Natural language processing Bible

Basic

서수원

Business Intelligence Lab.

산업경영공학과, 명지대학교

01

텍스트의 전처리

비정형 데이터 내의 오류

- 비정형 데이터 내의 오류
  - 비정형데이터
    - ✓ 일정한 규격이나 형태를 지니지 않고 구조화가 되어있지 않은 데이터를 뜻한다.
      - ❖ Ex) 음성정보, 시각정보등

- 비정형 데이터의 오류를 수정하는 과정을 전처리 과정이라고 한다.

구 분	내 용	종 류
정형 데이터	고정된 필드에 저장된 데이터	데이터베이스 스프레드 시트
반정형 데이터	고정된 필드에 저장되어 있지는 않지만 메타 데이터나 스키마를 포함하는 데이터	XML HTML
비정형 데이터	고정된 필드에 저장되어 있지 않은 데이터	텍스트 문서 이미지/동영상 등

텍스트 문서의 변환

- 텍스트 문서의 변환
  - 파일로부터 필요한 텍스트를 추출하는 것이 전처리의 첫번째 단계라고 한다.
  - 문서파일이나 웹사이트를 크롤링 하면 형식에 따라 사람이 내용을 파악할 수 없게 읽히게 된다.
  - 따라서 문서파일을 문서로 바꾸는 작업이 필요하다.(필요한 어휘만 남을 수 있도록 한다.)

- 텍스트 문서의 변환 방법
  - 특수문자를 제거한다.
  - 필요 없는 코딩이나 커맨드를 제거한다.
  - 문장의 경계를 인식한다.

텍스트 문서의 변환

ackground-size:20px;height:20px;padding:10px;width:20px}.AB4Wff{margin-left:16px}.AaVjTc a:link{display:block;color:#4285f4;font-weig
ht:normal}.AaVjTc td{padding:0;text-align:center}.YyVfkd{color:#202124;font-weight:normal;}.AaVjTc{margin:30px auto 30px}.SJajHc{back
ground:url(/images/nav_logo321.webp) no-repeat;overflow:hidden;background-position:0 0;height:40px;display:block}.NYbCr{cursor:pointe
r}</style><div class="g Ww4FFb tF2Cxc" lang="ko" style="width:652px" data-hveid="CD4QAA" data-ved="2ahUKEwjjweySluP4AhXdplYBHWrUD28QF
SgAegQIPhAA"> <div class="Z26q7c jGGQ5e YGXe8" data-header-feature="0" style
="grid-area:x5WNvb"> <a href="https://www.hankyung.com/entertainment/article/202205300591H" data-ved="2ahUKEwjjwey
SluP4AhXdplYBHWrUD28QFnoECCcQAQ" ping="/url?sa=t&source=web&rct=j&url=https://www.hankyung.com/entertainment/article/2022
05300591H&ved=2ahUKEwjjweySluP4AhXdplYBHWrUD28QFnoECCcQAQ">
<h3 class="LC20lb MBeuO DKV0Md">한지민보다 더 예쁘다는 친언니…인
기 많았다 - 한국경제</h3> <cite class="iUh30 qLRx3b tjvcx" role="text">https://www.hankyung.com<span class
="dyjrff qzEoUe" role="text"> › entertainment › article</cite> <cite class="i
Uh30 qLRx3b tjvcx" role="text">https://www.hankyung.com › entertainment › article</cit
e> <g-
popup jsname="V68bde" jscontroller="DPreE" jsaction="A05xBd:lYtByb;EOZ57e:WFrRFb;" jsdata="mVjAjf;_;ALzCdw"><div jsname="oYxtQd" clas
s="rlbAWc" aria-expanded="false" aria-haspopup="true" role="button" tabindex="0" aria-label="검색결과 옵션" jsaction="WFrRFb;keydown:
uYT2Vb"> </div
> </g-po
pup> <div class="Z26q7c" data-content-feature="0,1" style="margin-left:12px;grid-area:Yjbam;widt

In [63]:
```
# <h3 class="LC20lb MBeuO DKV0Md">
# 한지민보다 더 예쁘다는 친언니…인기 많았다 - 한국경제
# </h3>

re.findall(r'<h3 class="LC20lb [^"]+?">([^<]+?)</h3>', # Lazy
           resp.text)
```

Out[63]:
```
['한지민 - 나무위키',
 '한지민 - 위키백과, 우리 모두의 백과사전',
 '한지민보다 더 예쁘다는 친언니…인기 많았다 - 한국경제',
 '한지민 | 다음영화',
 '한지민 - Facebook',
 '#한지민 hashtag on Instagram • Photos and Videos',
 '한지민',
 '[인터뷰]&#39;한지민 쌍둥이 언니&#39;役 정은혜 &quot;꿈은 다 이뤘어요&quot;']
```

띄어쓰기 교정

- 실생활에서는 띄어쓰기가 제대로 되어 있지 않은 경우가 많다.
  - 따라서 문장의 가독성 및 의미혼용 방지를 위해 띄어쓰기를 교정하는 것이 필요하다

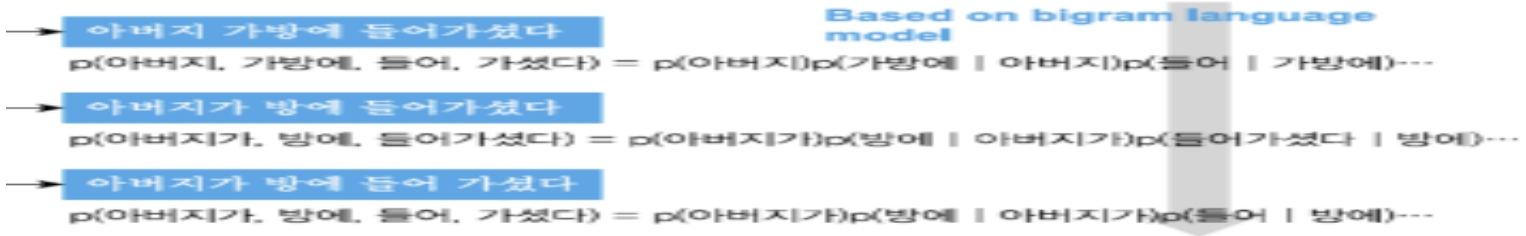
.

- 띄어쓰기 교정 방법
  - 규칙기반
    - ✓ 형태소 분석기를 사용하는 규칙기반의 분석법이다.
      - ❖ 장점 : 높은 정확도
      - ❖ 단점 : 한 규칙은 한 답변에서만 사용 가능하다.
        모든 경우의 수를 고려해서 규칙을 만들 수 없다.(높아지는 비용 및 유지보수 문제가 존재한다)

그림 4-4 규칙기반 띄어쓰기 교정 사용 예시

그림 4-5 규칙기반 띄어쓰기 교정기의 한계

  - 통계,확률기반
    - ✓ 언어 모델링 방법을 사용해 학습 말뭉치 내에서 수정 방향이 옳을 확률이 가장 높은 후보로 교정을 한다.
      - ❖ 장점 : 구현이 쉽고, 미등록어에 대해서도 분석이 가능하다
      - ❖ 단점 : 코퍼스의 영향을 크게 받음으로, 정확도 및 오류율이 높으며, 대량의 데이터가 필요하다.



철자 및 맞춤법 교정

- 철자교정
  - 띄어쓰기 교정과 유사하다.
    - ✓ 의미혼용 방지 및 정보전달을 위해 반드시 필요하다.
  - 철자 교정을 위해 '맞춤법 검사'를 시행 한다.
    - ✓ 텍스트 내 오류감지는 형태소 분석기를 이용한다.
- 오류를 감지하기 위해서는 어떤 오류들이 나타날 수 있는지 파악해야 한다.
- 오타로 인해 발생할 수 있는 오류
  - 삽입 : "뭐해?"를 "뭐핼?"처럼 추가 문자를 입력하는 오류이다.
  - 생략 : "안녕"을 "안녀" 처럼 문자를 생략하는 오류이다.
  - 대체 : "안녕"을 "안영"처럼 다른 문자를 입력하는 오류이다.
  - 순열 : "안녕"을 "녕안" 처럼 철자 순서가 뒤바뀌어 있는 오류이다.

.

철자 및 맞춤법 교정 방법

- 규칙기반 철자 및 맞춤법 교정방법
- 띄어쓰기 교정기의 규칙기반 방법과 장단점이 유사하다.
  - 어절은 어절보다 작은 단위(형태소)의 결합으로 구성된다.
  - 어절을 형태소로 분절하는 '형태소 분석기'를 사용하는 방식이 존재한다.

- 확률기반 절차 및 맞춤법 교정방법
  - Bayesian inference model
    - ✓ 베이즈 정리에 입각한 베이지안 추론이다.
    - ✓ 올바른 교정결과를.도출하기 위해 주어진 단어로부터 오타가 일어날 확률을 확률적으로 계산한다.(조건부확률)

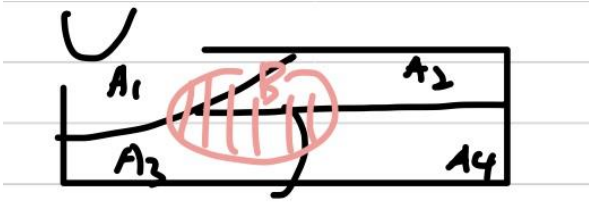
철자 및 맞춤법 교정 방법

- 베이즈 정리

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)$$
$$= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)$$

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \cdot P(A_j)}, \quad for\ all\ n,\ m\ s.t\ n \neq m,\ P(A_n \cap A_m) = 0$$



- 베이지안 추론

$$f(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{f(x)} = \frac{1}{f(x)} \pi(\theta) f(x|\theta) \quad f(x) = \int f(x|\theta)\pi(\theta)d\theta \ , \ where \iint f(x|\theta)\pi(\theta)dxd\theta = 1$$

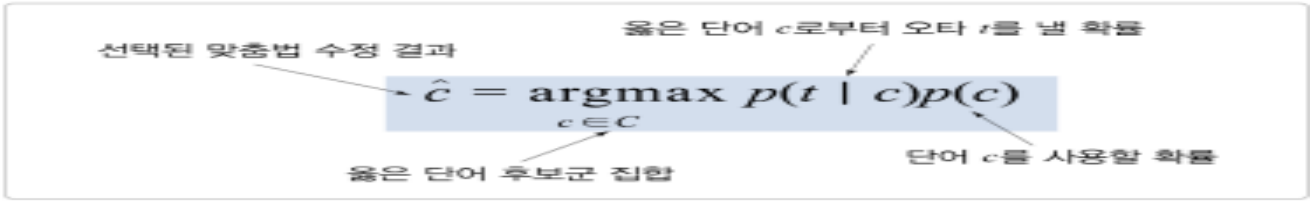


그림 4-9 Bayesian inference model 수식 설명

02

어휘 분석

형태소 분석

- 어휘 분석
  - 단어의 구조를 식별하고 분석을 통한 품사에 관한 단어수준의 연구이다.
- 형태소 분석
  - 형태소를 자연어의 제약 조건과 문법 규칙에 맞춰 분석 하는 것을 의미한다.

"컴퓨터를" = "컴퓨터" + "를" •••••••••••• (O) 각각 의미를 지님

"컴퓨터" =? "컴", "퓨", "터" •••••••••••• (X) 최소한의 의미를 지니지 않음

- 형태소 분석 절차
  - 단어에서 최소 의미를 포함하는 형태소 후보로 분리한다.
  - 형태론적 변형이 일어난 형태소를 원형으로 복원한다.
  - 단어와 사전들 사이의 결합 조건에 따라 옳은 분석 후보를 선택한다.

형태소 분석 절차

- 단어에서 최소 의미를 포함하는 형태소 후보로 분리
  - 형태소 분석의 처리 대상인 어절은 하나 이상의 형태소가 연결된 것이다.
  - 이를 형태소열 (Sequence of Morphemes)라고 부르기도 한다.

한국어(Korean)는 = 한국어 + (+ Korean +) + 는

그림 5-1 형태소열 예시

- 단어에서 최소 의미를 포함하는 형태소 후보로 분리
  - 형태소가 연결될 때, 형태소의 변형이 일어나기에, 형태소 원형의 복원이 필요하다.

나는 = 나 + 는

나는 = 날 + 는

형태소 분석 절차

- 형태론적 변형이 일어난 형태소의 원형 복원 및 형태소품사쌍 생성
  - 형태소는 하나 이상의 품사를 가질 수 있으므로, 형태소는 하나 이상의 형태소-품사 쌍으로 표현된다.
  - 형태소와 품사를 쌍으로 나타낸 것을 형태소품사쌍 이라고 한다.
    - ✓ 형태소품사쌍의 예시
    - ✓ 나 – (나_대명사),(나_명사),(나_동사),(나_보조용언)

- 단어와 사전들 사이의 결합 조건에 따라 옳은 분석 후보를 선택
  - “나는”에 대한 형태소품사쌍 열 후보군 중 선택한다.

“나_대명사 + 는_조사”	“나_대명사 + 는_어미”
“나_일반명사 + 는_조사”	“나_일반명사 + 는_어미”
“나_동사 + 는_조사”	“나_동사 + 는_어미”
“나_보조용언 + 는_조사”	“나_보조용언 + 는_어미”
“날_동사 + 는_조사”	“날_동사 + 는_어미”

그림 5-4 “한국어”의 형태소품사쌍열

한국어 형태소 분석 라이브러리

- 한국어 형태소 분석기의 오픈 라이브러리
  - KoNLPy - KUKoLex(고려대), 한나눔(Hannanum), 코모란(komoran), 미캡(mecab), 꼬꼬마(Kkma)
  - 각각 기준,성능,시간이 다르다.
  - 데이터에 맞는 분석기 활용능력이 필요하다.

| 표 5-2 | 형태소 분석기 라이브러리 별 결과[5]

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에들어가/N	아버지/NNG	아버지가방에들어가신다/NNP	아버지/NNG	아버지/Noun
이/J	가방/NNG		가/JKS	가방/Noun
시ㄴ다/E	에/JKM		방/NNG	에/Josa
	들어가/VV		에/JKB	들어가신/Verb
	시/EPH		들어가/VV	다/Eomi
	ㄴ다/EFN		신다/EP+EC	

품사 태깅

- 품사 태깅
  - 품사는 단어의 기능, 형태, 의미에 따라 나눈 것을 말한다.
  - 태깅이란 같은 단어에 대해 의미가 다를 경우(중의성이 존재) 해결하기 위한 부가 정보를 장착 하는 것을 말한다.

```
pprint(kkma.pos(a))

문장을 입력하세요:세종대왕님은 글을 만드셨습니다.
[('세종', 'NNG'),
 ('대왕', 'NNG'),
 ('님', 'XSN'),
 ('은', 'JX'),
 ('글', 'NNG'),
 ('을', 'JKO'),
 ('만들', 'VV'),
 ('시', 'EPH'),
 ('었', 'EPT'),
 ('습니다', 'EFN'),
 ('.', 'SF')]
```

Using KoNLPy Model

- 품사 태깅 접근법
  - 규칙 기반 접근법
  - 통계 기반 접근법
  - 딥러닝 기반 접근법

품사 태깅 접근법

- 규칙 기반의 접근법
  - 언어 정보에서 생성되는 규칙 형태를 적용하여 태깅을 수행한다.
  - 품사 사이 관계 외의 어절에 대해 높은 정확도를 나타내기 때문에, 통계 기반 접근법으로 다루지 못하는 부분에 대해 교정이 가능하다.
  - 긍정, 부정, 수정 정보를 이용하여 중의성을 해결하고 태깅을 부착하는 방법이다.
    - ✓ 긍정 정보 : 문장에서 선호되는 어휘 태그에 대한 언어 지식이다.
    - ✓ 부정 정보 : 문장에서 배제되는 어휘 태그에 대한 언어 지식이다.
    - ✓ 수정 정보 : 오류 교정 및 잘못된 정보 입력 시 수정될 정보에 대한 지식이다.

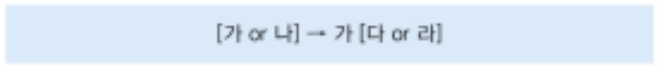


그림 5-8 긍정 정보 규칙 표현

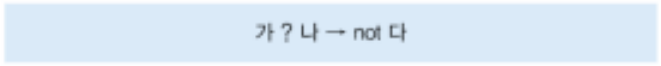


그림 5-9 부정 정보 규칙 표현

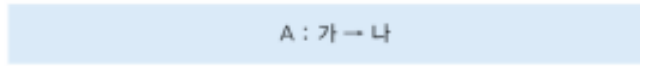


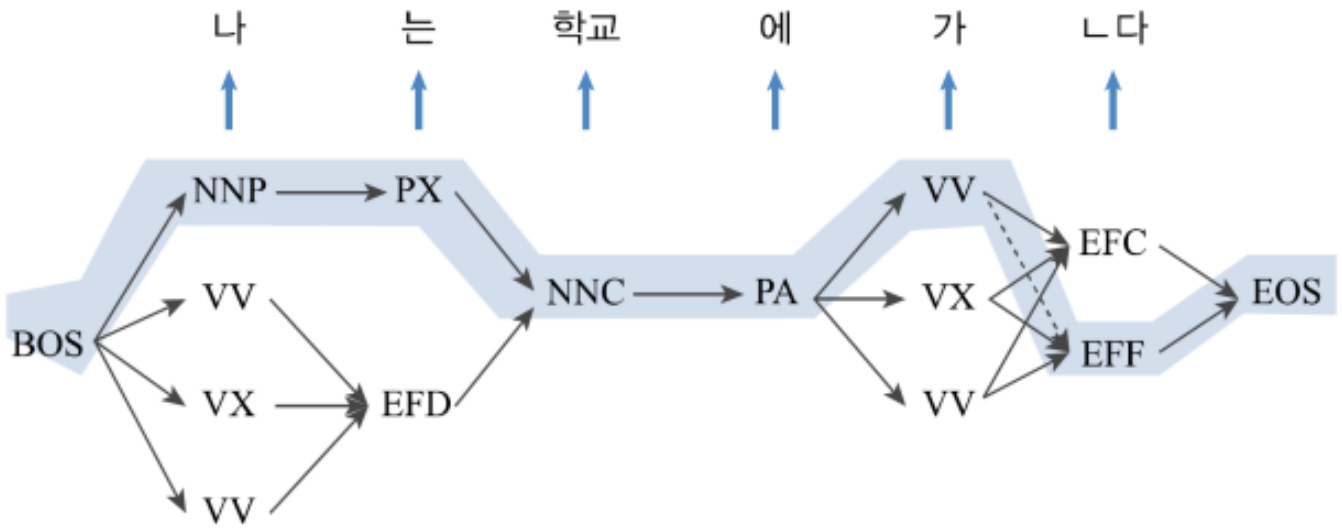
그림 5-10 수정 정보 규칙 표현

품사 태깅 접근법

- 통계 기반의 접근법(Hidden Markov Model)
  - 태그가 부착된 코퍼스 중 적합한 모델을 선정하고 코퍼스에서 추출된 정보를 이용한다.
  - 대량의 코퍼스에 태그가 부착되어야하는 단점이 있지만 주어지면 정보추출이 용이하고, 자동 추출이 가능하다.
  - 대표적인 방법은 어휘 확률을 이용하는 은닉 마르코프 모델(Hidden Markov Model)이 존재한다.

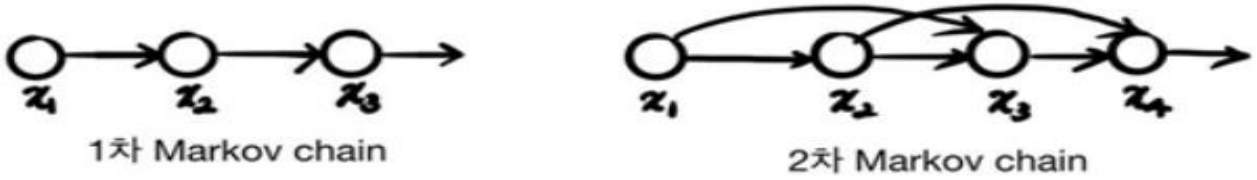
품사 태깅 접근법

- Hidden Markov Model(HMM)
  - 주어진 문장에서 형태소의 품사 태그 정보를 숨긴채로 확률 정보를 이용하여 가장 가능성이 높은 경로를 찾는다.



품사 태깅 접근법

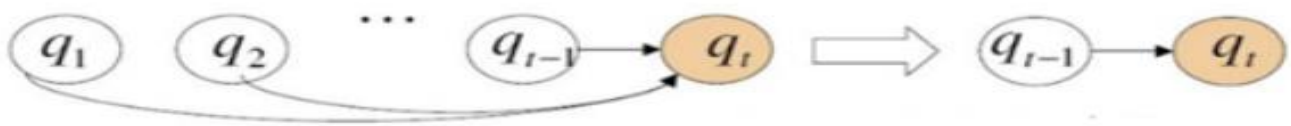
- 마르코프 모델(Morkov Model)
  - 현재의 observation은 바로 이전의 state에 의해서만 결정된다.



  - 예) 날씨예측 문제 (state : 날씨, observation : {rainy,cloudy,sunny},가정 : 오늘 날씨는 어제 날씨에만 영향을 받는다.)

1st order Markov assumption

$$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \cdots) = P(q_t = j \mid q_{t-1} = i)$$



품사 태깅 접근법

- 마르코프 모델(Morkov Model)

		Tomorrow		
		Rainy	Cloudy	Sunny
Today	Rainy	0.4	0.3	0.3
	Cloudy	0.2	0.6	0.2
	Sunny	0.1	0.1	0.8

- HMM
  - 우리가 알고싶은 state(상태) 가 숨겨져 있는 상황을 의미한다.
    - ✓ 날씨 예측을 하고 싶은데 과거 날씨를 알 수 없는 상황 이라면, 사람들이 우산을 가지고 다니는지 아닌지로 observation(관측치)를 구할 수 있다.
  - HMM에는 3가지 문제가 존재
    - ✓ 확률평가 문제
    - ✓ 최적의 상태열을 찾는 문제
    - ✓ 파라미터 추정의 문제

품사 태깅 접근법

- HMM(최적의 상태열을 찾기)
  - 하나의 최적 상태열의 경로를 찾는 방법은 사후확률P(q|O,λ)을 최대화 하는 것을 경로로 설정할 수 있다.
    - ✓ 사후확률을 forward-backward변수로 정의한다.

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \frac{P(O, q_t = S_i \mid \lambda)}{P(O \mid \lambda)} = \frac{P(O, q_t = S_i \mid \lambda)}{\sum_{i=1}^{N} P(O, q_t = S_i \mid \lambda)}$$

    - ✓ 이는 forward prob.와 backward prob.의 곱과 같다. 따라서 밑과 같이 정의 가능하다.

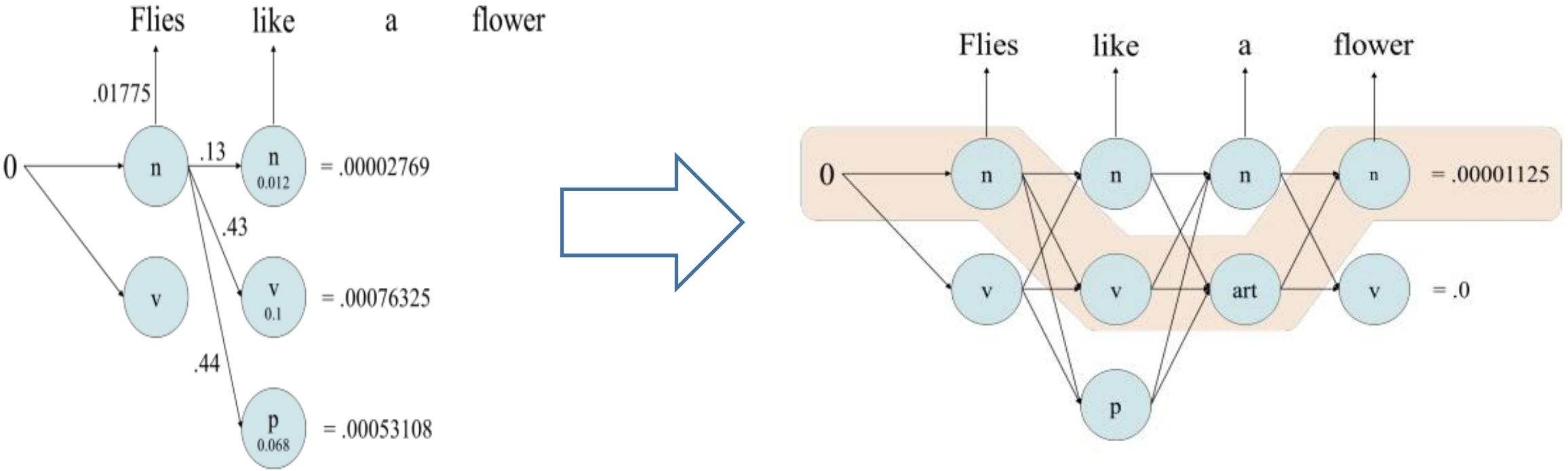
$$\gamma_t(i) = \frac{P(O, q_t = S_i \mid \lambda)}{\sum_{i=1}^{N} P(O, q_t = S_i \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

    - ✓ 이를 최대화 하는 상태를 찾아 모든 t에 대해 모으면 최적의 상태열을 구할 수 있다.

$$q_t^* = \arg\max\left[\gamma_t(i)\right] \quad ※ \ \forall t = 1 \ldots T$$

품사 태깅 접근법

- HMM(최적의 상태열을 찾기)



품사 태깅 접근법

- 딥러닝 기반의 접근법
  - 데이터로부터 특징을 자동으로 학습한다.
  - 폭넓은 문맥 정보를 다룰 수 있다.
  - 모델에 적합한 출력을 다루기가 간단하다.

03

실습

한국어 형태소 분석

```
# konlpy 관련 패키지 import
from konlpy.tag import Okt
from konlpy.tag import Kkma
from konlpy.tag import Hannanum
from konlpy.tag import Komoran
from konlpy.tag import Twitter
```

```
kkma = Kkma()
okt = Okt()
komoran = Komoran()
hannanum = Hannanum()
twitter = Twitter()
```

```
# konlpy 중 Kkma는 문장 분리가 가능 (다른 라이브러리는 되지 않음)
print ("kkma 문장 분리 : ", kkma.sentences('네 안녕하세요 반갑습니다.'))
```

```
kkma 문장 분리 :  ['네 안녕하세요', '반갑습니다.']
```

```
# konlpy 의 라이브러리 형태소 분석 비교
print("okt 형태소 분석 :", okt.morphs(u"집에 가면 감자 좀 쪄줄래?"))
print("kkma 형태소 분석 : ", kkma.morphs(u"집에 가면 감자 좀 쪄줄래?"))
print("hannanum 형태소 분석 : ", hannanum.morphs(u"집에 가면 감자 좀 쪄줄래?"))
print("komoran 형태소 분석 : ", komoran.morphs(u"집에 가면 감자 좀 쪄줄래?"))
print("twitter 형태소 분석 : ", twitter.morphs(u"집에 가면 감자 좀 쪄줄래?"))
```

```
okt 형태소 분석 : ['집', '에', '가면', '감자', '좀', '쪄줄래', '?']
kkma 형태소 분석 :  ['집', '에', '가', '면', '감자', '좀', '찌', '어', '주', 'ㄹ래', '?']
hannanum 형태소 분석 :  ['집', '에', '가', '면', '감', '자', '좀', '찌', '어', '줄', '래', '?']
komoran 형태소 분석 :  ['집', '에', '가', '면', '감자', '좀', '찌', '어', '주', 'ㄹ래', '?']
twitter 형태소 분석 :  ['집', '에', '가면', '감자', '좀', '쪄줄래', '?']
```

한국어 품사태깅

```
# konlpy 관련 패키지 import
from konlpy.tag import Okt
from konlpy.tag import Kkma
from konlpy.tag import Hannanum
from konlpy.tag import Komoran
from konlpy.tag import Twitter
```

```
kkma = Kkma()
okt = Okt()
komoran = Komoran()
hannanum = Hannanum()
twitter = Twitter()
```

```
# konlpy 의 라이브러리 품사태깅 비교
print("okt 품사태깅 :", okt.pos(u"집에 가면 감자 좀 쪄줄래?"))
print("kkma 품사태깅 : ", kkma.pos(u"집에 가면 감자 좀 쪄줄래?"))
print("hannanum 품사태깅 : ", hannanum.pos(u"집에 가면 감자 좀 쪄줄래?"))
print("komoran 품사태깅 : ", komoran.pos(u"집에 가면 감자 좀 쪄줄래?"))
print("twitter 품사태깅 : ", twitter.pos(u"집에 가면 감자 좀 쪄줄래?"))
```

```
okt 품사태깅 : [('집', 'Noun'), ('에', 'Josa'), ('가면', 'Noun'), ('감자', 'Noun'), ('좀', 'Noun'), ('쪄줄래', 'Verb'), ('?', 'Punctuation')]
kkma 품사태깅 :  [('집', 'NNG'), ('에', 'JKM'), ('가', 'VV'), ('면', 'ECE'), ('감자', 'NNG'), ('좀', 'MAG'), ('찌', 'VV'), ('어', 'ECS'), ('주', 'VXV'), ('ㄹ래', 'EFQ'), ('?', 'SF')]
hannanum 품사태깅 :  [('집', 'N'), ('에', 'J'), ('가', 'P'), ('면', 'E'), ('감', 'P'), ('자', 'E'), ('좀', 'M'), ('찌', 'P'), ('어', 'E'), ('줄', 'P'), ('래', 'E'), ('?', 'S')]
komoran 품사태깅 :  [('집', 'NNG'), ('에', 'JKB'), ('가', 'VV'), ('면', 'EC'), ('감자', 'NNP'), ('좀', 'MAG'), ('찌', 'VV'), ('어', 'EC'), ('주', 'VX'), ('ㄹ래', 'EF'), ('?', 'SF')]
twitter 품사태깅 :  [('집', 'Noun'), ('에', 'Josa'), ('가면', 'Noun'), ('감자', 'Noun'), ('좀', 'Noun'), ('쪄줄래', 'Verb'), ('?', 'Punctuation')]
```