Natural Language Processing

A Quick Tour of Traditional NLP

KyungTae Lim

Contents

Story telling of AI in 1960~2000

- AI는 어찌되었든 인간의 말을 기계가 알아들어야 AI라고 할 수 있지

  - 그럼 인간의 언어를 이해하는 AI를 개발하려면 무엇이 필요할까?

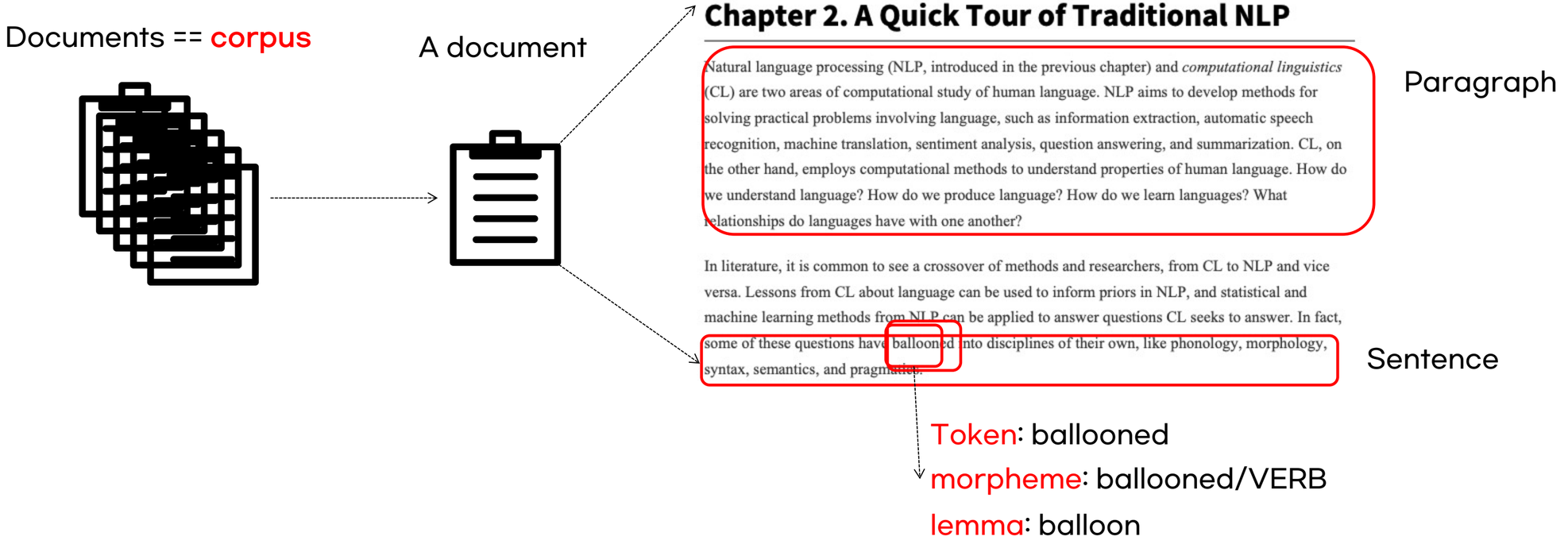
  - 바로 교과서!

  - 책으로부터 기계를 학습 시키는 거지!

Let's think about how we can teach languages to a machine

- Then, how is the textbook organized?

  - We are going to talk about those things a bit

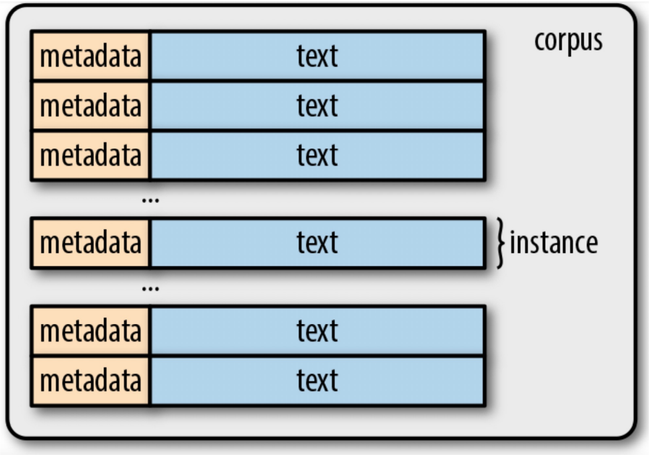


http://morphological.org/?words=ballooned

1980년대 AI연구자 입장에서
교과서의 구성이 어떻게 생겼는지 살펴보자

2.1 Corpora, Tokens, and Types

- Corpus, Corpora (말뭉치 in Korean)
  - A group of documents
  - All NLP tasks, whether classical or modern, start with text data called corpus
  - Generally, it includes raw text and metadata.
  - In the field of ML, text with metadata attached is called a sample or data point
  - A collection of samples, which is a corpus, is called a dataset.



그래 말뭉치는 알겠는데 그걸로 기계를 어떻게 가르쳐주지? 이 녀석들은 단어도 모르는데?

단어를 정의해 보자!

2.1 Corpora, Tokens, and Types

- Tokenization (토큰화)
  - The process of dividing text into tokens
  - For example, there are six tokens in “Maria frapis la verda sorcistino”
  - Agglutinative languages like Turkish cannot be adequately divided by whitespace

    → Need better solution!!

    · Agglutinative language (교착어) :

    where word meaning is determined by roots

    and affixes (particles (조사)), such as Korean.

Turkish	English
kork(-mak)	(to) fear
korku	fear
korkusuz	fearless
korkusuzlaş (-mak)	(to) become fearless
korkusuzlaşmış	One who has become fearless
korkusuzlaştır(-mak)	(to) make one fearless
korkusuzlaşstırıl(-mak)	(to) be made fearless
korkusuzlaştırılmış	One who has been made fearless
korkusuzlaştırılabil(-mek)	(to) be able to be made fearless
korkusuzlaştırılabilecek	One who will be able to be made fearless
korkusuzlaştırabileceklerimiz	Ones who we can make fearless
korkusuzlaştırabileceklerimizden	From the ones who we can make fearless
korkusuzlaştırabileceklerimizdenmiş	I gather that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesine	As if that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesineyken	when it seems like that one is one of those we can make fearless

2.1 Corpora, Tokens, and Types

- Tokenization
  - To tokenize tweets like the following, hashtags, mentions, smileys like :-)

    and URLs must be recognized as a single unit.
  - The criteria for tokenization differ for each case,

    and can have a greater impact on accuracy than expected
  - Most open-source NLP packages provide basic tokenization

```
import spacy
nlp = spacy.load('en_core_web_sm')
text = "Mary, don't slap the green witch"
print([str(token) for token in nlp(text.lower())])
```

```
['mary', ',', 'do', "n't", 'slap', 'the', 'green', 'witch']
```

2.1 Corpora, Tokens, and Types

- Token types

  - Unique tokens that appear in a corpus

  - The set of all types in a corpus is called a vocabulary or lexicon.

  - Words are classified into content words and stop words.

  - Stop words, such as articles and prepositions,

    are used mostly for grammatical purposes to supplement content words.

그래 단어는 알겠는데 공부해보니 여러
복합단어들이 있네? 이런건 어떻게 처리하지?

단어를 붙여보자

2.2 Unigrams, Bigrams, Trigrams, ..., N-grams

- N-grams
  - A sequence of fixed-length (n) consecutive tokens in text.
  - A unigram consists of one token, while a bigram consists of two tokens.
  - If a partial word conveys useful information itself, then character n-grams can be generated.
    · For example, the suffix "-ol" in "methanol" indicates a type of alcohol.

```python
def n_grams(text, n):
    '''
    takes tokens or text, returns a list of n-grams
    '''
    return [text[i:i+n] for i in range(len(text)-n+1)]

cleaned = ['mary', ',', "n't", 'slap', 'green', 'witch', '.']
print(n_grams(cleaned, 3))
```

```
[['mary', ',', "n't"], [',', "n't", 'slap'], ["n't", 'slap', 'green'], ['slap', 'green', 'witch'], ['gre
en', 'witch', '.']]
```

단어를 정의하다 보니 단어가 너무많아.. 기계가 다 기억하기가 어렵데

단어를 간소화 시켜보자

2.3 Lemmas and Stems

- Lemma (표제어)

  - Lemmas are root forms of words (

  - The verb 'fly', can be transformed into several words depending on the inflection,

    such as 'flow', 'flew', 'flies', 'flown', 'flowing', etc. → “fly” is the lemma of the above words

  - Lemma extraction is a technique of reducing the dimensionality

    of vector representation by replacing tokens with their lemma.

  - Stemming is a reduction technique used instead of lemma extraction, which cuts off

ex. ‘geese’ → 표제어 추출 : ‘goose’, 어간 추출 : ‘gees’

2.3 Lemmas and Stems

- Stemming

  - It a reduction technique used instead of lemma extraction

  - It cuts off the end of a word using manually created rules to reduce it to a common form called a stem.

  - For example, 'geese' would become 'goose' with lemma extraction, or 'gees' with stemming.

2.3 Lemmas and Stems

- An example of lemmatization

  - question: what is the lemma of “has”?

Input[0]

```
import spacy
nlp = spacy.load('en')
doc = nlp(u"he was running late")
for token in doc:
    print('{} --> {}'.format(token, token.lemma_))
```

Output[0]

```
he --> he
was --> be
running --> run
late --> late
```

단어에서 문장으로 확장해보자

단어는 배웠는데 각 단어의 문장 내 역할을 정확히 알아야 기계가 사람의 명령을 이해 할것같아!

2.5 Categorizing Words: POS Tagging

- Token Classification

  - The concept of assigning labels to documents can be extended to words or tokens.

  - An example of a word classification task is part-of-speech tagging.

```
1 import spacy
2 nlp = spacy.load('en_core_web_sm')
3 doc = nlp(u"Mary slapped the green witch.")
4 for token in doc:
5     print('{} - {}'.format(token, token.pos_))
```

```
Mary - PROPN
slapped - VERB
the - DET
green - ADJ
witch - NOUN
. - PUNCT
```

2.6 Chunking and Named Entity Recognition

- Chunking, Partial Parsing

  - Assigning labels to text segments that are distinguished by several consecutive tokens.

  - E.g.) A sentence “Mary slapped the green witch” can be distinguished noun and verb phrases

    [NP Mary] [VP slapped] [the green witch]

  - The aim is to derive higher-level units composed of grammatical elements

    such as nouns, verbs, and adjectives

  - It also includes recognizing named entities

    such as people, places, companies, and drug names.

각 단어의 문장 내 역할은 알겠는데 아직 문장의 의미를 파악하기엔 부족해..

문장의 의존구조를 분석해 의미를 파악해보자

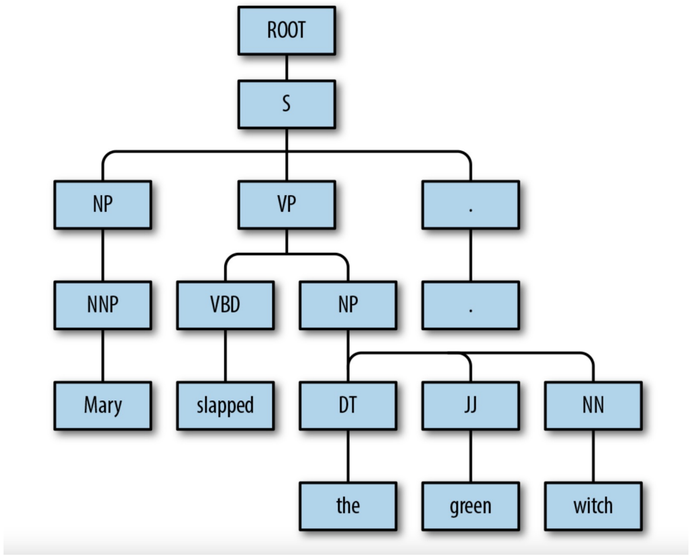
2.7 Structure of Sentences

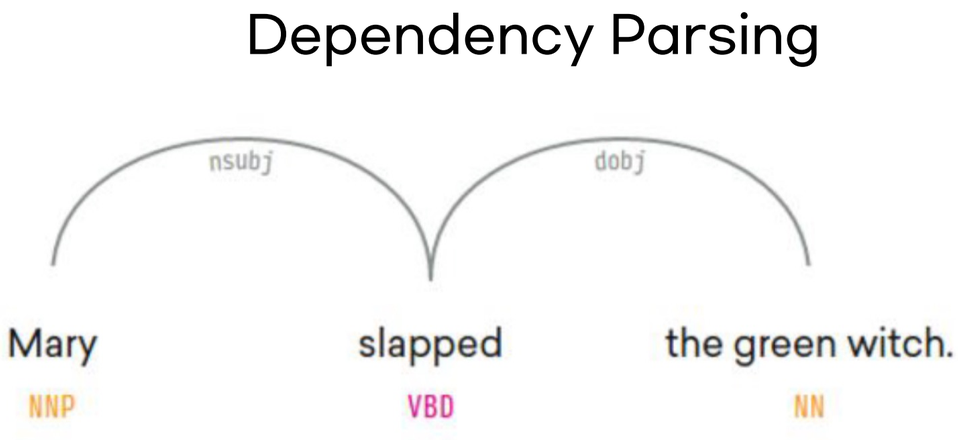
- Parsing

  - Unlike partial parsing that identifies clause-level units, parsing involves understanding the relationship between clauses.

  - It is possible to draw a diagram by analyzing a sentence.

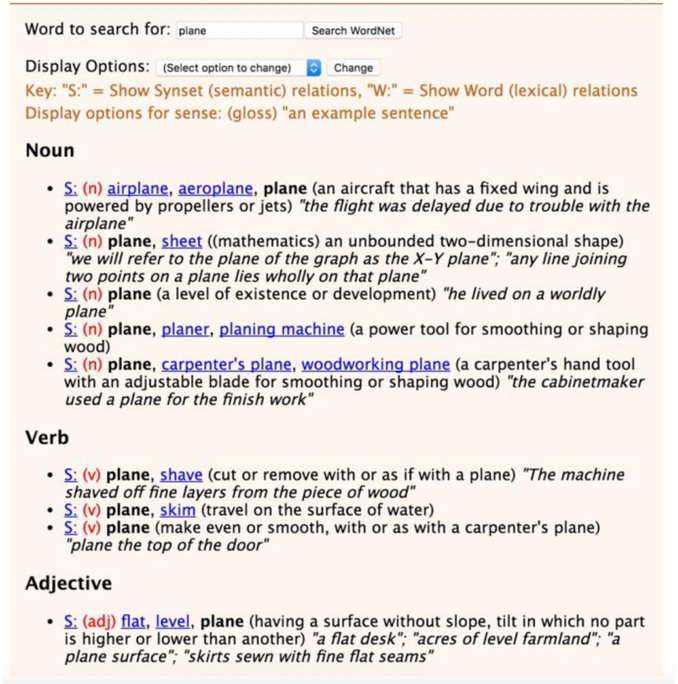
  - A parsing tree shows how grammatical elements in a sentence are hierarchically related.





2.8 Word Senses and Semantics

- Meaning of Word: the meaning represented by each word.
- WordNet
  - A vocabulary project being conducted at Princeton University.
  - The goal is to collect the relationships and meanings of almost all English words.
  - It is useful even with modern approaches.



Word to search for: plane Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) airplane, aeroplane, **plane** (an aircraft that has a fixed wing and is powered by propellers or jets) *"the flight was delayed due to trouble with the airplane"*
- S: (n) **plane**, sheet ((mathematics) an unbounded two-dimensional shape) *"we will refer to the plane of the graph as the X-Y plane"; "any line joining two points on a plane lies wholly on that plane"*
- S: (n) **plane** (a level of existence or development) *"he lived on a worldly plane"*
- S: (n) **plane**, planer, planing machine (a power tool for smoothing or shaping wood)
- S: (n) **plane**, carpenter's plane, woodworking plane (a carpenter's hand tool with an adjustable blade for smoothing or shaping wood) *"the cabinetmaker used a plane for the finish work"*

Verb

- S: (v) **plane**, shave (cut or remove with or as if with a plane) *"The machine shaved off fine layers from the piece of wood"*
- S: (v) **plane**, skim (travel on the surface of water)
- S: (v) **plane** (make even or smooth, with or as with a carpenter's plane) *"plane the top of the door"*

Adjective

- S: (adj) flat, level, **plane** (having a surface without slope, tilt in which no part is higher or lower than another) *"a flat desk"; "acres of level farmland"; "a plane surface"; "skirts sewn with fine flat seams"*

2.4 Categorizing Sentences and Documents

- Document Classification

  - One of the early applications in the field of NLP.

  - TF and TF-IDF representations are useful for classifying large text chunks like documents or sentences.

  - It is a supervised learning-based problem that includes:

    topic labeling,

    sentiment prediction of reviews,

    spam email filtering,

    language identification

Traditional chatGPT was

Q: 세종대왕이 다른 이름은 뭐야?

- if(“다른” and “이름” in question_token_list):

  if(PERSON in “document_token_list.NER):

    …

Sejong the Great

66 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

"Sejong" redirects here. For the city, see Sejong City. For other uses, see Sejong (disambiguation).
"Sejong of Joseon" redirects here. Not to be confused with Sejo of Joseon.

Sejong of Joseon (15 May 1397 - 8 April 1450), personal name **Yi Do** (Korean: 이도; Hanja: 李祹), widely known as **Sejong the Great** (Korean: 세종대왕; Hanja: 世宗大王), was the fourth ruler of the Joseon dynasty of Korea. Initially titled **Grand Prince Chungnyeong** (Korean: 충녕대군; Hanja: 忠寧大君), he was born as the third son of King Taejong and Queen Wongyeong. In 1418, he was designated as heir after his eldest brother, Crown Prince Yi Je, was stripped of his status. Today, King Sejong is regarded as one of the greatest leaders in Korean history.

Despite ascending to the throne after his father's voluntary abdication in 1418, Sejong was a mere figurehead; Taejong continued to hold the real power and govern the country up until his death in 1422. Sejong was the sole monarch for the next 28 years, although after 1439 he became increasingly ill,[2] and starting from 1442, his eldest son, Crown Prince Yi Hyang (the future King Munjong), acted as regent.

Text	Lemma	POS	Tag	Dep
text	text	NOUN	NN	nsubj
=	=	PUNCT	.	punct
"	"	PUNCT	``	punct
Sejong	Sejong	PROPN	NNP	appos
of	of	ADP	IN	prep
Joseon	Joseon	PROPN	NNP	pobj
(	(	PUNCT	-LRB-	punct
15	15	NUM	CD	nummod
May	May	PROPN	NNP	appos
1397	1397	NUM	CD	appos
–	–	PUNCT	:	punct
8	8	NUM	CD	nummod
April	April	PROPN	NNP	appos
1450	1450	NUM	CD	nummod
)	)	PUNCT	-RRB-	punct
,	,	PUNCT	,	punct
personal	personal	ADJ	JJ	amod
name	name	NOUN	NN	conj
Yi	Yi	PROPN	NNP	compound
Do	Do	PROPN	NNP	appos
(	(	PUNCT	-LRB-	punct
Korean	korean	ADJ	JJ	amod

Text	NER
Sejong	PERSON
Joseon	GPE
15 May 1397	DATE
8 April 1450	DATE
Yi	PERSON
Korean	NORP
Hanja	PERSON
Sejong the Great	PERSON
Korean	NORP
Hanja	PERSON
fourth	ORDINAL
the Joseon dynasty	DATE
Korea	GPE

Thank you!