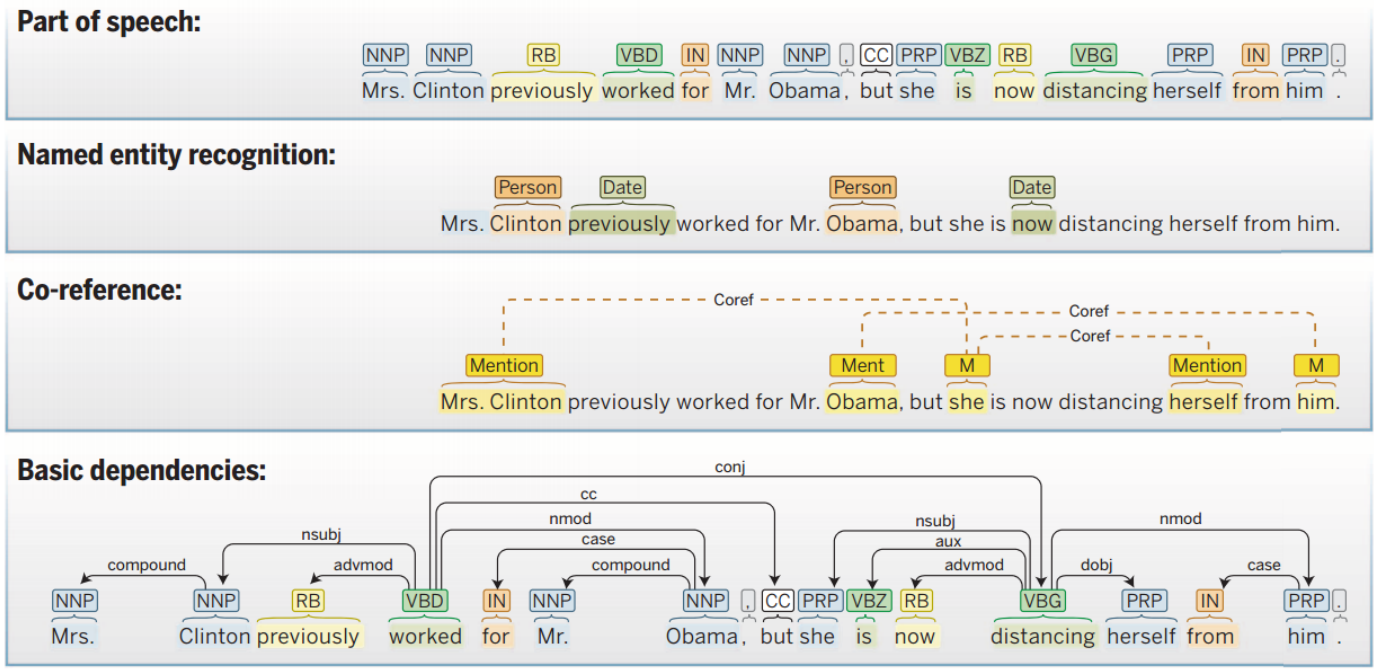


Lecture 2: Text Preprocessing

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Introduction to NLP

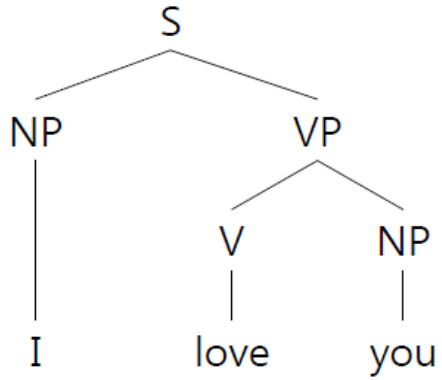
02 Lexical Analysis

03 Syntax Analysis

04 Other Topics in NLP

Syntax Analysis

- Syntax Analysis
  - ✓ Process of analyzing a string of symbols conforming to the rules of a formal grammar

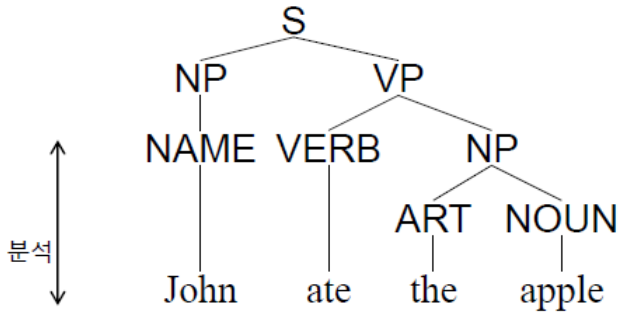


- Parser
  - ✓ An algorithm that computes a structure for an input string given a grammar
  - ✓ All parsers have two fundamental properties
    - Directionality: the sequence in which the structures are constructed (e.g., top-down or bottom-up)
    - Search strategy: the order in which the search space of possible analysis explored (e.g., depth-first, breadth-first)

Syntax Analysis

- Parsing Representation
  - ✓ Tree vs List

Tree Representation



List Representation

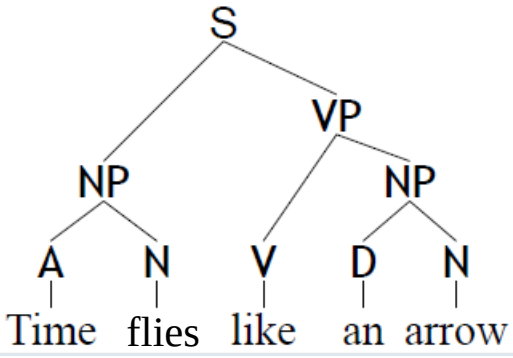
```
(S   (NP (NAME  John))
     (VP (VERB   ate)
         (NP  (ART  the)
              (NOUN  apple)) ) )
```

  - ✓ Meaning
    - S (Sentence) consists of NP (Noun Phrase) and VP (Verb Phrase)
    - NP consists of Name (John)
    - VP consists of VERB (ate) and the other NP
    - NP consists of ART (the) and Noun (apple)

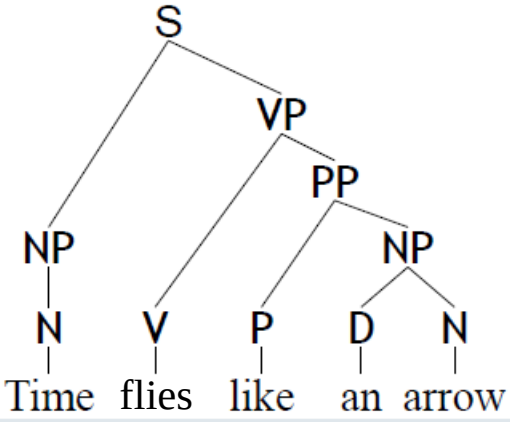
Syntax Analysis

- Not a single parsing tree due to language ambiguity
- Lexical ambiguity
  - ✓ One word can be used for multiple parts of speech
  - ✓ Lexical ambiguity causes structural ambiguity

G : S → NP VP
NP → D N | A N | N
VP → V | VP NP | VP PP
PP → P NP



Input Sentence :
Time flies like an arrow

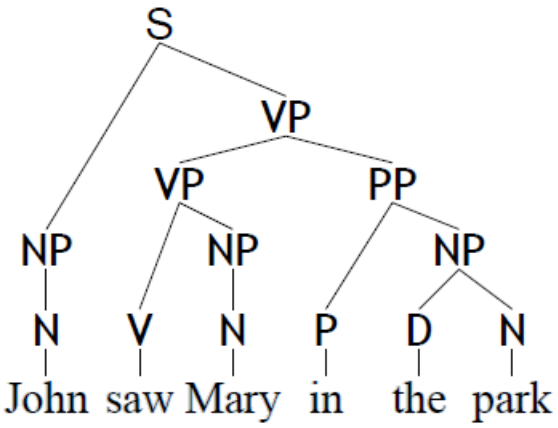


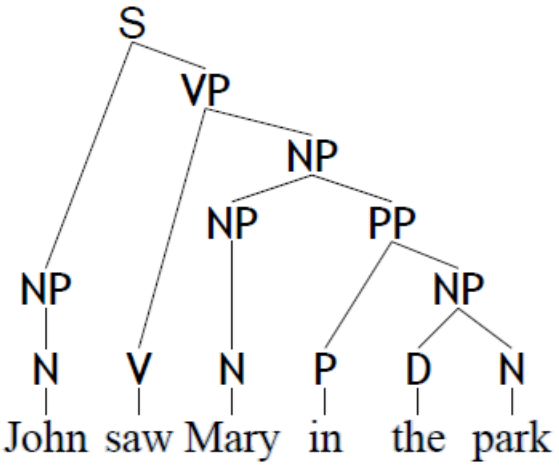
Syntax Analysis

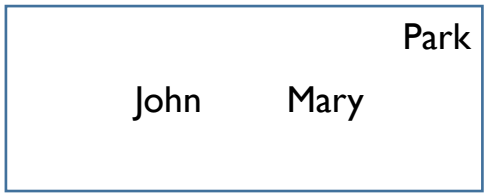
- Structural Ambiguity
  - ✓ One sentence can be understood in different ways

G : S → NP VP
NP → N | D N | NP PP
VP → V NP | VP PP
PP → P NP

Input Sentence :
John saw Mary in the park.







AGENDA

01 Introduction to NLP

02 Lexical Analysis

03 Syntax Analysis

04 Other Topics in NLP

Language Modeling

- Probabilistic Language Model
  - ✓ Assign a probability to a sentence (not POS tags, but the sentence itself)
- Applications
  - ✓ Machine Translation
    - P(**high** wind tonight) > P(**large** wind tonight)
  - ✓ Spell correction
    - The office is about fifteen **minuets** from my house
    - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
  - ✓ Speech recognition
    - P(I saw a van) >> P(eyes awe of an)
  - ✓ Summarization, question-answering, etc.

Language Modeling

- Probabilistic Language Modeling
  - ✓ Compute the probability of a sentence or sequence of words

$$P(W) = P(w_1, w_2, w_3, ..., w_n)$$

  - ✓ Related task: probability of an upcoming word

$$P(w_5|w_1, w_2, w_3, w_4)$$

    - ex) I love you more than I can _____. (swim? say?)
- How to compute P(W)
  - ✓ What is P(its, water, is, so, transparent, that)?
  - ✓ Chain Rules of Probability:

$$P(w_1, w_2, w_3, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, ..., w_{n-1})$$

$$P(\text{its water is so trasparent"}) = P(\text{its}) \times P(\text{water}|\text{its}) \times P(\text{is}|\text{its water}) \times P(\text{so}|\text{its water is}) \times P(\text{transparent}|\text{its water is so})$$

Language Modeling

- Markov Assumption
  - ✓ Consider only k previous words when estimating the conditional probability

$$P(w_1 w_2 ... w_n) \approx \prod_i P(w_i | w_{i-k} ... w_{i-1}) \quad P(w_i | w_1 w_2 ... w_{i-1}) \approx \prod_i P(w_i | w_{i-k} ... w_{i-1})$$

  - ✓ Simplest case: Unigram model

$$P(w_1 w_2 ... w_n) \approx \prod_i P(w_i)$$

  - ✓ An example of automatically generated sentences from a unigram model

```
fifth, an, of, futures, the, an, incorporated, a,
a, the, inflation, most, dollars, quarter, in, is,
mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the
```

Language Modeling

- Bigram model
  - ✓ Condition on the previous word

$$P(w_i|w_1 w_2 \ldots w_{i-1}) \approx P(w_i|w_{i-1})$$

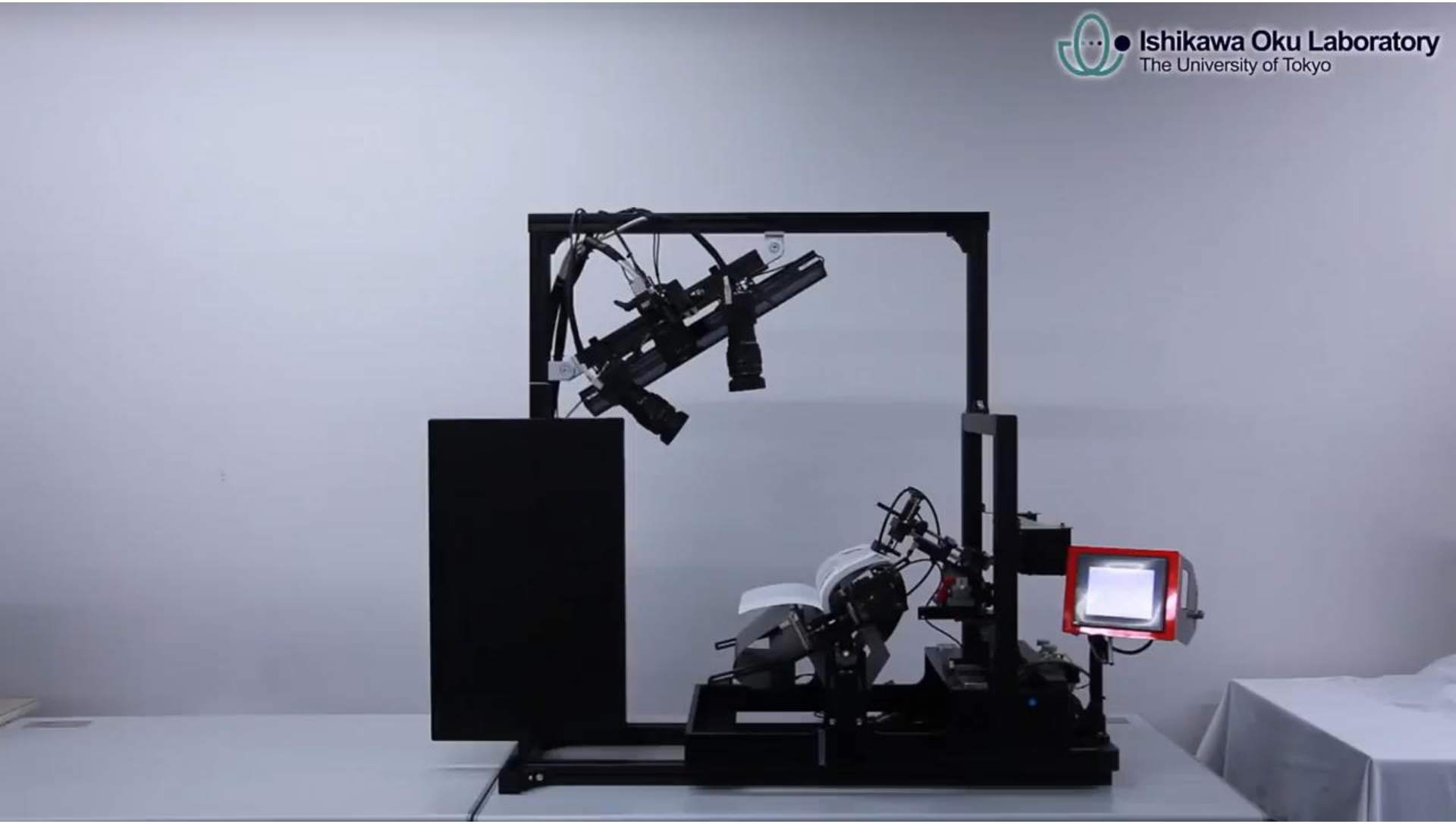
texaco, rose, one, in, this, issue, is, pursuing, growth, in,
a, boiler, house, said, mr., gurria, mexico, 's, motion,
control, proposal, without, permission, from, five, hundred,
fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november

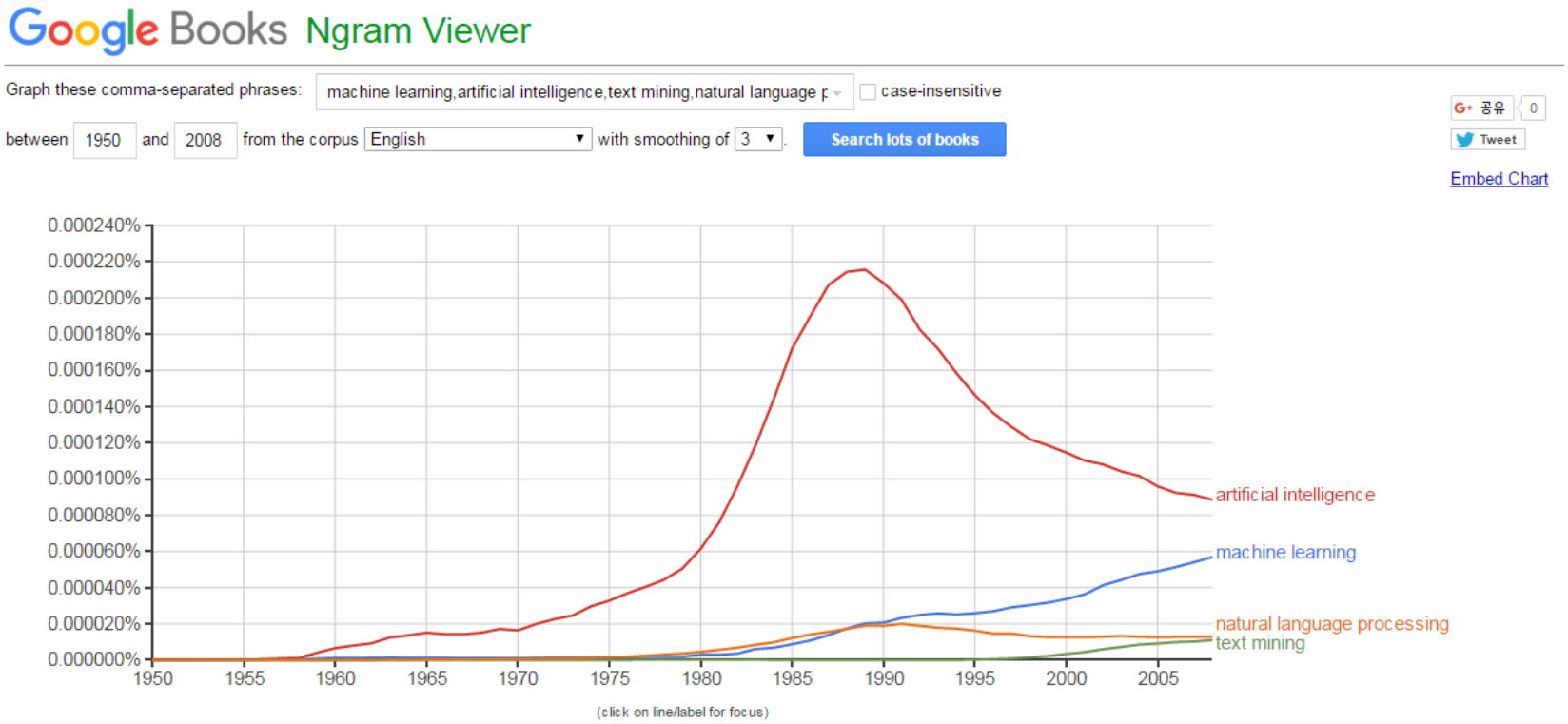
- N-gram models
  - ✓ Can extend to trigrams, 4-grams, 5-grams
    - In sufficient model of language because language has long-distance dependencies
    - "The computer when I had just put into the machine room on the fifth floor crashed."
  - ✓ We can often get away with N-gram models

Language Modeling



Language Modeling

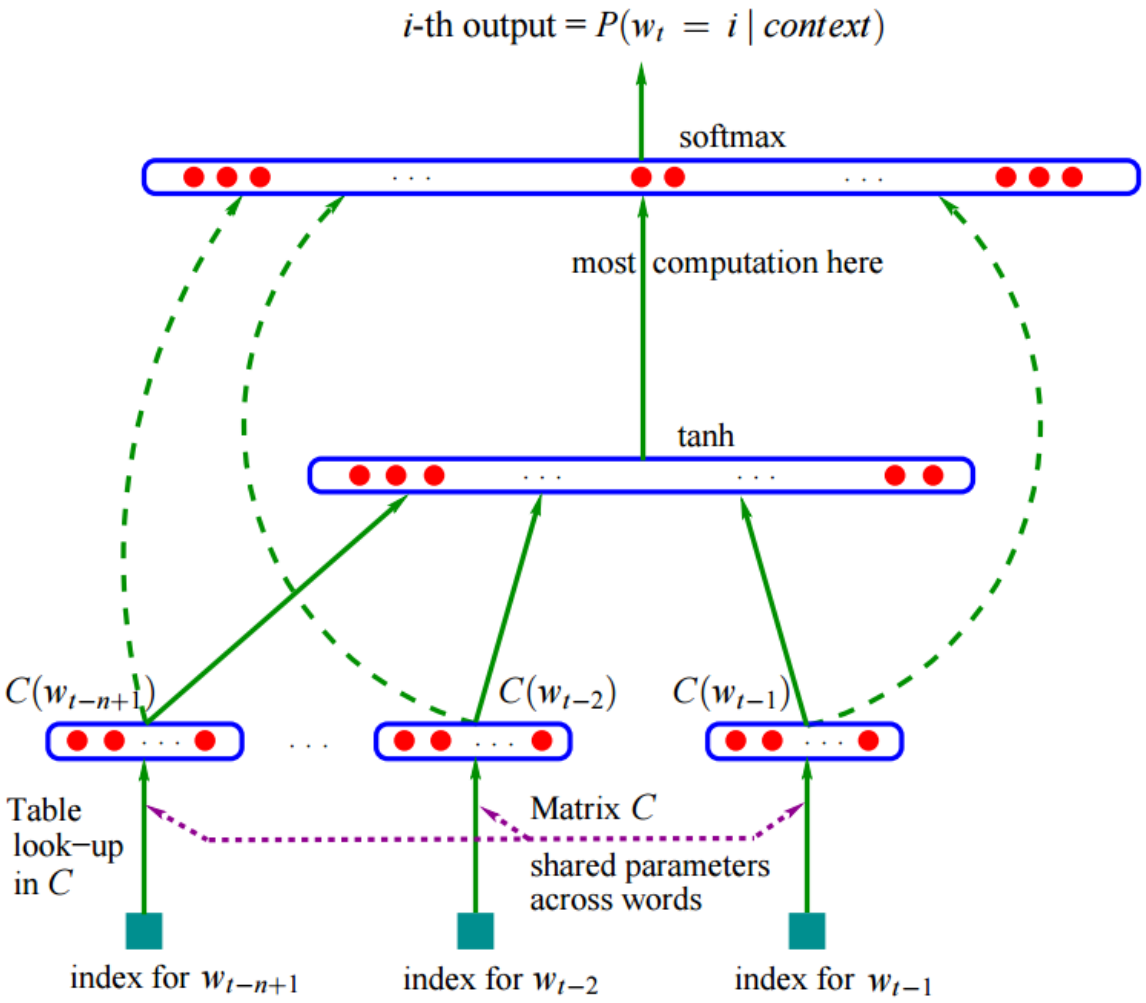
- Google Books N-Gram
  - ✓ 1,024 billion words & 1.1 billion 5-grams that appeared at least 40 times (2006)



Language Modeling

Bengio et al. (2003)

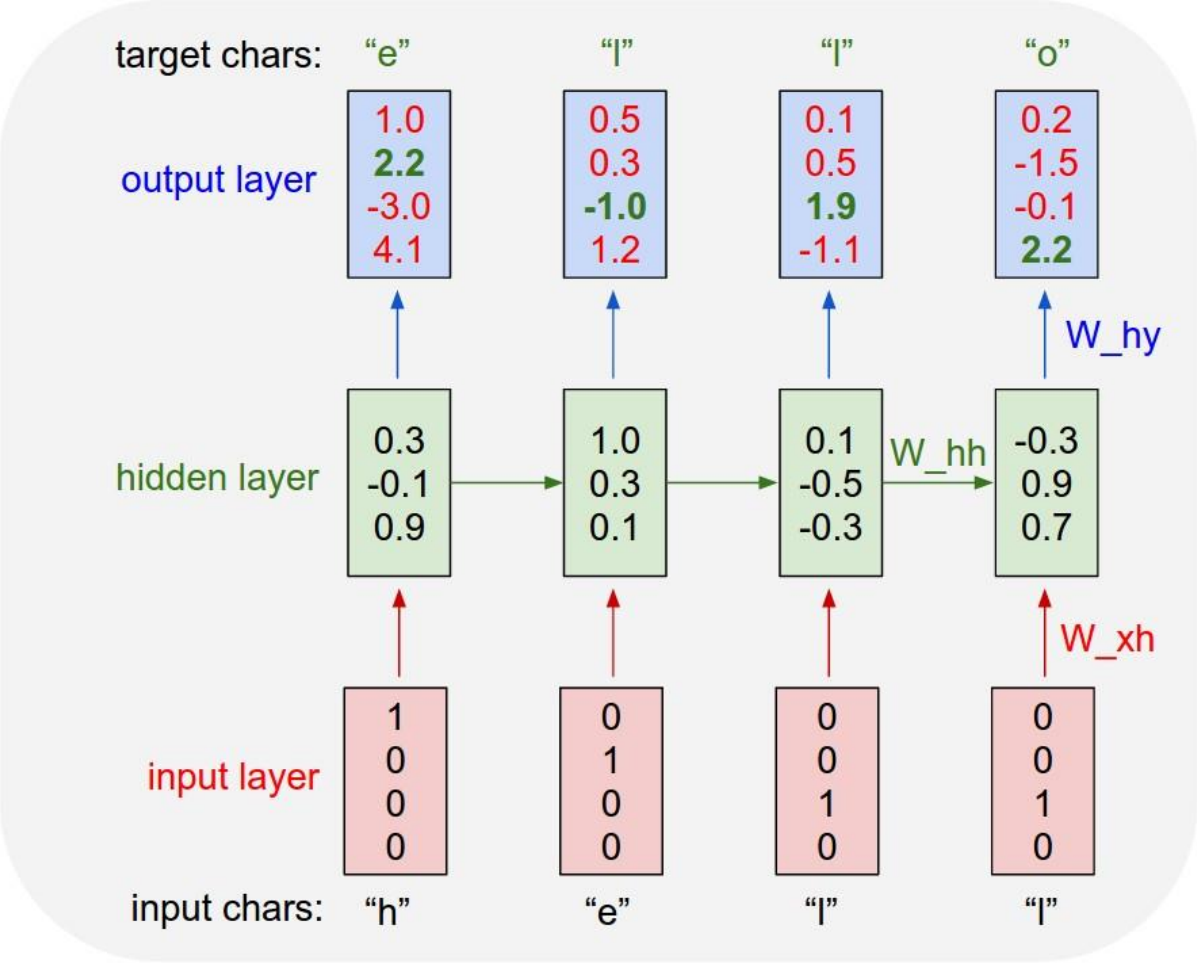
- Neural Network-based Language Model



Language Modeling

Mikolov et al. (2010)

- Recurrent Neural Network (RNN)-based Language Model
  - ✓ A simplified RNN structure for character-level language model



http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Language Modeling

- Recurrent Neural Network (RNN)-based Language Model
  - ✓ Character-level RNN vs Word-level RNN

Char-RNN

```
ESCALUS:
What is our honours, such a Richard story
Which you mark with bloody been Thilld we'll adverses:
That thou, Aurtructs a greques' great
Jmander may to save it not shif theseen my news
Clisters it take us?
Say the dulterout apy showd. They hance!

AnBESS OF GUCESTER:
Now, glarding far it prick me with this queen.
And if thou met were with revil, sir?

KATHW:
I must not my naturation disery,
And six nor's mighty wind, I fairs, if?

Messenger:
My lank, nobles arms;
```

https://github.com/hunkim/word-rnn-tensorflow

Word-RNN

```
LEONTES:
Why, my Irish time?
And argue in the lord; the man mad, must be deserved a spirit as drown the warlike Pray him, how seven in.

KING would be made that, methoughts I may married a Lord dishonour
Than thou that be mine kites and sinew for his honour
In reason prettily the sudden night upon all shalt bid him thus again. times than one from mine unaccustom'd sir.

LARTIUS:
O,'tis aediles, fight!
Farewell, it himself have saw.

SLY:
Now gods have their VINCENTIO:
Whipt fearing but first I know you you, hinder truths.

ANGELO:
This are entitle up my dearest state but deliver'd.

DUKE look dissolved: seemeth brands
That He being and
full of toad, they knew me to joy.
```

Language Modeling

Lee (2017)

- Recurrent Neural Network (RNN)-based Language Model
  - ✓ Character-level RNN (Korean)
    - 이광수 장편소설「무정」(총 323,660 음절, 1,680개 단어)
    - 특징: 1917년 작품이라 한자어가 많이 쓰였음, 큰따옴표와 줄바꿈을 포함한 대화체 문장이 많으며, 중고교생 대상으로 읽히는 작품이라 중간중간 괄호 속에 편집자 주석이 끼어 있음

형식은, 아뿔싸! 내가 어찌하여 이러한 생각을 하는가, 내 마음이 이렇게 약하던가 하면서 두 주먹을 불끈 쥐고 전신에 힘을 주어 이러한 약한 생각을 떼어 버리려 하나, 가슴속에는 이상하게 불길이 확확 일어난다. 이때에,
"미스터 리, 어디로 가는가" 하는 소리에 깜짝 놀라 고개를 들었다. (중략) 형식은 얼마큼 마음에 수치한 생각이 나서 고개를 돌리며, "아직 그런 말에 익숙지를 못해서....." 하고 말끝을 못 맺는다.
"대관절 어디로 가는 길인가? 급지 않거든 점심이나 하세그려."
"점심은 먹었는걸."
"그러면 맥주나 한잔 먹지."
"내가 술을 먹는가."
(중략)
"요— 오메데토오(아— 축하하네). 이이나즈케(약혼한 사람)가 있나 보네그려. 음 나루호도(그러려니). 그러구도 내게는 아무 말도 없단 말이야. 에, 여보게" 하고 손을 후려친다.

Language Modeling

Lee (2017)

Iter 0 :
랫萬게좁뉘쁠름끈玄른작밭裸觀갈나맡文플조바늠헝伍下잊볕홀툽뤘혈調記운피悲렙司狼독벗칼둡걷착날完잣老엇낫業4改'촉수릎낯깽잊쯤죽道넌友련친씌았융타雲채發造거크휘탁亨律與命텐암먼헝평琵헤落유 리벤産이馨텐

Iter 1300 : 를 옷 사가 려만다밤 말어변 대니 심로 려이, 순 과 이을 죄사글를 . 사람을 영채와 이니아베을 니러, 다가 달고 면 를 아잘 하 기 성구을 을 실튿으루 아잠 고 이 그와 매못 더 (띄어쓰기)

Iter 4900 : 를 왔다내 루방덩이종 은 얼에는 집어흔영채는 아무 우선을 에서가며 건들하아버전는 애양을 자에 운 모양이 랐다. 은 한다선과 '마는 .식세식가들어 ,
형식다
"내었다.있이 문 (줄바꿈)

Iter 100000 : 면서 치현분들더 중 한통 선교잤다.
"처럼 우셨다시가..... 것이 말사도? 여자려겠습니다" 하는 마음(裸生)은 이런 적드렸**다.** 그 말이 얼굴이 딸로 나고 얼굴이 마음불 하고 따라 선

Iter 300000 : 씻었다. 선형은 형식의 형식은 빛이 가슴을 오고 걸현감에는 일이 는 눈과 의고 아이얗어 알으로 자기의 구원을 내어려가 여러 짓을 쾌처게 안아 말고였는 악한 순간에 속으로 두 학교에

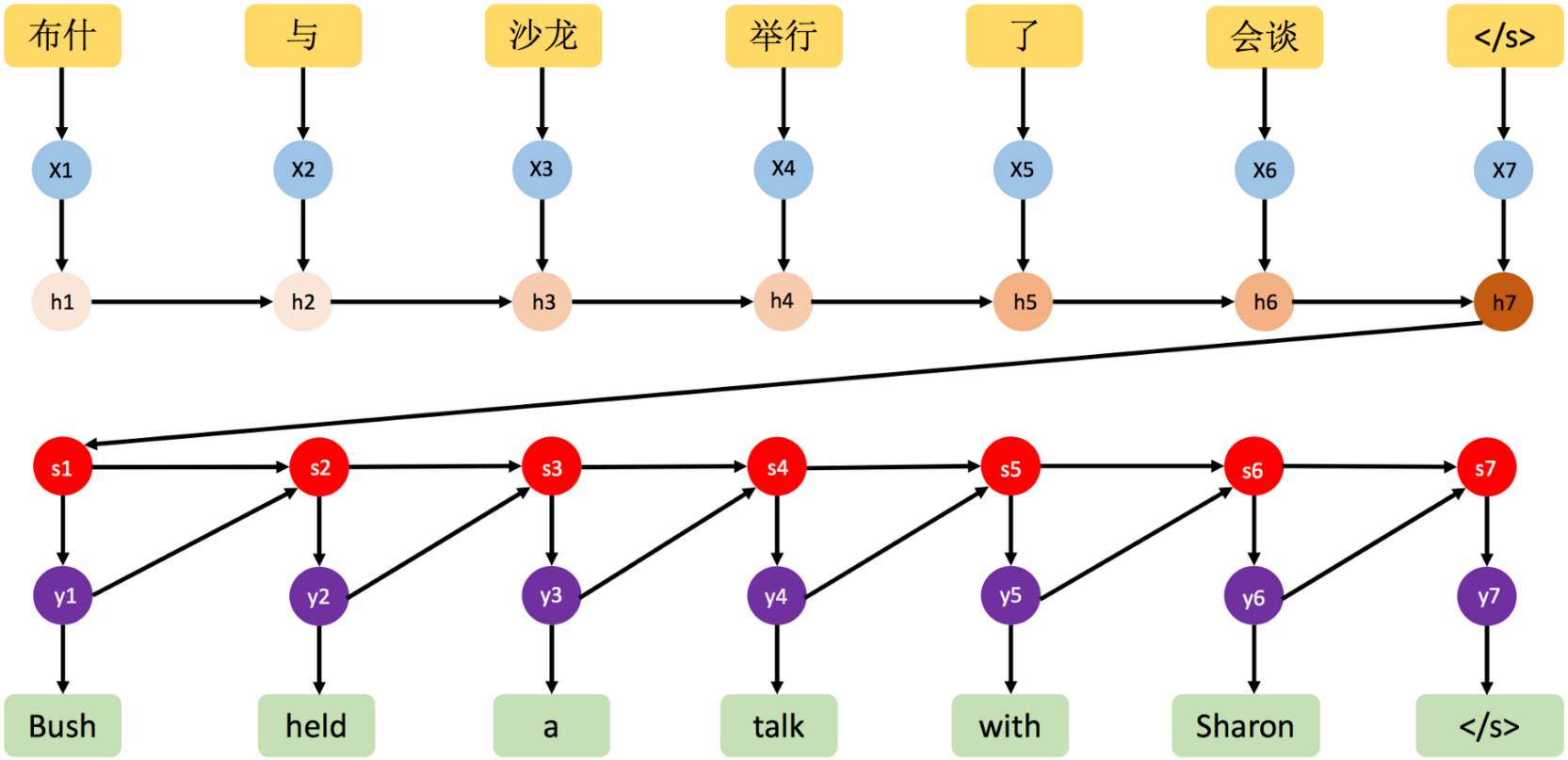
Iter 500000 : 본다. 성학과 평양으로 새로도 처음의 타던 공격하였다. '영채의 꽁인의 생각을 하면 때에 기생의 이는 것 보더니 나는 듯이 제인 소세건과 영채의 모양이를 대하였다. 형식을 생각하여

Iter 750000 : 으로 유안하였다. 더할까 하는 세상이 솔이요, 알고 게식도 들어울는 듯하였다. 태에그려 깔깔고 웃는 듯이 흔반다. 우선형은 사람을 어려보낸다.
"그려가?" (간접 인용)
한다. 영채는 손을 기쁘

Iter 1000000 : 에 돌내면서,
"여러 넣어오습데다. 그 말대 아무도 좀 집림과 시오 백매, 저는 열녀더러, 기런 소년이가 아니라."
"어리지요."
노파도 놀라며,
"저희마다가 말없습니까."
"아니 (대화체)

Language Modeling

- Sequence to Sequence (Seq2Seq) Learning



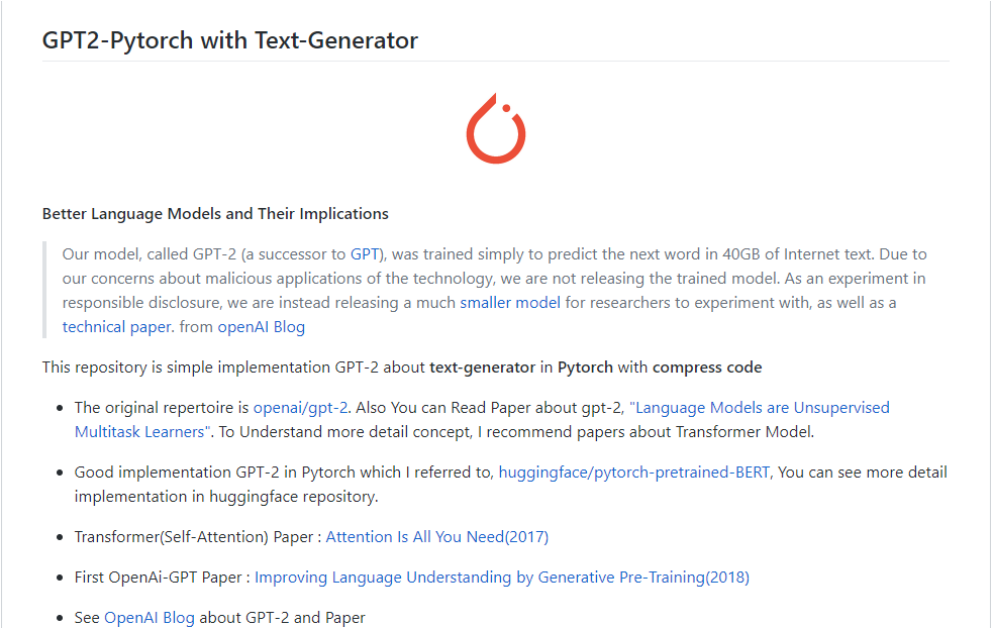
(Sutskever et al., 2014)

https://medium.com/@Synced/history-and-frontier-of-the-neural-machine-translation-dc981d25422d

Language Modeling

- Performance Improvements
  - ✓ GPT-2 (Open AI): Too Good to open the source code??

Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper.



GPT2-Pytorch with Text-Generator

Better Language Models and Their Implications

> Our model, called GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text. Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper. from openAI Blog

This repository is simple implementation GPT-2 about **text-generator** in **Pytorch** with **compress code**

- The original repertoire is openai/gpt-2. Also You can Read Paper about gpt-2, "Language Models are Unsupervised Multitask Learners". To Understand more detail concept, I recommend papers about Transformer Model.
- Good implementation GPT-2 in Pytorch which I referred to, huggingface/pytorch-pretrained-BERT, You can see more detail implementation in huggingface repository.
- Transformer(Self-Attention) Paper : Attention Is All You Need(2017)
- First OpenAi-GPT Paper : Improving Language Understanding by Generative Pre-Training(2018)
- See OpenAI Blog about GPT-2 and Paper

https://github.com/graykode/gpt-2-Pytorch?fbclid=IwAR0mHAR1cEPpuJ7QT9TBir-37_32tkQnpMjvsu57qPacEZz2-zbTm3Iibj8

Language Modeling

- Performance Improvements
  - ✓ GPT-2 (Open AI): Too Good to open the source code??

System prompt (human-written)

```
In a shocking finding, scientist discovered a herd of unicorns living in a remote,
previously unexplored valley, in the Andes Mountains. Even more surprising to the
researchers was the fact that the unicorns spoke perfect English.
```

Model completion (machine-written, 10 tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."


ANY
questions?