

# **DAY3**

## **자연어처리의 이해**

# 강의 순서

- 자연어처리 기술의 이해
  - 자연어처리 기술 개요
  - 형태소분석과 품사태깅
  - 언어데이터 구축 방법
  - 구문분석, 패턴매칭, 의미 분석
- 자연어처리 시스템 사례
  - 소셜미디어 분석
- 실습

자연어처리 기술의 이해

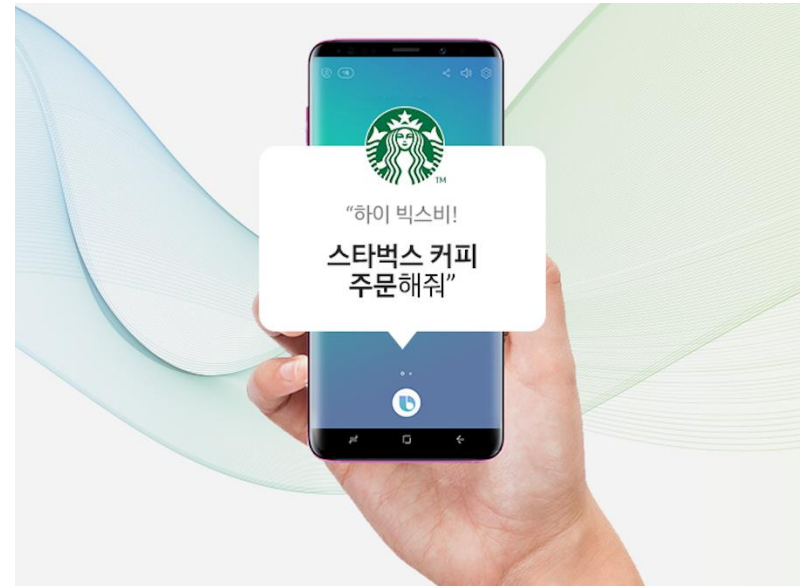
# 일상에서 만나는 NLP

“따뜻한 카페라떼 주문해줘”

따뜻하다 → 음료속성(Hot)  
카페라떼 → 메뉴명(카페라떼)

“그란데 라떼 두잔 주문해줘”  
그란데 → 음료속성(그란데  
사이즈)  
라떼 → 메뉴명(카페라떼)  
두잔 → 주문건수 (2건)

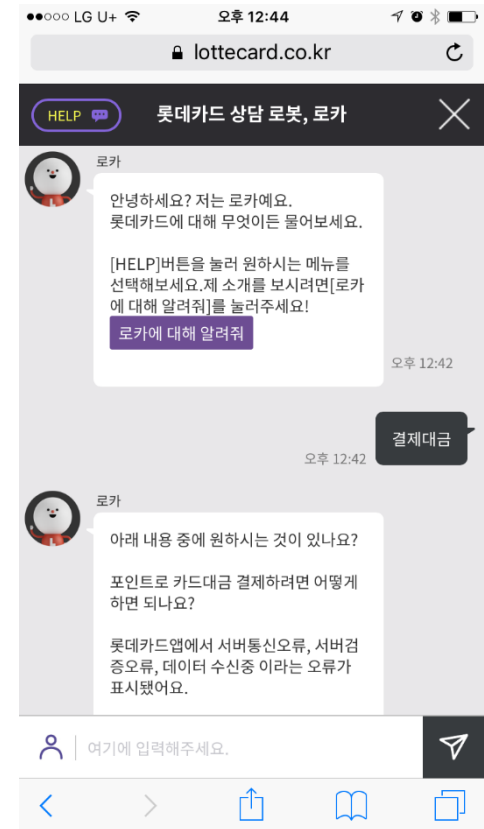
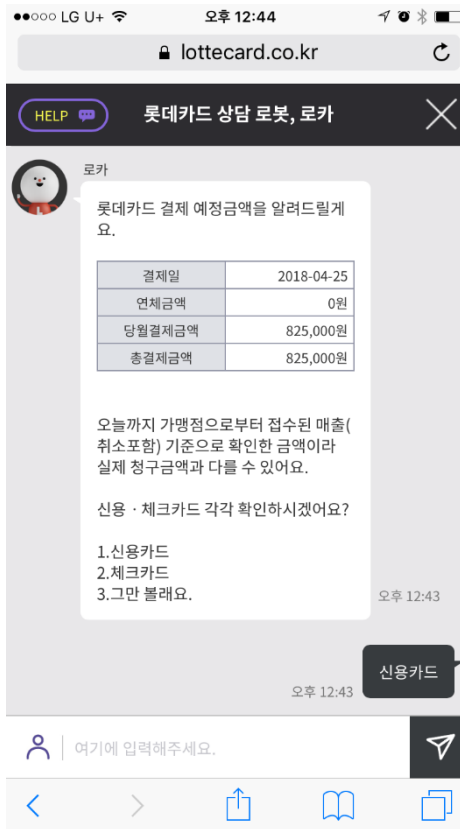
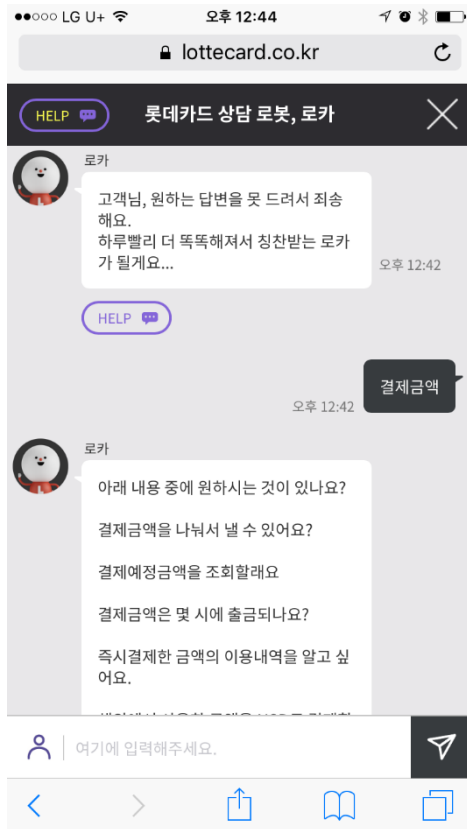
“앗, 잠시만요! 따뜻한 카페라떼  
취소하고...”  
???



빅스비 x 스타벅스 음성주문

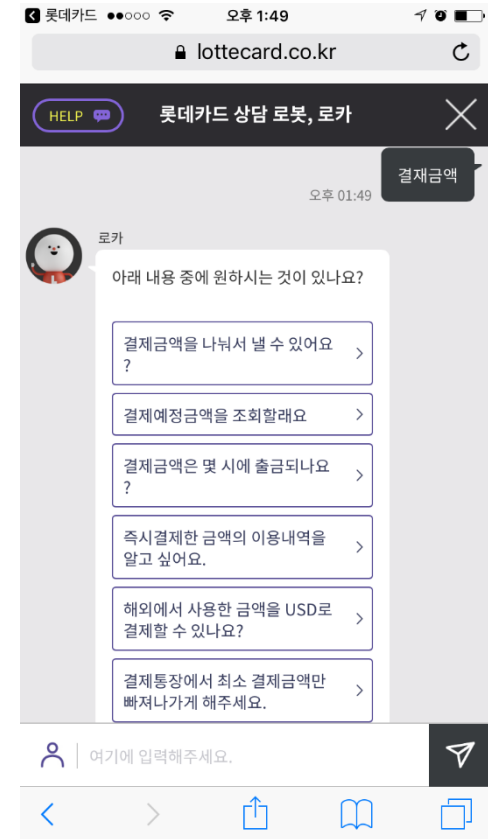
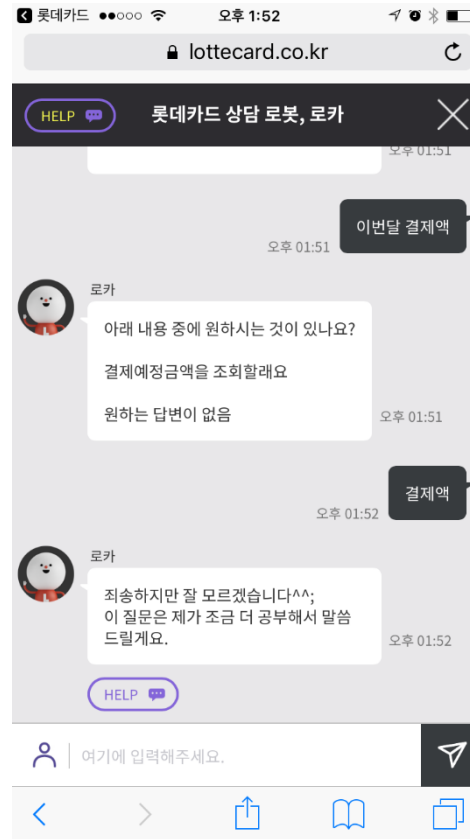
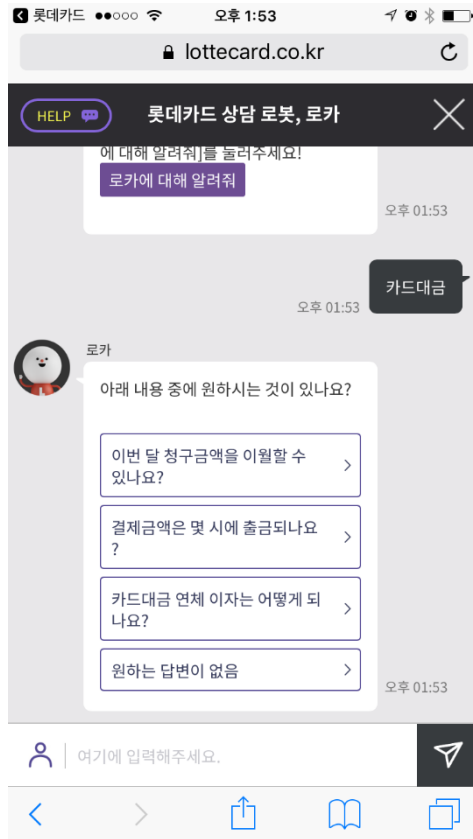
# 일상에서 만나는 NLP

## 롯데카드 상담 로봇, 로카 - “결제금액” 문의



# 일상에서 만나는 NLP

## 동의어 처리 – 결제금액, 카드대금, 결제액



# 일상에서 만나는 NLP

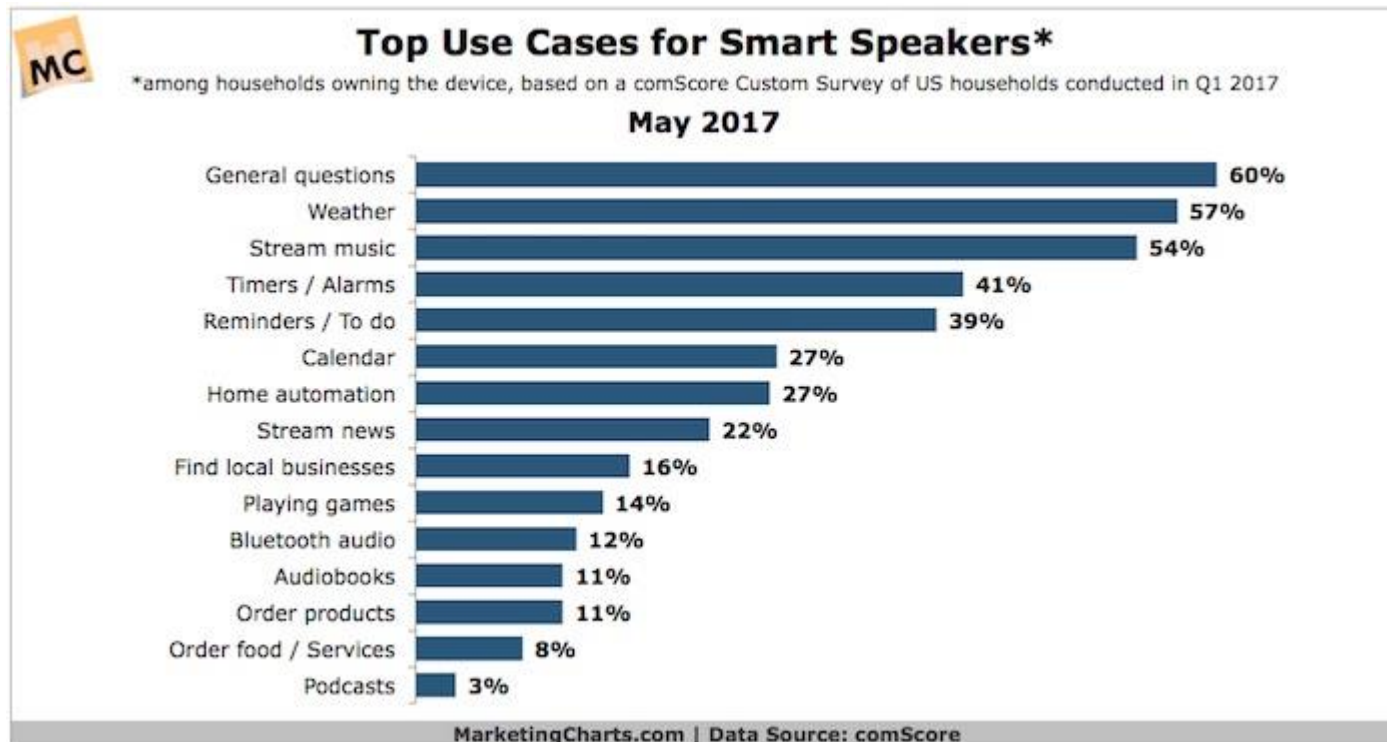
## Smart Speakers

- Amazon Alexa
- Google Assistant
- Siri
- Cortana
- Bixby
- SKT, KT, LGU+
- Naver, Kakao



# Smart Speakers – Top Use Cases

간단한 질문, 날씨, 음악감상, 타이머/알람,  
IoT, 쇼핑, 음식주문





# SELF-SERVICE의 시대



# 콜센터의 진화

- [Fujitsu AI Call Center Demo](#)

FUJITSU AI Zinrai Use Case: Call center

Query  
I'm together with my mom, but she's in a wheelchair so...

Additional keywords  
reservation restroom barrier-free taxi  
bump elevator parking

Search

Search results

1	86 %	I'd like to reserve a <b>barrier-free</b> hotel.
2	65 %	Are the <b>restrooms</b> <b>wheelchair</b> accessible?
3	62 %	Can I go to restaurants in a <b>wheelchair</b> ?

This enables the operator to respond in a shorter time.

1:14 / 2:51

FUJITSU

# Personal Assistant의 진화

- [Google I/O 2018 Duplex Demo](#)



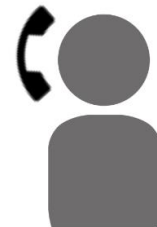
Google AI

Google Duplex



Hi, how can I help?

Google Assistant



(이미지 출처)

<https://blog.systoolsgroup.com/latest-google-assistant-duplex-update/>

# 자연어처리 기술 개요

# NLP란 무엇인가

- NLP = Computer Science + Linguistics + Artificial Intelligence
  - 컴퓨터를 통해 인간의 언어를 처리하고 이용하려는 연구 분야
  - 전산 언어학 (Computational Linguistics)
- Goal
  - 인간의 언어로 디지털 디바이스(SW, HW)와 *interaction*하여 원하는 *task*을 수행하게 하는 것
  - 자연어로 된 대량의 콘텐츠를 분석하여 *business insight*를 획득하는 것

# NLP 요소기술

- 전처리
- 형태소 분석
- 품사 태깅
- 개체명 인식
- 패턴 매칭
- 구문 분석
- 의미 분석
- 감성 분석
- 문장 생성

# 자연어 vs 인공어

- 자연어
  - 인간이 일상적으로 사용하는 언어
    - 한국어, 영어, 불어, 일본어, 중국어 등
- 인공어
  - 특정 목적을 위해 인위적으로 만든 언어
    - 에스페란토어 – 다른 민족 간에 대화 소통을 위해 인위적으로 만든 국제 공용어
    - 프로그래밍 언어 – Python, Java, C++ 등

# NLP 적용 분야

- 정보검색, 질의응답, 기계번역, 챗봇, 스마트스피커, 문서분류, 텍스트 분석(소셜미디어, VOC)





# NLP 적용 분야

- 정보검색
  - 입력된 키워드에 따라 관련 정보를 담은 문서를 검색
  - 웹페이지를 수집, 색인
- Q&A 서비스
  - 자연어로 질의 응답
  - 자연어로 질문을 하면 해답을 담은 문서를 검색
  - *아직까지는* 질문 문장의 의도 분석이 어려워 기대하는 수준의 결과를 제공하지 못함

# NLP 적용 분야

- 기계번역
  - SW로 언어간 장벽을 해결
  - 짧은 시간에 대량의 번역이 가능
  - 초기의 규칙 기반 기계번역, 통계 기반 기계번역  
시스템에선 도메인 커버리지, 품질의 한계로 인해  
시장의 기대에 못 미침
  - 최근 구글, 네이버 등에서 신경망 기반 기계번역을  
적용하면서 품질이 대폭 향상

# Google 번역 (translate.google.com)

United States officials and analysts have called on the North to submit a full inventory of its nuclear program for verification and to start dismantling its nuclear and missile facilities.



미국 당국자들과 분석가들은 북한이 검증을 위해 핵 계획의 전체 목록을 제출하고 핵 시설과 미사일 시설을 해체하기 시작할 것을 촉구했습니다.

# NLP 적용 분야

- 텍스트 분석
  - 소셜미디어 분석
  - VOC 분석 (고객서비스센터)
- 스마트스피커/인공지능 비서
  - 자연어로 원하는 태스크를 표현 (음성, 텍스트)
  - 생활 정보 제공, 상품 구매, TV/오디오 작동
- 챗봇
  - Superbot (아마존 알렉사)
  - Domain-specific bot (고객센터, 배달음식 주문, 대화형 정보검색 등)

# NL Interface 개념도

- Input의 modal은 말(voice), 글(text)
- Output은 의도한 바가 달성된 state 혹은 answer



# NLP SW의 결과물

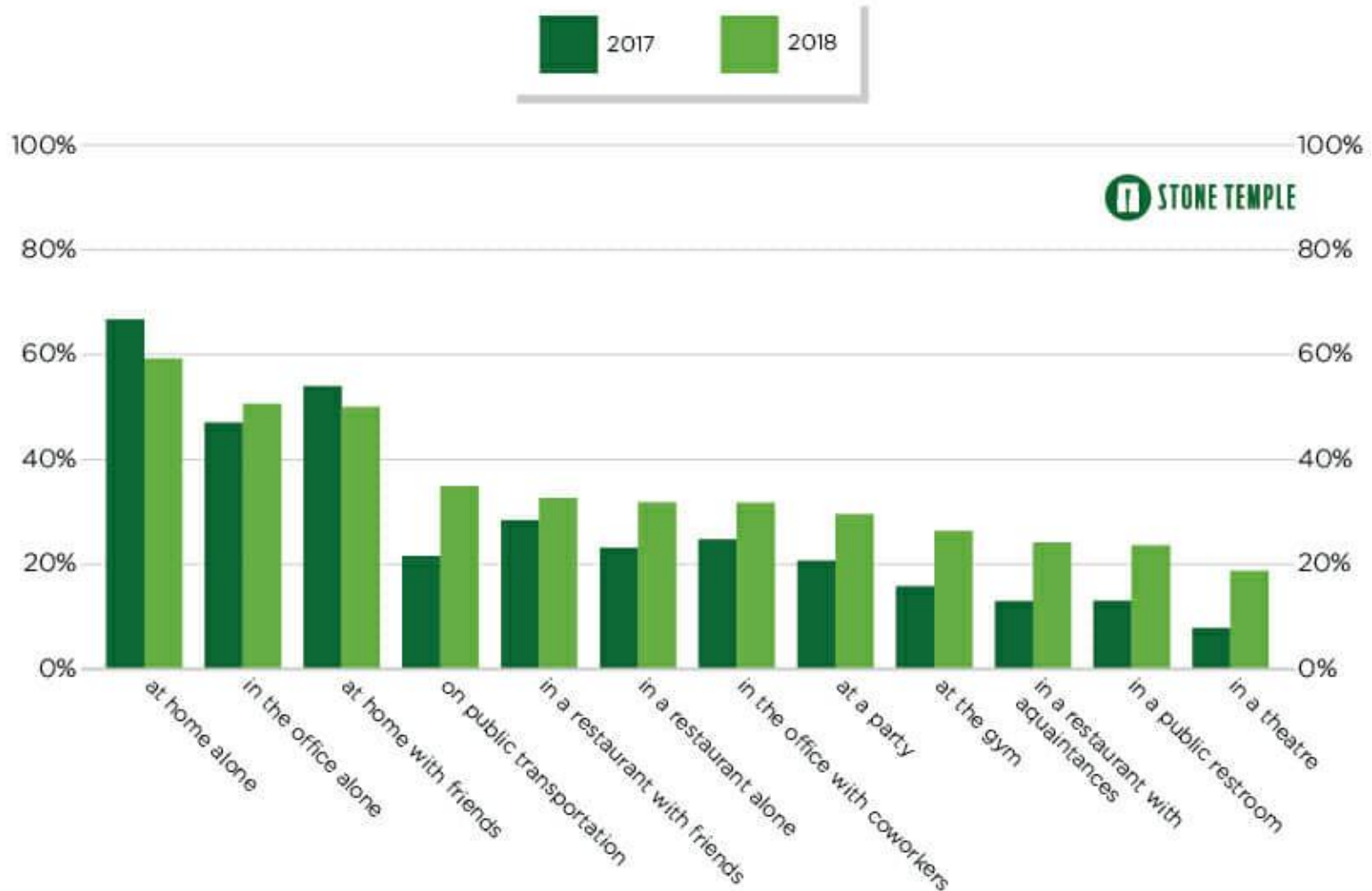
- 정보 검색
  - 원하는 정보를 포함한 문서를 제공
- 기계 번역
  - 원문과 동일한 의미를 가지는 타겟 문장을 생성
- 챗봇
  - 간단한 대화체 입력으로부터 사용자 의도를 파악
  - 원하는 액션을 수행하거나 응답을 제공
- 비정형 텍스트 분석
  - 비즈니스 인사이트

# Voice Input vs Text Input

- voice input은 Speed, Accuracy에 약점
- voice input만이 가능한 경우에는 유용함
- Library-Drive problem
  - Users can compose messages any way they want, and consume any way they want.

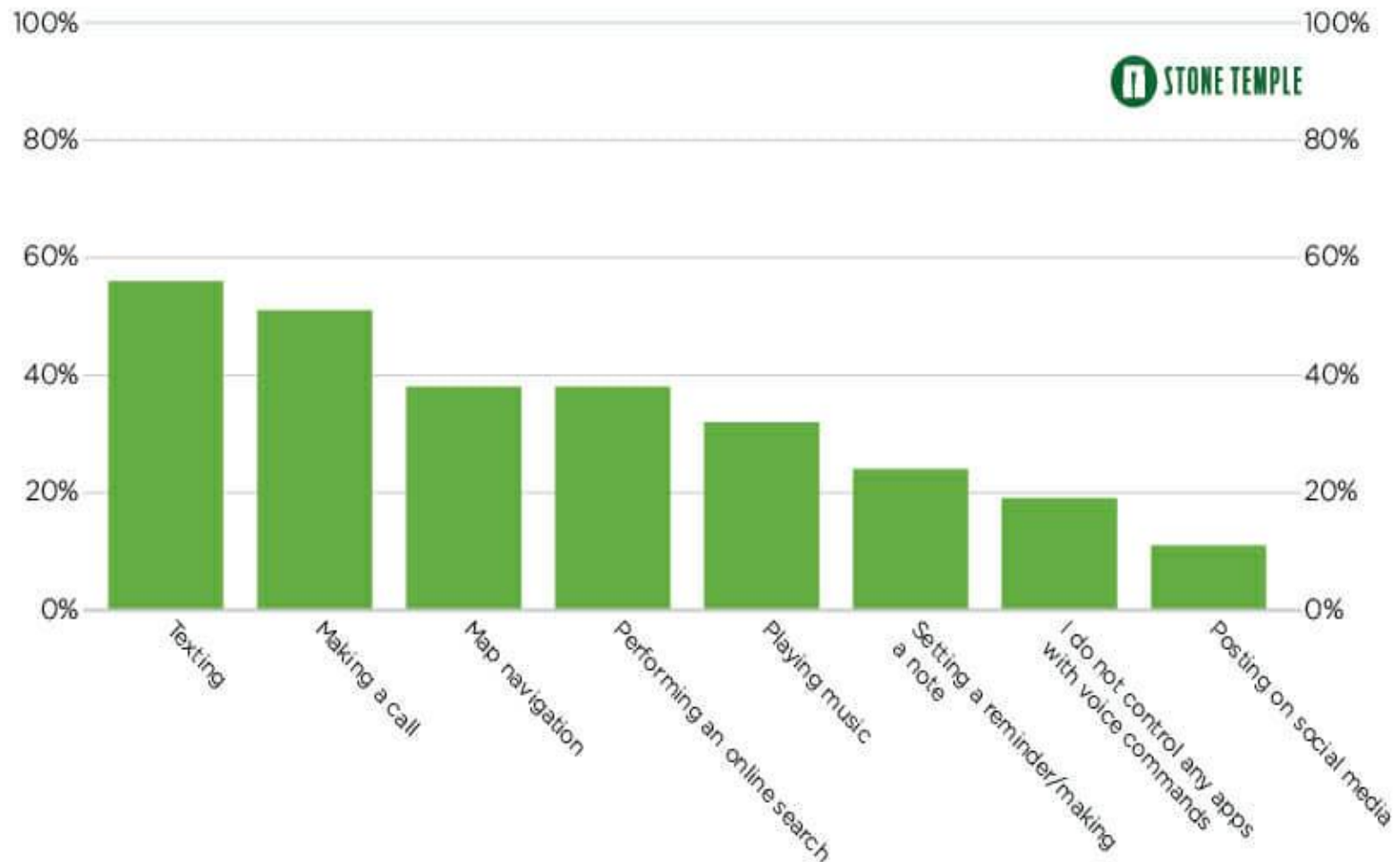
# 음성 인터페이스 사용 증가 추세

## In What Environments Do People Use Voice Search?

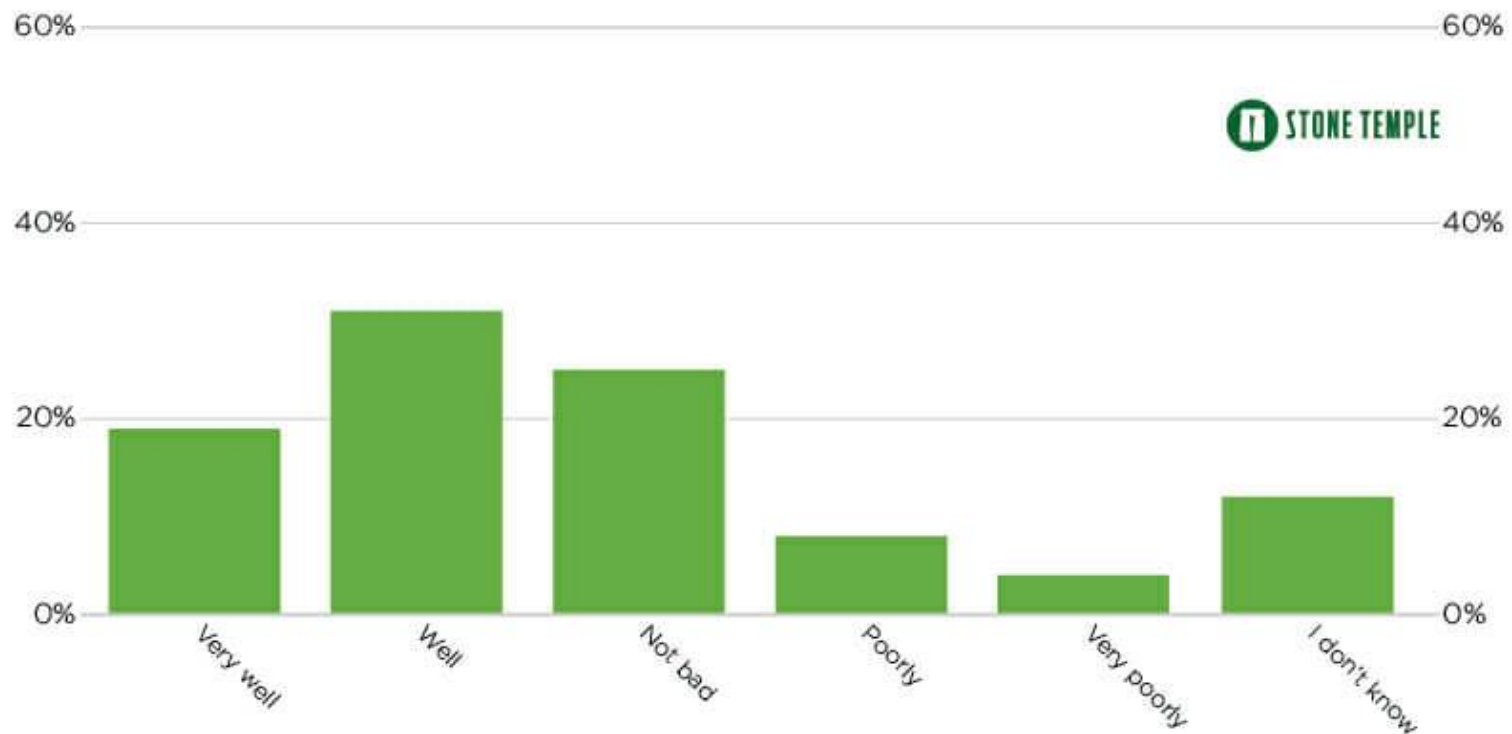




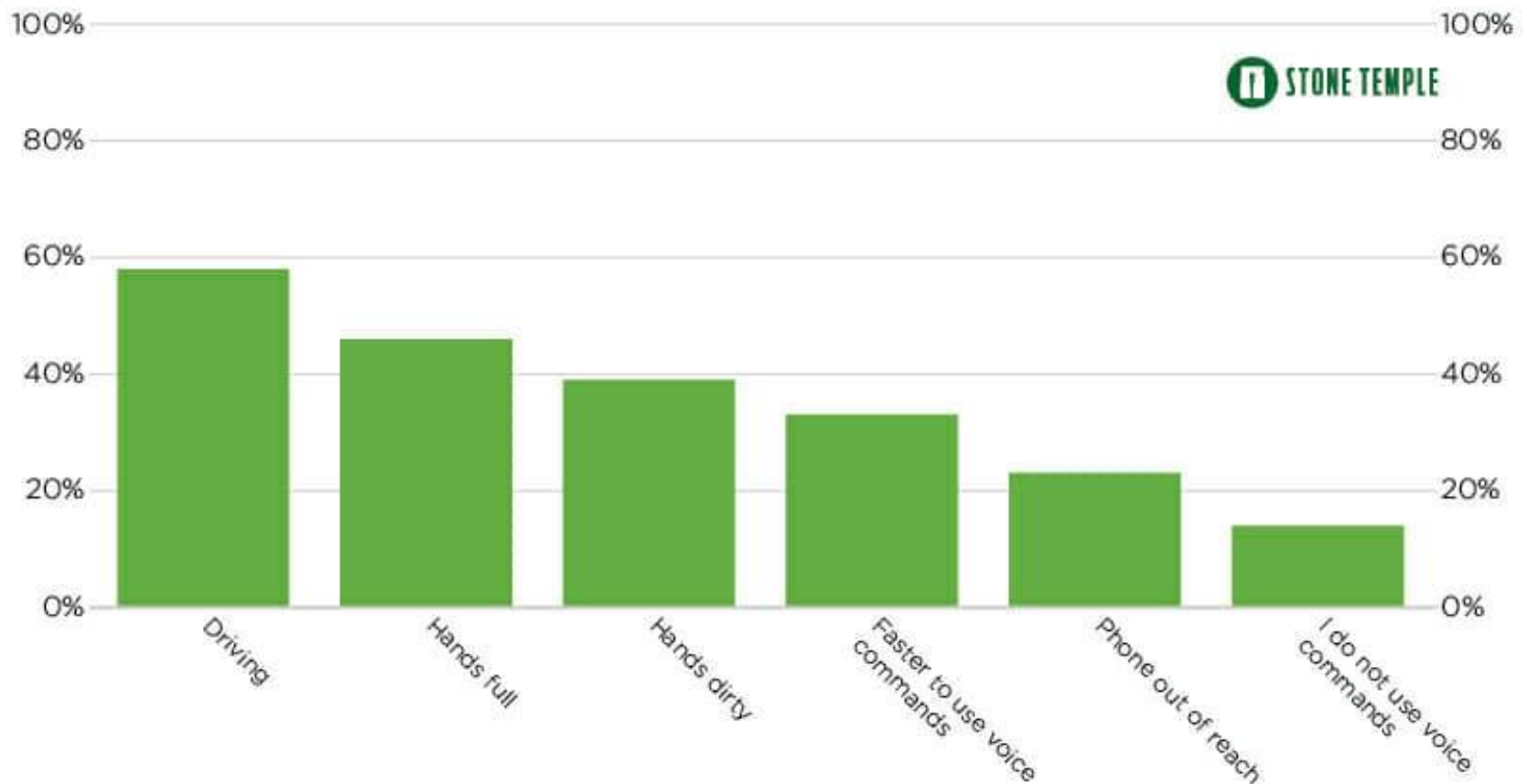
# Which of These Applications Have You Controlled With Voice?



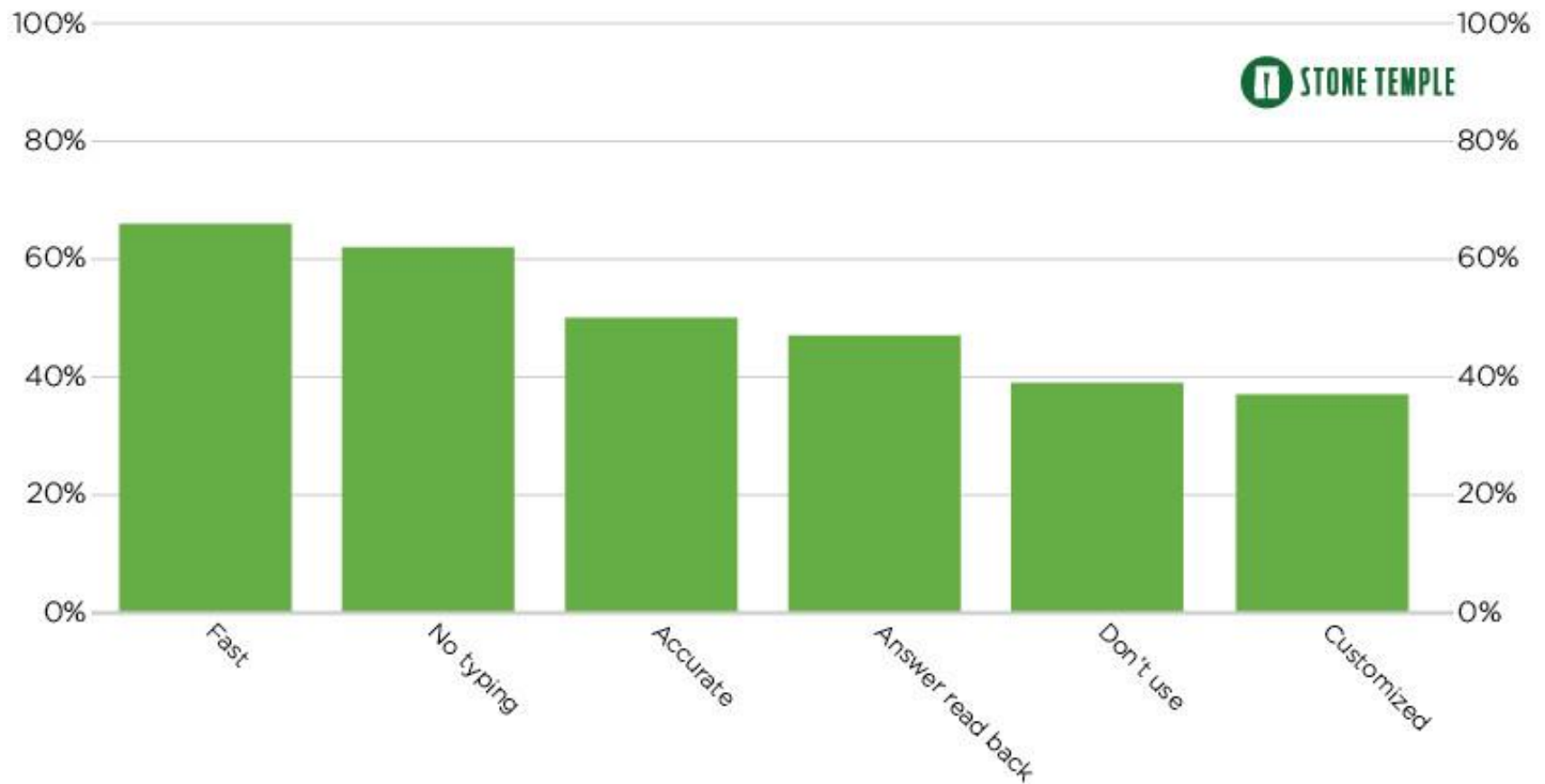
# How Well Do Your Built-In Personal Assistants on Your Phone Such as Google Assistant, Cortana, and Siri Understand You?



# In Which of These Situations Would You Be More Likely to Use Voice Commands on Your Smartphone Instead of Your Hands?



# Reasons People Like Voice Search



형태소분석과 품사태깅

# 한국어 문장의 구조

춘향이가      프로그래밍을      시작했다.  
<주어>      <목적어>      <서술어>

- (     ) 춘향이가 (     ) 프로그래밍을 (     ) 시작했다.  
(     ) 춘향이가 (     ) 프로그래밍을 (     ) (     ) 시작했다.  
(     ) 춘향이가 (     ) (     ) 프로그래밍을 시작했다.

춘향이가      의자에      앉았다.  
<주어>      <부사어>      <서술어>

- (     ) 춘향이가 (     ) 의자에 (     ) 앉았다.

# 한국어의 품사

- 품사
  - 단어를 그 문법적 성질에 따라 분류한 갈래
- 체언 - 명사, 대명사, 수사
- 용언 - 동사, 형용사
- 수식언 - 관형사, 부사
- 독립언 - 감탄사
- 관계언
  - 조사 (격조사, 접속조사, 서술격조사)

# 격조사

- 명사에 붙어서 그 명사가 문장에서 일정한 자격을 가지게 하는 조사
- 주격, 목적격, 서술격, 보격, 부사격 조사

춘향이가 저고리를 향단이에게 주었다.

춘향이가 스타가 되었다.

방자가 동생이다.

- 서술격 조사 (copula)
  - ~이다 (영어의 be, become에 해당)
  - 체언이나 체언 구실을 하는 말 뒤에 붙어 서술어 자격을 가지게 하는 격조사



# 보조사

- 격조사 뒤 또는 부사 뒤에 붙어 뜻을 보조해주는 조사  
대표적인 보조사로는 '-는/-도/-만'

춘향이는 저고리도 향단이에게만 주었다.

춘향이도 저고리만 향단이에게는 주었다.

춘향이만 저고리는 향단이에게도 주었다.

향단이는 떡을 빨리는 먹는다.

향단이는 떡을 빨리도 먹는다.

향단이는 떡을 빨리만 먹는다.

# 어간과 어미

- 용언
  - 어간 + 어말어미
  - 어간 + 선어말어미 + 어말어미
- 선어말어미
  - 시제 -았/-었
  - 추정 -겠
  - 높임 -시
- 어말어미
  - 종결어미 -는다, -느냐, -자, -어라, ...
  - 접속어미(연결어미) -고, -나, -지만, ...
  - 전성어미 -음, -기 ; -은, -는, -을, -던 ; -니까, -다가

# 형태소 분석

- 한 어절 내에 있는 모든 **형태소(사전 표제어)**를 분리  
사무실에서부터였다고는 → 사무실+에서부터+이+있+다고는  
영장전담판사실 → 영장+전담+판사+실
- 용언(동사, 형용사)의 **원형을 복원**
- 형태소들 간의 결합관계가 적합한지 검사
  - 체언과 조사의 결합 제약
  - 용언과 어미의 결합 제약
  - 조사, 어미 사이의 결합 제약
- 복합어 분석, 미등록어 추정 (신조어, 외국어 등)

# 형태소 결합 법칙 예시

- 명사
- 명사 + 격조사
- 명사 + 격조사 + 보조사
- 명사 + 접미사
- 접두사 + 명사
- 접두사 + 명사 + 접미사
- 명사 + 명사
- 명사 + 명사 + 조사
- 명사 + 명사 + 명사 + 접미사 // **영장+전담+판사+실**
- 동사 + 어미
- 동사 + ~~전~~성어미 + 조사 // **감+기+는**
- 동사 + 선어말어미 + 어말어미 // **읽 + 었 + 다**

# 형태소 분석 예

- “나는”

- ① 나(대명사, I) + 는(조사)
- ② 날다(동사, fly) + 는(어미)
- ③ 나다(동사, born) + 는(어미)

- (연습) “감기는”

- ①
- ②
- ③

# 형태소 분석의 모호성 morphological ambiguity

- 동일한 표층형 (surface form) 어절이 여러가지 형태소 결합으로 분석 가능한 문제

“감기는”

- ① 감기/명사 + 는/조사
- ② 감기/동사 + 는/어미
- ③ 감/동사 + 기/전성어미 + 는/조사

- 복합명사 분해 수준

“영장전담판사실”

- ① 영장+전담+판사+실
- ② 영장전담+판사+실
- ③ 영장전담+판사실
- ④ 영장전담판사+실
- ⑤ 영장전담판사실

# 품사 태깅 (Parts-of-Speech Tagging)

- 형태소 분석의 모호성을 해결
- 해당 문맥에 맞는 품사태그를 선택하는 문제
- 이후 단계 모듈의 정확도를 좌우
- 통계적 기법에 의한 해결
  - 대량의 품사 태깅된 말뭉치를 구축
  - 통계적으로 확률이 높은 품사 태그를 선택

대분류	소분류	세분류
체언	명사	일반명사 NNG
		고유명사 NNP
		의존명사 NNB
	대명사	대명사 NP
	수사	수사 NR
외국어		외국어 F
용언	동사	동사 VV
	형용사	형용사 VA
	보조용언	보조용언 VX
	지정사	긍정지정사 VCP
		부정지정사 VCN
수식언	관형사	관형사 MM
	부사	일반부사 MAG
		접속부사 MAJ
독립언	감탄사	감탄사 IC
관계언	격조사	주격조사 JKS
		보격조사 JKC
		관형격조사 JKG
		목적격조사 JKO
		부사격조사 JKB
		호격조사 JKV
		인용격조사 JKQ

세종말뭉치  
품사태그 (1/2)

품사태그 :  
형태소 분석의  
기준이 되는  
세분화된  
품사 체계

참고 :  
형태소 태깅  
말뭉치 작성용  
품사태그 세트  
(한국정보통신기  
술협회 표준안)



대분류	소분류	세분류
관계언	보조사	보조사 JX
	접속조사	접속조사 JC
의존형태	어미	선어말어미 EP
		종결어미 EF
		연결어미 EC
		명사형전성어미 ETN
		관형형전성어미 ETM
	접두사	체언접두사 XPN
	접미사	명사파생접두사 XSN
		동사파생접두사 XSV
		형용사파생접미사 XSA
	어근	어근 XR
기호	마침표, 물음표, 느낌표	SF
	쉼표, 가운데점, 콜론, 빗금	SP
	따옴표, 괄호표, 줄표	SS
	줄임표	SE
	불임표(물결, 숨김, 빠짐)	SO
	한자	SH
	기타 기호(논리 수학기호, 화폐기호 등)	SW
	명사추정범주	NF
	용언추정범주	NV
	숫자	SN
	분석불능범주	NA

세종말뭉치  
품사태그 (2/2)

Tag	Description	Example
CC	Coordinating conjunction	and, or, but
CD	Cardinal number	five, three, 13%
DT	Determiner	the, a, these
EX	Existential <i>there</i>	<i><u>there</u> were six boys</i>
FW	Foreign word	mais
IN	Preposition or subordinating conjunction	of, on, before, unless
JJ	Adjective	nice, easy
JJR	Adjective, comparative	nicer, easier
JJS	Adjective, superlative	nicest, easiest
LS	List item marker	
MD	Modal	may, should
NN	Noun, singular or mass	tiger, chair, laughter
NNS	Noun, plural	tigers, chairs, insects
NNP	Proper noun, singular	Germany, God, Alice
NNPS	Proper noun, plural	two Teslas
PDT	Predeterminer	<u>both</u> her children
POS	Possessive ending	's
PRP	Personal pronoun	me, you
PRP\$	Possessive pronoun	my, your
RB	Adverb	extremely, loudly
RBR	Adverb, comparative	better
RBS	Adverb, superlative	best

영어 품사태그  
예시  
(Penn Treebank  
[https://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) )

Tag	Description	Example
RP	Particle	about, off, up
SYM	symbol	%
TO	infinitival <i>to</i>	what <u>to</u> do?
UH	interjection	oh, oops, gosh
VB	verb, base form	think
VBD	verb, past tense	she <u>thought</u>
VBG	verb, gerund or present participle	<u>playing</u> is fun
VCN	verb, past participle	a <u>sunken</u> ship
VBP	verb, non-3 <sup>rd</sup> person singular present	I <u>think</u>
VBZ	verb, 3 <sup>rd</sup> person singular present	she <u>thinks</u>
WDT	wh-determiner	which, whatever, whichever
WP	wh-pronoun	what, who, whom
WP\$	possessive wh-pronoun	whose, whosever
WRB	wh-adverb	where, when

영어 품사태그

예시

(Penn Treebank

[https://www.ling.upenn.edu/courses/Fall\\_2003/ling001/penn\\_treebank\\_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) )

# Sample pos-tagged text

United States/NNP officials/NNS and/CC  
analysts/NNS have/VB called/VBN on/IN  
the/DT North/NNP to/TO submit/VB a/DT  
full/JJ inventory/NN of/IN its/PRP\$ nuclear/JJ  
program/NN for verification and to start  
dismantling its nuclear and missile facilities.

# Sample pos-tagged text

United States/NNP officials/NNS and/CC  
analysts/NNS have/VB called/VBN on/IN the/DT  
North/NNP to/TO submit/VB a/DT full/JJ  
inventory/NN of/IN its/PRP\$ nuclear/JJ  
program/NN for/IN verification/NN and/CC to/TO  
start/VB dismantling/VBG its/PRP\$ nuclear/JJ  
and/CC missile/NN facilities/NNS ./SYM

# HMM에 의한 품사 태깅 예

- Hidden Markov Model 접근법
- 예문 "Birds like flowers."
  - 형태소 사전  
bird : Noun(N)  
like : Verb(V), Preposition(P)  
flower : Noun(N)
  - 가능한 품사 태그열  
(N, V, N)  
(N, P, N)

# HMM에 의한 품사 태깅 예

- $\Pr(N, V, N)$   
 $= \Pr(V|N) * \Pr(N|V) * \Pr(\text{birds}|N) * \Pr(\text{like}|V) * \Pr(\text{flowers}|N)$
- $\Pr(N, P, N)$   
 $= \Pr(P|N) * \Pr(N|P) * \Pr(\text{birds}|N) * \Pr(\text{like}|P) * \Pr(\text{flowers}|N)$

$\Pr(V | N) : 0.4$   $\Pr(N | V) : 0.5$   $\Pr(N | P) : 0.6$

$\Pr(\text{birds} | N) : 0.04$

$\Pr(\text{like} | V) : 0.06$   $\Pr(\text{like} | P) : 0.05$

$\Pr(\text{flowers} | N) : 0.05$

$W_i$  는  $i$ 번째 단어,  $P_i$  는  $W_i$  의 품사태그일 때

$\Pr(W_i | P_i) = \text{freq}(W_i, P_i) / \text{freq}(P_i)$

$\Pr(P_i | P_{i-1}) = \text{freq}(P_{i-1} P_i) / \text{freq}(P_{i-1})$

# 개체명 인식 (Named Entity Recognition)

- 누가, 무엇(누구)을, 언제, 어디서, 얼마나(수량, 크기, 비율 등), ... 에 대한 정보
  - 사람, 동식물, 장소, 국가, 조직, 브랜드명 등 고유명사
  - 날짜, 시간, 화폐(금액), 주소, 전화번호, 이메일, 결제 방식 등
- Why NER?
  - Information Extraction, Question Answering, Coreference Resolution, Smart Speakers, Chatbots



# 개체명 타입 예 (OntoNotes)

Entity Name Types	Description
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language

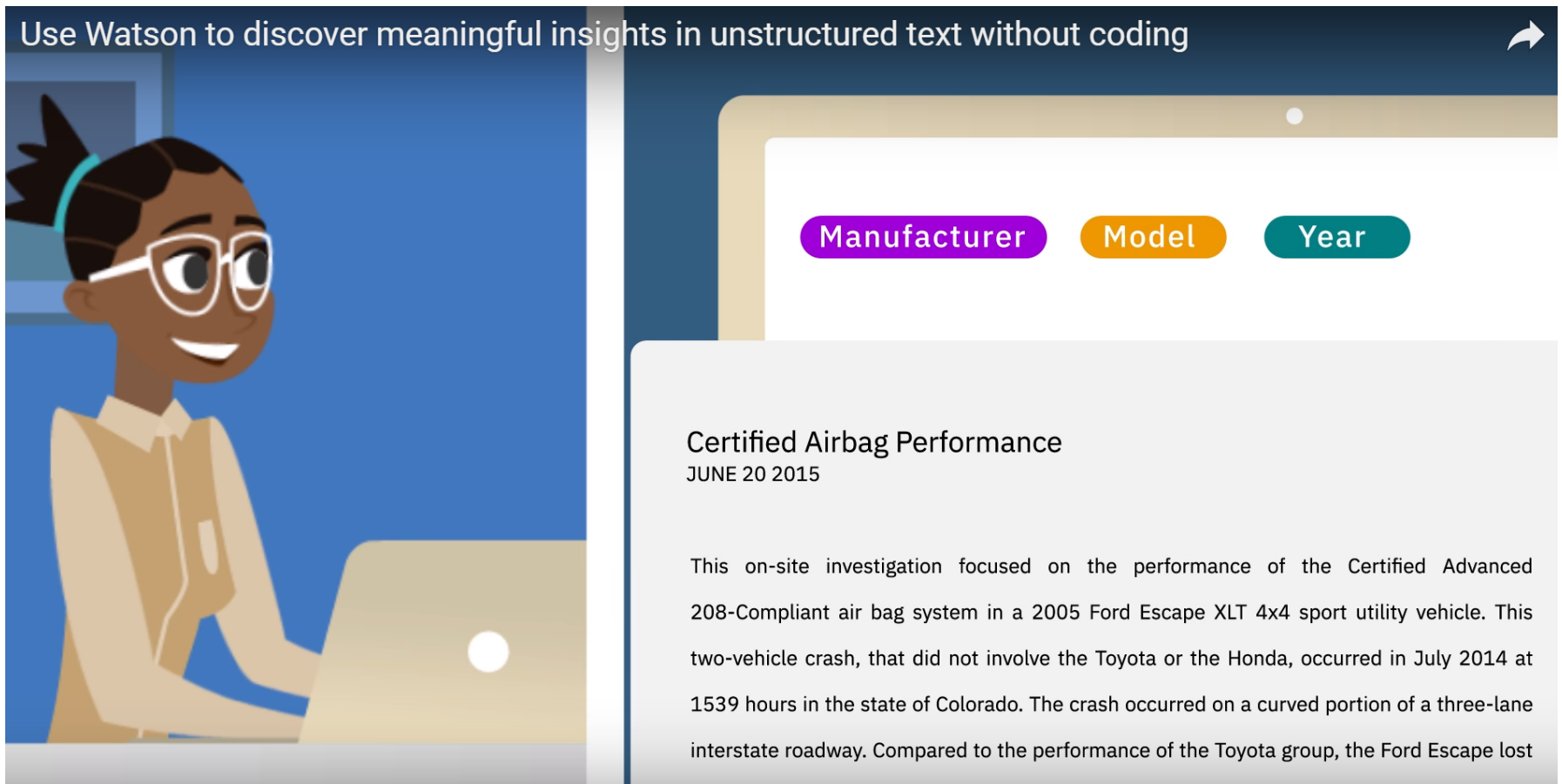
# 개체명 타입 예 (OntoNotes)

Value Types	Description
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second"
CARDINAL	Numerals that do not fall under another number

- OntoNotes project
  - A corpus of large scale, accurate, and integrated annotation of multiple levels of the *shallow* semantic structure in text (POS, names, predicate-argument, syntax, word senses, coreference)
  - Annotated English 1.5M words, Chinese 800K words, Arabic 300K words

# 개체명 태깅

IBM Watson Knowledge Studio (<https://youtu.be/byqpojcfDZM>)



Use Watson to discover meaningful insights in unstructured text without coding

Manufacturer Model Year

### Certified Airbag Performance


JUNE 20 2015

This on-site investigation focused on the performance of the Certified Advanced 208-Compliant air bag system in a 2005 Ford Escape XLT 4x4 sport utility vehicle. This two-vehicle crash, that did not involve the Toyota or the Honda, occurred in July 2014 at 1539 hours in the state of Colorado. The crash occurred on a curved portion of a three-lane interstate roadway. Compared to the performance of the Toyota group, the Ford Escape lost

# 개체명 태깅

## Annotation by Domain Experts

Use Watson to discover meaningful insights in unstructured text without coding



208-compliant air bag system in a 2005 **Ford** Escape XLT 4x4 sport utility vehicle. The two-vehicle crash, that did not involve the **Toyota** or the **Honda**, occurred in July 2014 at 1539 hours in the state of Colorado. The crash occurred on a curved portion of a three-lane interstate roadway. Compared to the performance of the **Toyota** group, the **Ford** Escape lost control on an interstate highway and struck a concrete barrier on the right side of the roadway. In relation to the performance of the **Honda** systems, the impact resulted in sufficient longitudinal deceleration of the Escape to command the deployment of the front air bag system and actuation of the driver's seat belt pretensioner. The vehicle rotated over from the initial wall impact and was subsequently struck by a 2013 **BYD** Qin pulling a single trailer. The restrained 48-year-old male driver of the **Ford** Escape appears to have sustained a minor facial injury. **Honda** came in a close second in performance in 2014, followed by **Toyota**. **Ford** reports a 10% increase in performance over 2015.

Owned by

# 개체명 인식 결과

Person

Manufacturer

Car Model

Place

가격 4000만원 전기차 '테슬라 모델3' 공개... 국내 예약자도 여럿

테슬라는 31일(현지 시각) 미국 로스엔젤레스에서 신형 전기차 '모델3'를 공개했다. 이 차는 테슬라의 엔트리 모델로, 메르세데스-벤츠 C클래스를 비롯해 BMW 3시리즈, 아우디 A4 등 독일 프리미엄 세단과 경쟁하게 될 것으로 예상된다.

...

테슬라의 CEO 엘론머스크는 "모델3의 엔트리 트림은 정지상태에서 시속 100km까지 채 6초가 안걸리는 우수한 가속 성능을 갖췄다"면서 "완충 시 최대 주행 가능 거리도 346km에 달한다"라고 설명했다. 또, "전 모델이 테슬라의 급속 충전 시스템을 지원하며 오토파일럿 하드웨어도 기본 적용됐다"라고 덧붙였다.

모델3의 생산은 내년 말부터 진행되며 (모델3의) 가격은 3만5000달러(약 4000만원)부터 시작한다. 이는 현재 판매 중인 모델S와 모델X의 절반 수준이다.

# 개체명 관련 자료

- 공공 인공지능 오픈 API (<http://aiopen.etri.re.kr/>)
- 표준 개체명 태그세트 및 태깅 말뭉치  
(한국정보통신기술협회)
- Stanford NER
  - <https://nlp.Stanford.edu/ner/>
- Apache OpenNLP NER
  - <https://opennlp.apache.org/>  
The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

# 형태소 분석의 이슈

- 전처리
  - 띄어쓰기 오류, 맞춤법 오류
- 형태소 분석의 모호성
- 미등록어 처리
  - 신조어, 복합명사, 전문용어, 축약어, 외국어 등
  - 신조어 추정 규칙 필요
  - 언어별로 추정 규칙이 다름
  - 품사 추정 (주로 명사, 고유명사로 추정)
- 지속적인 사전 업데이트
  - 사용자사전, 신조어, 고유명사(개체명), 동의어 사전 등
  - **사전구축 프로세스와 관리 툴 필요**(현업에서 중요)

# 전처리 (Preprocessing)

- Why Preprocessing?
  - 분석된 결과에 대한 신뢰도 확보 위해
  - Cleansing (분석 대상이 아닌 입력문서를 제거)
  - Normalization (분석의 기본 단위인 단어를 정규화)
- 전처리 종류
  - 스팸 필터링
  - 중복문서 제거
  - 맞춤법 교정
    - 띄어쓰기 오류, 오타 처리
  - 동의어 처리
    - 동의어 처리가 필요한 경우, 필요 없는 경우
  - 축약어 처리
    - 축약어와 동일한 뜻의 원래 단어(대표형, 표준형)로 변환



# 띄어쓰기 처리 예시

0 entries loaded...

> 문제는 대내외 불확실성, 소비심리 위축 등에 따른 내수 둔화, 구조 조정 영향으로 당분간 청년 눈높이에 맞는 양질의 일자리 부족 문제가 지속할 수 있다는 점이다.

문제는 대내외 불확실성, 소비심리 위축 등에 따른 내수 둔화, 구조 조정 영향으로 당분간 청년 눈높이에 맞는 양질의 일자리 부족 문제가 지속할 수 있다는 점이다.

>

> 정부는 불충분한 고용 기회로 실업이 장기화하고 구직 활동이 위축되는 등 어려운 청년 고용 여건이 지속할 것으로 보인다고 판단했다.

정부는 불충분한 고용 기회로 실업이 장기화하고 구직 활동이 위축되는 등 어려운 청년 고용 여건이 지속할 것으로 보인다고 판단했다 .

>

> 이를 해결하기 위한 대책 마련이 필요했지만, 대통령이 탄핵되고 새 정부가 들어서지 않은 어수선한 상황에서 새로운 일자리 창출 대책을 만들어 내기엔 부담이 컸다.

이를 해결하기 위한 대책 마련이 필요했지만, 대통령이 탄핵되고 새 정부가 들어서지 않은 어수선한 상황에서 새로운 일자리 창출 대책을 만들어 내기엔 부담이 컸다 .

>

> 하지만 고용 시장이 악화 일로를 걷는 상황에서, 정부가 주는 시그널은 필요했다. 구직 단념자가 더 늘어나지 않도록 기존 대책을 활용해 맞춤형 지원을 강화하겠다는 취지로 보완 방안을 내놓은 것이다.

하지만 고용 시장이 악화 일로를 걷는 상황에서, 정부가 주는 시그널은 필요했다. 구직 단념자가 더 늘어나지 않도록 기존 대책을 활용해 맞춤형 지원을 강화하겠다는 취지로 보완 방안을 내놓은 것이다.

> □

# NLP 모듈의 품질평가

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$= \frac{\text{True Positive}}{\text{Total Predicted Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$= \frac{\text{True Positive}}{\text{Total Actual Positive}}$$

$$F1 = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

조화 평균 (Harmonic Mean)

언어 데이터 구축 방법

# 사용자 사전

- 도메인 사전
- 고유명사, 복합명사, 전문용어, 외국어, 축약어
- 특수문자, 숫자가 조합된 제품명
  - Node.js, 아티브 NT905S3T-KDBS
- 신조어
  - 신규 항목을 추가할 때는 반드시 회귀테스트  
(예) 나홀로족, 리즈시절, 스몰웨딩,  
fake news, airpocalypse

# 미등록어 (Unknown Word) 처리

- 사전에 없는 단어
  - 외국어로 된 인명, 지명
  - 신조어이나 사전에 등록되기 이전
- 사전에 없더라도 품사 태그를 추정하여 부착해야 이후 단계로 처리를 진행할 수 있음
  - 주로 명사 또는 고유명사로 추정
  - partial string이 사전에 등록된 단어일 경우 추정을 도와주는 힌트가 된다.  
주로 마지막 단어가 정보를 많이 준다.  
예) OO인 OO자 OO호텔 OO시

# 사전의 종류

- 시스템 사전 (기본 어휘 사전)
- 신조어 사전
- 동의어 사전
- 반의어 사전
- 감성어 사전
- 도메인 사전
  - 커머스, 제조, 법률, 특허, 의학, 과학, ...
- 오타 교정 사전
- 패턴 사전 (응용 영역별로 설계)

# 사전 구축 및 업데이트 프로세스

- 미등록어 후보 추출
  - 수집된 텍스트를 형태소 분석하는 과정에서 분석에 실패하여 품사 태깅이 안되는 단어 추출
  - 자동으로 품사 추정된 단어를 미등록어 후보 파일로 추출
- 사전에 등록할 후보를 선정
  - 언어데이터 구축 전문가가 미등록어 후보 파일에서 시스템적으로 유의미한 단어를 선별
- 회귀 테스트
  - 기존 형태소 사전(정답셋)과 신규 단어 집합(테스트셋)을 비교하여 side effect가 발생할 수 있는 후보 단어가 있는지 테스트하는 과정을 거쳐 최종 사전을 빌드
- 사전 배포

# 말뭉치 (Corpus)

- 언어 연구를 위해 특정 목적을 가지고 언어의 표본을 추출한 집합
- 유형
  - 원시 말뭉치
  - 품사 태깅된 말뭉치
  - 개체명 인식을 위한 말뭉치
  - 구문분석/의미분석을 위한 말뭉치
  - 단일 언어 말뭉치 vs 다중 언어 말뭉치



# 세종 말뭉치

- 국어정보화 중장기 발전계획
- 21세기 세종계획 국어기초자료 구축 사업 (1998~2007)
  - 현대국어 기초 말뭉치 개발
  - 말뭉치 구축 및 활용에 필요한 방법론 및 표준화 연구
  - 말뭉치를 활용한 SW 개발

# 세종 말뭉치

종류	규모(어절수)	비고
세종말뭉치 통합본 (기존 말뭉치)	1억 2,000만	세종계획 이전 구축 국어정보처리기반구축 사업(1994-1997) 7,000만 국립국어연구원 구축 5,000만
원시 말뭉치	6,200만	원문 그대로 유지
형태소분석 말뭉치	1,500만	'표준국어대사전'의 표제어를 기준으로 형태소 분석 및 태깅
형태의미분석 말뭉치	1,250만	동음이의어 형태에 의미표지를 부착 표준국어대사전의 의미어깨번호 사용
구문분석 말뭉치	80만	구문구조 분석

대분류	소분류	세분류
체언	명사	일반명사 NNG
		고유명사 NNP
		의존명사 NNB
	대명사	대명사 NP
	수사	수사 NR
외국어		외국어 F
용언	동사	동사 VV
	형용사	형용사 VA
	보조용언	보조용언 VX
	지정사	긍정지정사 VCP
		부정지정사 VCN
수식언	관형사	관형사 MM
	부사	일반부사 MAG
		접속부사 MAJ
독립언	감탄사	감탄사 IC
관계언	격조사	주격조사 JKS
		보격조사 JKC
		관형격조사 JKG
		목적격조사 JKO
		부사격조사 JKB
		호격조사 JKV
		인용격조사 JKQ

세종말뭉치  
품사태그 (1/2)

품사태그 :  
형태소 분석의  
기준이 되는  
세분화된  
품사 체계

대분류	소분류	세분류
관계언	보조사	보조사 JX
	접속조사	접속조사 JC
의존형태	어미	선어말어미 EP
		종결어미 EF
		연결어미 EC
		명사형전성어미 ETN
		관형형전성어미 ETM
	접두사	체언접두사 XPN
	접미사	명사파생접두사 XSN
		동사파생접두사 XSV
		형용사파생접미사 XSA
	어근	어근 XR
기호	마침표, 물음표, 느낌표	SF
	쉼표, 가운데점, 콜론, 빗금	SP
	따옴표, 괄호표, 줄표	SS
	줄임표	SE
	불임표(물결, 숨김, 빠짐)	SO
	한자	SH
	기타 기호(논리 수학기호, 화폐기호 등)	SW
	명사추정범주	NF
	용언추정범주	NV
	숫자	SN
	분석불능범주	NA

세종말뭉치  
품사태그 (2/2)

# 공공 인공지능 오픈 API / 데이터

- 국가 R&D 과제를 통해 개발한 언어처리 기술 및 데이터를 공개 (aiopen.etri.re.kr)
- 언어분석 및 음성인식 API, 언어분석용 말뭉치 등



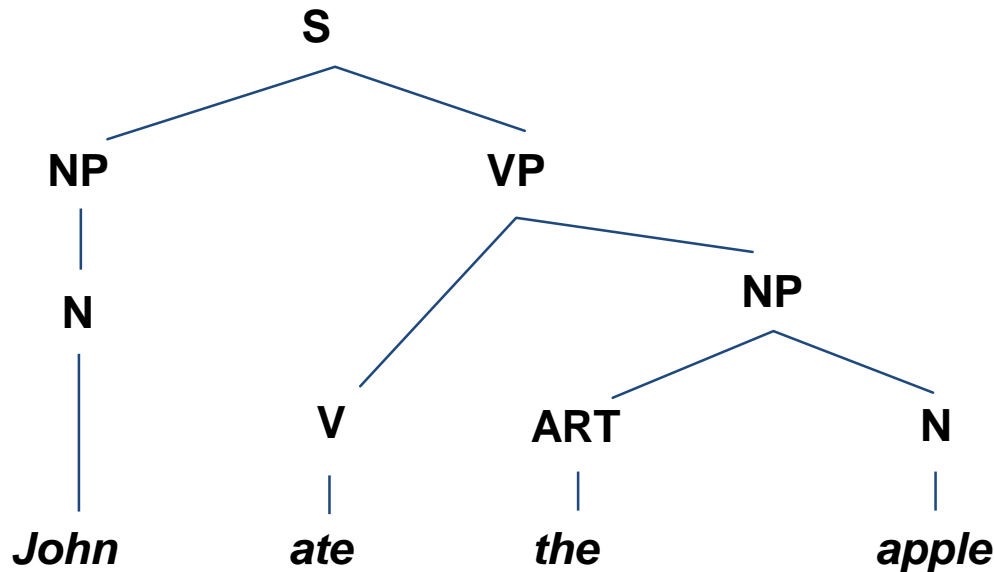
# 오픈소스SW vs 상용SW

- 오픈소스 SW
  - 정확도 / 사전 (관리 툴) / 속도 / 안정성
  - 응용분야를 위한 사전을 직접 구축해야 함
  - 형태소 분석과 품사 태깅을 모두 제공하는 패키지
  - 도메인 사전, 신조어, 동의어, 불용어 사전 등
  - 품사 태깅 정확도, 속도, 안정성 이슈 해결
- 상용 SW
  - NLP 모듈에 대한 라이선싱 필요
  - 프로젝트 단위로 외주 (검색, 분석, 소셜미디어분석, 챗봇 등)

구문분석, 패턴매칭, 의미  
분석

# 구문 분석 (Syntax Analysis)

- 문장의 구조를 분석하여 계산 가능한 내부 형태로 표현하는 것



S : Sentence

NP : Noun

Phrase

VP : Verb

Phrase

N : Noun

V : Verb

ART : Article

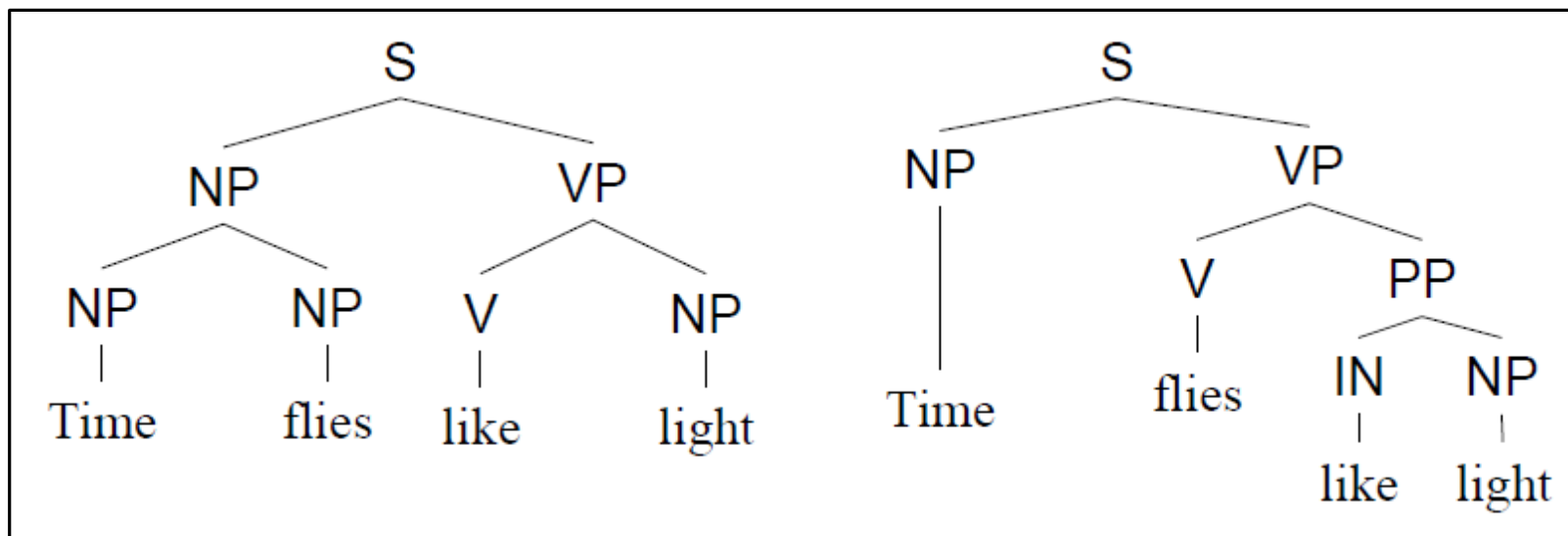


# 구문 분석의 모호성

- 하나의 입력문장이 여러가지 구조로 분석 가능

flies (명사-파리, 동사-날다)

like (동사-좋아하다, 전치사-~와 비슷한)



# 명사구 청킹

- 명사구 청킹 패턴 예

{명사}+ 명사

대명사 {명사}+

{관형사}+ {명사}+

'+' : 1번 이상 반복 가능함을 의미

# 명사구 청킹

- 패턴 적용 예

어제 저녁에 우리 회사 직원을 보았다.

-> [NP 어제/명사 저녁/명사]

-> [NP 우리/대명사 회사/명사 직원/명사]

오늘 저 두 사람을 처음 만났다.

-> [NP 저/관형사 두/관형사 사람/명사]

# 패턴 매칭

- '출생지'에 대한 구문 패턴 예

**태어나다 :**

A/명사 + 는/조사 + B/명사 + 에서/조사 + 태어나다

**탄생하다 :**

A/명사 + 는/조사 + B/명사 + 에서/조사 + 탄생하다

→ 출생지(A, B) : A의 출생지는 B이다

이때 A는 '사람', B는 '장소' 속성을 갖는 단어

# 패턴 매칭

“이순신은 서울 건천동에서 태어났다.”

- 품사 태깅 결과

이순신/명사 은/조사

서울/명사 건천동/명사 에서/조사

태어나/동사 았/선어말어미 다/어말어미

- 명사구 청킹

서울 건천동 -> [서울 건천동]

- 구문 패턴 매칭

‘태어나다’의 구문 패턴을 매칭

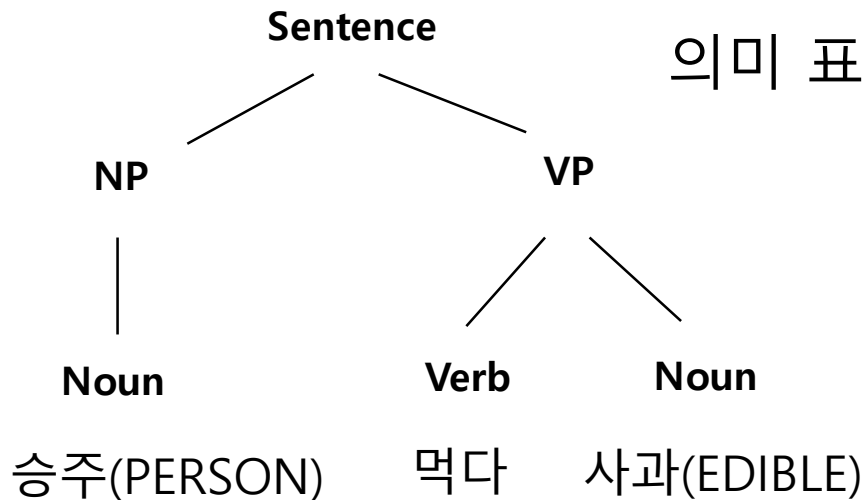
→ 출생지(이순신, 서울 건천동)

# 의미 분석 (Semantic Analysis)

- 문장에 표현된 각 단어가 의미하는 개념과 그 개념들 간의 관계를 분석

입력문 "승주가 사과를 먹는다"

의미 표현 예시

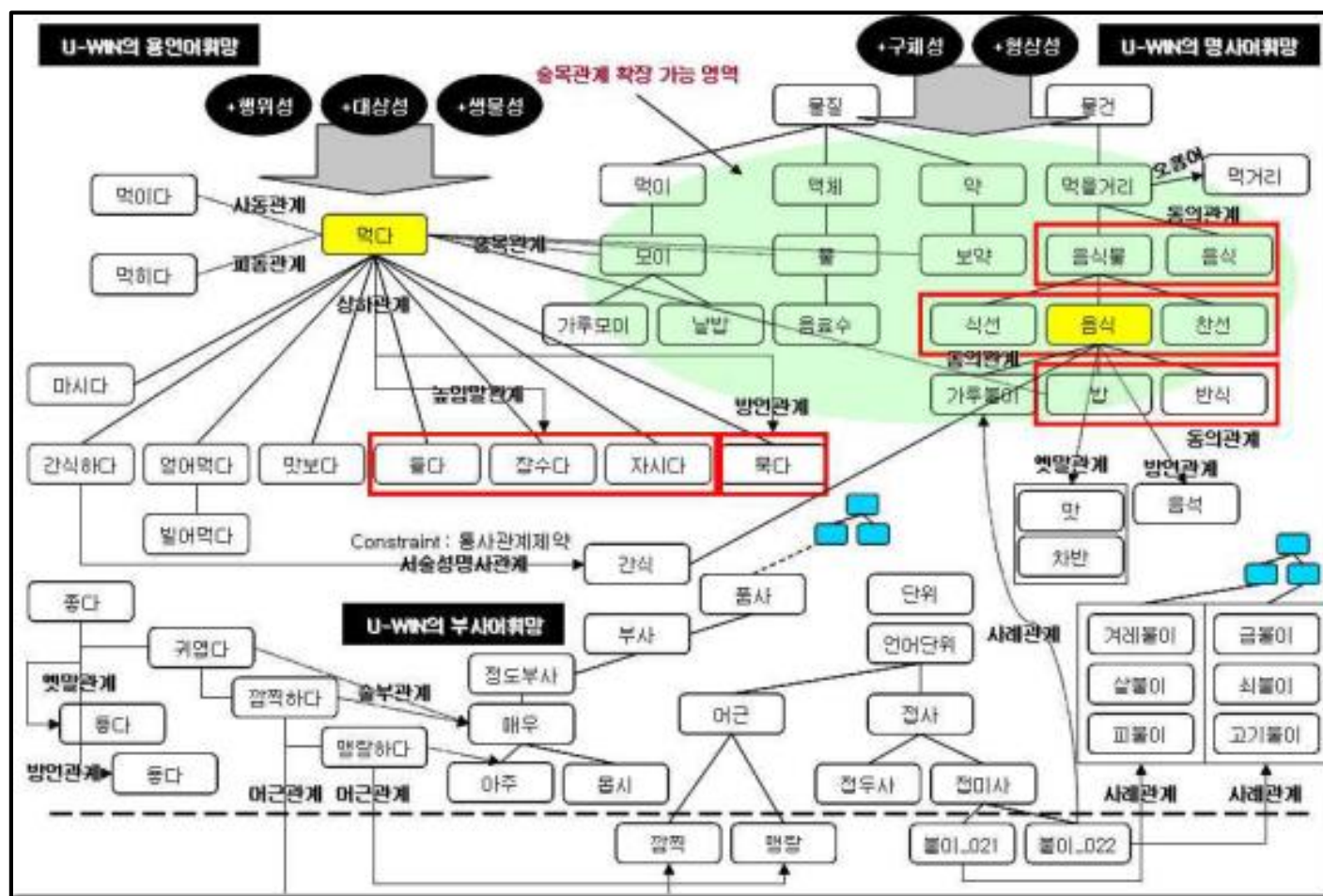


("먹다" : V, Meaning=EAT  
(Subject : N, +PERSON,+ANIMAL)  
(Object : N, +EDIBLE))

("승주" : N, +PERSON)  
("사과" : N, +EDIBLE)

# 어후의 의미 네트워크

- 어휘를 의미 관계에 따라 네트워크로 표현  
예) 울산대 한국어 어휘지도(UWordMap)



# 의미 분석의 이슈

- 다의어 (多義語, Polysemy)
  - 은행 -> bank, ginkgo nut
  - 다리 -> bridge, leg
- 동의어 (同義語, Synonym)
  - 책방, 서점
  - 음식점, 식당, 레스토랑



# 중의성 (重義性)

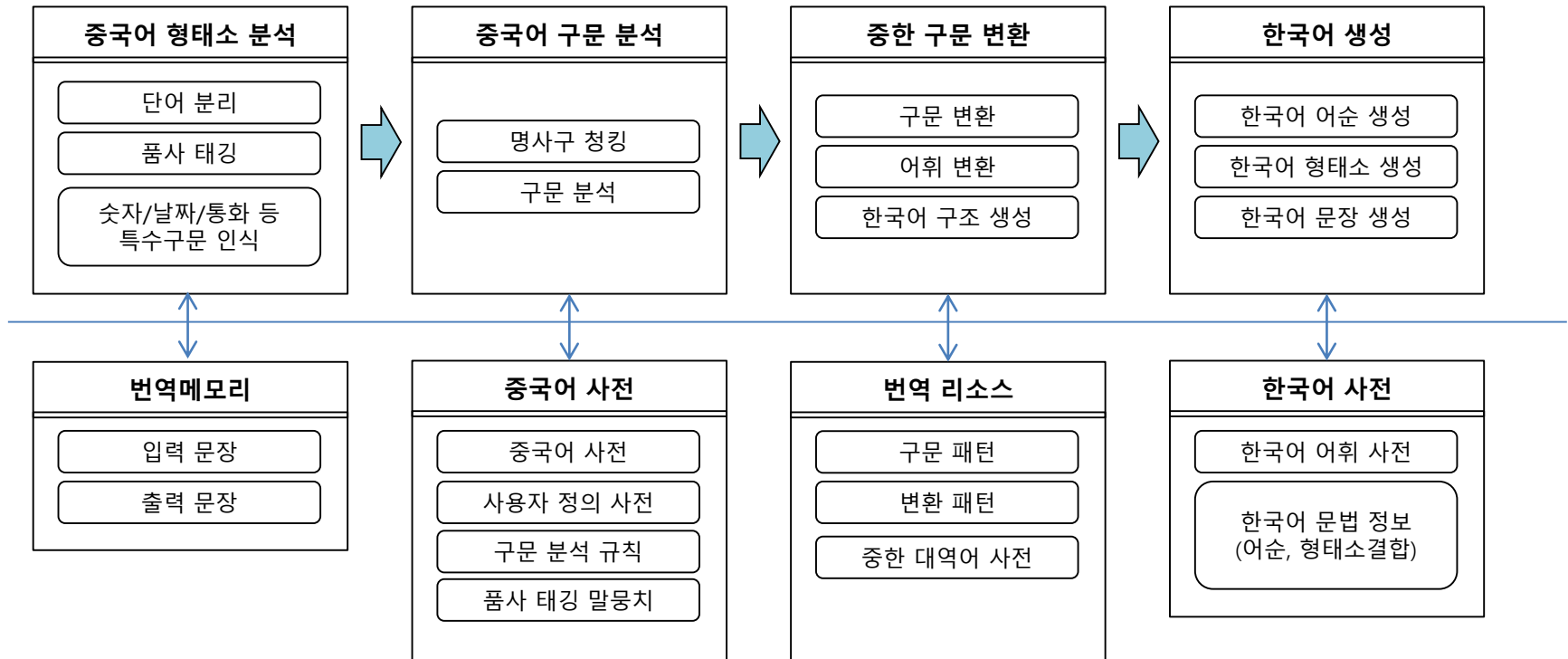
- 정보 검색에서 중의성
  - 동형이의어를 의미별로 분류해서 색인해야 하는 문제가 발생
  - 예) 은행 (bank), 은행 (ginkgo nut)
- 기계번역에서 중의성
  - 한 단어나 구문이 두 가지 이상의 의미를 가질 때 올바른 번역어를 선택해야 하는 문제
  - 예) 진정한 사과를 받았다 vs 맛있는 사과를 받았다
  - 예) 그는 말이 많다.
    - > He has many horses.
    - > He is talkative.
    - > He has many words. ??

# 화용 분석(Pragmatic Analysis)

- 문장이 실세계와 가지는 연관관계 분석
  - 실세계의 지식과 상식의 표현이 필요함
  - 지시, 화법, 간접화법 등의 분석
- 대명사의 지시 대상 (coreference)
  - "그것 좀 주문해 줘"
  - > "<치킨, 피자, 햄버거> 좀 주문해 줘"
  - "지난번에 주문했던 걸로 할게"
  - > "<2017\_03\_25>에 주문했던 걸로 할게"
- 상대방에게 행동을 요구하는 언어 행위, 간접적인 의사표현
  - "거실이 좀 덥네~" (에어컨을 틀어줘)

# 규칙기반 NLP시스템 사례

## 중한 기계번역 시스템 (Rule-Based MT, RBMT)



# 문장 번역 처리단계 예(중한)

(S (VP\* (PP-ppmod (PO\* (base "关于") +SB +locobj +absobj +ZsentPP)  
 (NP-obj (DNP-dnp (NP-xde (NR\* (base "中国") +Nproper1 +Rnat))  
 (DE\* (base "的") ))  
 (NP\* (NN\* (base "对外政策") +Nabs))))  
 (VP\* (ADVP (AD\* (base "也") ))  
 (VP\* (V\* (V\* (VV\* (base "出现") +Vnpobj +Vrecom +Vpostqp +Vexist))  
 (AS (base "了") ))  
 (NP-obj (DT (base "各种") )  
 (NP\* (CP-adn (ADJP-xde (A\* (AJ\* (base "不同") )))  
 (DE\* (base "的") ))  
 (NP\* (NN\* (base "声音") +Nabs))))))  
 (PU (base "。") +SE +nonSB))

구문 분석

关于/PO 中国/NR 的/DE 对外政策/NN 也/AD 出现/VV  
 了/AS 各种/DT 不同/AJ 的/DE 声音/NN 。/PU

품사 태깅

关于中国的对外政策也出现了各种不同的声音。

언어간 변환

(S (VP\* (NP-adv (NP-mod (NR\* 중국 <2>))  
 (NP\* (NN\* 대외\_정책/에\_관해 <4>)))  
 (VP\* (VP\* (VP\* (VV\* 나타나[tense\_past] <6>))  
 (NP-com (DT-mod 각종 <8>)  
 (NP\* (ADJP-adn (AJ\* 다르[mode\_adn] <9>))  
 (NP\* (NN\* 소리/가 <11>))))))  
 (PU . <12>)))

번역문 생성

중국의 대외 정책에 관해  
 각종 다른 의견이 나타났다.

# 중한번역 사전

- 형태소 분석 사전 예시

关于/PO:1.0

中国/NR:1.0

...

也/AD:0.993772 SP:0.002348

出现/NN:0.066269 VV:0.933731

- 번역 사전 예시

<te> 关于/PO @ 에\_관해/PO

<tu> = (pobj \$xvp hd 的/DE) @ \$xvp[e:ㄴ\_것 p:에\_대한]

<tu> = (pobj \$xnp hd 的/DE) @ \$xnp[p:에\_대한]

<tu> = (pobj \$xvp) @ \$xvp[mode:nmz p:에\_관해]

<tu> = (pobj \$np) @ \$np[p:에\_관해]

# 규칙기반 기계번역의 이슈

- NLP의 각 단계에서 발생하는 오류의 누적
  - 형태소분석과 품사태깅의 모호성
  - 구문분석의 모호성의 해결
  - 의미 모호성의 해결
  - 생성된 문장 내 의미적 오류와 비문 발생
- 2016년 이전
  - RBMT, Statistical MT, Hybrid MT (RBMT + Statistical MT)
- 2016년 이후
  - Neural Machine Translation으로 번역품질 대폭 향상
  - 구글은 자사 NMT를 적용하여 위키피디아와 뉴스의 샘플 문장을 번역해 본 결과 기존 시스템 대비 번역오류를 55%~85% 가량 줄였다고 설명

# 자연어처리 시스템 사례

소셜미디어분석

# 소셜미디어 분석 니즈

- SNS, 블로그, 제품리뷰 등 VOC의 급증
- 대량의 온라인 대화로부터 비즈니스적으로 유의미한 데이터를 추출하려는 니즈
- 위기 이슈 예방, 관리 니즈
- 긍정, 부정 이분법적인 분석에서 다양한 감성에 대한 분석 요구로 발전



# 소셜미디어분석 프로세스 (예시)

온라인 미디어분석 서비스 (<http://www.pulsek.com>)

## 펄스K 분석 프로세스



# 소셜미디어분석 시스템의 기능

## 펄스K 주요 기능

소셜  
인지도

이슈  
키워드

소셜  
호감도

영향력  
분석

비교  
분석

누구나 쉽게 이슈 발견 및 분석 서비스 활용 가능

소셜 인지도

연급 추이를 통한 이슈 발생, 성장, 절정, 소멸

이슈 키워드

연관 이슈어 · 감성어를 통한 이슈 상세 분석

소셜 호감도

온라인 미디어내 긍·부정 반응 및 인식 분석

영향력분석

영향력자 파악 및 수치화된 여론 동향 확인

비교 분석

경쟁력 측정 및 미디어별 이슈어 분석



시각화

그래프 / 워드클라우드 / 다이어그램



상세 분석

이슈어랭크 / 엑셀다운로드 / 원문링크



API

비교 분석 / 소셜스코어 / 인플루언서

# 소셜미디어분석 시스템의 기능

- 소셜 인지도
  - 언급 추이 트래킹
  - 이슈의 발생, 성장, 절정, 소멸 파악
- 이슈 키워드
  - 연관된 이슈 분석
- 소셜 호감도 (감성 분석)
  - 온라인 대화 상의 긍부정 반응 및 인식 분석
- 인플루언서 분석
  - 영향력자 파악 및 수치화된 여론 동향 확인
- 경쟁사 비교 분석
  - 경쟁력 측정

# 소셜미디어분석의 활용

- 브랜드의 인지도, 평판, 경쟁력 분석
- 고객의 미디어 이용 행태 파악
- 온라인 위기 이슈에 대응
- 소비자 트렌드 파악
- VOC 분석
  - PR/마케팅/기획/고객만족 부서 등 전 부문
  - 지속가능한 비즈니스 경쟁력 확보를 위해서 필수적인 툴이라는 인식이 확대됨

# 텍스트마이닝 활용 분야

- **Customer Insight**
  - VOC Analysis, Social Media Analysis
- **Productivity**
  - Document Classification, Summary
- **Question Answering**
  - Information Extraction

# 실습 가이드

# 한국어 형태소 분석기 mecab-ko

- 실습파일 다운로드
- mecab, mecab-ko-dic 설치
  - mecab.zip을 C:\mecab 에 압축해제
- mecab-python 인스톨
  - mecab\_python-0.996\_ko\_0.9.2\_msvc-cp36-cp36m-win\_amd64.whl
  - > pip install [mecab\\_python-0.996\\_ko\\_0.9.2\\_msvc-cp36-cp36m-win\\_amd64.whl](#)

# 한국어 형태소 분석

- 실습 개요
  - mecab 형태소 분석기를 이용, 한국어 형태소 분석 결과를 확인
  - 사용자 사전을 추가
- 실습 코드
  - 01\_mecab\_test.ipynb
  - MeCab.Tagger() : mecab 형태소 분석기 생성
  - tagger.parse() : 형태소 분석 결과 반환
  - mecab\_split.py : 실행 결과 중 토큰과 품사정보만을 추출하는 함수



# Mecab user dictionary

- 사용자 사전 추가
- User-dic 디렉토리 내 csv 파일
  - 파일형식

표층형	0	0	0	품사 태그	의미 부류	종성 유무	읽기	타입	첫번째 품사	마지막 품사	표현
-----	---	---	---	-------	-------	-------	----	----	--------	--------	----

- Powershell 실행 (마우스 우측버튼, 관리자 권한으로 실행)
  - > cd C:\mecab
  - Set-ExecutionPolicy RemoteSigned 실행  
또는
  - Set-ExecutionPolicy unrestricted 실행 (윈도우 10일 때)
  - > .\tools\add-userdic-win.ps1

# Mecab 사전 형식

- 표충형, 품사태그, 의미분류, 종성, 읽기, 타입
- 표충형 : surface form
- 품사태그 : 품사 표지
- 의미분류 : 인명, 지명 또는 \*
- 종성 : 단어의 마지막 음절의 받침 유무로 T, F
- 읽기 : 발음 (단어의 발음을 입력)
- 타입 : inflected, compound, preanalysis

# Mecab-ko-dic 사이즈

- 총 80여만 단어
- 일반명사 (NNG) 20만여개
- 고유명사 (NNP) 2,371개
- 동사(VV) 7,300여개
- 형용사(VA) 2,300여개
- 부사(MAG) 14,000여개

# Mecab user dictionary

- 이체한도 증액 어떻게 하나요  
디폴트 분석 결과 (사용자 사전 없을 때)
  - 이 MM,~명사,F,이,\*,\*,\*,\*
  - 체한 NNG,\*,T,체한,\*,\*,\*,\*
  - 도 JX,\*,F,도,\*,\*,\*,\*
  - 증액 NNG,\*,T,증액,\*,\*,\*,\*
  - 어떻게 MAG,\*,F,어떻게,\*,\*,\*,\*
  - 하나 NR,\*,F,하나,\*,\*,\*,\*
  - 요 VCP+EC,\*,F,요,Inflect,VCP,EC,0|/VCP/\*+오/EC/\*
  - EOS

# Mecab user dictionary 추가

- nng.csv
  - 이체한도 NNG
  - 외화통장 NNG
  - 외환통장 NNG
  - ...
- 이체한도 증액 어떻게 하나요?
  - 이체한도 NNG (일반명사)
  - 증액 NNG (일반명사)
  - 어떻게 MAG (일반부사)
  - 하 VV (동사)
  - 나요 EF (종결어미)
  - ? SF (문장부호)

End of Document