Text Analytics

Ch1&2

서수원

Business Intelligence Lab.

산업경영공학과, 명지대학교

01

Text Analytics Introduction part1

Background

- Background
  - 새로 생성되는 데이터의 80%이상이 비정형 데이터이다.
  - 이전과 같이 단순히 검색 하는 것에서, 지금은 새로운 지식을 탐구 하는 것이 매우 중요해졌다.

Topic: Storage

Follow via:

Within two years, 80% of all medical data will be unstructured

Summary: The data storage load for medical organizations is going to skyrocket.

By Denise Amrich for ZDNet Health | April 9, 2013 -- 01:09 GMT (18:09 PDT)

Follow @denisearnrich

We are all aware of the growth in medical health records because of changes in regulation. You would think, then, that most of the storage requirements would be for managing and storing those health records.

But the following IBM video states that 80 percent of health data stored in 2015 will be unstructured data, data that doesn't fit into nice rows and columns. Since nearly all medical records subject to the Affordable Care Act are, essentially, form data, that means that a huge amount of the storage required by the medical world isn't a result of that particular set of regulations.

In fact, IBM says that the body of medical knowledge doubles every five years. As imaging improves and M2M grows, more mobile data is captured. More intelligent sensors are deployed, both to home and hospital-bound patients. The data storage load for medical organizations is simply going to skyrocket.

This, of course, leads to another huge problem: Confidentiality and security. The Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health Act (HITECH) both require record-keeping safeguards. As medical data explodes in volume, that data protection challenge gets greater and greater.

The following video from IBM explores some of the scope of the problem. It's a marketing video, so keep that in mind, but it gives you a good understanding of the sorts of challenges and solutions that we'll be looking at for medical data storage.

http://www.zdnet.com/within-two-years-80-percent-of-medical-data-will-be-unstructured-7000013707/

How to manage unstructured data for business benefit

Stephen Pritchard, Contributor

Businesses across all industries are gathering and storing more and more data on a daily basis. But when it comes to assessing the benefits and challenges of big data, sometimes it is easy to overlook one key point: Most of the business information in use today does not reside in a standard relational database.

An often-cited statistic is that 80% of business data is unstructured, be it in word processor, spreadsheet and PowerPoint files, audio, video, sensor and log data, or external data such as social media feeds.

Exploiting unstructured data

This raises some issues for organisations that need to process and exploit unstructured data. Some big data tools, primarily those based on Hadoop, are designed from the ground up to manage and analyse unstructured information. Other, more conventional business intelligence (BI) and data warehousing technologies may not be.

Business intelligence (BI) and data warehousing suppliers have been adding support for unstructured data management to their tool sets, and some IT organisations have built their own platforms for converting unstructured data into structured records, for example, through knowledge management systems. But that can be a time-consuming and expensive process.

http://www.computerweekly.com/feature/How-to-manage-unstructured-data-for-business-benefit

Background

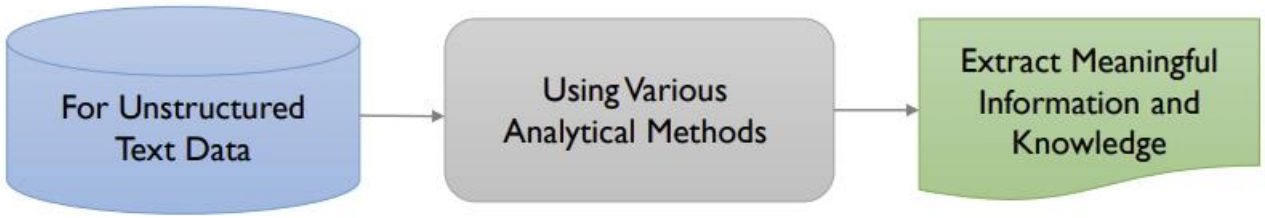
- arXiv
  - 수학, 물리학, 천문학 등의 논문을 수집하는 사이트이다.
  - 위 사이트에 의하면 새로 생성되는 논문의 수는 연간 약 3000개가 넘는다.

- 기계 학습은 지식기반 추론을 능가한다.

- 신경망에 대한 연구가 많아졌다.
  - 1990년대 까지는 신경망에 대한 논문을 받아주지도 않는 경우가 많았다.
  - 2000년대 들어서 신경망에 대한 논문의 등재수가 폭발적으로 증가 하였다.

Definition

- Definition of Text Analytics
  - 텍스트 데이터를 가지고 의미있는 정보나 지혜를 추출 한다.



Applications

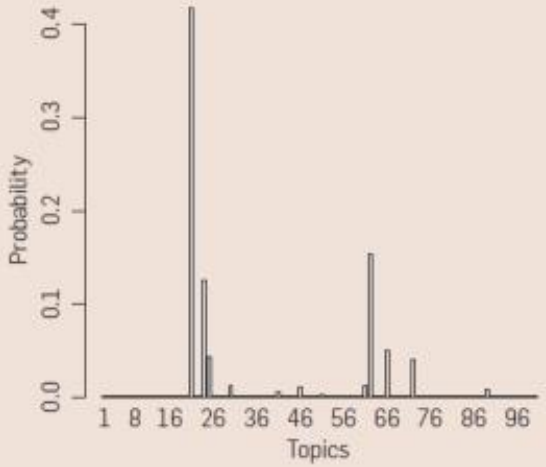
- Applications of Text Analytics
  - 정보 요약에 용의하다.
  - 정보 시각화에 매우 좋다.
  - 정보 추출에도 좋다.
    - ✓ Document clustering, Topic Extraction, Document Categorization/Classification, Recommendation등 여러 분야에 사용이 가능하다.

Applications

Applications

- Topic Extraction
  - ✓ Analyze documents and extract latent topics in the corpus

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Applications

Lee et al. (201

Document Categorization/Classification

✓ Sentiment Analysis

Player	Line	Score	승	패배	피안타	피홈런	볼넷	실점	평균 자책점	이닝당 삼진	이닝당 볼넷	삼진 /볼넷	방어율
웨버	NC 웨버는 7이닝 동안 1안타 3볼넷 10탈삼진으로 무실점 호투했다.	0.991	1	0	0	0	1	0	0	1	1	1	0
이재학	이재학 '10승', 3년만에 달라진 위상 증명하다	0.986	1	0	0	1	0	1	0	0	0	1	0
이태양	'이태양 완벽투' NC, LG 상대로 창단 첫 스윕	0.966	1	0	0	0	0	0	0	0	0	1	0
에릭	한편, 이날 7이닝 3실점으로 잘 던지며 7경기 만에 국내 첫 승을 거둔 에릭은 "굉장히 흥분된다.	0.803	1	0	1	0	0	0	0	0	0	0	0
찰리	어이없는 실책 2개가 나오자 찰리는 흔들리기 시작했다.	0.506	0	1	1	1	0	1	0	1	0	1	1
찰리	찰리가 대량 실점을 했지만 그 과정에서 3루수 모창민과 1루수 테임즈의 실책이 동반되면서 자책은 1점 밖에 되지 않았다.	0.506	0	1	1	1	0	1	0	1	0	1	1
원종현	NC 원종현이 패전투수가 됐다.	0.293	0	1	0	0	0	1	1	1	0	1	1
원종현	세 번째 투수로 등판한 원종현 역시 1사 후 8번 김성현에게 솔로포를 맞았다.	0.102	0	0	0	1	1	0	1	1	1	0	0
이재학	NC 선발 이재학은 8이닝 8피안타(2홈런) 5탈삼진 3볼넷 2실점을 기록했다.	0.079	0	0	1	1	0	1	1	0	0	0	1
이혜천	이혜천이 올라온 뒤 한꺼번에 5점을 내주면서 맥빠진 경기가 되고 말았다.	0.066	0	0	1	1	0	1	0	0	1	0	0

Challenges

- Challenges of Text Analytics
  - 차원수가 너무 크다.
    - ✓ 한국어의 경우 단어의 수가 약 100만개를 넘어 가는데, 이는 차원수가 100만 이상임을 의미하고, 이는 인공지능이 제대로된 분석을 하는데 상당한 어려움이 존재하게 한다.
  - 문장의 의미 파악이 어렵다.
    - ✓ “장명준은 오버워치를 하다 지도교수에게 들켰다.” “지도교수는 장명준 학생이 오버워치 하는 것을 목격했다.” 사람이 봤을땐 같은 의미의 문장 이지만, 인공지능은 전혀 다른 문장으로 인식하게 된다.
  - 중의어가 존재한다.
    - ✓ APPLE을 회사명으로 인식할지, 과일 이름으로 인식할지 혼란이 온다.

A simplified process of Text Analytics

- A simplified process of Text Analytics
  - Step1.어떤 데이터를 모을지 정한다.
  - Stpe2. 데이터를 전처리 한다.
  - Step3. 특징을 추출한다.
  - Stpe4. 알고리즘을 이용하여 결과를 내고 평가한다.

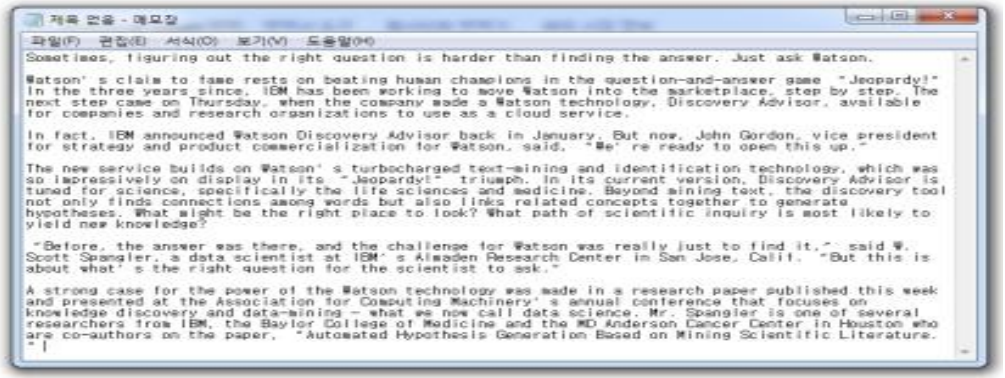
02

Text Analytics Introduction part2

Text preprocessing Level 0 : Text

- Text preprocessing Level 0 : Text
  - 목적에 맞는 소스를 이용 해야한다.(논문, 기사,책)
  - 크롤링&웹 스크랩핑을 통해 얻은 정보중 불필요한 정보를 제거 해야한다.
  - 중요한 정보를 가지고 있는 메타 데이터는 제거하지 않고, 불필요한 정보만 제거 하는것이 핵심이다.

- Remove figures, advertisements, html syntax, hyperlinks, etc.



Text preprocessing Level 1 : Sentence

- Text preprocessing Level 1 : Sentence
  - 올바르게 문장 단위로 나누는 것도 매우 중요하다.
    - ✓ 요약 시스템의 경우 올바르게 문장을 나누는게 핵심이다.
    - ✓ 분류기 별로 성능 및 문장 나누는 능력이 매우 상이하다.
      - ❖ 서로 다른 종류의 분류기를 같이 사용 해볼것을 추천한다.

Mark Sentence Boundaries

Detects sentence units. Easy case:

- often, sentences end with ".", "!", or "?"

Hard (or annoying) cases:

- difficult when a "." do not indicate an EOS:
  "MR. X", "3.14", "Y Corp.", ...
- we can detect common abbreviations ("U.S."), but what if a sentence ends with one?
  "...announced today by the U.S. The..
- Sentences can be *nested* (e.g., within quotes)

강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다. 회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 넥넥버거 골목길로 쭉 올라가야 하는데 다들 넥넥버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습. 강남 토끼정은 4층 건물 독채로 이루어져 있습니다. 역시 토끼정 본 점 답죠? ㅎㅅㅎ 건물은 크지만 간판이 없기 때문에 지나칠 수 있으니 조심하세요 강남 토끼정의 내부 인테리어. 평일 저녁이었지만 강남역 맛집 답게 사람들이 많았어요. 전체적으로 편안하고 아늑한 공간으로 꾸며져 있었습니다ㅎㅎ 한 가지 아쉬웠던 건 조명이 너무 어두워 눈이 침침했던.. 저희는 3층에 자리를 잡고 음식을 주문했습니다. 총 5명이서 먹고 싶은 음식 하나씩 골라 다양하게 주문했어요 첫 번째 준비된 메뉴는 토끼정 고로케와 깻잎 불고기 사라다를 듬뿍 올려 먹는 맛있는 밥입니다. 여러가지 메뉴를 한 번에 시키면 준비되는 메뉴부터 가져다 주더라구요. 토끼정 고로케 금방 튀겨져 나와 겉은 바삭하고 속은 촉촉해 맛있었어요! 깻잎 불고기 사라다는 불고기, 양배추, 버섯을 볶아 깻잎을 듬뿍 올리고 우엉 튀김을 곁들여 밥이랑 함께 먹는 메뉴입니다. 사실 전 고기를 안 먹어서 무슨 맛인지 모르겠지만.. 다들 엄청 잘 드셨습니다ㅋㅋ 어건 제가 시킨 촉촉한 고로케와 크림스튜우동. 강남 토끼정에서 먹은 음식 중에 이게 제일 맛있었어요!!! 크림소스를 원래 좋아하기도 하지만, 느끼하지 않게 부드럽고 담담한 스튜와 짭짤한 우동면이 너무 잘 어울려 계속 손이 가더라구요. 사진을 보니 또 먹고 싶습니다 간사이 풍 연어 지라시입니다. 일본 간사이 지방에서 많이 먹는 떠먹는 초밥(지라시스시)이라고 하네요. 밑에 와사비 마요밥 위에 연어들이 담겨져 있어 코끝이 찡할 수 있다고 적혀 있는데, 난 와사비 맛1도 모르겠던데..ㅋ 와사비를 안 좋아하는 저는 불행인지 다행인지 연어 지라시를 매우 맛있게 먹었습니다ㅋㅋㅋ 다음 메뉴는 달짝지근한 숯불 갈비 덮밥입니다! 간장 양념에 구운 숯불 갈비에 양파, 깻잎, 당갈 반숙을 터트려 비벼 먹으면 그 맛이 크..(물론 전 안 먹었지만..다른 분들이 그렇다고 하더라구요ㅋㅋㅋㅋㅋㅋㅋ) 마지막 메인 메뉴 양송이 크림수프와 숯불떡갈비 밥입니다. 크림리조또를 베이스로 위에 그루통과 숯불로 구운 떡갈비가 올라가 있어요! 크림스튜 우동 만큼이나 대박 맛있습니다... ㅠㅠㅠㅠㅠㅠ (크림 소스면 다 좋아하는 거 절대 아닙니다ㅋㅋㅋㅋㅋㅋ) 강남 토끼정 요리는 다 맛있지만 크림소스 요리를 참 잘하는 거 같네요 요건 몰만 마시기 아쉬워 사진 뉴자몽과 밀키소다 딸기통통! 유자와 자몽의 맛을 함께 느낄 수 있는 뉴자몽은 상큼함 그 자체였어요. 하지만 저는 딸기통통 밀키소다가 더 맛있었습니다ㅎㅎ 밀키소다는 토끼정에서만 만나볼 수 있는 메뉴라고 하니 한 번 드셔보시길 추천할게요!! 강남 토끼정은 강남역 맛집답게 모든 음식들이 대체적으로 맛있었어요! 건물 위치도 강남 대로변에서 조금 떨어져 있어 내부 인테리어처럼 아늑한 느낌도 있었구요ㅎㅎ 기회가 되면 다들 꼭 들러보세요~

순번	정답	자바	오픈소스 한국어 처리기	한나눔
1	강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다.	강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다.	강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다.	강남역 맛집으로 소문난 강남 토끼정에 다녀왔습니다 .
2	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 넥넥버거 골목길로 쭉 올라가야 하는데 다들 넥넥버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 넥넥버거 골목길로 쭉 올라가야 하는데 다들 넥넥버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습.	회사 동료 분들과 다녀왔는데 분위기도 좋고 음식도 맛있었어요 다만, 강남 토끼정이 강남 넥넥버거 골목길로 쭉 올라가야 하는데 다들 넥넥버거의 유혹에 넘어갈 뻔 했답니다 강남역 맛집 토끼정의 외부 모습 .
3	다만, 강남 토끼정이 강남 넥넥버거 골목길로 쭉 올라가야 하는데 다들 넥넥버거의 유혹에 넘어갈 뻔 했답니다			
4	강남역 맛집 토끼정의 외부 모습.			
5	강남 토끼정은 4층 건물 독채로 이루어져 있습니다.	강남 토끼정은 4층 건물 독채로 이루어져 있습니다.	강남 토끼정은 4층 건물 독채로 이루어져 있습니다.	강남 토끼정은 4층 건물 독채로 이루어져 있습니다 .

Text preprocessing Level 2 : Token

- Text preprocessing Level 2 : Token

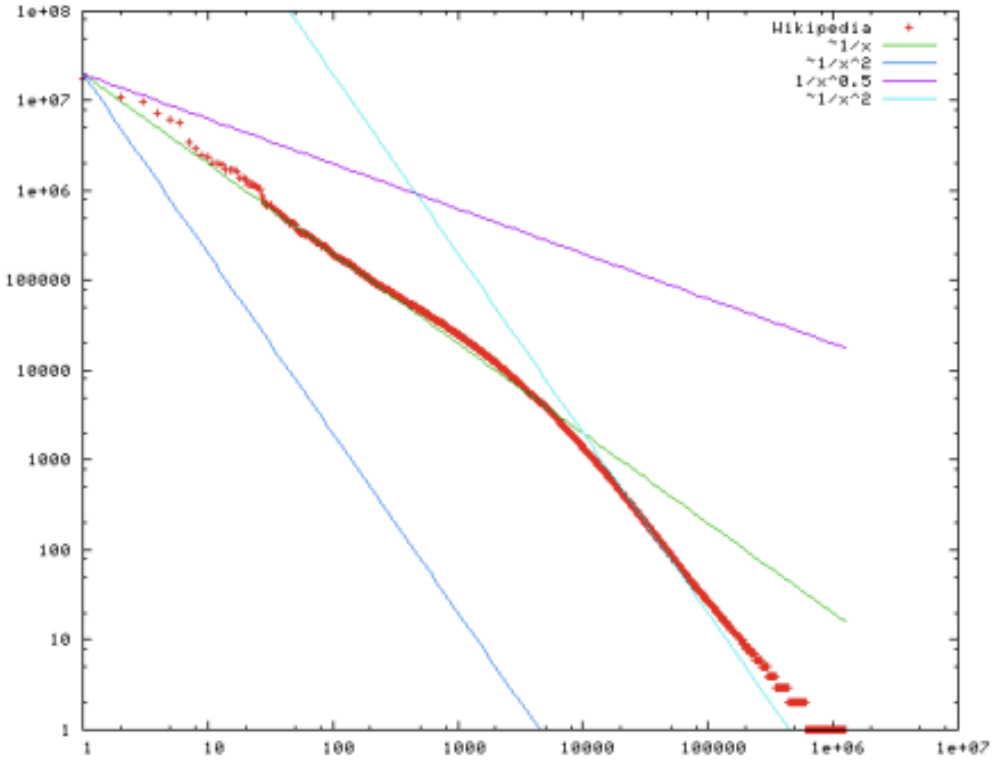
  의미있는 토큰을 추출 하는것을 때와 어렵다(단어 스잡 떠어쓰기)

100 common words in the Oxford English Corpus

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

http://en.wikipedia.org/wiki/Most_common_words_in_English

Word frequency distribution in Wikipedia



http://upload.wikimedia.org/wikipedia/commons/b/b9/Wikipedia-n-zipf.png

Text preprocessing Level 2 : Token

- Text preprocessing Level 2 : Token
  - Stop – words
    - ✓ 아무 의미 없는 단어를 뜻한다.
      - ❖ 습니다, 로써….. a an
      - ❖ Stop - words를 제거 하면 문법적으로는 의미가 없어지지만, 의미는 살아 있음으로 차원수는 줄이고 의미는 가져올수 있다.
  - 차원을 줄인다
    - ✓ 품사, 시제가 달라질 경우, 같은 의미의 단어는 같은 형태로 구성해야 한다.
    - ✓ Stemming과 Lemmatization이 있다.
      - ❖ Stemming : 서로 다른 형태의 단어들을 정규형태(가장 공통된 음절, 알파벳)로 변형 한다.
      - ❖ Lemmatization: 단어의 품사를 보존 하면서 원형을 찾는다.

Text preprocessing Level 2 : Token

[Original text]

Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region.

http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf

[After removing stop words]

Information Systems Asia Web provides research IS-related commercial materials interaction research sponsorship interested corporations focus Asia Pacific region

Word	Stemming	Lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

Text Transformation

- Transformation
  - 문서를 어떻게 연속형의 숫자 벡터로 표현 할 것인가를 생각한다.
  - 많은 알고리즘은 연속형으로 변형 해줘야 사용 가능하다.
  - Bag – of – words: 고전적인 방식이다. 하나의 문서에 사용된 단어의 빈도를 가지고 벡터화 한다.

S1: Jon likes to watch movies. Mary likes too.

S2: John also likes to watch football game.

Word	S1	S2
John	1	1
Likes	2	1
To	1	1
Watch	1	1
Movies	1	0
Also	0	1
Football	0	1
Games	0	1
Mary	1	0
too	1	0

Text Transformation

- Transformation
  - Word weighting
    - ✓ TF-IDF
      - ❖ 단어 중요도에 가중치를 준다.
      - ❖ TF : 단어가 문서에 얼마나 반복적으로 사용이 되었는지에 대한 것이다.
      - ❖ IDF : 단어가 코퍼스 내 에서 얼마나 자주 사용 되었나에 대한 것이다.(the)

$$TF - IDF(w) = tf(w) \times \log\left(\frac{N}{df(w)}\right)$$

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Text Transformation

- Transformation
  - One-hot-vector representation
    - ✓ 단어가 등장하는 부분만 1의값으로 표현 한다.
    - ✓ 단점은 두 단어 사이에 유사성이 보존이 안된다는 점이다.
      - ❖ 벡터의 내적이 항상 0이 나오기 때문이다.

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

$$(w^{hotel})^{\top} w^{motel} = (w^{hotel})^{\top} w^{cat} = 0$$

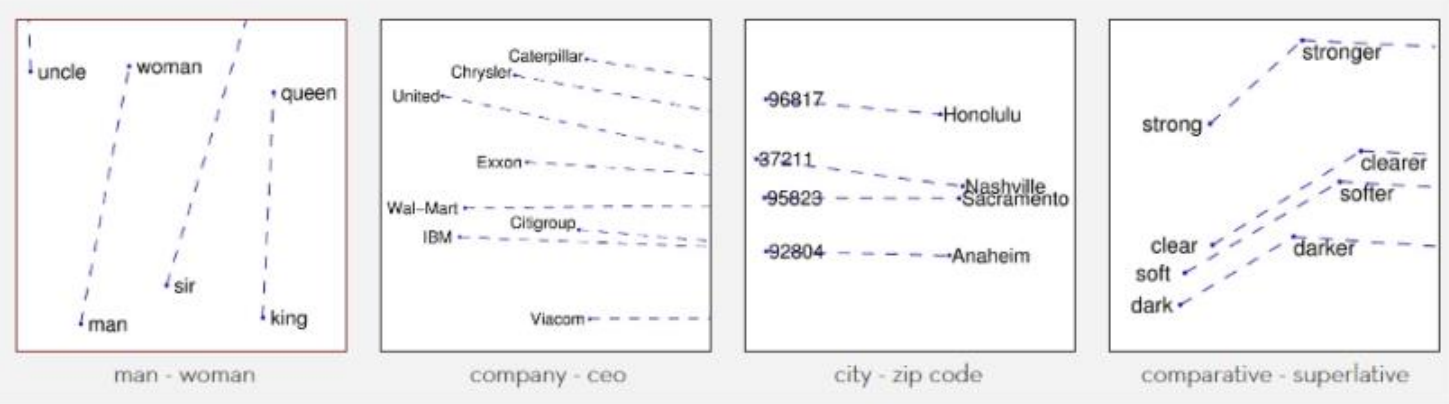
Text Transformation

- Transformation
  - Word vectors : distributed representation
    - ✓ 단어를 n차원의 실수 공간에 맵핑 해보자는 의미이다.

$$W : \text{words} \rightarrow \mathbb{R}^n$$

$$W(\text{“cat”}) = (0.2,\ \text{-}0.4,\ 0.7,\ \ldots)$$

$$W(\text{“mat”}) = (0.0,\ 0.6,\ \text{-}0.1,\ \ldots)$$



Text Transformation

- Pre-trained Word Model
  - 거대 모델을 활용해서 여러 연구에 적용이 가능하다.

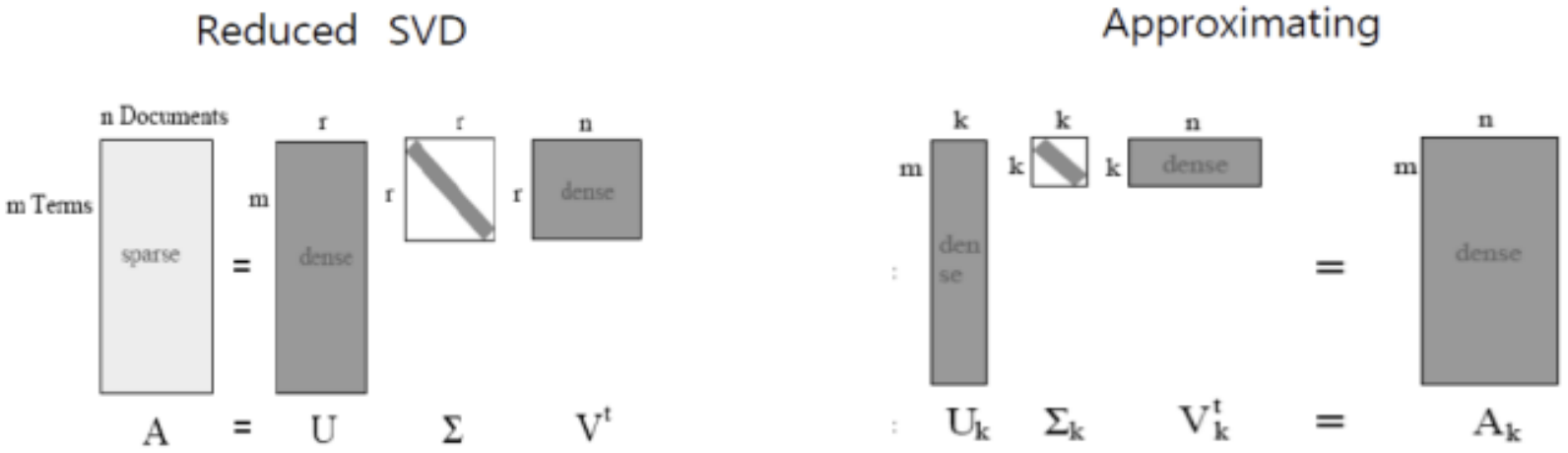
Feature Selection/Extraction

- Feature Subset selection
  - 특정 목적에 맞는 최적 변수 집합을 선택 하는 것을 의미한다.
    - ✓ 각각의 특징에 점수를 매길수 있다.

- Information gain: $\sum_{F=W,\overline{W}} P(F) \sum_{C=pos,neg} P(C|F) \log \frac{P(C|F)}{P(C)}$
- Cross-entropy: $P(W) \sum_{C=pos,neg} P(C|W) \log \frac{P(C|W)}{P(C)}$
- Mutual information: $\sum_{C=pos,neg} P(C) \log \frac{P(W|C)}{P(W)}$
- Weight of evidence: $\sum_{C=pos,neg} P(C)P(W) \left| \log \frac{P(C|W)(1-P(C))}{P(C)(1-P(C|W))} \right|$
- Odds ratio: $\log \frac{P(W|pos) \times (1-P(W|neg))}{(1-P(W|pos)) \times P(W|neg)}$
- Frequency: $Freq(W)$

Feature Selection/Extraction

- Feature subset extraction
  - 주어진 데이터로부터 새로운 변수를 추출한다.
  - 원래 데이터는 보존 하면서 최대한 적은양의 데이터셋을 구축 한다.
  - Extraction후에는 항상 차원이 축소되어야 한다.
  - Latent Semantic Analysis를 이용한다.
    - ✓ 잠재 의미 분석이다.
    - ✓ 쉽고 빠르게 단어의 잠재적인 의미를 끌어낼수 있다.



Feature Selection/Extraction

- Topic Modeling
  - 원래 목적은 코퍼스를 관통하는 주요 주제를 파악 하는 것이다.
    - ✓ 문서별로 주제의 비중을 파악한다.
    - ✓ 주제별로 단어의 비중을 파악한다.

(a) Per-document topic proportions (θ_d)

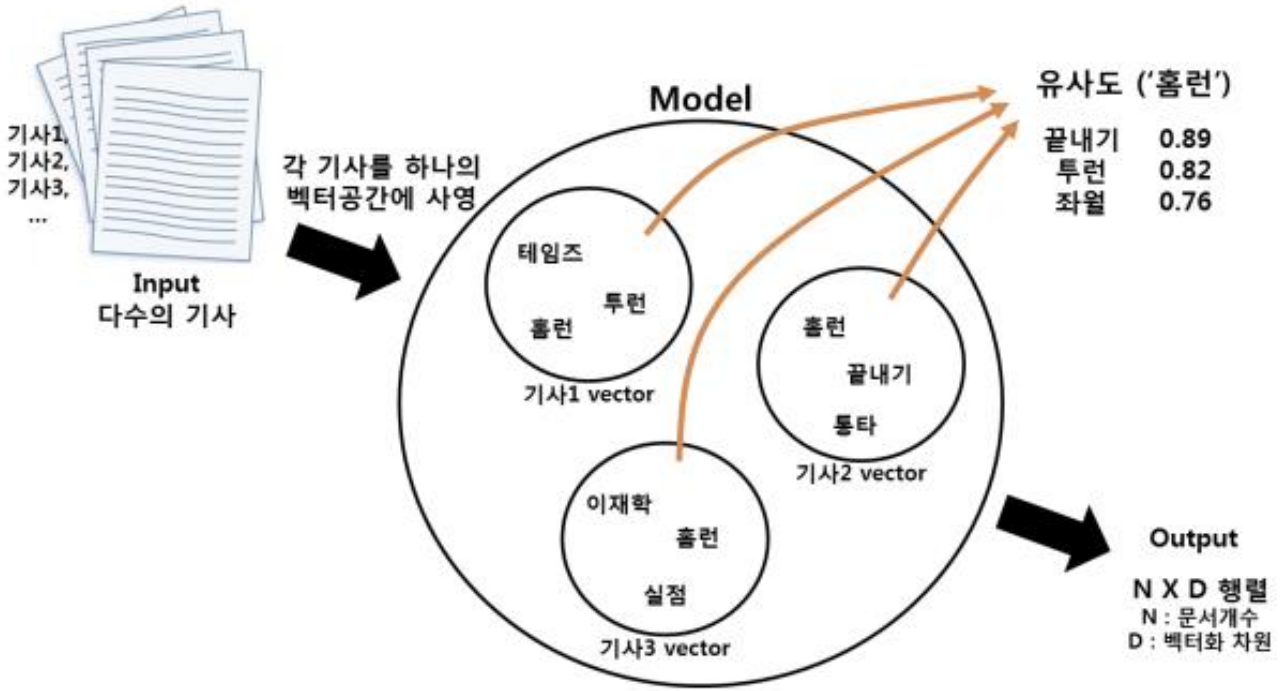
	Topic 1	Topic 2	Topic 3	…	Topic K	Sum
Doc 1	0.20	0.50	0.10	…	0.10	1
Doc 2	0.50	0.02	0.01	…	0.40	1
Doc 3	0.05	0.12	0.48	…	0.15	1
…	…	…	…	…	…	1
Doc N	0.14	0.25	0.33	…	0.14	1

(b) Per-topic word distributions (ϕ_k)

	Topic 1	Topic 2	Topic 3	…	Topic K
word 1	0.01	0.05	0.05	…	0.10
word 2	0.02	0.02	0.01	…	0.03
word 3	0.05	0.12	0.08	…	0.02
…	…	…	…	…	…
word V	0.04	0.01	0.03	…	0.07
Sum	1	1	1	1	1

Feature Selection/Extraction

- Document to vector
  - Word to vector의 확장판이다.
  - 단어 차원이 아닌 문서 차원에서 distributed representation 을 하는 방식이다.



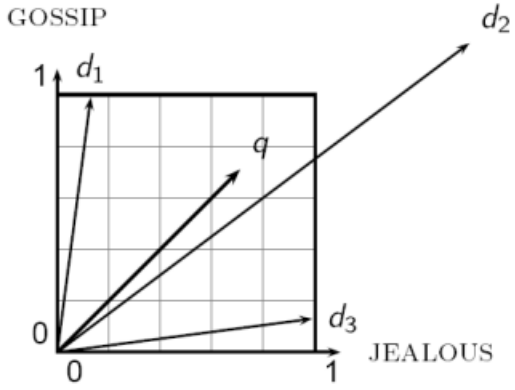
Learning & Evaluation

- Document similarity
  - 코사인 유사도를 사용한다.

▪ Which two documents are more similar?

Doc.	Word 1	Word 2	Word 3
Document 1	1	1	1
Document 2	3	3	3
Document 3	0	2	0

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$



Learning & Evaluation

- Learning Task
  - Classification
    - ✓ 자동으로 레이블, 카테고리를 예측 해준다.
      - ❖ 예시로는 spam filtering, sentiment analysis가 있다.
  - Clustering & Visualization
    - ✓ 문서내에서 중요한 토픽을 빠르게 파악하게 해준다.
    - ✓ 중요 키워드의 연관성 및 주제별 유사성을 시각화를 통해 볼수있다.
  - Information Extraction/Retrieval
    - ✓ 유용한 정보를 뽑아낸다.
    - ✓ QnA가 대표적인 예시이다.

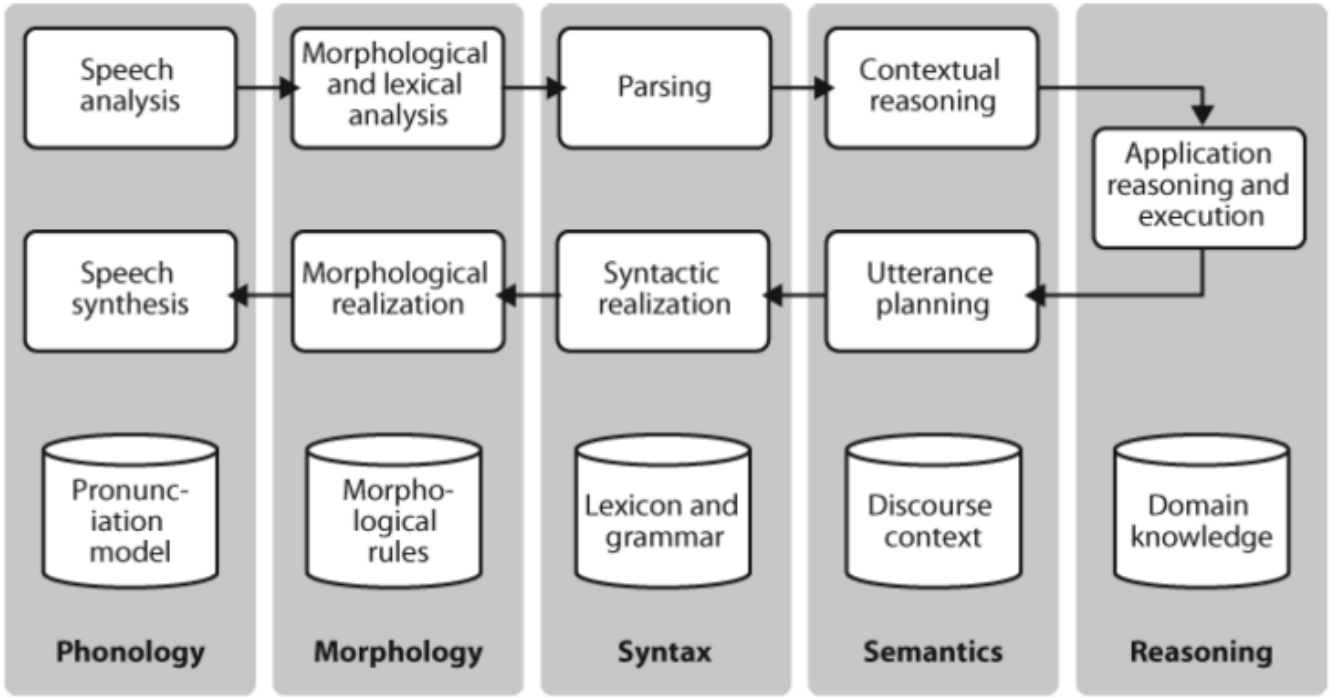
03

Text Preprocessing Part1

Natural Language Processing

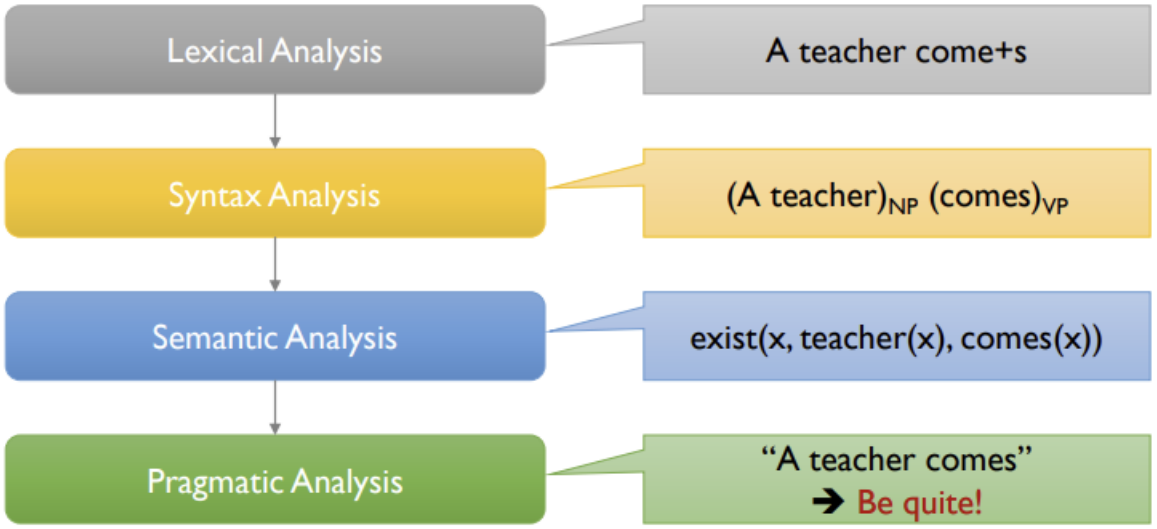
- Classical categorization of NLP
  - 음운론,형태론,구문론,의미론,논증 으로 구분 가능하다.

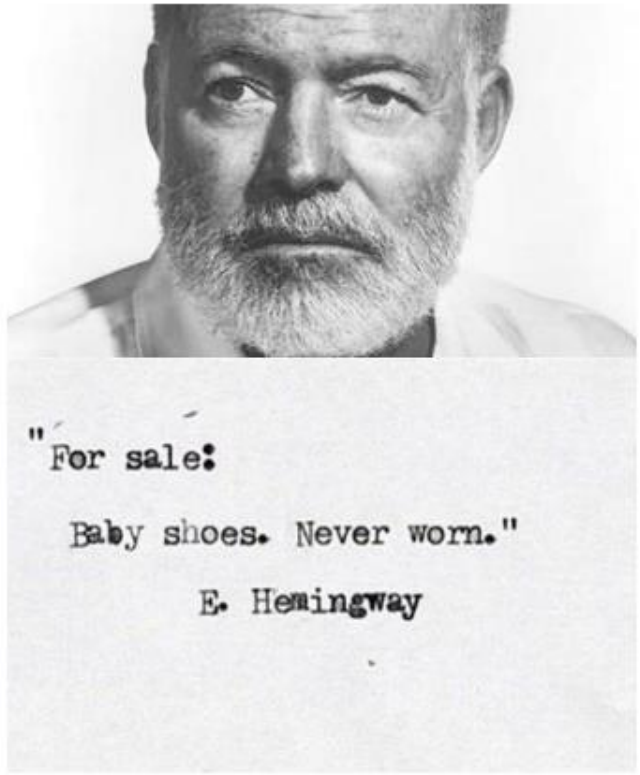
• Natural language processing sequence



Natural Language Processing

- Classical categorization of NLP
  - 음운론,형태론,구문론,의미론,논증 으로 구분 가능하다.





Natural Language Processing

- NLP가 어려운 점
  - 컴퓨터는 명확한 언어를 취급한다.
  - 현실의 단어는 모호하고, 의미가 불분명 한 경우가 많으며, 새로운 의미의 단어가 계속 생성되기 때문에 NLP가 어렵다.

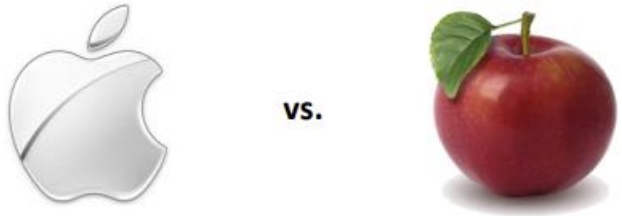
```
from __future__ import absolute_import
from __future__ import division
from __future__ import print_function

import re
import tensorflow as tf


def create_optimizer(loss, init_lr, num_train_steps, num_warmup_steps, use_tpu):
  """Creates an optimizer training op."""
  global_step = tf.train.get_or_create_global_step()

  learning_rate = tf.constant(value=init_lr, shape=[], dtype=tf.float32)

  # Implements linear decay of the learning rate.
  learning_rate = tf.train.polynomial_decay(
      learning_rate,
      global_step,
      num_train_steps,
      end_learning_rate=0.0,
      power=1.0,
      cycle=False)
```



Research Trends in NLP

- Research Trends in NLP
  - 고전적 방법 : Rule-based 접근법을 의미한다.
    - ✓ 융통성이 없고 신조어에 약하다.
  - 통계적 방법 : 히든 마르코프 모델, svm, clustering등의 기법을 사용하는 것을 의미한다.
    - ✓ Rule-based에 비해 좀더 유연해 졌지만, 코퍼스가 필요하다.
    - ✓ 현실에서는 두개의 방식을 사용 하는 것이 더 좋다.
  - 현재의 방법 : 신경망을 이용한 딥러닝적 접근법을 의미한다.
    - ✓ Pretrained 된 모델을 사용 하는 것이 최신 트랜드 라고 볼 수 있다.

Research Trends in NLP

- Research Trends in NLP

1. Universal Models
2. Massive Multi-task Learning
3. Beyond the Transformer
4. Prompting
5. Efficient Methods
6. Benchmarking
7. Conditional Image Generation
8. ML for Science
9. Program Synthesis
10. Bias
11. Retrieval Augmentation
12. Token-free Models
13. Temporal Adaptation
14. The Importance of Data
15. Meta-learning

04

Text Preprocessing Part2

Lexical Analysis

- Lexical Analysis
  - 어휘분석이다
  - 의미가 있는 토큰을 만들고, 형태소 태깅을 하고, 개체명인식 등을 하는 것이 Lexical Analysis의 목표이다.

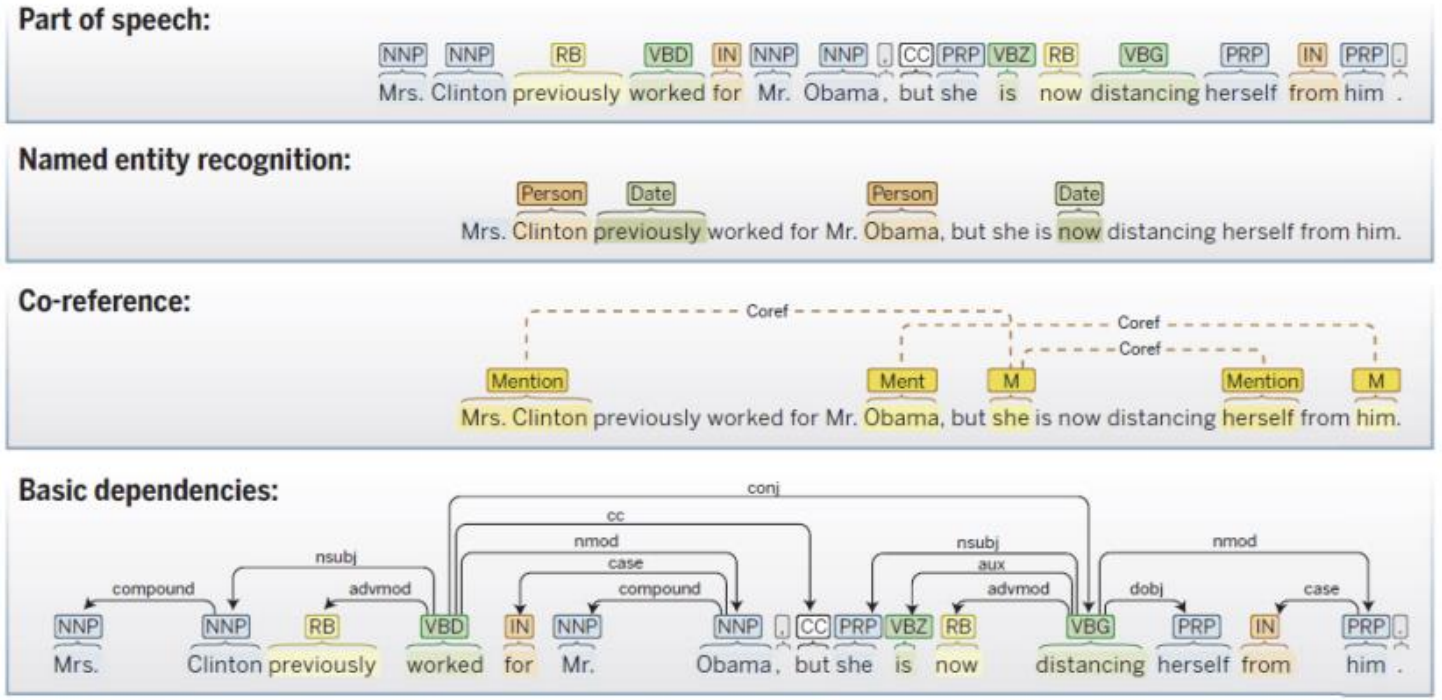


Fig. 1. Many language technology tools start by doing linguistic structure analysis. Here we show output from Stanford CoreNLP. As shown from top to bottom, this tool determines the parts of speech of each word, tags various words or phrases as semantic named entities of various sorts, determines which entity mentions co-refer to the same person or organization, and then works out the syntactic structure of each sentence, using a dependency grammar analysis.

Lexical Analysis 1 : Sentence Splitting

- Lexical Analysis 1 : Sentence Splitting
  - 어휘분석의 가장 기초가 되는 것은 문장 구분이다.
  - 토픽모델링 같은 일부 Text Mining기술에선 별로 안 중요 하지만, 대부분의 NLP에선 매우 중요한 과정이다.

Mark Sentence Boundaries

Detects sentence units. Easy case:

- often, sentences end with ".", "!", or "?"

Hard (or annoying) cases:

- difficult when a "." do not indicate an EOS:
  "MR. X", "3.14", "Y Corp.", ...
- we can detect common abbreviations ("U.S."), but what if a sentence ends with one?
  "...announced today by the U.S. The..."
- Sentences can be *nested* (e.g., within quotes)

Lexical Analysis 2 : Tokenization

- Lexical Analysis 2 :Tokenization
  - Text는 토큰화 작업을 거쳐야 한다.
  - 토근화 또한 매우 어렵다.

✓ What to do with hyphens?

- database vs. data-base vs. data base

✓ What to do with "C++","A/C",":-)","…","ㅋㅋㅋㅋㅋㅋㅋㅋㅋ"?

✓ Some languages do not use whitespace (e.g., Chinese)

2013年5月，习主席在视察成都战区时，郑重提出在适当时候召开全军政治工作会议，并明确提出到古田召开这次会议，以更好弘扬我党我军的光荣传统和优良作风。6月，总政治部向中央军委提交《关于筹备召开全军政治工作会议的请示》，提出要通过召开会议形成一个指导性文件。习主席随即批示同意，明确要求这个文件要充分体现深厚的历史积淀和政治意蕴，能够管一个时期，起到历史性作用。

Consistent tokenization is important for all later processing steps.

Lexical Analysis 3: Morphological Analysis

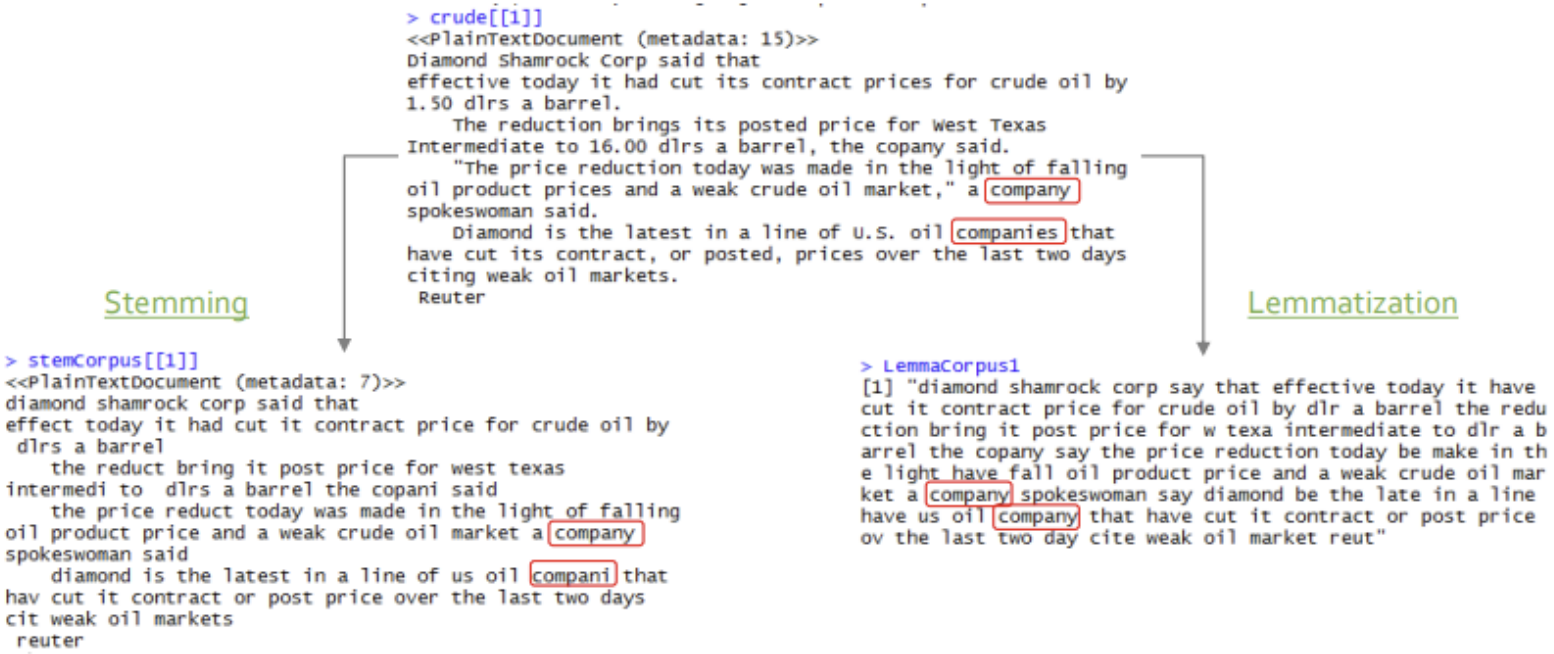
- Lexical Analysis 3 : Morphological Analysis
  - 형태를 분석한다.
  - Stemming 기법과 Lemmatization기법이 있다.

- Stemming vs. Lemmatization

Word	Stemming	Lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

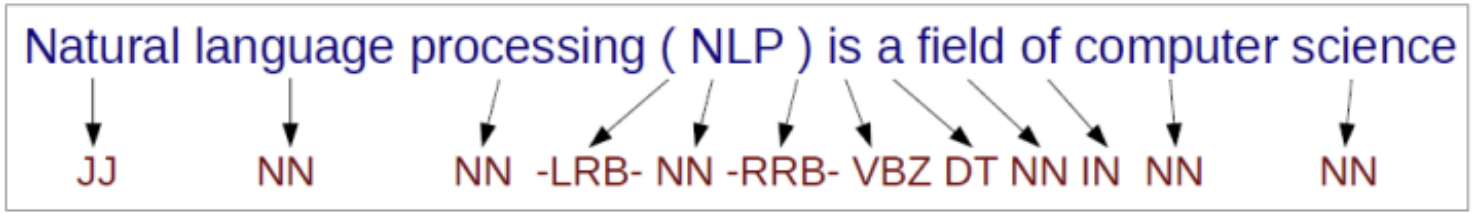
Lexical Analysis 3: Morphological Analysis

- Lexical Analysis 3 : Morphological Analysis
  - 형태를 분석한다.
  - Stemming 기법과 Lemmatization기법이 있다.



Lexical Analysis 4: Part-of-Speech(POS) Tagging

- Lexical Analysis 4 : POS Tagging
  - 형태소 분석을 의미한다.
  - 상황에 맞게 형태소를 잘 분석 해야 한다.
  - 문장이 주어지면(X) POS(Y)값을 나타내야 한다.
  - 같은 코퍼스 내 에서 정확도가 뛰어나다.
  - 알고리즘으론 Decision Trees, HMM,SVM등이 사용된다.



Fundamentals

POS-Tagging generally requires:

Training phase where a manually annotated corpus is processed by a machine learning algorithm; and a

Tagging algorithm that processes texts using learned parameters.

Performance is generally good (around 96%) when staying in the same domain.

Lexical Analysis 4: Part-of-Speech(POS) Tagging

- Lexical Analysis 4 : POS Tagging
  - Pointwise prediction : 각각의 단어를 분류기를 통해 분류한다(SVM, Maximum Entropy Model등)
  - Probabilistic models:
    - ✓ Generative sequence models : 가장 가능성 있는 태그를 순차적으로 가져온다.(HMM)
    - ✓ Discriminative sequence models : 한번에 모든 태그를 예측한다.(Conditional Random Filed)
  - Neural network-based models

Lexical Analysis 4: Part-of-Speech(POS) Tagging

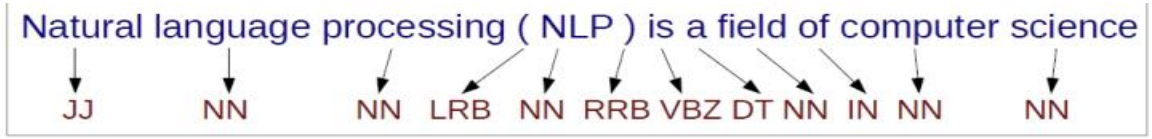
- Lexical Analysis 4 : POS Tagging
  - Maximum Entropy Model

$$p(t|C) = \frac{1}{Z(C)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(C,t)\right) \qquad p(t_1,...,t_n|w_1,...,w_n) \approx \prod_{i=1}^{n} p(t_i|w_i)$$

- f_i is a feature
- λi is a weight (large value implies informative features)
- Z(C) is a normalization constant ensuring a proper probability distribution
- Makes no independence assumption about the features

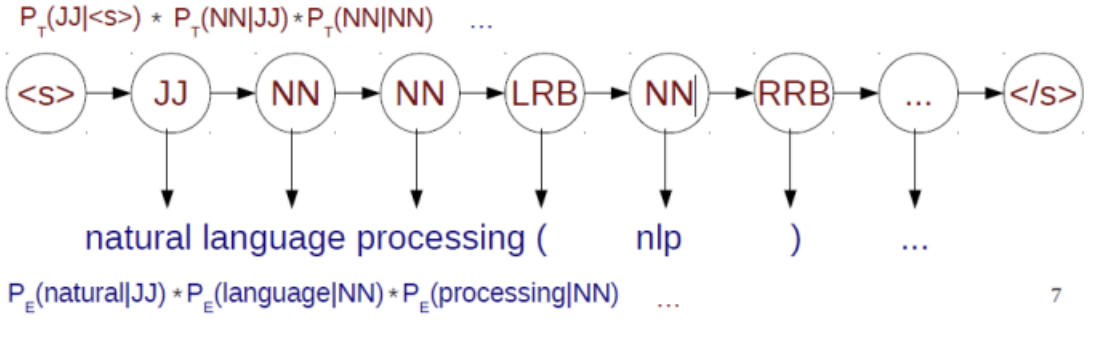
Lexical Analysis 4: Part-of-Speech(POS) Tagging

- Lexical Analysis 4 : POS Tagging
  - Probabilistic Model for Pos Tagging



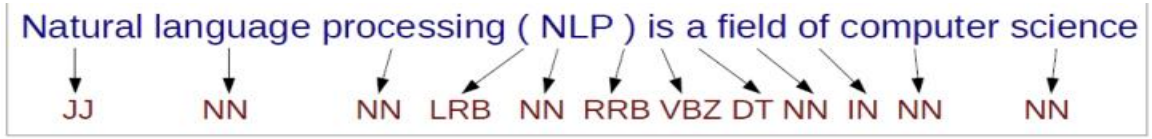
$$\underset{Y}{\operatorname{argmax}} P(Y|X)$$

✓ Generative Sequence Model



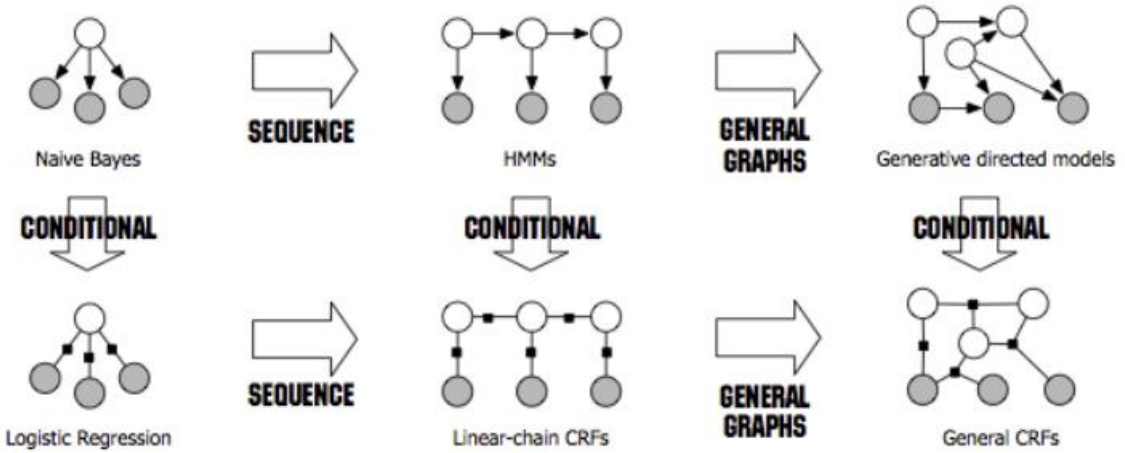
Lexical Analysis 4: Part-of-Speech(POS) Tagging

- Lexical Analysis 4 : POS Tagging
  - Probabilistic Model for Pos Tagging



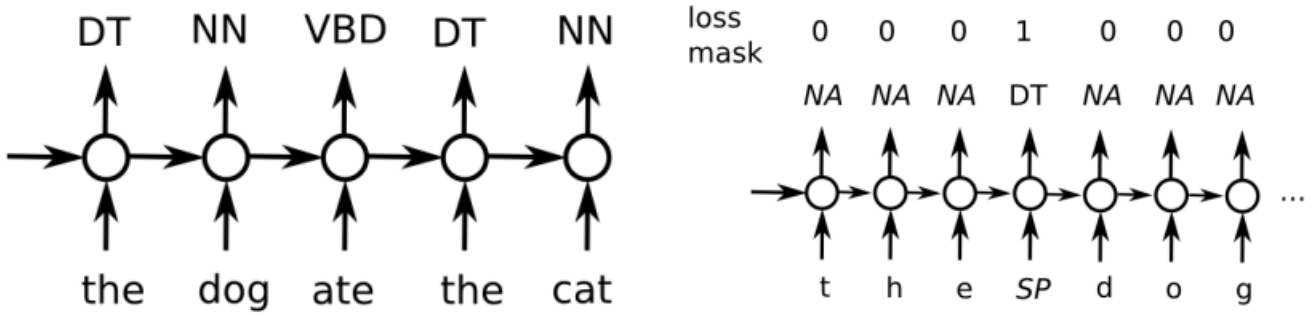
$$\underset{Y}{\operatorname{argmax}}\, P(Y|X)$$

✓ Discriminative Sequence Model : CRF
- ❖ POS에선 여전히 SOTA이다.



Lexical Analysis 4: Part-of-Speech(POS) Tagging

- Lexical Analysis 4 : POS Tagging
  - Neural network-based models : Recurrent neural networks
    - ✓ RNN의 순환구조를 활용한 방식이다.
    - ✓ 과거의 학습 가중치를 통해 현재의 학습에 반영하기도 하고, 시간에 종속 된다는 특징이 있다.



.

Lexical Analysis 5: Named Entity Recognition

- Lexical Analysis 5 : NER
  - 개체명을 인식하는 것을 의미한다.
  - Dictionary / Rule-based
    - ✓ List lookup : 먼저 만들어 놓은 리스트를 가지고 인식한다.
      - ❖ 장점 : 쉽고 빠르고 정확하다
      - ❖ 단점 : 애매한 단어에 대해 정확히 인식이 어렵다(APPLE)
  - Model-based
    - ✓ MITIE
    - ✓ CRF++

BERT for Multi NLP Tasks

- Google Transformer

.

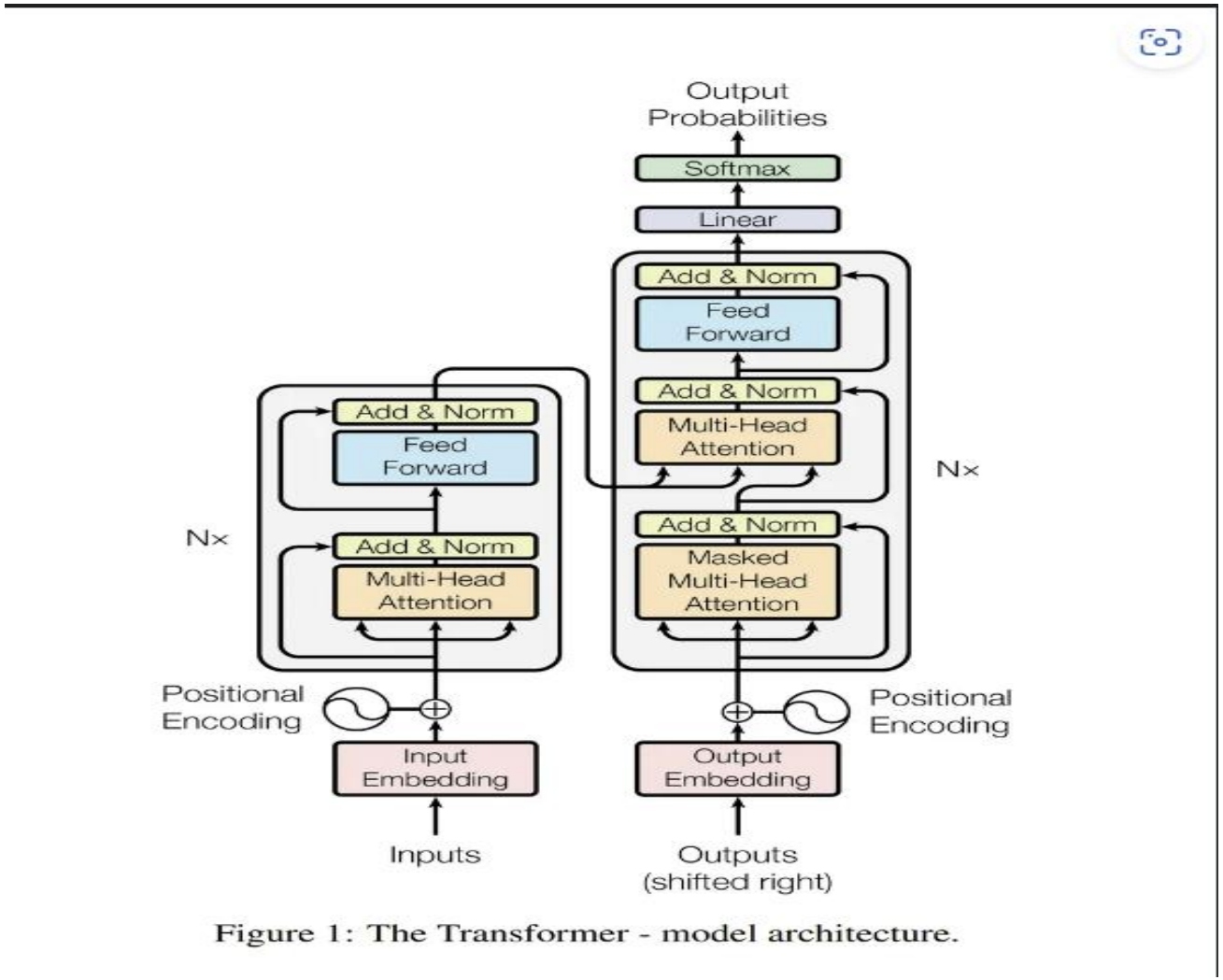


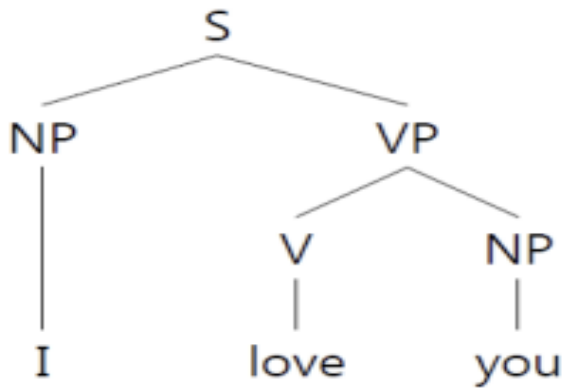
Figure 1: The Transformer - model architecture.

05

Text Preprocessing Part3

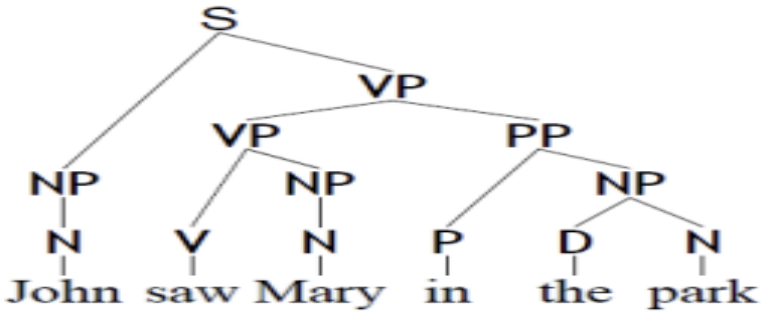
Syntax Analysis

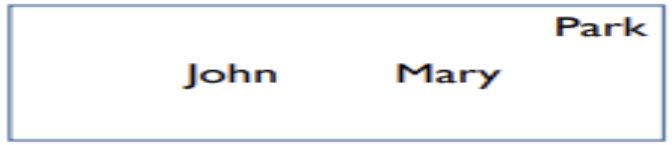
- Syntax Analysis
  - 문장 구조를 파악하기 위한 구문분석을 의미한다.
  - Parser
    - ✓ 입력 토큰에 내재된 자료구조를 빌드하고 문법을 검사한다.
    - ✓ 모든 파서는 두가지 특징이 있다.
      - ❖ Directionality : 구조물이 만들어 지는 순서에 관한 특징을 의미한다.(탑다운 바텀 업)
      - ❖ Search strategy : 어떻게 분석할 것인지에 대한 특징을 의미한다.(좌우, 하나씩 깊게)

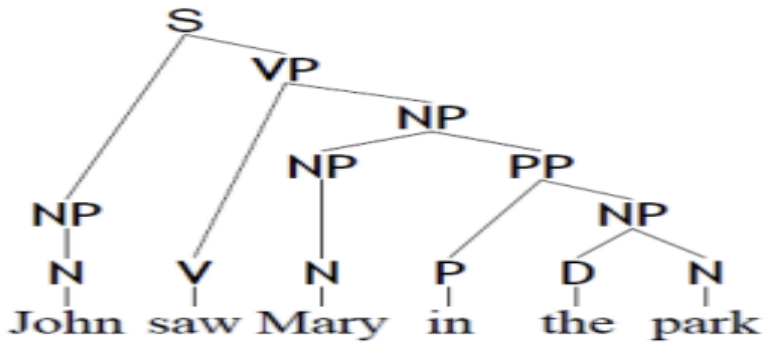


Syntax Analysis

- Syntax Analysis
  - Parser
    - ✓ Parsing Representation
      - ❖ 트리 vs 리스트 방식이 있다.
      - ❖ 어휘의 모호성 때문에 하나의 문장에도 여러 개의 파싱트리가 존재한다.
      - ❖ 구조적 모호성 또한 마찬가지로 여러 개의 파싱트리가 존재하게 된다.







Language Modeling

- Probabilistic Language Model
  - 문장에 확률을 넣어 문장을 계산한다. Ex) P(high wind)> P(large wind tonight)
    - ✓ 번역, 맞춤법 검사, 음성인식 ,요약 등에 활용 된다.
- Probabilistic Language Modeling
  - 문장의 단어들에 대한 확률을 계산한다. 앞뒤 문맥을 같이 확인 한다.

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

  - 마르코프 가정을 활용한 N그램 모델이 있다.

Language Modeling

- 딥러닝 기반 언어모델
  - Neural Network-based Language Model
  - Recurrent Neural Network –based Language Model
  - Sequence to Sequence Learning