Attention Mechanism in NLP

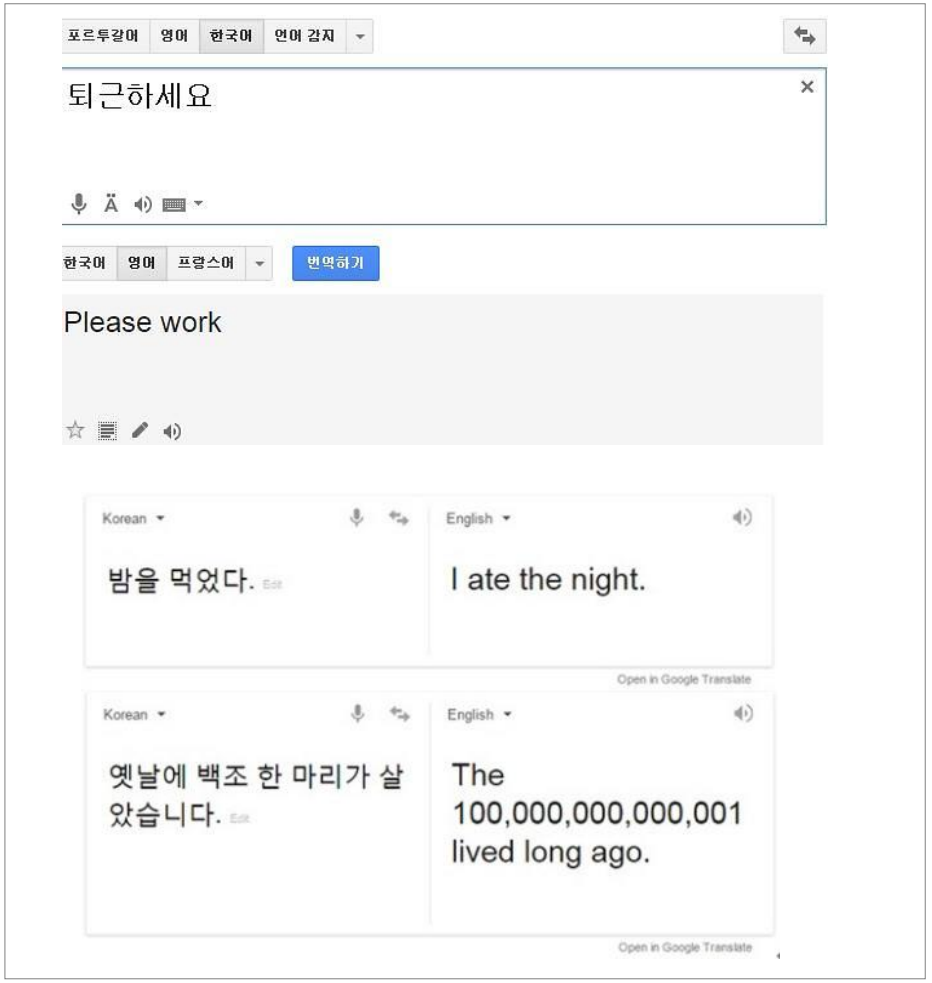
– Machine Translation 영역에서 Neural Network의 효율성과 효과성 확보 –

목차

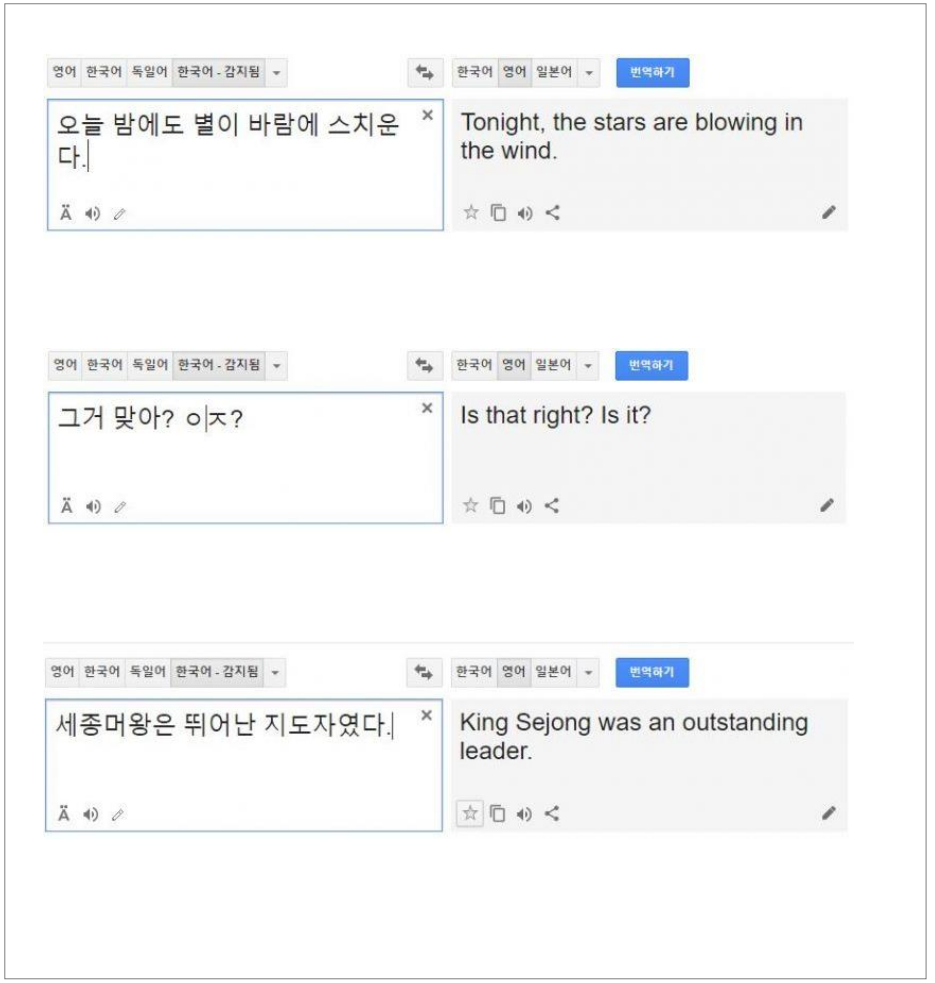
I. History of Machine Translation

2016년 이후 구글 신경망 기계번역(GNMT)의 도입으로 번역 정확도 55%에서 85%로 향상

과거



현재



I. History of Machine Translation

사용 가능한 Computing Power가 늘어남에 따라 Machine Translation 또한 점진적 자동화 달성

Rule-Based

반자동 번역

- 언어 별 규칙 정의
  - 형태소 등 수동으로 정의

제한점

- 언어 별 사전 필요
  - 유지보수 제한
- 언어 별 구문 및 어순 정형화 필요
- Target 언어에 맞게 품사 및 순서 변경
  - 조사 등 언어특성의 부적절한 연계

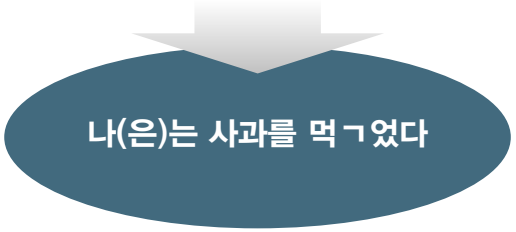
Statistical Method

통계 모델 기반 번역

- 번역/언어/재배열 모델로 구분
- Corpus 기반 기계학습
  - 사람의 개입 최소화
- Word/Phrase 기반 번역 가능

제한점

- 적절한 양의 Corpus 확보 문제
- 예측 불가능한 번역 성능
  - 버전 또는 도메인 별 번역 품질 변동
- 어족 간 번역품질 제한
  - 어순이 다를 경우 번역품질 저하



Neural Network

문장을 별도의 가공 없이 활용

- Corpus 불필요
  - 문장 또는 문단 단위 번역 가능
- Word Embedding
  - Word2Vec
  - Doc2Vec
- RNN 등

제한점

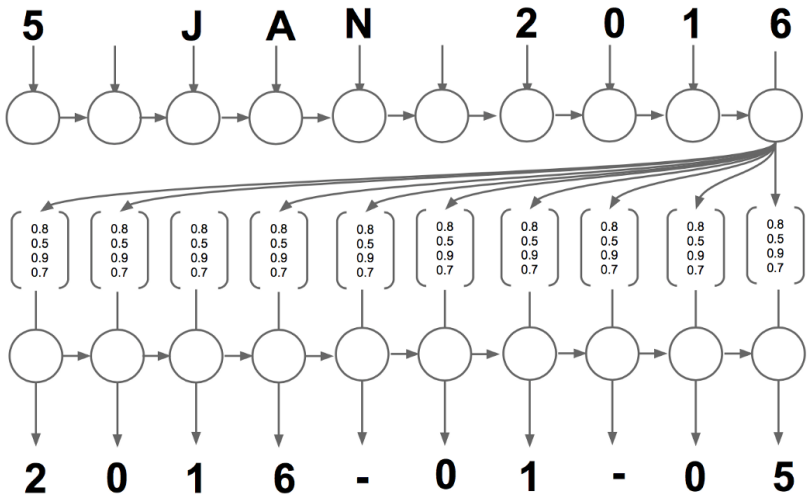
- Source-Target 문장의 Coupling 필요
- 연산자원 확보 문제
- 문장이 길어질 경우 번역품질 저하

* Byeong il Go, GNMT로 알아보는 신경망 기반 기계번역, 2016 – https://www.slideshare.net/ByeongilKo/gnmt-69817390

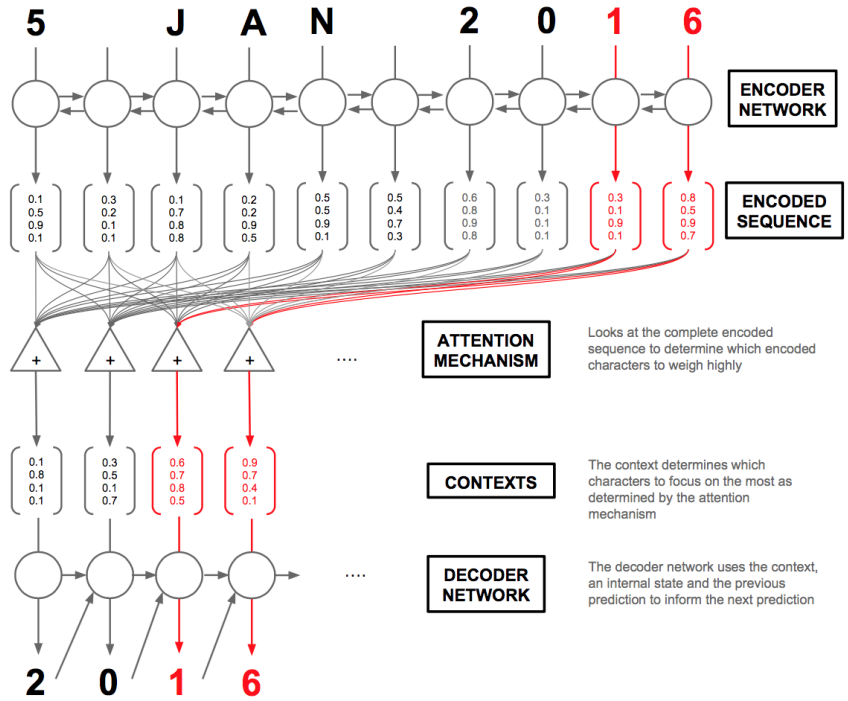
II. What is Attention?

Target Sequence의 Element를 예측할 때 마다 Source의 Weight를 변경

Seq2Seq



Seq2Seq with Attention



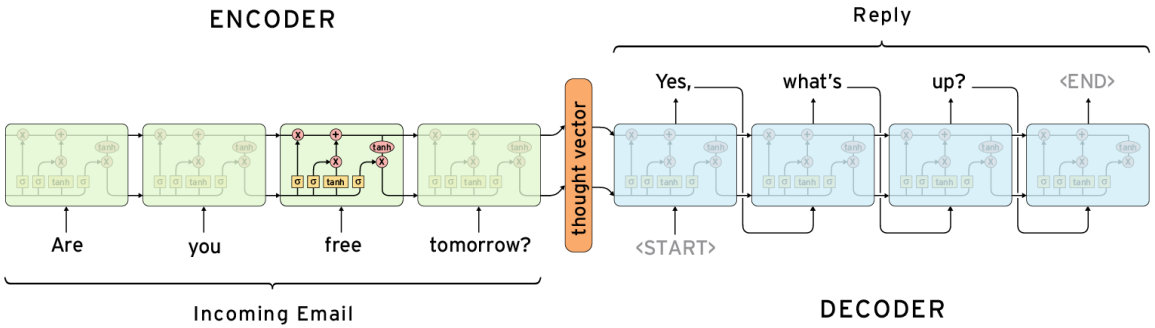
III. Why Attention?

Source Data에서 필요한 부분만을 바라보는 직관적인 방법으로 효율성과 효과성을 모두 달성

Cho, et al.,『Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation』, 2014

Seq2Seq(Encoder–Decoder)

- Encoder로 Source를 하나의 Vector로 압축
- Decoder는 압축된 Vector를 Hidden State로 수용
  - Hidden State에 각 Cell의 정보가 누적
- Training과 Inference의 구분
  - Training: Good Morning, Everybody!
    여러분, 좋은 아침입니다!(정답 필요)
  - Inference: Good Morning, Everybody!

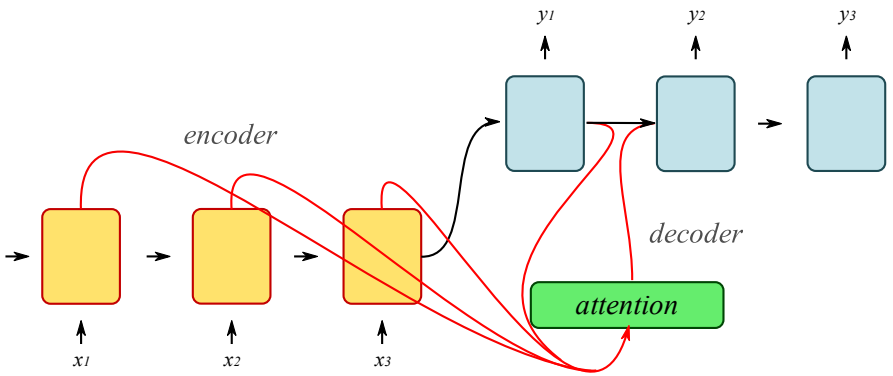


Dzmitry Bahdanau, et al., 『Neural Machine Translation by Jointly Learning to Align and Translate』, 2014

Seq2Seq with Attention

- 효율성의 문제
  :Target 문장의 단어를 번역할 때 마다
  Source 문장 전체를 반영
- 효과성의 문제
  : 학습 데이터보다 긴 문장이 들어오면 품질 악화

"Encoding 시 중요한 부분에 집중"



IV. How to Attend?

Encoder와 Decoder의 Hidden State 유사도를 측정하여 Softmaxed Weight Vector로 치환

Formula

RNN

- Input Sequence $X = (x_1, x_2, \dots, x_{T_x})$
- Hidden State $h_t = f(x_t, h_{t-1})$
- Vector $c = q(\{h_1, h_2, \dots, h_{T_x}\})$
- $q(\cdot)$: non-linear function
- Predicted Words $Y = (y_1, y_2, \dots, y_{T_y})$
- Decoder $p(y) = (\prod_{t=1}^{T} p(y_t | \{y_1, y_2, \dots y_{t-1}\}, c)$
- Each Conditional Probability of y_t

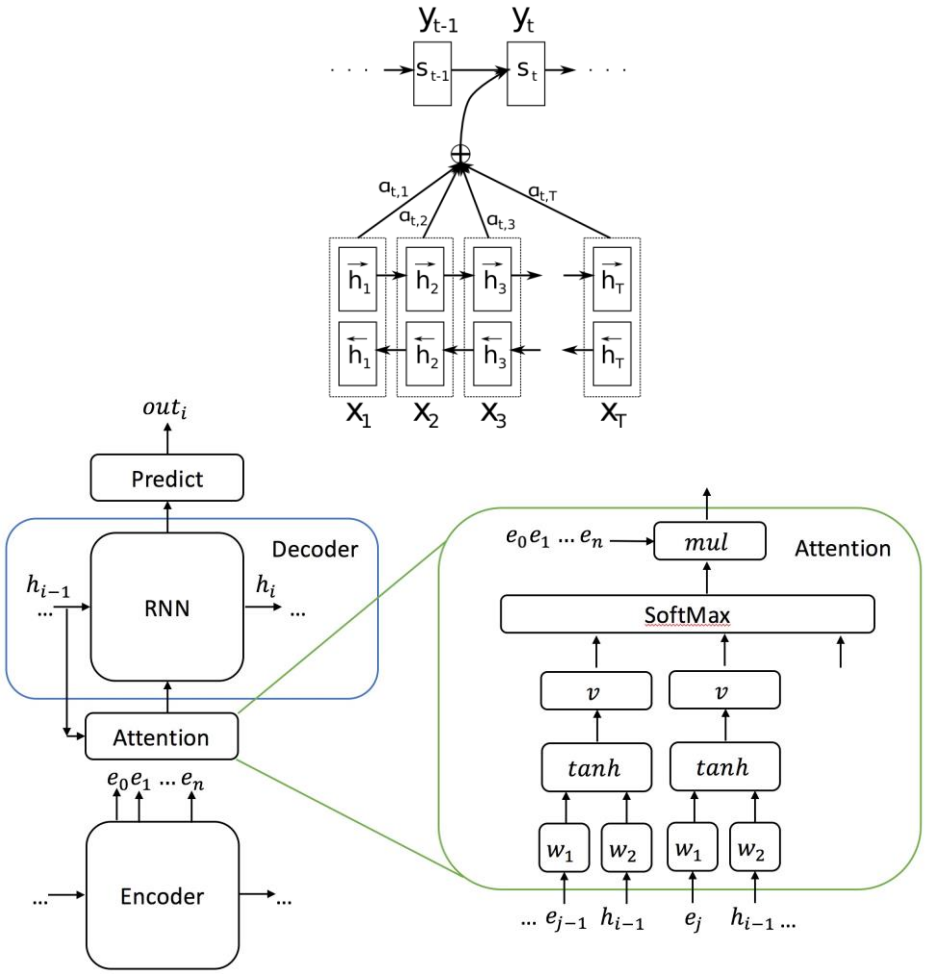
  $p(y_t | \{y_1, y_2, \dots y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$

Attention Decoder

- Each Conditional Probability of y_i

  $p(y_i | \{y_1, y_2, \dots y_{i-1}\}, X) = g(y_{i-1}, s_i, c_i)$
- Hidden State of Decoder $s_i = f(s_{i-1}, y_{i-1}, c_i)$
- Hidden State of Encoder $h_j = f(x_j, h_{j-1})$
- Context Vector $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$
- Weight $\alpha_{ij} = \sum_{k=1}^{T_x} \frac{\exp(e_{ij})}{\exp(e_{ik})}$ (softmaxed)
- $e_{ij} = a(s_{i-1}, h_j)$ (Decoder s_{i-1}와 Encoder h_j 유사도 측정$(-1 \sim +1)$)
- $a(\cdot)$: non-linear function (tanh)

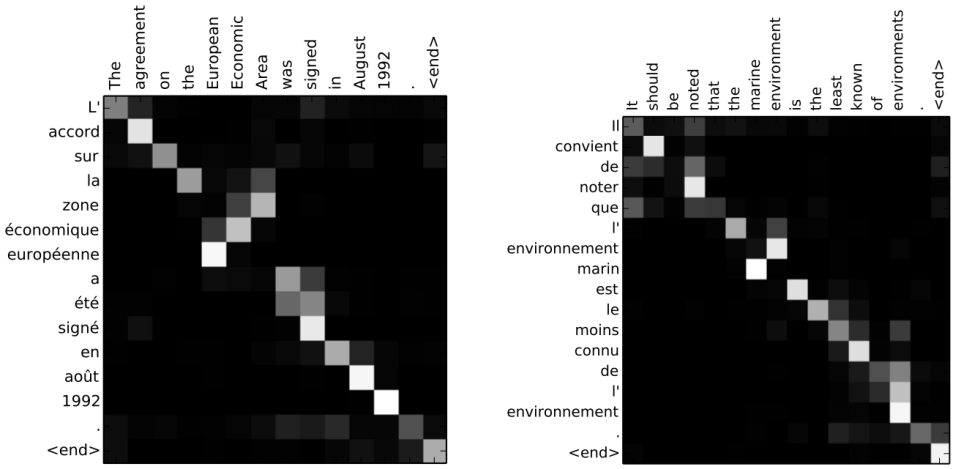
Example

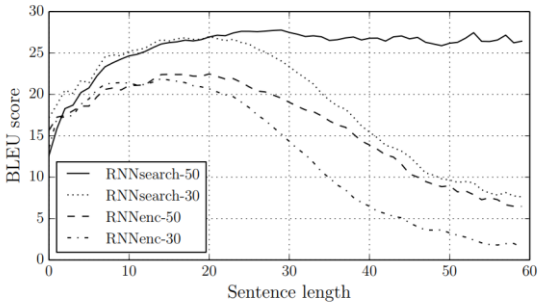


V. Results

양적, 질적인 면에서 번역 품질 또한 기존 대비 상승

Quantitative Analysis





Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

Qualitative Analysis

As an example, consider this source sentence from the test set:

> *An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.*

The RNNencdec-50 translated this sentence into:

> *Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.*

Let us consider another sentence from the test set:

> *This kind of experience is part of Disney's efforts to "extend the lifetime of its series and build new relationships with audiences via digital platforms that are becoming ever more important," he added.*

The translation by the RNNencdec-50 is

> *Ce type d'expérience fait partie des initiatives du Disney pour "prolonger la durée de vie de ses nouvelles et de développer des liens avec les lecteurs numériques qui deviennent plus complexes.*

End of Document