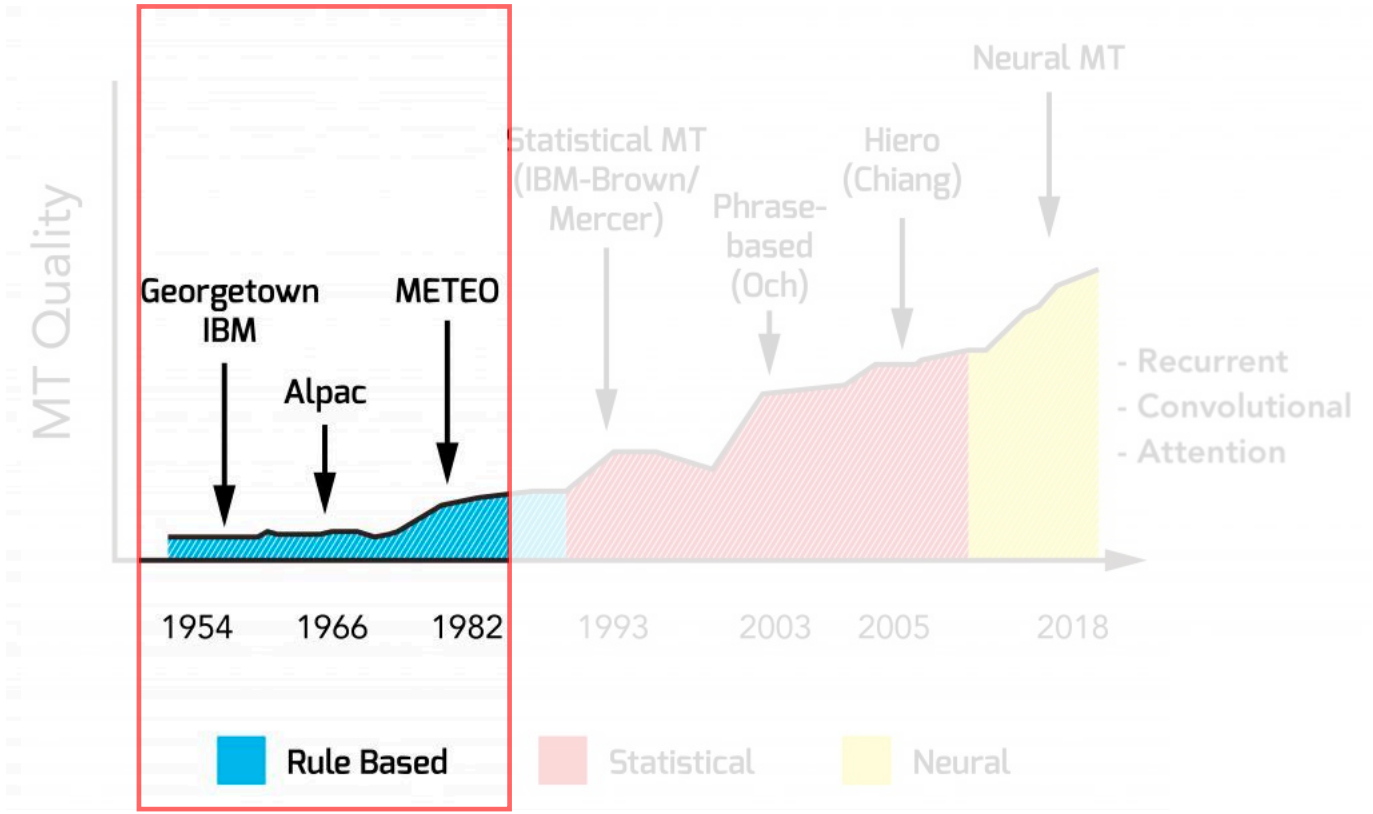
- 나는 학교에 가는 중이다.

↓

I am going to school

What is Neural Machine Translation

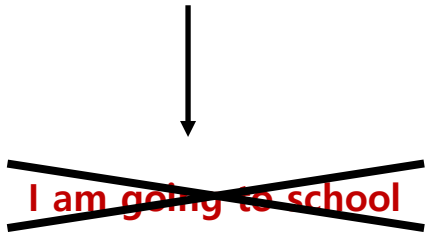
신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

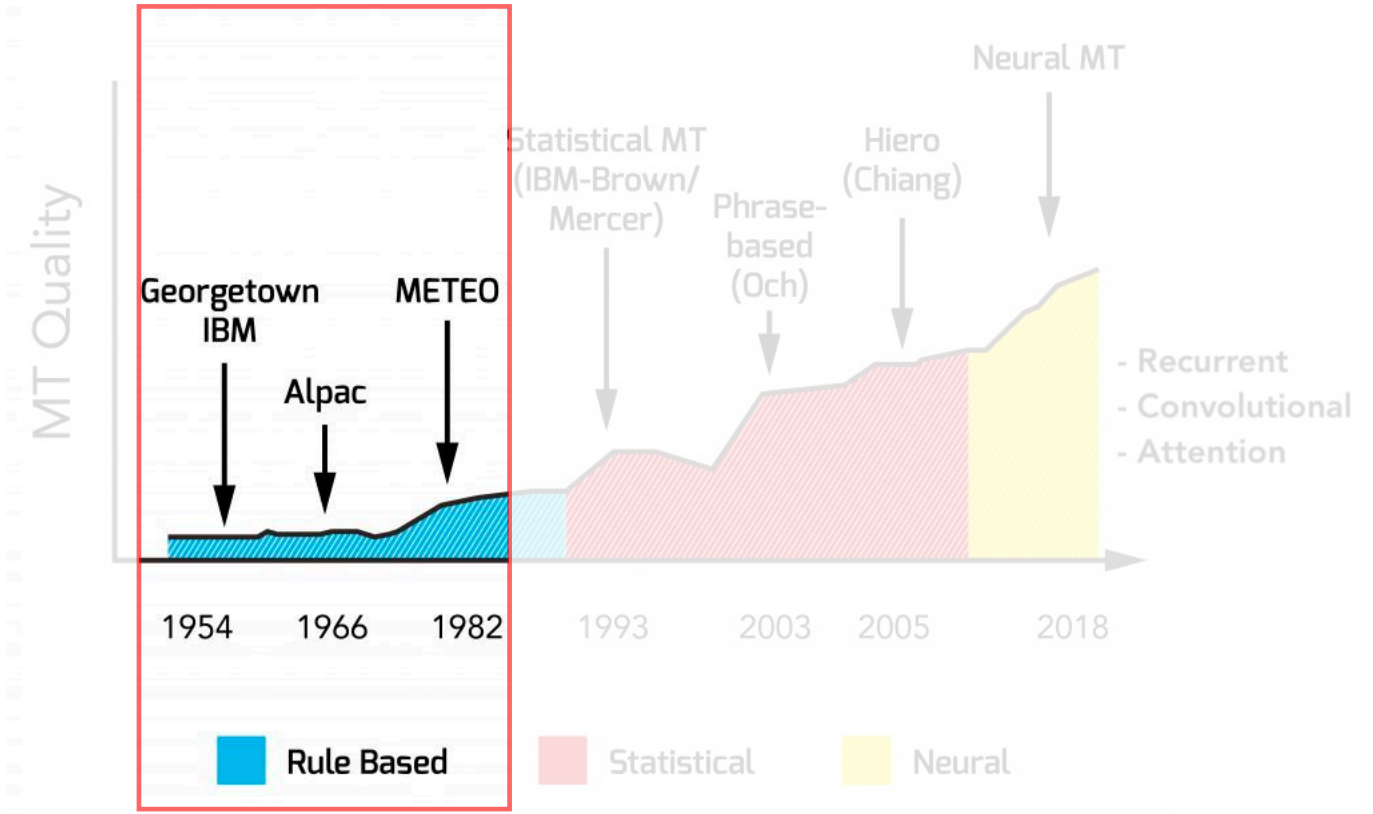
Source:

- 나는 학교에 가는 중이다.



What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Rule Based Machine Translation (RBMT)

Source:

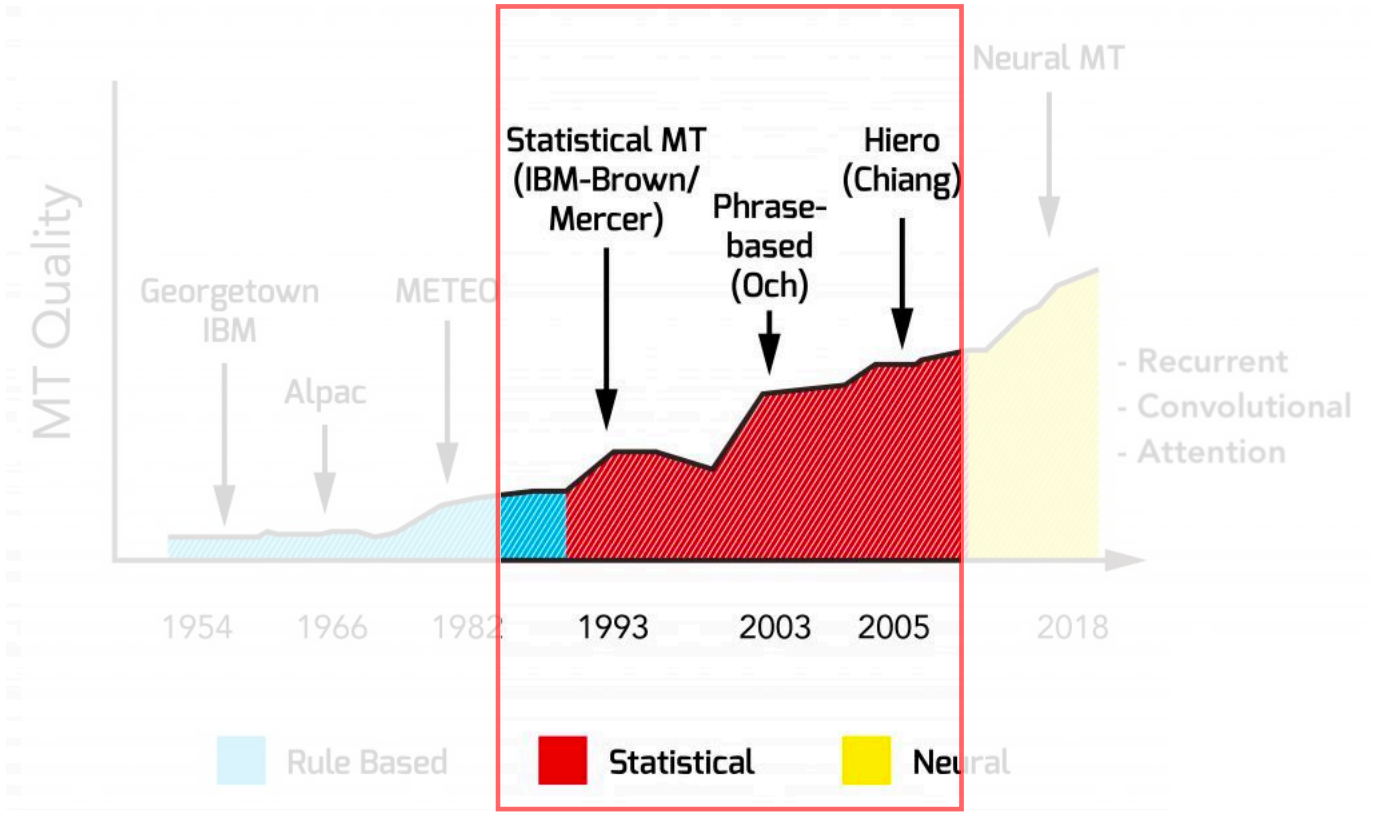
- 나는 학교에 가는 중이다.

↓

I am school going

What is Neural Machine Translation

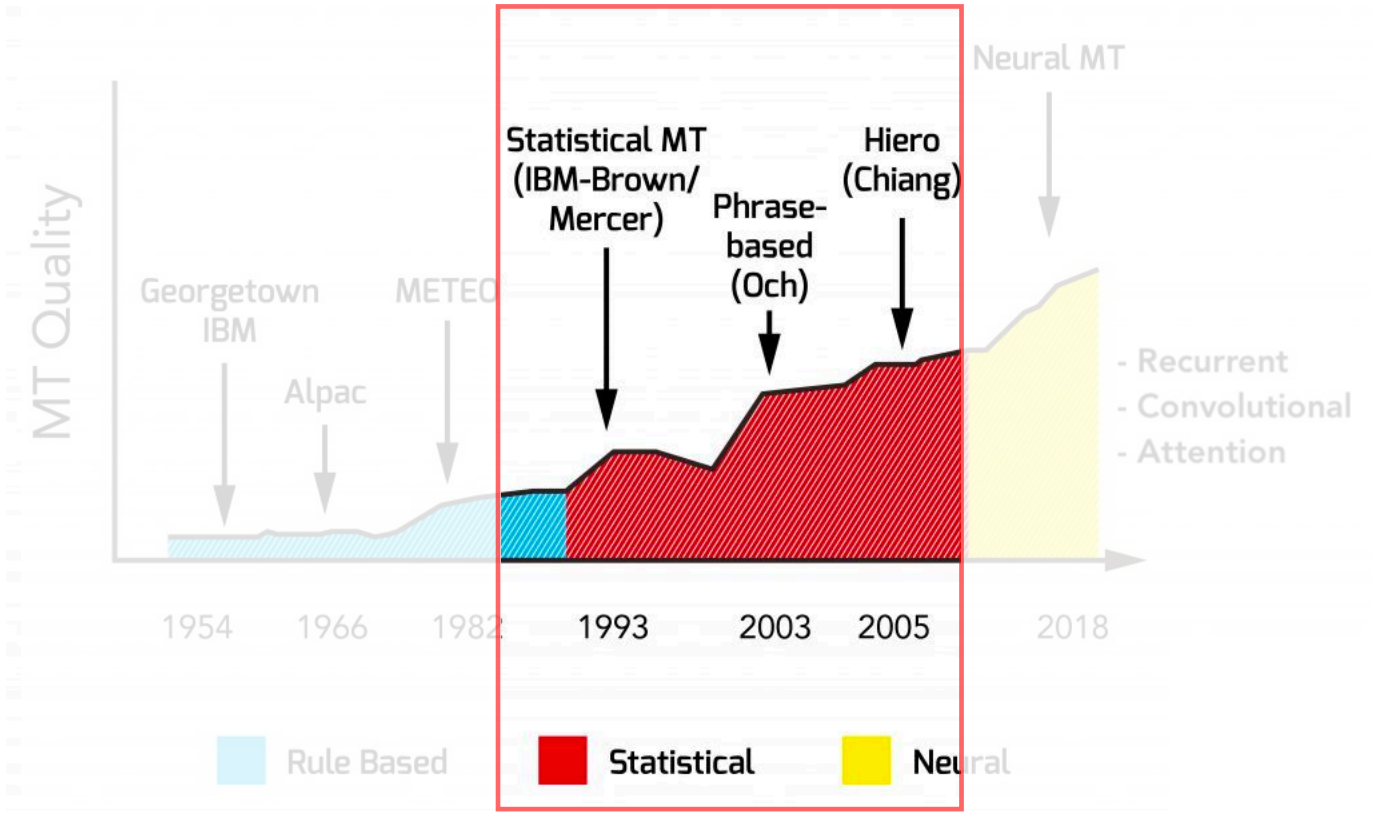
신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Statistical Machine Translation (SMT)

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

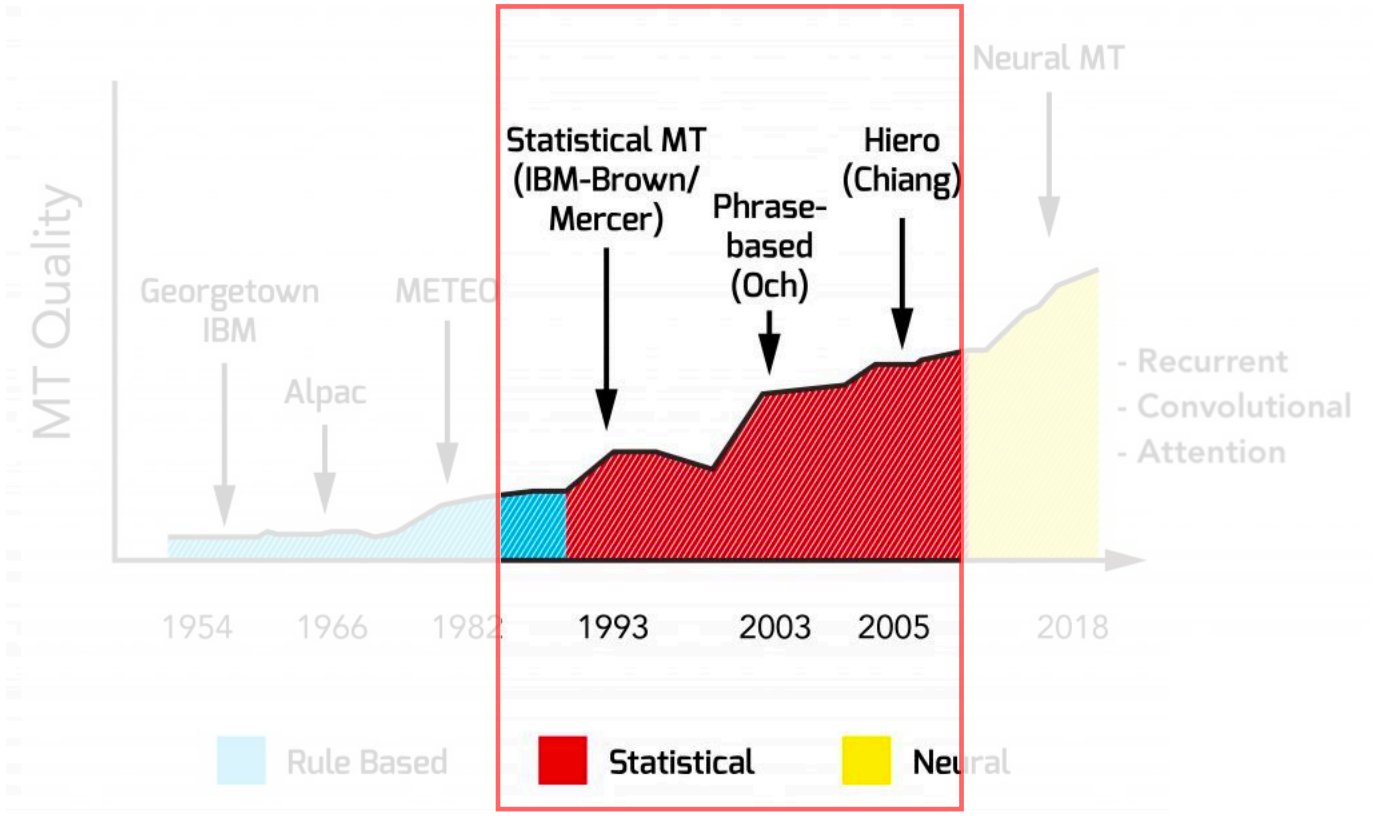


Statistical Machine Translation (SMT)

- Text 데이터를 통해 계산된 확률 모델을 사용해 번역

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Statistical Machine Translation (SMT)

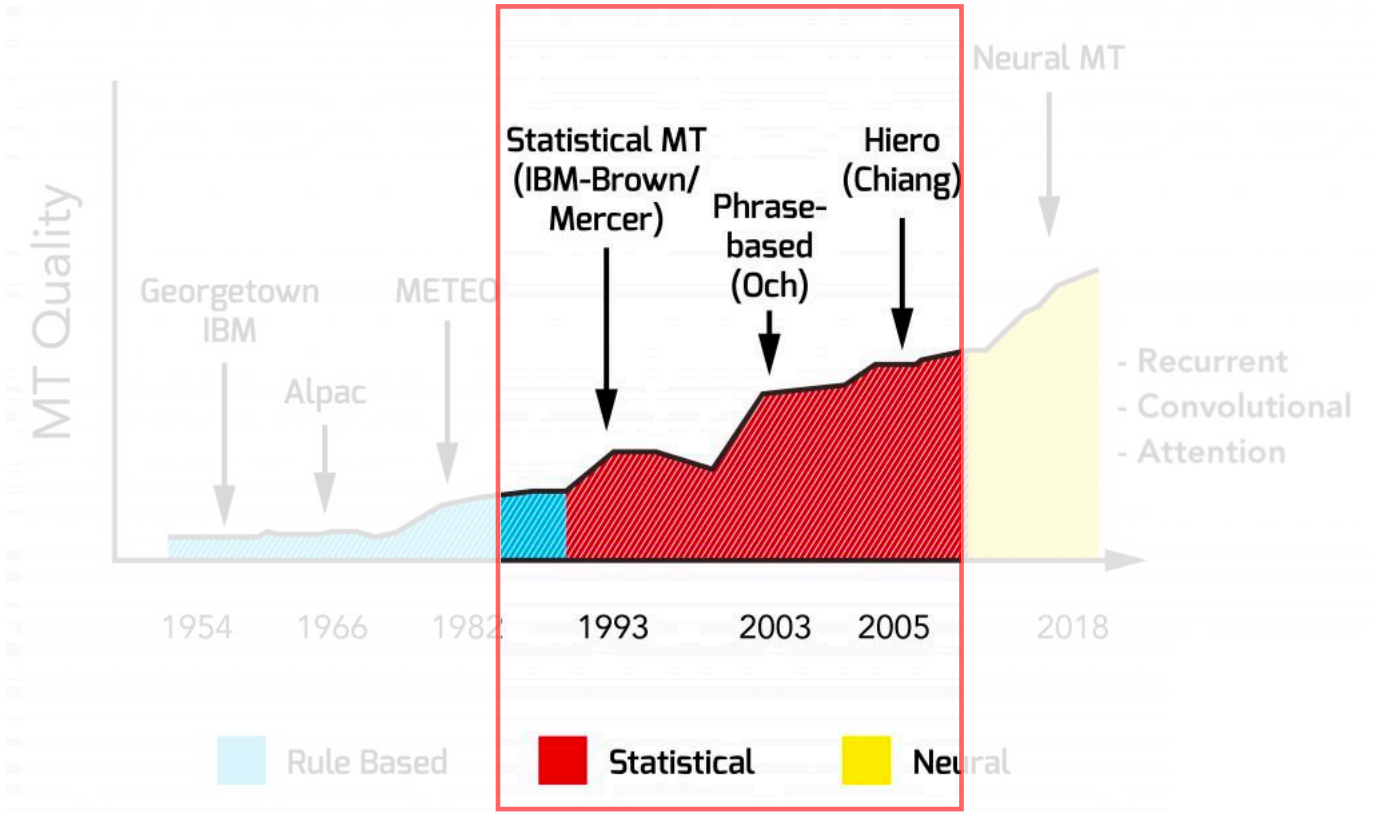
Source:

- 나는 학교에 가는 중이다. $(= S)$

$$'\mathrm{I}' = \arg\max_{w} P(w|S)$$

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Statistical Machine Translation (SMT)

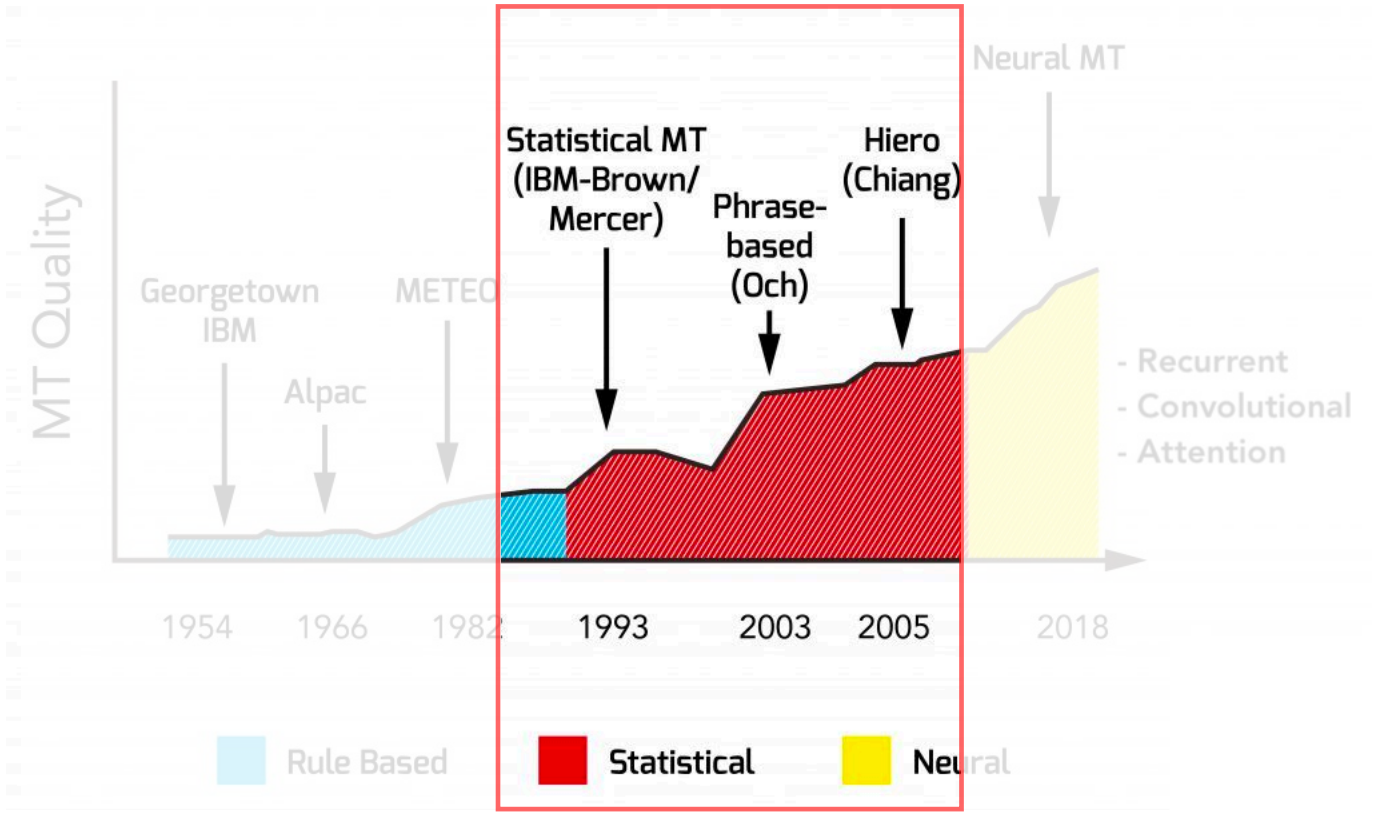
Source:

- 나는 학교에 가는 중이다. $(= S)$

$$'\text{am}' = \arg\max_{w} P(w|S, 'I')$$

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task



Statistical Machine Translation (SMT)

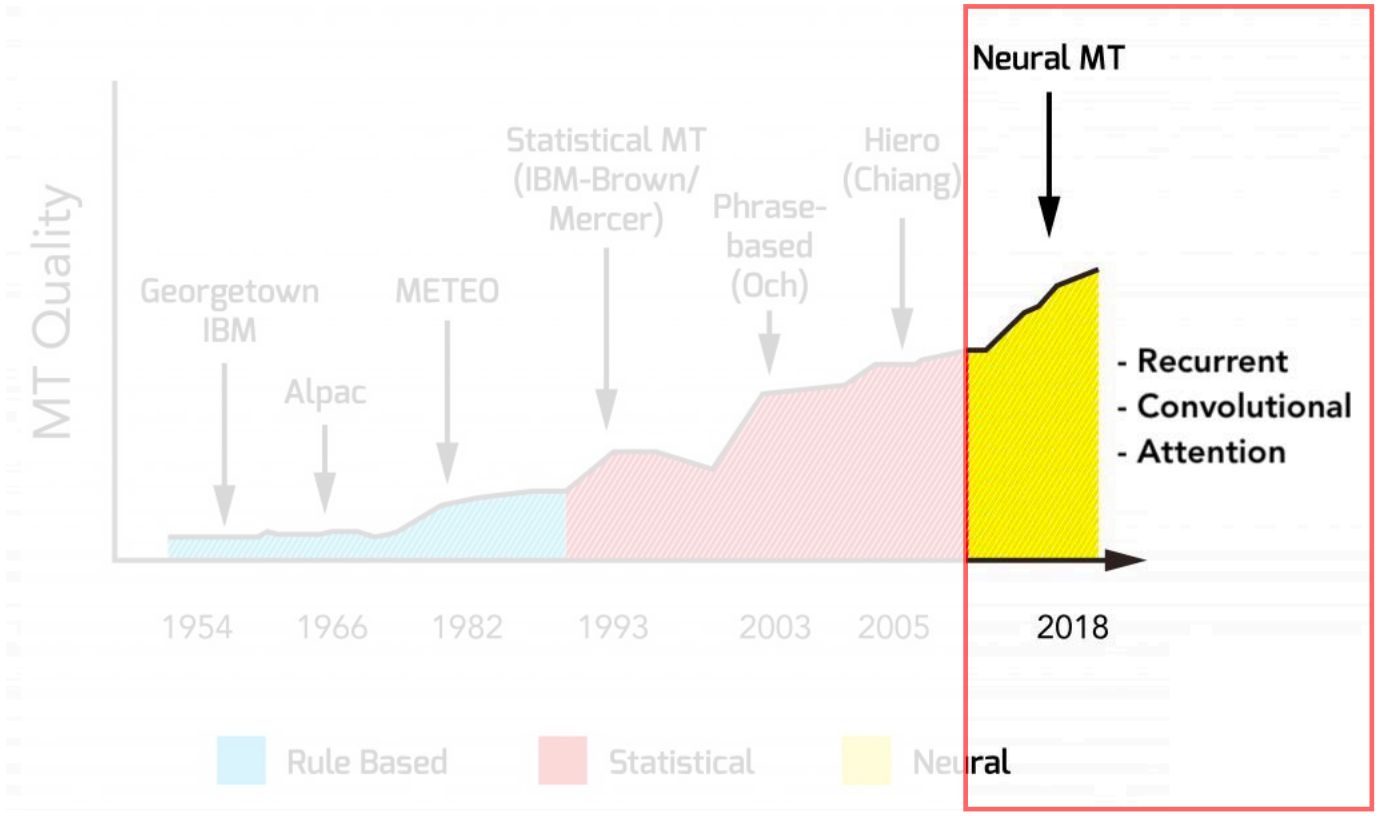
Source:

- 나는 학교에 가는 중이다. $(= S)$

$$'\text{going}' = \arg\max_{w} P(w|S, 'I', 'am')$$

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

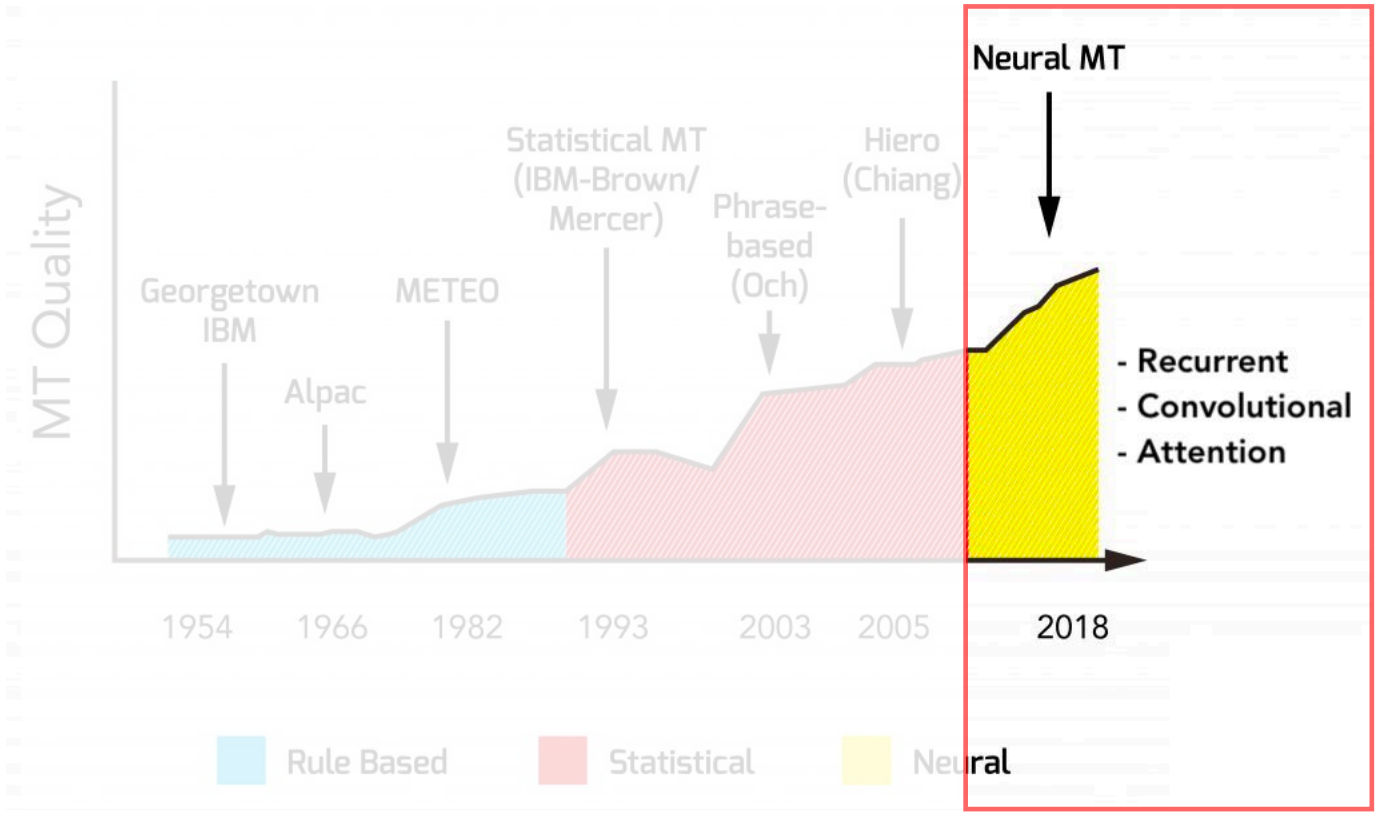


Neural Machine Translation (NMT)

- Neural Network를 사용해 Text Data를 학습해 주어진 문장을 번역
- 2014년 Sequence to Sequence 모델로 Neural Machine Translation 개념이 등장

What is Neural Machine Translation

신경망을 활용해 Source Language의 Text를 Target Language Text로 번역하는 Task

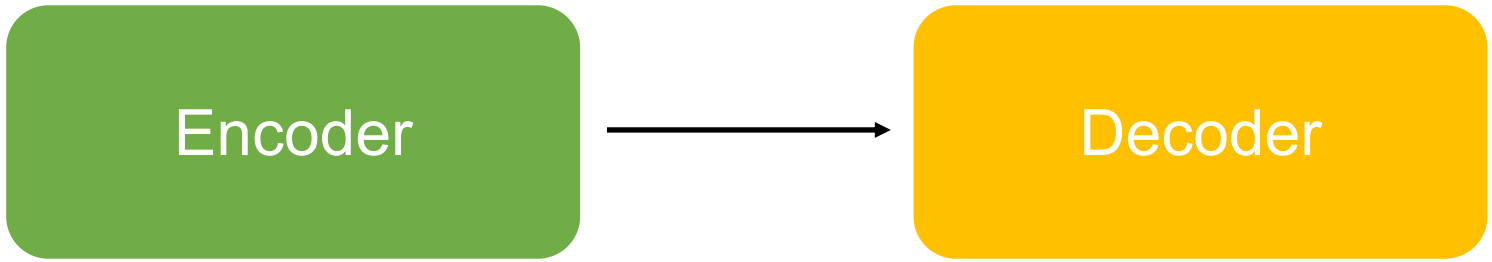


Neural Machine Translation (NMT)

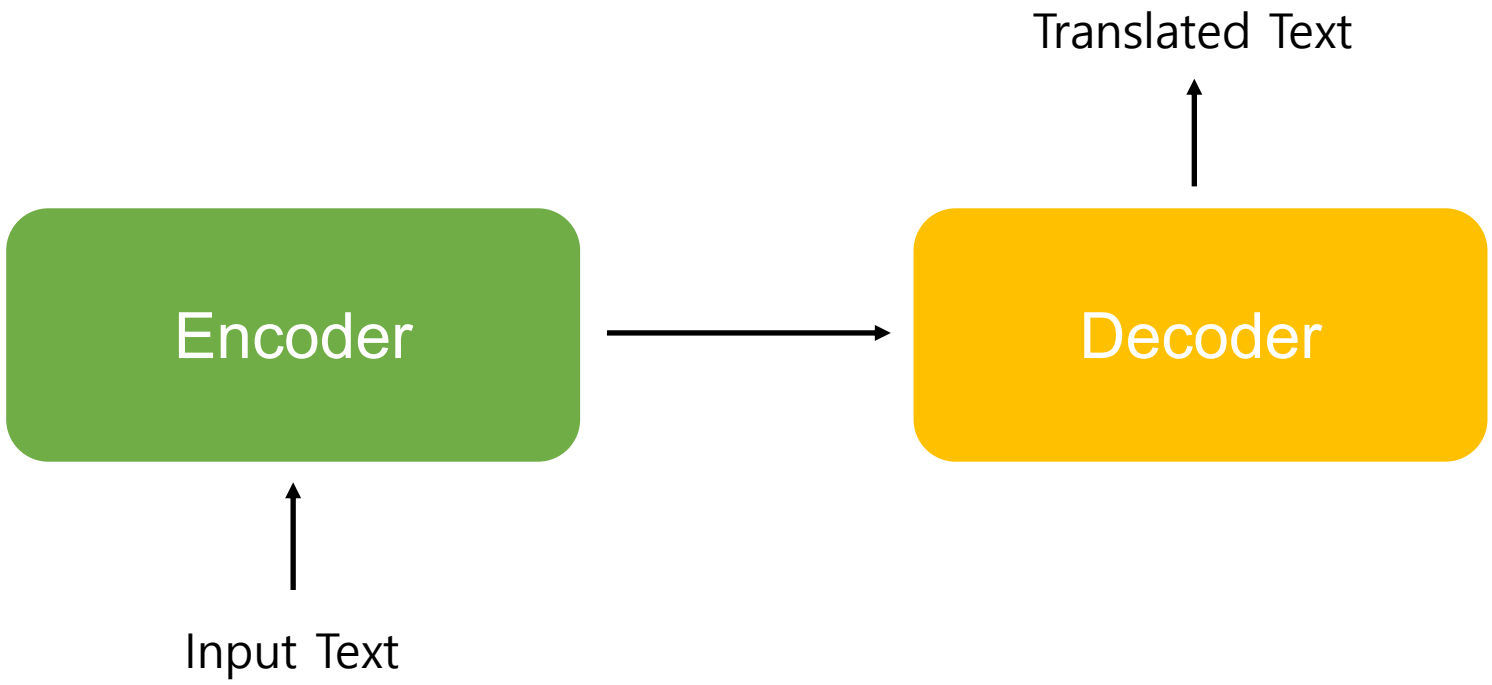
- Sequence to Sequence Learning with Neural Networks
- Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

Neural Machine Translation Architecture

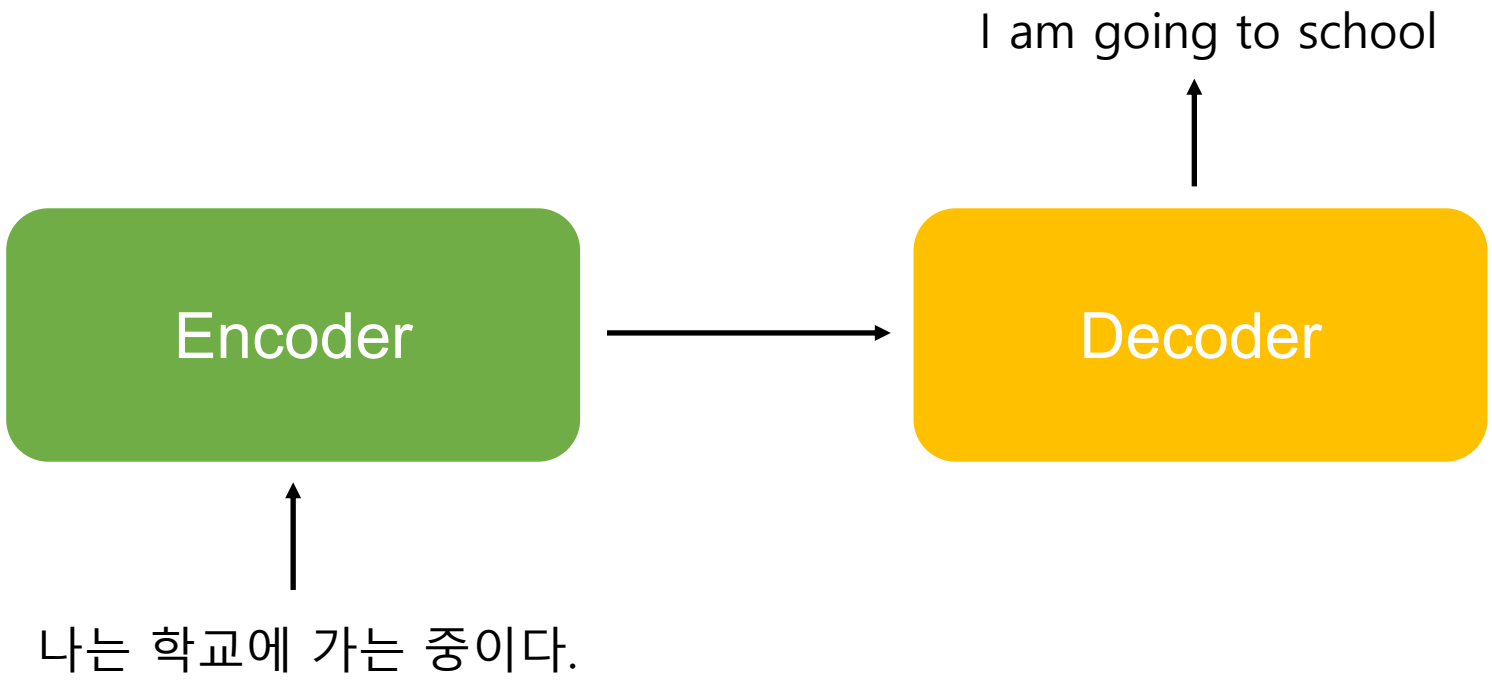
Neural Machine Translation Architecture



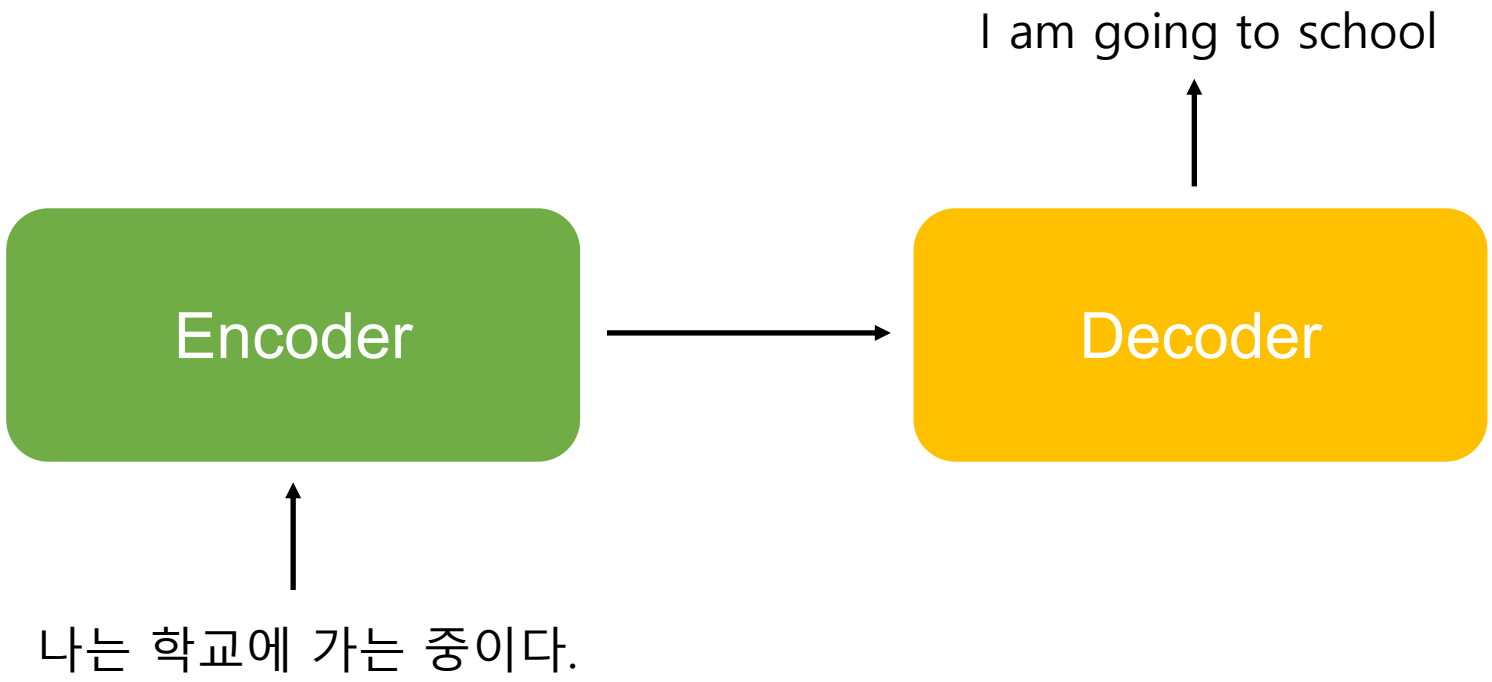
Neural Machine Translation Architecture



Neural Machine Translation Architecture

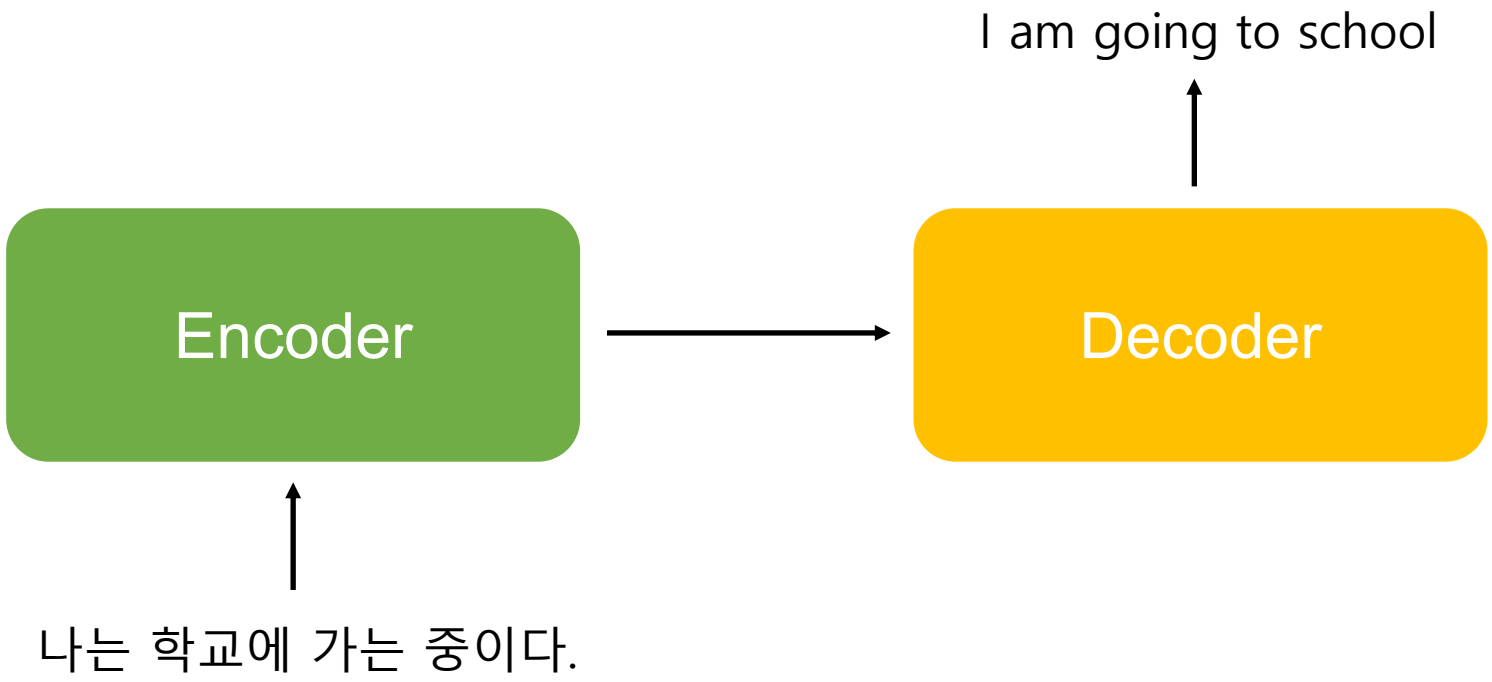


Neural Machine Translation Architecture



Encoder–Decoder Architecture

Neural Machine Translation Architecture



Sequence to Sequence Model

Neural Machine Translation

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre**
Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau
Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk**
Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

- Cho et al. (2014)

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre**
Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau
Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk**
Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

- Cho et al. (2014)
- Encoder-Decoder 구조를 처음 제시

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre**
Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau
Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk**
Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

- Cho et al. (2014)
- Encoder-Decoder 구조를 처음 제시
- NMT의 시발점이 된 논문

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre** Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk** Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

- Cho et al. (2014)
- Encoder-Decoder 구조를 처음 제시
- NMT의 시발점이 된 논문
  - 하지만 해당 논문에서는 해당 신경망 모델을 기계 번역에 직접 적용하지 않음

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre** Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk** Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

- Cho et al. (2014)
- Encoder-Decoder 구조를 처음 제시
- NMT의 시발점이 된 논문
  - 하지만 해당 논문에서는 해당 신경망 모델을 기계 번역에 직접 적용하지 않음
  - 기존의 가장 높은 성능을 보이던 Phrase-based SMT 모델에 적용시킴

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

Kyunghyun Cho
Bart van Merriënboer **Caglar Gulcehre** Université de Montréal
firstname.lastname@umontreal.ca

Dzmitry Bahdanau Jacobs University, Germany
d.bahdanau@jacobs-university.de

Fethi Bougares **Holger Schwenk** Université du Maine, France
firstname.lastname@lium.univ-lemans.fr

Yoshua Bengio Université de Montréal, CIFAR Senior Fellow
find.me@on.the.web

Abstract

In this paper, we propose a novel neural network model called RNN Encoder–Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

1 Introduction

Deep neural networks have shown great success in various applications such as objection recognition (see, e.g., (Krizhevsky et al., 2012)) and speech recognition (see, e.g., (Dahl et al., 2012)). Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP). These include, but are not limited to, language modeling (Bengio et al., 2003), paraphrase detection (Socher et al., 2011) and word embedding extraction (Mikolov et al., 2013). In the field of statistical machine translation (SMT), deep neural networks have begun to show promising results. (Schwenk, 2012) summarizes a successful usage of feedforward neural networks in the framework of phrase-based SMT system.

Along this line of research on using neural networks for SMT, this paper focuses on a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system. The proposed neural network architecture, which we will refer to as an *RNN Encoder–Decoder*, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair. The encoder maps a variable-length source sequence to a fixed-length vector, and the decoder maps the vector representation back to a variable-length target sequence. The two networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence. Additionally, we propose to use a rather sophisticated hidden unit in order to improve both the memory capacity and the ease of training.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically evaluated on the task of translating from English to French. We train the model to learn the translation probability of an English phrase to a corresponding French phrase. The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table. The empirical evaluation reveals that this approach of scoring phrase pairs with an RNN Encoder–Decoder improves the translation performance.

We qualitatively analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance. The further analysis of the model reveals that the RNN Encoder–Decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase.

Neural Machine Translation

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

1 Introduction

Deep Neural Networks (DNNs) are extremely powerful machine learning models that achieve excellent performance on difficult problems such as speech recognition [13, 7] and visual object recognition [19, 6, 21, 20]. DNNs are powerful because they can perform arbitrary parallel computation for a modest number of steps. A surprising example of the power of DNNs is their ability to sort N N-bit numbers using only 2 hidden layers of quadratic size [27]. So, while neural networks are related to conventional statistical models, they learn an intricate computation. Furthermore, large DNNs can be trained with supervised backpropagation whenever the labeled training set has enough information to specify the network's parameters. Thus, if there exists a parameter setting of a large DNN that achieves good results (for example, because humans can solve the task very rapidly), supervised backpropagation will find these parameters and solve the problem.

Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality. It is a significant limitation, since many important problems are best expressed with sequences whose lengths are not known a-priori. For example, speech recognition and machine translation are sequential problems. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a

Neural Machine Translation

- Sutskever et al. (2014, Google)

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

1 Introduction

Deep Neural Networks (DNNs) are extremely powerful machine learning models that achieve excellent performance on difficult problems such as speech recognition [13, 7] and visual object recognition [19, 6, 21, 20]. DNNs are powerful because they can perform arbitrary parallel computation for a modest number of steps. A surprising example of the power of DNNs is their ability to sort N N-bit numbers using only 2 hidden layers of quadratic size [27]. So, while neural networks are related to conventional statistical models, they learn an intricate computation. Furthermore, large DNNs can be trained with supervised backpropagation whenever the labeled training set has enough information to specify the network's parameters. Thus, if there exists a parameter setting of a large DNN that achieves good results (for example, because humans can solve the task very rapidly), supervised backpropagation will find these parameters and solve the problem.

Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality. It is a significant limitation, since many important problems are best expressed with sequences whose lengths are not known a-priori. For example, speech recognition and machine translation are sequential problems. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a

Neural Machine Translation

- Sutskever et al. (2014, Google)
- 본격적인 NMT 모델의 시작

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

1 Introduction

Deep Neural Networks (DNNs) are extremely powerful machine learning models that achieve excellent performance on difficult problems such as speech recognition [13, 7] and visual object recognition [19, 6, 21, 20]. DNNs are powerful because they can perform arbitrary parallel computation for a modest number of steps. A surprising example of the power of DNNs is their ability to sort N N-bit numbers using only 2 hidden layers of quadratic size [27]. So, while neural networks are related to conventional statistical models, they learn an intricate computation. Furthermore, large DNNs can be trained with supervised backpropagation whenever the labeled training set has enough information to specify the network's parameters. Thus, if there exists a parameter setting of a large DNN that achieves good results (for example, because humans can solve the task very rapidly), supervised backpropagation will find these parameters and solve the problem.

Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality. It is a significant limitation, since many important problems are best expressed with sequences whose lengths are not known a-priori. For example, speech recognition and machine translation are sequential problems. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a

Neural Machine Translation

- Sutskever et al. (2014, Google)
- 본격적인 NMT 모델의 시작
- RNN 을 사용해 Input 과 output을 modeling함

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

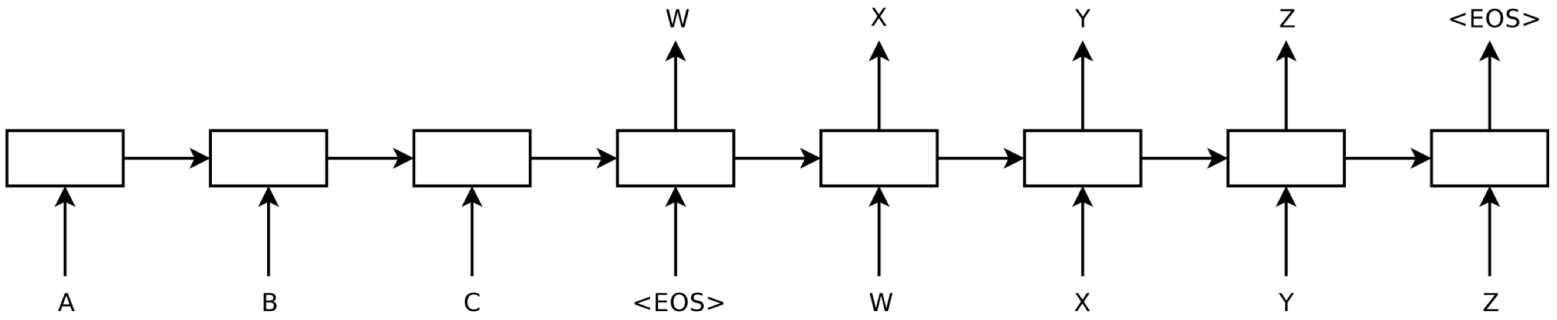
Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

1 Introduction

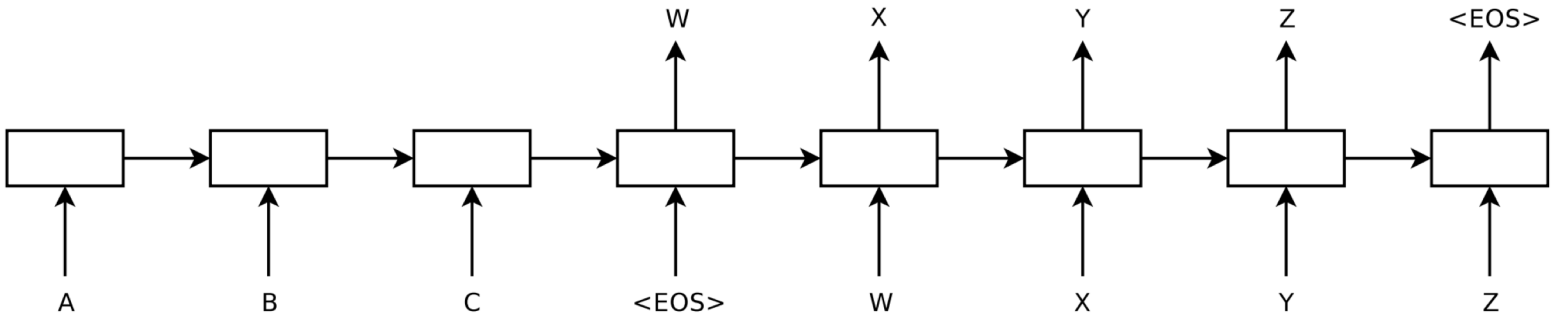
Deep Neural Networks (DNNs) are extremely powerful machine learning models that achieve excellent performance on difficult problems such as speech recognition [13, 7] and visual object recognition [19, 6, 21, 20]. DNNs are powerful because they can perform arbitrary parallel computation for a modest number of steps. A surprising example of the power of DNNs is their ability to sort N N-bit numbers using only 2 hidden layers of quadratic size [27]. So, while neural networks are related to conventional statistical models, they learn an intricate computation. Furthermore, large DNNs can be trained with supervised backpropagation whenever the labeled training set has enough information to specify the network's parameters. Thus, if there exists a parameter setting of a large DNN that achieves good results (for example, because humans can solve the task very rapidly), supervised backpropagation will find these parameters and solve the problem.

Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality. It is a significant limitation, since many important problems are best expressed with sequences whose lengths are not known a-priori. For example, speech recognition and machine translation are sequential problems. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a

Sequence to Sequence Learning with Neural Networks



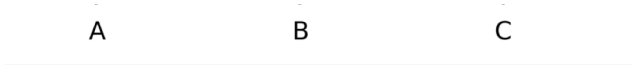
Sequence to Sequence Learning with Neural Networks



$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

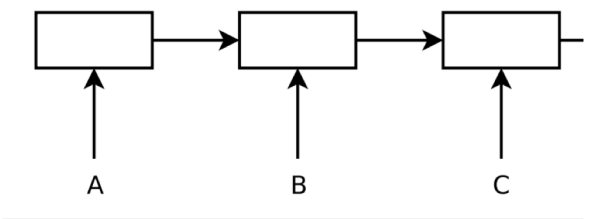
Sequence to Sequence Learning with Neural Networks

Input



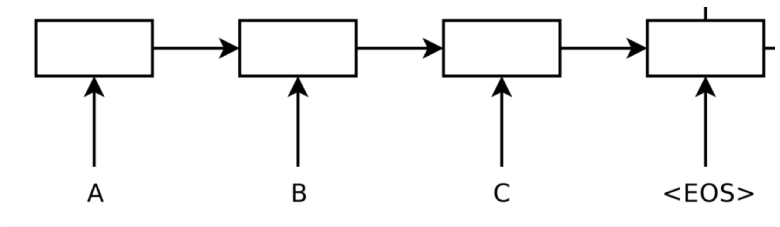
Sequence to Sequence Learning with Neural Networks

Input Sentence LSTM



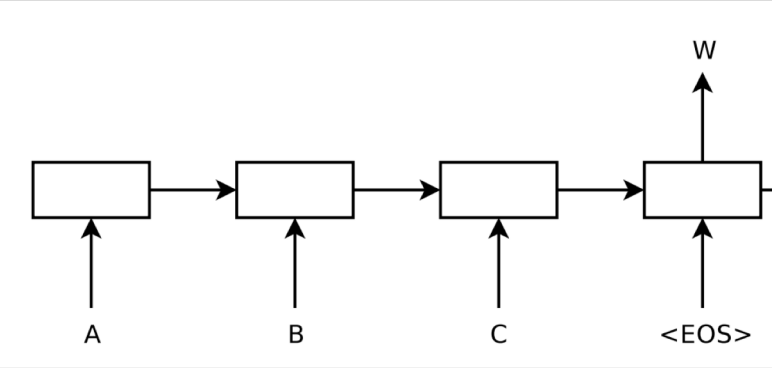
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (first step)



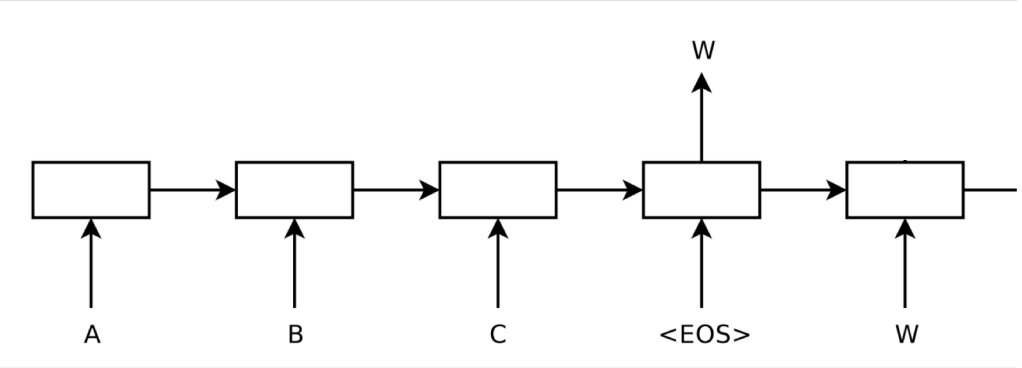
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (first step)



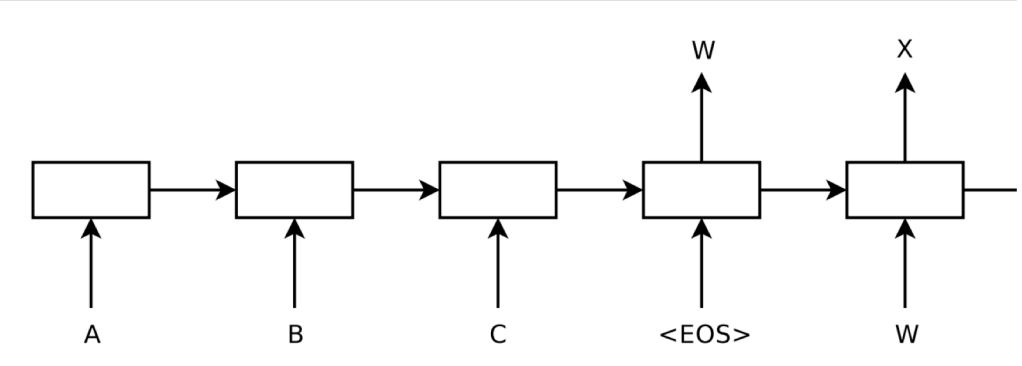
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (second step)



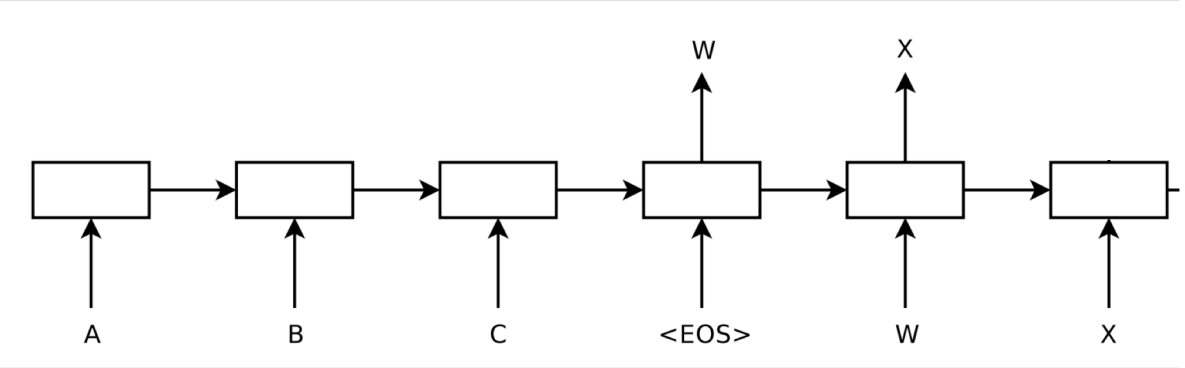
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (second step)



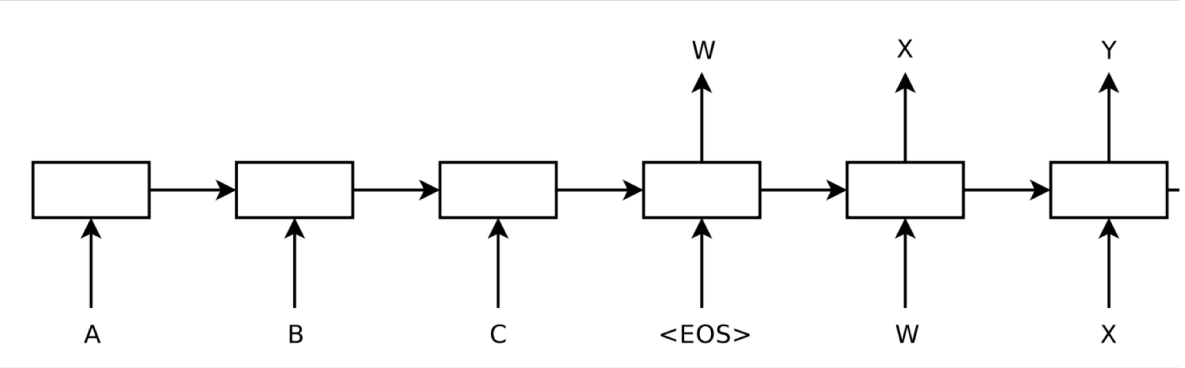
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (third step)



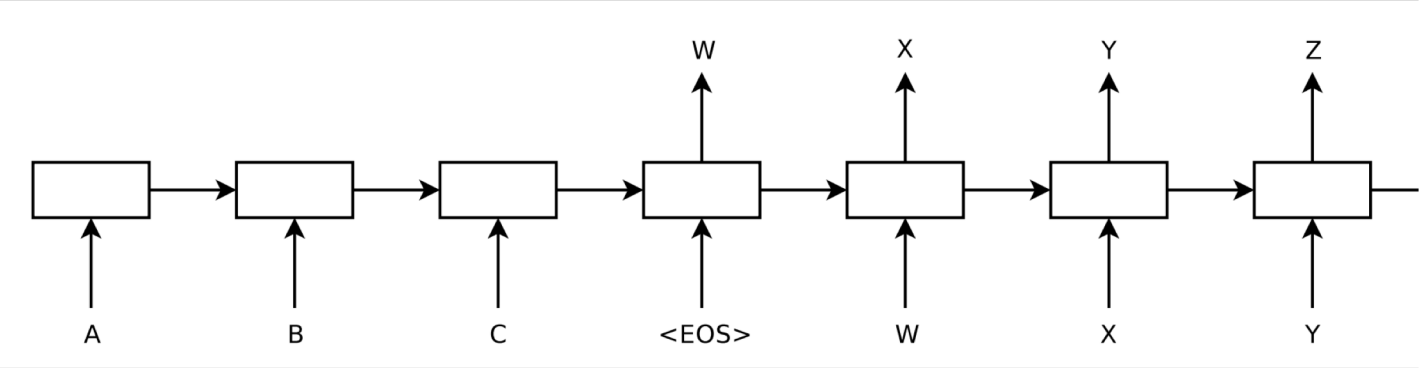
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (third step)



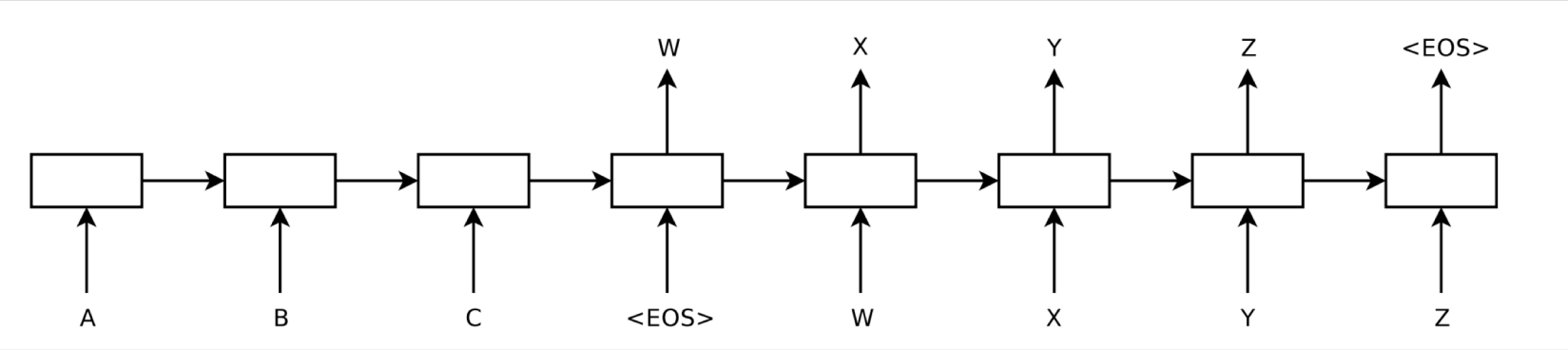
Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (fourth step)

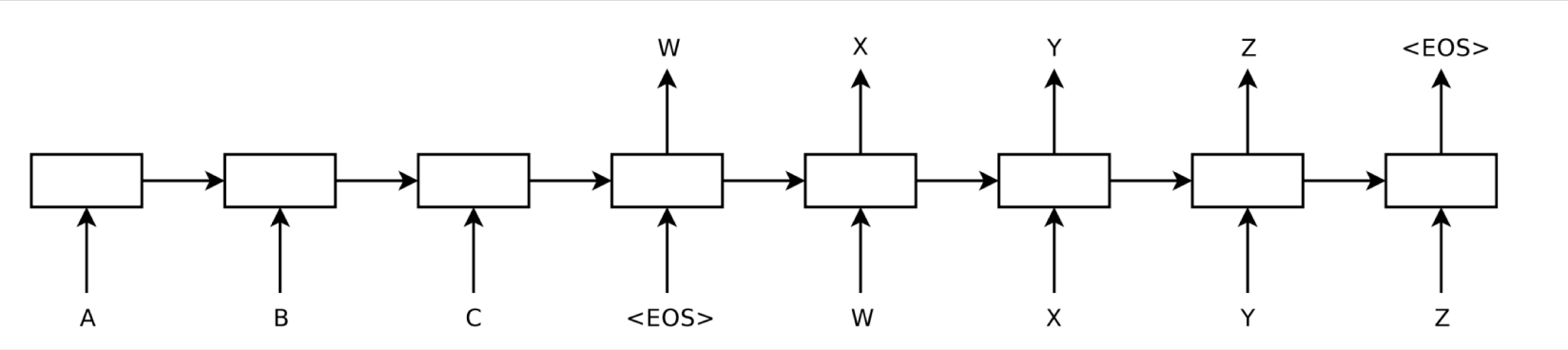


Sequence to Sequence Learning with Neural Networks

Output Sentence LSTM (fifth step)

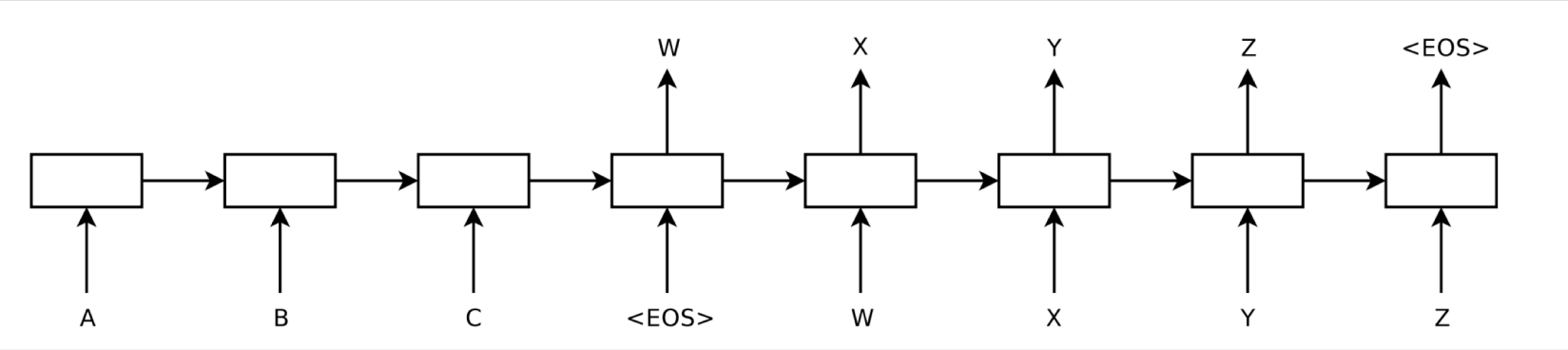


Sequence to Sequence Learning with Neural Networks



Important things of model

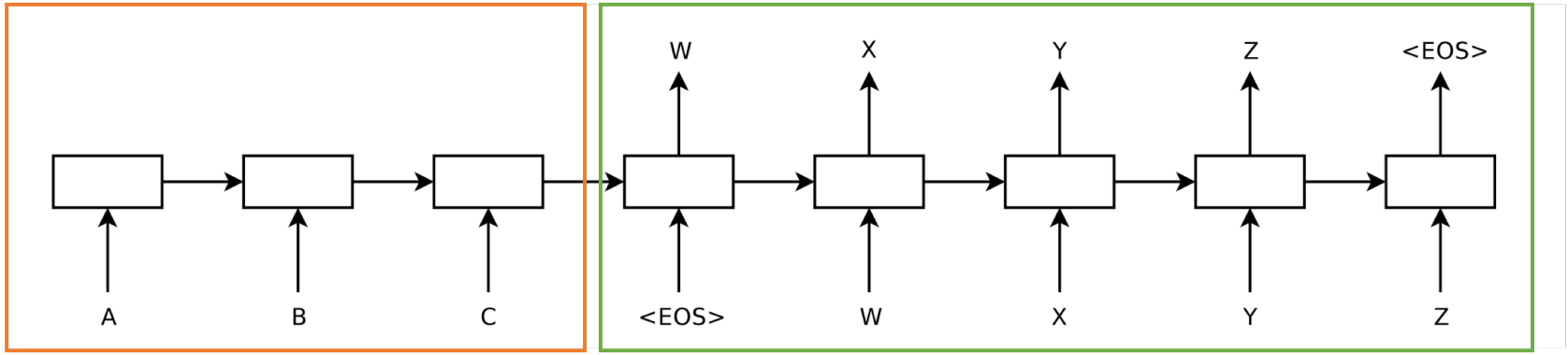
Sequence to Sequence Learning with Neural Networks



Important things of model

- Using different two LSTM

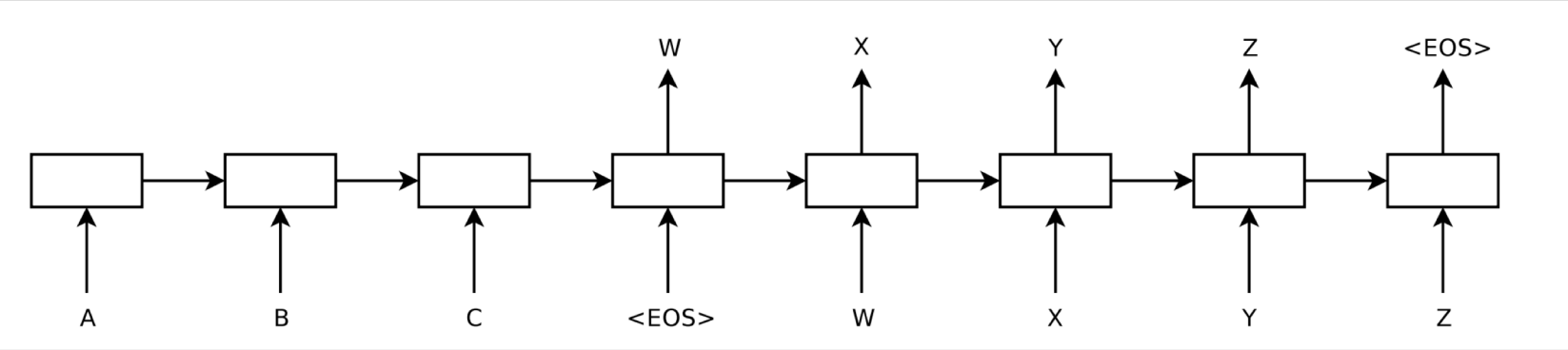
Sequence to Sequence Learning with Neural Networks



Important things of model

- Using different two LSTM

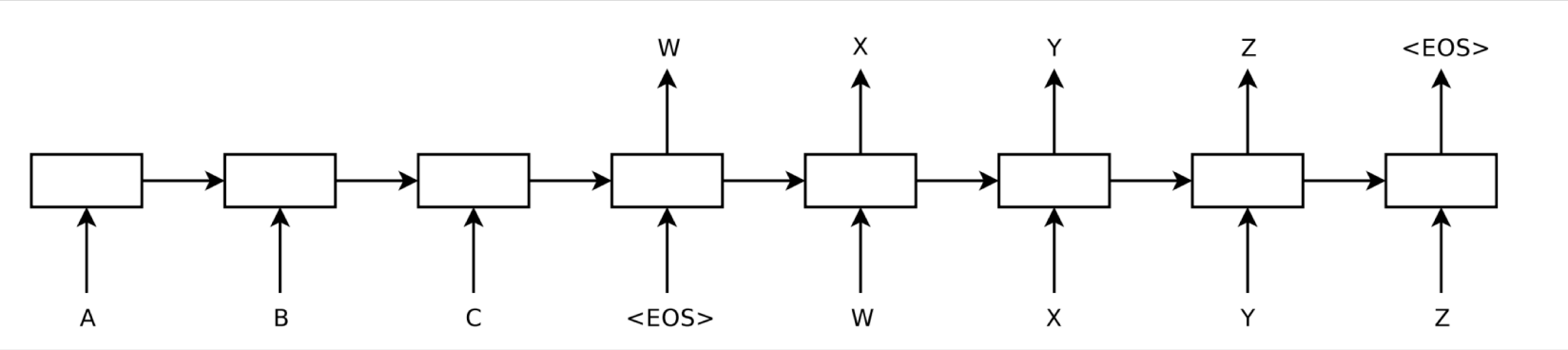
Sequence to Sequence Learning with Neural Networks



Important things of model

- Using different two LSTM
- Deep LSTM(4 layer)

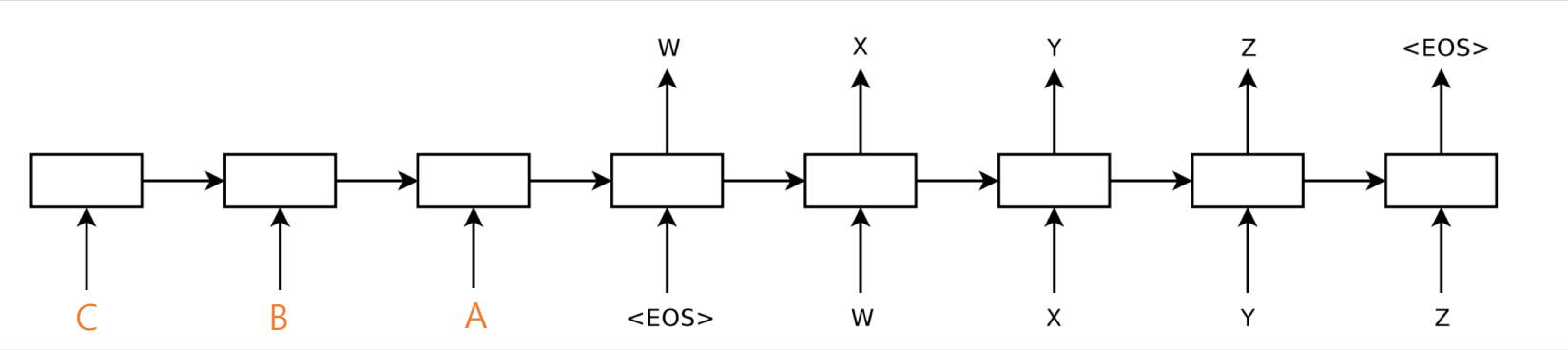
Sequence to Sequence Learning with Neural Networks



Important things of model

- Using different two LSTM
- Deep LSTM(4 layer)
- Reverse the order of words of the input sentences

Sequence to Sequence Learning with Neural Networks



Important things of model

- Using different two LSTM
- Deep LSTM(4 layer)
- Reverse the order of words of the input sentences

How do we evaluate translation quality

How do we evaluate Translation quality

How do we evaluate Translation quality

아침을 너무 오래 먹어서 수업에 지각했다.

How do we evaluate Translation quality

아침을 너무 오래 먹어서 수업에 지각했다.

I was late for class as I ate breakfast too long

How do we evaluate Translation quality

아침을 너무 오래 먹어서 수업에 지각했다.

I was late for class as I ate breakfast too long

I ate breakfast so long that I was late for class.

How do we evaluate Translation quality

아침을 너무 오래 먹어서 수업에 지각했다.

I was late for class as I ate breakfast too long

I ate breakfast so long that I was late for class.

Because I ate breakfast too long, I was late for class.

How do we evaluate Translation quality

아침을 너무 오래 먹어서 수업에 지각했다.

I was late for class as I ate breakfast too long.

I ate breakfast so long that I was late for class.

Because I ate breakfast too long, I was late for class.

How do we evaluate???

BLEU Score

BLEU Score

Bil**l**ingual **E**valuation **U**nderstudy

BLEU Score

Bil**l**ingual **E**valuation **U**nderstudy

기계 번역된 Text의 품질을 평가하기 위한 알고리즘 중 하나.

BLEU Score

Bilingual **E**valuation **U**nderstudy

기계 번역된 Text의 품질을 평가하기 위한 알고리즘 중 하나.

Proceedings of the 40th Annual Meeting of the Association for
Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.[1]

1 Introduction

1.1 Rationale

Human evaluations of machine translation (MT) weigh many aspects of translation, including *adequacy*, *fidelity* , and *fluency* of the translation (Hovy, 1999; White and O'Connell, 1994). A comprehensive catalog of MT evaluation techniques and their rich literature is given by Reeder (2001). For the most part, these various human evaluation approaches are quite expensive (Hovy, 1999). Moreover, they can take *weeks* or *months* to finish. This is a big problem because developers of machine translation systems need to monitor the effect of *daily* changes to their systems in order to weed out bad ideas from good ideas. We believe that MT progress stems from evaluation and that there is a logjam of fruitful research ideas waiting to be released from the evaluation bottleneck. Developers would benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. We propose such an evaluation method in this paper.

1.2 Viewpoint

How does one measure translation performance? *The closer a machine translation is to a professional human translation, the better it is.* This is the central idea behind our proposal. To judge the quality of a machine translation, one measures its closeness to one or more reference human translations according to a numerical metric. Thus, our MT evaluation system requires two ingredients:

1. a numerical "translation closeness" metric
2. a corpus of good quality human reference translations

We fashion our closeness metric after the highly successful *word error rate* metric used by the speech recognition community, appropriately modified for multiple reference translations and allowing for legitimate differences in word choice and word order. The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.

In Section 2, we describe the baseline metric in detail. In Section 3, we evaluate the performance of BLEU. In Section 4, we describe a human evaluation experiment. In Section 5, we compare our baseline metric performance with human evaluations.

[1] So we call our method the bilingual evaluation understudy, BLEU.

BLEU Score

Bilingual **E**valuation **U**nderstudy

기계 번역된 Text의 품질을 평가하기 위한 알고리즘 중 하나.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and **Wei-Jing Zhu**
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.[1]

1 Introduction

1.1 Rationale

Human evaluations of machine translation (MT) weigh many aspects of translation, including *adequacy*, *fidelity* , and *fluency* of the translation (Hovy, 1999; White and O'Connell, 1994). A comprehensive catalog of MT evaluation techniques and their rich literature is given by Reeder (2001). For the most part, these various human evaluation approaches are quite expensive (Hovy, 1999). Moreover, they can take *weeks* or *months* to finish. This is a big problem because developers of machine translation systems need to monitor the effect of *daily* changes to their systems in order to weed out bad ideas from good ideas. We believe that MT progress stems from evaluation and that there is a logjam of fruitful research ideas waiting to be released from the evaluation bottleneck. Developers would benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. We propose such an evaluation method in this paper.

1.2 Viewpoint

How does one measure translation performance? *The closer a machine translation is to a professional human translation, the better it is.* This is the central idea behind our proposal. To judge the quality of a machine translation, one measures its closeness to one or more reference human translations according to a numerical metric. Thus, our MT evaluation system requires two ingredients:

1. a numerical "translation closeness" metric
2. a corpus of good quality human reference translations

We fashion our closeness metric after the highly successful *word error rate* metric used by the speech recognition community, appropriately modified for multiple reference translations and allowing for legitimate differences in word choice and word order. The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.

In Section 2, we describe the baseline metric in detail. In Section 3, we evaluate the performance of BLEU. In Section 4, we describe a human evaluation experiment. In Section 5, we compare our baseline metric performance with human evaluations.

[1] So we call our method the bilingual evaluation understudy, BLEU.

BLEU Score

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

BLEU Score

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU Score

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

더 높은 점수 가져야 함

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU Score

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

→ 더 높은 점수 가져야 함

How?

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU Score

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

→ 더 높은 점수 가져야 함

N-Gram precision

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU Score

N-Gram precision

BLEU Score

N-Gram precision

BLEU Score

N-Gram precision

주어진 Text에서 연속된 N개의 Sequence를 하나의 item으로 봄

BLEU Score

N-Gram precision

주어진 Text에서 연속된 N개의 Sequence를 하나의 item으로 봄

Example)

I was too late for class today.

BLEU Score

N-Gram precision

주어진 Text에서 연속된 N개의 Sequence를 하나의 item으로 봄

Example)

I was too late for class today. ⟶

Uni-Gram(1-Gram)

=> I / was / too / late / for / class / today

BLEU Score

N-Gram precision

주어진 Text에서 연속된 N개의 Sequence를 하나의 item으로 봄

Example)

I was too late for class today. ⟶

Uni-Gram(1-Gram)

=> I / was / too / late / for / class / today

Bi-Gram(2-Gram)

=> I, was / was, too / too, late / late, for / for , class / class, today

BLEU Score

N-Gram precision

주어진 Text에서 연속된 N개의 Sequence를 하나의 item으로 봄

Example)

I was too late for class today. ⟶

Uni-Gram(1-Gram)

=> I / was / too / late / for / class / today

Bi-Gram(2-Gram)

=> I, was / was, too / too, late / late, for / for , class / class, today

Tri-Gram(3-Gram)

=> I, was, too / was, too, late / too, late, for / late, for, class, / for, class, today

BLEU Score

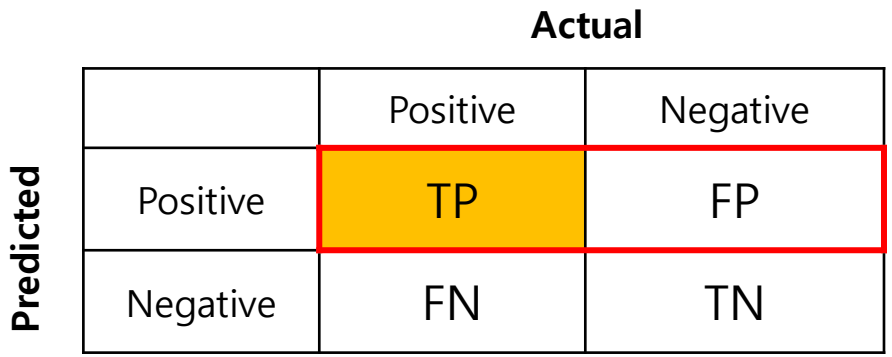
N-Gram precision

BLEU Score

N-Gram precision

$$Precision = \frac{TP}{TP + FP}$$

Classifier가 Positive로 예측한 값 중 실제 Positive 비율



BLEU Score

N-Gram precision

Candidate(Translated Text)의 N-Gram이 Reference와 얼마나 겹치는지 계산한 Precision

BLEU Score

N-Gram precision

Example.

Candidate 1: **It is a guide to action which ensures that the military always obeys the commands of the party.**

Candidate 2: **It is to insure the troops forever hearing the activity guidebook that party direct.**

Reference 1: **It is a guide to action that ensures that the military will forever heed Party commands.**

Reference 2: **It is the guiding principle which guarantees the military forces always being under the command of the Party.**

Reference 3: **It is the practical guide for the army always to heed the directions of the party.**

BLEU Score

N-Gram precision

Example.

Candidate 1: **It is a guide to action which ensures that the military always obeys the commands of the party.**

Candidate 2: **It is to insure the troops forever hearing the activity guidebook that party direct.**

Reference 1: **It is a guide to action that ensures that the military will forever heed Party commands.**

Reference 2: **It is the guiding principle which guarantees the military forces always being under the command of the Party.**

Reference 3: **It is the practical guide for the army always to heed the directions of the party.**

BLEU Score

Uni-Gram precision

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU Score

Uni-Gram precision

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

- Total: 18
- N-Gram Precision $= \frac{17}{18}$

BLEU Score

N-Gram precision

Example.

Candidate 1: **It is a guide to action which ensures that the military always obeys the commands of the party.**

Candidate 2: **It is to insure the troops forever hearing the activity guidebook that party direct.**

Reference 1: **It is a guide to action that ensures that the military will forever heed Party commands.**

Reference 2: **It is the guiding principle which guarantees the military forces always being under the command of the Party.**

Reference 3: **It is the practical guide for the army always to heed the directions of the party.**

BLEU Score

N-Gram precision

Example.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

→

- Total = 14
- N-Gram Precision = $\frac{7}{14}$

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

비교 추가

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: **the the the the the the the**

Reference 1: **The cat is on the mat**

Reference 2: **There is a cat on the mat**

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: **the the the the the the the** → **Totally Wrong!!!**

Reference 1: **The cat is on the mat**

Reference 2: **There is a cat on the mat**

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: the the the the the the the

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: the the the the the the the

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

- Total = 7
- N-Gram Precision $= \frac{7}{7} = 1 = 100\%$

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: the the the the the the the

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

- Total = 7
- N-Gram Precision $= \frac{7}{7} = 1 = 100\%$

Perfect??

BLEU Score

Issue of N-Gram precision

Example.

Candidate 1: the the the the the the the

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

- Total = 7
- N-Gram Precision $= \frac{7}{7} = 1 = 100\%$

"Reasonable" word over-generating

BLEU Score

Solution for ”Reasonable” word over-generating

BLEU Score

Solution for ”Reasonable” word over-generating

- Clipping maximum reference count
- Using different “N” in N-gram precision

BLEU Score

Solution for ”Reasonable” word over-generating

- Clipping maximum reference count
- Using different “N” in N-gram precision

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

Example.

Candidate 1: **the the the the the the the**

Reference 1: **The cat is on the mat**

Reference 2: **There is a cat on the mat**

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

Example.

Candidate 1: the the **the the the the the**

Reference 1: The **cat is on** the **mat**

Reference 2: **There is a cat on the mat**

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

Example.

Candidate 1: the the **the the the the the**

Reference 1: The **cat is on** the **mat**

Reference 2: **There is a cat on the mat**

- Total = 7
- Modified N-Gram Precision $= \frac{2}{7} = 0.28 \ldots = 28.57\%$
- Standard N-Gram Precision $= \frac{7}{7} = 1 = 100\%$

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = avg(P_1, P_2, \dots, P_n)$$

BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = avg(P_1, P_2, \dots, P_n)$$

- 큰 N에 대해서 Precision이 높을 수록 더 정확한 번역

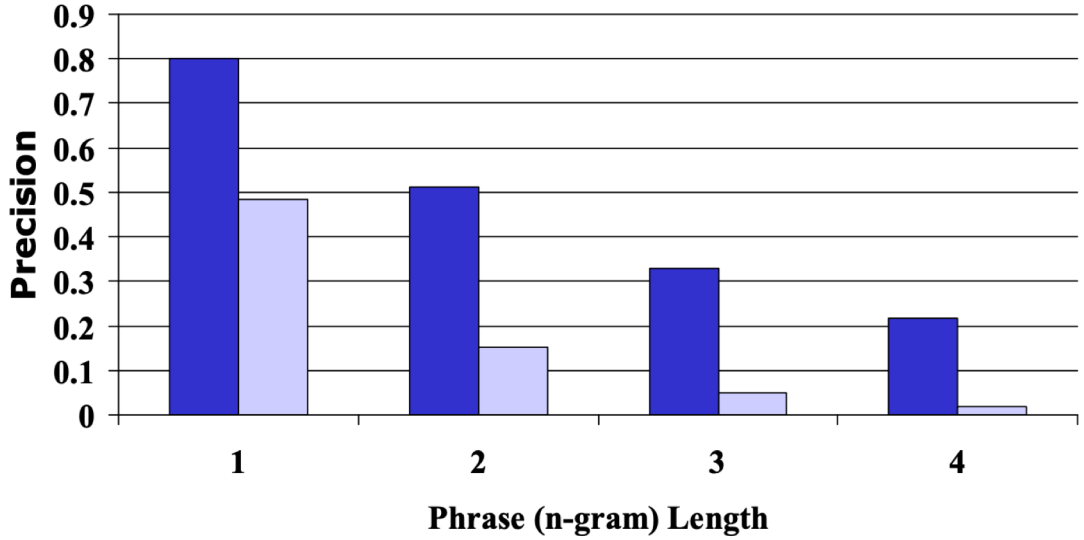
BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = avg(P_1, P_2, \dots, P_n)$$

- 큰 N에 대해서 Precision이 높을 수록 더 정확한 번역



- Blue => Human
- Light blue => Machine

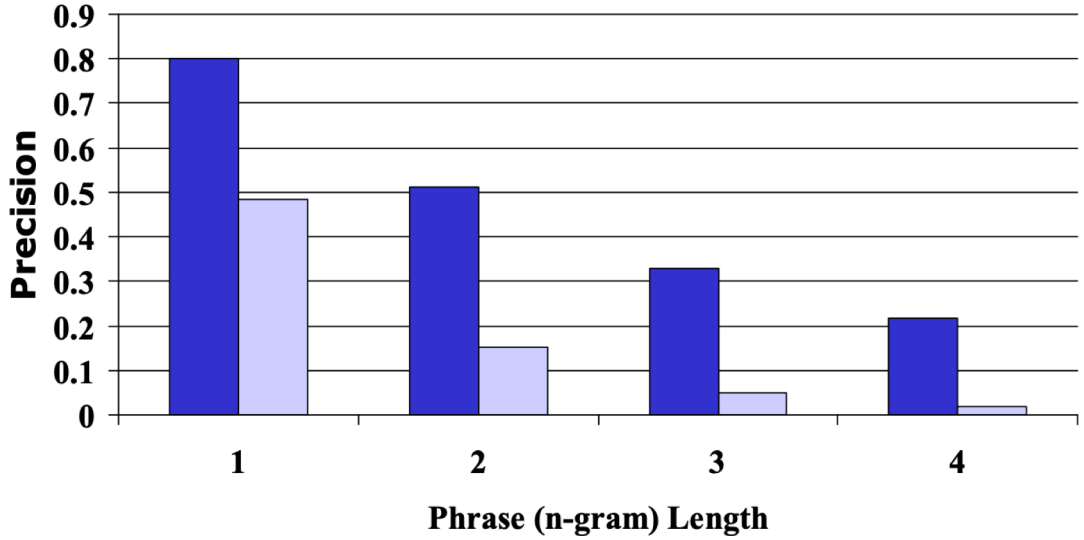
BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = avg(P_1, P_2, \dots, P_n)$$

- 큰 N에 대해서 Precision이 높을 수록 더 정확한 번역
- N이 커질수록 지수적(exponential)으로 줄어든다



- Blue => Human
- Light blue => Machine

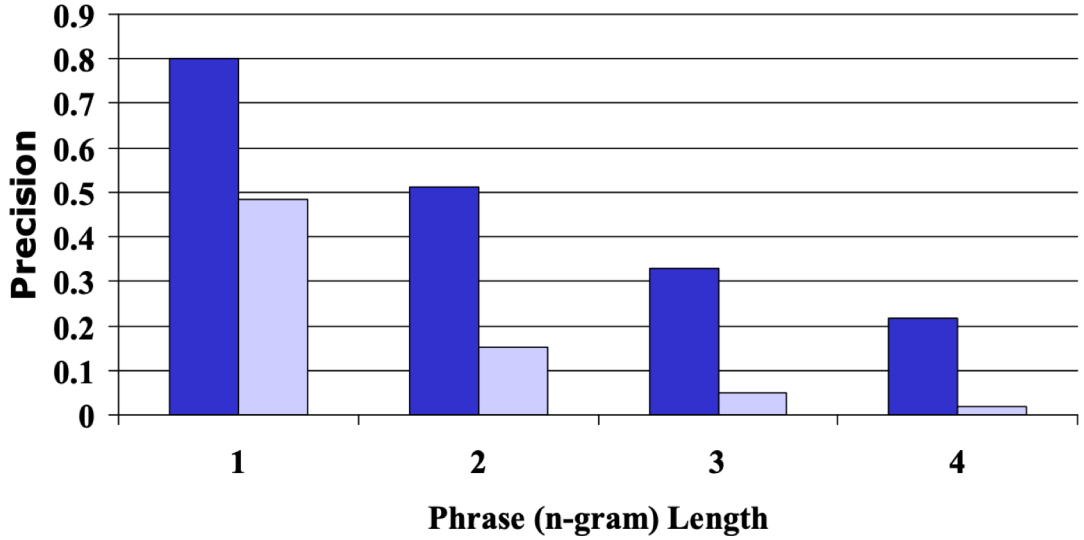
BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = avg(P_1, P_2, \dots, P_n)$$

- 큰 N에 대해서 Precision이 높을 수록 더 정확한 번역
- N이 커질수록 지수적(exponential)으로 줄어든다.
  - 큰 N에 대응할 수 있도록 기하 평균 사용



- Blue => Human
- Light blue => Machine

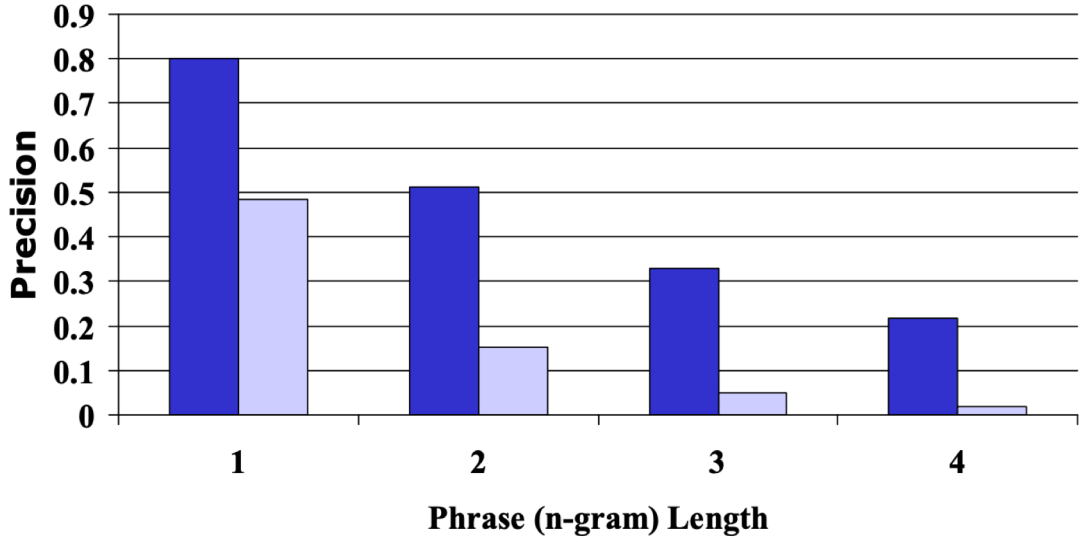
BLEU Score

Solution for "Reasonable" word over-generating

- Clipping maximum reference count
- Using different "N" in N-gram precision

$$Precision = (P_1 \cdot P_2 \cdots P_n)^{\frac{1}{n}}$$

- 큰 N에 대해서 Precision이 높을 수록 더 정확한 번역
- N이 커질수록 지수적(exponential)으로 줄어든다.
  - 큰 N에 대응할 수 있도록 기하 평균 사용
  - 경우에 따라 산술 평균도 사용할 수 있음



- Blue => Human
- Light blue => Machine

BLEU Score

Second Issue of N-Gram precision

BLEU Score

Second Issue of N-Gram precision

Example.

Candidate 1: **of the**

Reference 1: **It is a guide to action that ensures that the military will forever heed Party commands.**

Reference 2: **It is the guiding principle which guarantees the military forces always being under the command of the Party.**

Reference 2: **It is the practical guide for the army always to heed directions of the party.**

BLEU Score

Second Issue of N-Gram precision

Example.

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 2: It is the practical guide for the army always to heed directions of the party.

BLEU Score

Second Issue of N-Gram precision

Example.

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 2: It is the practical guide for the army always to heed directions of the party.

→

- $P_1 = \frac{2}{2} = 1$
- $P_2 = \frac{1}{1} = 1$

BLEU Score

Second Issue of N-Gram precision

Example.

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 2: It is the practical guide for the army always to heed directions of the party.

→

- $P_1 = \frac{2}{2} = 1$
- $P_2 = \frac{1}{1} = 1$

Something wrong!

BLEU Score

Second Issue of N-Gram precision

Example.

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 2: It is the practical guide for the army always to heed directions of the party.

- $P_1 = \frac{2}{2} = 1$
- $P_2 = \frac{1}{1} = 1$

Something wrong!

Brevity Penalty

BLEU Score

Brevity Penalty

BLEU Score

Brevity Penalty

- Candidate의 길이가 Reference 보다 짧으면 Penalty를 줌

BLEU Score

Brevity Penalty

- Candidate의 길이가 Reference 보다 짧으면 Penalty를 줌
  - BP = 1 (Candidate이 짧지 않을 때)
  - BP < 1 (Candidate이 짧을 때)

BLEU Score

Brevity Penalty

- Candidate의 길이가 Reference 보다 짧으면 Penalty를 줌
  - BP = 1 (Candidate이 짧지 않을 때)
  - BP < 1 (Candidate이 짧을 때)

BLEU Score

Brevity Penalty

- Candidate의 길이가 Reference 보다 짧으면 Penalty를 줌
  - BP = 1 (Candidate이 짧지 않을 때)
  - BP < 1 (Candidate이 짧을 때)

길이가 짧아질수록 지수적(exponential)으로 penalty를 받도록 함.

BLEU Score

Brevity Penalty

- Candidate의 길이가 Reference 보다 짧으면 Penalty를 줌
  - BP = 1 (Candidate이 짧지 않을 때)
  - BP < 1 (Candidate이 짧을 때)

길이가 짧아질수록 지수적(exponential)으로 penalty를 받도록 함.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

BLEU Score

Finally

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

BLEU Score

Finally

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

- c: candidate
- r: reference
- w_n: weight (Uniform weight: $\mathrm{w_n} = \frac{1}{\mathrm{N}}$)

BLEU Score

Finally (Log)

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

BLEU Score

Disadvantage of BLEU

BLEU Score

Disadvantage of BLEU

BLEU doesn't consider meaning

BLEU doesn't directly consider sentence structure

BLEU doesn't handle morphologically rich languages well

Another method

- NIST
- ROUGE
- Perplexity
- WER
- F-Score
- METEOR
- TER
- etc...

Machine Translation Datasets

WMT Dataset

Machine Translation Datasets

WMT Dataset

- Workshop on Statistical Machine Translation
- Bilingual Datasets
  - English-French and French-English
  - English-Portuguese and Portuguese-English
  - English-Spanish and Spanish-English
  - English-German and German-English
  - English-Chinese and Chinese-English

Machine Translation Datasets

WMT Dataset

- Workshop on Statistical Machine Translation
- Bilingual Datasets
  - English-French and French-English
  - English-Portuguese and Portuguese-English
  - English-Spanish and Spanish-English
  - English-German and German-English
  - English-Chinese and Chinese-English

Source	un homme est debout près d' une série de jeux vidéo dans un bar .
Iteration 0	a man is seated near a series of games video in a bar .
Iteration 1	a man is standing near a closeup of other games in a bar .
Iteration 2	a man is standing near a bunch of video video game in a bar .
Iteration 3	a man is standing near a bunch of video games in a bar .
Reference	**a man is standing by a group of video games in a bar .**
Source	une femme aux cheveux roses habillée en noir parle à un homme .
Iteration 0	a woman at hair roses dressed in black speaks to a man .
Iteration 1	a woman at glasses dressed in black talking to a man .
Iteration 2	a woman at pink hair dressed in black speaks to a man .
Iteration 3	a woman with pink hair dressed in black is talking to a man .
Reference	**a woman with pink hair dressed in black talks to a man .**
Source	une photo d' une rue bondée en ville .
Iteration 0	a photo a street crowded in city .
Iteration 1	a picture of a street crowded in a city .
Iteration 2	a picture of a crowded city street .
Iteration 3	a picture of a crowded street in a city .
Reference	**a view of a crowded city street .**