Lexical Analysis
(어휘 분석)

Morphological Analysis (형태소 분석)

역할 및 필요성

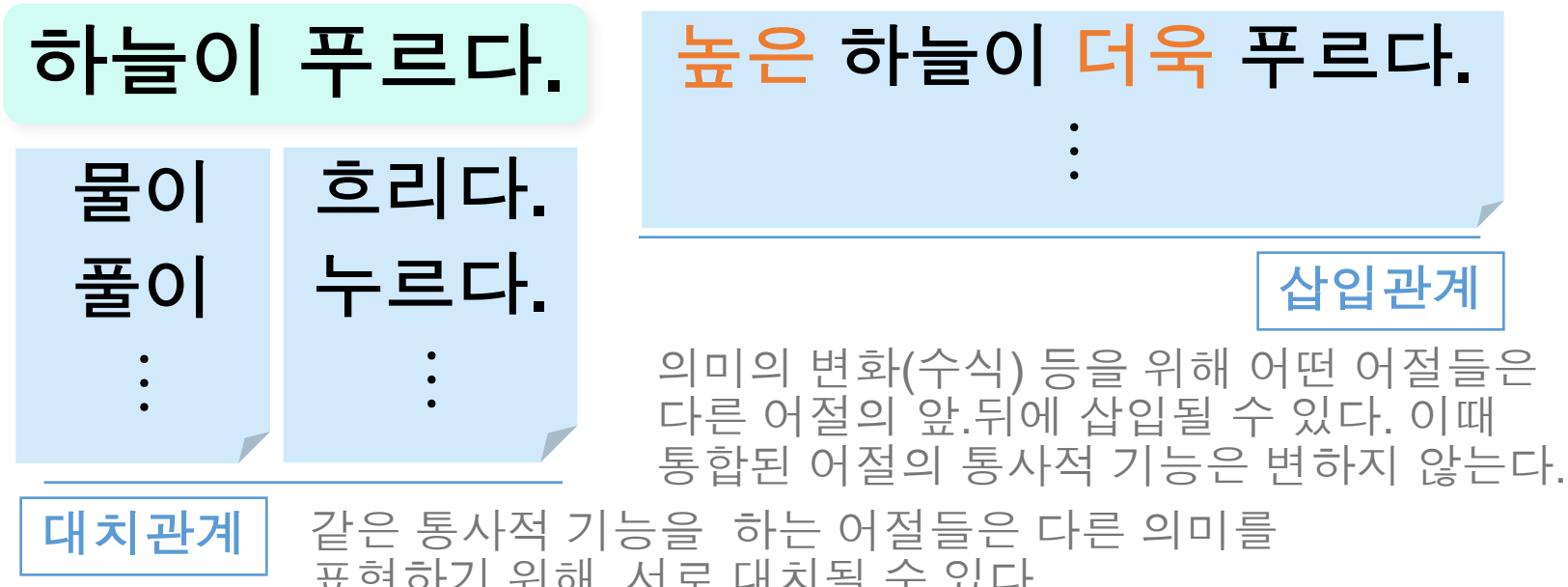
- 역할
  - 어절의 가능한 모든 형태소 분석열을 생성
  - 한국어 정보처리를 위한 필수적 역할

나는	수염이	나는	이유를	물었다
나/N+는/J 나/V+는/E 날/V+ㄴ/E	수염/N+이/J	나/N+는/J 나/V+는/E 날/V+ㄴ/E	이유/N+를/J	물/V+었/PE+다/E 묻/V+었/PE+다/E

- 필요성
  - 문장 이해를 위해서 어절에 대한 정보가 필요함
  - 생성 가능한 어절의 수가 무한하여 어절 단위로 전자 사전을 구성하는 것이 불가능
  - 형태소 단위의 분석과 분석 결과의 조합을 통하여 어절의 정보를 분석할 수 있음
  - 신조어(미등록어) 분석

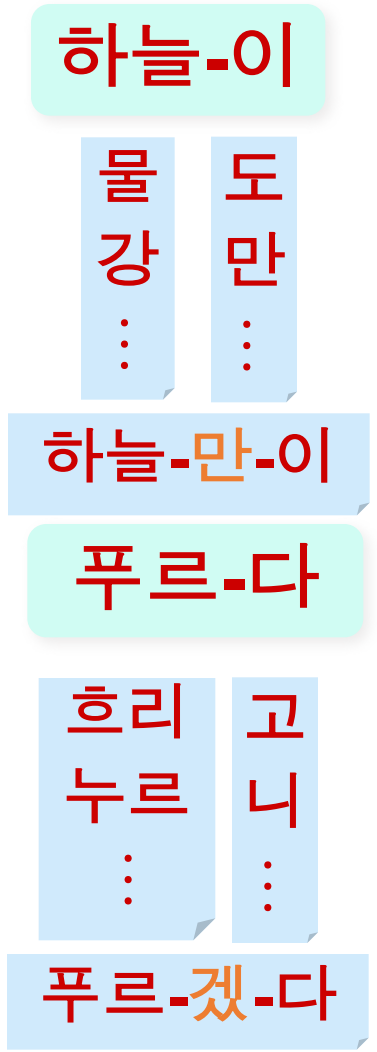
어절, 형태소, 단어 (1/4)

- 어절(Eojeol)
  - 문장 구성의 단위:
    - **대치관계**와 **삽입관계**에 의해 마디 지어지는 한 덩어리의 말
  - 띄어쓰기 단위:
    - 공백에 의해 구별되는 문자열



어절, 형태소, 단어 (2/4)

- 형태소(Morpheme)
  - 더 이상 분해될 수 없는 최소의 뜻 단위
    - 분해하면 의미를 잃어버림.
  - 어절의 구성 요소
    - 대치관계와 삽입관계에 의해 분리되는 문자열.
- 형태(Morph)와 이형태(Allomorph)
  - **형태**: 형태소가 어절로 실현될 때의 모습.
    - 하늘-이, 철수-가 (유종성, 무종성)
    - 읽-었-다, 가-았-다(>갔다), 하-였-다 (음성모음, 양성모음, ?)
  - **이형태**: 하나의 형태소가 환경에 따라 모습을 달리할 때 그것들을 그 형태소의 이형태라 함.
    - 이/가 : 음운론적 환경으로 제약된 이형태
    - 었/았/였 : 형태론적 환경으로 제약된 이형태



어절, 형태소, 단어 (3/4)

- 형태소의 갈래: 자립성의 여부
  - 자립(free)형태소: 자립성이 있어 독립하여 사용될 수 있는 형태소.
    예) 하늘, 철수, ...
  - 의존(bound)형태소: 자립성이 없고 다른 말에 의존해야만 함.
    예) 푸르, 읽, 이, 가, 를, 다, 었, ...
- 형태소의 갈래: 의미의 허실
  - 실질(full)형태소: 구체적인 대상이나 동작, 상태와 같은 어휘적 의미를 표시 = 어휘(lexical)형태소
    예) 하늘, 철수, 푸르, 읽, ...
  - 형식(empty)형태소: 실질형태소에 붙어 말과 말 사이의 관계나 기능을 형식적으로 표시 =문법(grammatical)형태소
    예) 이, 가, 를, 다, 었, ...

어절, 형태소, 단어 (4/4)

- 단어(Word)
  - 자립하여 쓰일 수 있는 말의 단위, 최소 자립 형식
  - 자립성과 분리성에 의해 구별되는 문자열
    - 자립할 수 있는 말이나 자립형태소와 쉽게 분리되는 말.
    - 앞 뒤에 숨의 끊어짐(휴지, pause)가 올 수 있는 말.
  - 홀로 자립하는 말
    - 체언(명사, 대명사, 수사), 수식언(관형사, 부사), 독립언(감탄사)
  - 자립 형태소와 쉽게 분리되는 말
    - 관계언(조사)
  - 의존 형태소끼리 어울려서 자립하는 말
    - 용언(동사, 형용사)

철수-가 동화-를 잘 읽었다.

형태론이란? (1/2)

- 언어학적 관점
  - 형태론(Morphology)은 형태소와 형태소 배열(Arrangement)에 대한 연구.
- 전산학적 관점
  - 표층 정보와 어휘층 정보 사이의 대응 관계에 대한 연구.
    - 두-층 대응(Two-level mapping)
  - 표층(surface-level) 정보: 어절에 해당하는 문자열
  - 어휘층(lexical-level) 정보: 형태소에 대한 정보의 결합

표층 정보	어휘층 정보	
flies	fly+s	Noun+Pl
	fly+s	Verb+3P
나는	나(I)+는	대명사+조사
	나(sprout)+는	동사어간+어미
	날(fly)+는	동사어간+어미

형태론이란? (2/2)

- 계산 형태론(Computational Morphology)
  - 전산학적 관점에서의 형태론
  - 자연어처리(Natural Language Processing)의 기초
  - 계산 언어학(Computational Linguistics)의 한 분야

- **형태론(Morphology)** : 어절과 형태소의 관계
- **구문론(Syntax)** : 문장과 어절의 관계
- **의미론(Semantics)** : 하나의 문장(형태소, 단어, 어절, 구, 절등)의 의미
- **담화론(Discourse)** : 하나 이상의 문장들(예, 문단)의 의미
- **화용론(Pragmatics)** : 문장(발화)들이 나타내는 실제적 의미(의도)

형태론의 두 주요 쟁점

- 형태자소론(Morphographemics)
  - 두 형태소의 경계에서 발생하는 철자의 변형

  예) get+s = gets, go+s = goes, try+s = tries
    - 영어 동사의 3인칭 단수형

  예) 막(block)+은 = 막은, 갈(grind)+은 = 간, 가(go)+은 = 간
    - 한국어 동사의 과거시제 관형형
- 형태통사론(Morphotactics)
  - 어떤 형태소 배열이 올바른 어절을 형성하는가?

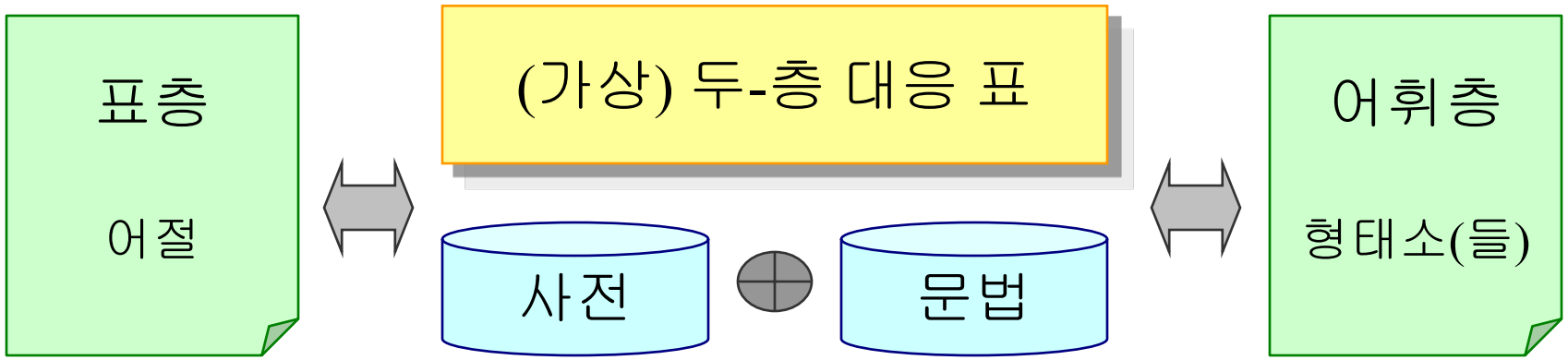
  예) fragment-al, *employment-al
    - 영어 명사(-ment)의 형용사화(-al)

  예) 동작(action)+ 하+다, *정보(information)+하+다

  예) 완전(completeness)+하+다, *결과(result)+하+다
    - 한국어 명사의 동사화, 형용사화

사전과 문법

- 사전(Lexicon)
  - 형태소에 대한 정보의 집합
    - 형태(form): 형태소의 어휘 문자열, 형태자소론적 자질 정보
    - 기능(function): 품사(part-of-speech), 형태통사론적 자질 정보
    - 의미(meaning): 의미(sense), 형태의미론적 자질 정보
- 문법(Grammar)
  - 형태자소규칙과 형태통사규칙을 포함하는 규칙의 집합



왜 형태소 분석이 필요한가?

- 정보량의 감소
  - 규칙 활용하는 동사의 모든 형태는 규칙에 의해 계산될 수 있기 때문에 사전에 저장할 필요가 없음
- 어절 등록의 편리성
  - 어절의 모든 가능한 형태를 사전에 등록할 필요 없이 단지 그 어절을 구성하는 형태소들만 등록
  - 예) 우리는, 우리들은, 나무는, 나무들은, 하늘은, …
    - 우리, 는, 들, 은, 나무, 하늘, ...
- 신조어(Neologism)의 인식
  - 새로 만들어진 신조어를 사전과 문법에 의해서 쉽게 인식할 수 있다.
  - 예) reprogrammability = re+program+able+ity
  - 예) 한국청년봉사대 = 한국+청년+봉사대

형태소 분석과 생성 (1/7)

- 형태소 분석 (Morphological Analysis)
  - 표층 어절을 어휘층 형태소로 분해하는 작업
  - 표층에서 어휘층으로의 대응

  예) ‘나는’의 형태소 분석

  1. 나_대명사 + 는_조사
  2. 나_동사 + 는_어미
  3. 날_동사 + 는_어미
- 형태소 생성 (Morphological Generation)
  - 어휘층 형태소들로부터 표층 어절을 합성(생성)하는 작업
  - 어휘층에서 표층으로의 대응 (형태소 분석의 역과정)

  예) ‘분석_명사 + 하_동사화접미사 + 어_어미’의 형태소 합성

  1. 분석하여
  2. 분석해

형태소 분석과 생성 (2/7)

- 형태소 분석의 요소
  - 사전
    - 표제어(형태소) + 형태소 정보(품사, 자질)
  - 문법
    - 형태자소 규칙: 음운현상을 표현
    - 형태통사 규칙: 형태소 결합의 합법성을 표현
  - 분석 알고리즘
    - **전처리**: 한글 코드 변환, 한글 및 기호 분리
    - **형태자소 규칙 적용**: 형태 분리, 원형 복원
    - **사전 탐색**: 형태소 정보 부여, 미등록 형태소 추정
    - **어절의 형태소분석 정보 생성**: 형태소 정보 결합
    - **형태통사 규칙 적용**: 비합법적인 형태소 결합 제거

형태소 분석과 생성 (3/7)

- 한국어 형태소 분석의 과정
  - 1. 한글 코드 변환
    - 형태소변형을 처리하는 데는 조합형한글코드가 편리
    - 완성형, 조합형의 상호 변환
  - 2. 한글 및 기호 분리
    - 형태자소규칙은 한글문자열에만 적용되므로 한글문자열과 기타문자열(영문자, 숫자, 기호 등)을 분리한다.

    예) 클린턴(clinton)은 = 클린턴+(+clinton+)+은

    * (_기호 , clinton_영어 ,)_기호

    예) 나는 = 나는
  - 3. 형태 분리(morph segmentation)
    - 한글문자열을 형태로 단순 분리

    예) 클린턴 = 클린턴, 클린터+ㄴ ; 은 = 은, 으+ㄴ

    예) 나는 = 나+는, 나+느+ㄴ, 나느+ㄴ

형태소 분석과 생성 (4/7)

- 한국어 형태소 분석의 과정 (계속)
  - 4. 원형 복원(root form restoration)
    - 형태변형이 발생된 형태소의 원형을 복원

    예) (무종성) + ㄴ = ㄹ + ㄴ('ㄹ'탈락 규칙 현상)
    클린터+ㄴ = 클린털+ㄴ; 으+ㄴ = 을+ㄴ
    나+는 = 날+는; 나+느+ㄴ = 나+늘+ㄴ; 나느+ㄴ = 나늘+ㄴ
  - 5. 형태소 정보 부여 및 미등록 형태소 추정
    - 각 형태소의 품사와 자질 정보를 사전 탐색으로 부여하고, 사전에 등록되지 않은 형태소의 품사 및 자질 정보는 추정

- ㄴ_조사^어미
- 은_조사^어미^명사
- 는_조사^어미
- 을_조사^어미^명사
- 나_동사^대명사
- 날_동사^명사
- 늘_동사^부사

- 클린턴_명사?
- 클린터_명사?
- 클린털_비형태소
- 나늘_비형태소
- 나느_비형태소
- 으_비형태소
- 느_비형태소

형태소 분석과 생성 (5/7)

- 한국어 형태소 분석의 과정 (계속)
  - 6. 형태소 분석 결과 생성
    - 등록되거나 추정된 각 형태소의 정보를 결합하여 분석결과 생성

```
클린턴_명사?+(_기호+clinton_영어+)_기호+은_조사
클린턴_명사?+(_기호+clinton_영어+)_기호+은_어미
클린턴_명사?+(_기호+clinton_영어+)_기호+은_명사
클린터_명사?+ㄴ_조사+(_기호+clinton_영어+)_기호+은_조사
클린터_명사?+ㄴ_조사+(_기호+clinton_영어+)_기호+은_어미
클린터_명사?+ㄴ_조사+(_기호+clinton_영어+)_기호+은_명사
```

```
나_동사+는_조사              나_동사+는_어미
나_대명사+는_조사            나_대명사+는_어미
날_동사+는_조사              날_동사+는_어미
나_동사+늘_동사+ㄴ_조사      나_동사+늘_동사+ㄴ_어미
나_대명사+늘_동사+ㄴ_조사    나_대명사+늘_동사+ㄴ_어미
```

형태소 분석과 생성 (6/7)

- 한국어 형태소 분석의 과정 (계속)
  - 7. 비합법적인 형태소결합 제거
    - 형태통사규칙을 적용하여 비합법적인 형태소결합 제거

비합법적인 형태소 결합에 대한 형태통사규칙과 예

• 기호+어미	)_기호+은_어미
• 조사+기호	는_조사+(_기호
• 동사+조사	나_동사+는_조사, 날_동사+는_조사 늘_동사+ㄴ_조사
• 대명사+어미	나_대명사+는_어미
• 동사+동사	나_동사+늘_동사
• 대명사+동사	나_대명사+늘_동사

•클린턴_명사?+(_기호+clinton_영어+)_기호+은_조사
•클린턴_명사?+(_기호+clinton_영어+)_기호+은_명사

의미적으로 비합법적

•나_동사+는_어미
•나_대명사+는_조사
•날_동사+는_어미

형태소 분석에서 다루어야 할 현상들

- 부호, 영문자, 한자, 숫자 등의 분리
  - 가을인가/?, 1998/년
- 어절 내의 형태소 분리
  - 가을/이/ㄴ가
- 불규칙 활용 등의 음운현상 처리 및 원형 복원:
  - 고마우+어서= 고맙+어서, 하야+ㄴ=하얗+ㄴ
- 미등록 형태소 추정 : 클린턴_명사?
- 파생 접사 처리
  - 세계+화+하+자
- 복합어 처리
  - 선진+한국, 미래+지향, 분석+결과
- 미등록어(고유명사) 인식
  - 미항공우주연구소에서는 = **미+항공+우주+연구소**+에서+는
  - 예술의 전당에서 = **예술+의 전당**+에서

품사 태깅
(Part-of-Speech Tagging)

품사 태깅

- 품사 태깅이란?
  - 단어(어절)의 형태론적 중의성을 해소하여 올바른 품사를 할당하는 작업
  - 단어(어절) 중의성 해소 작업

나는	수염이	나는	이유를	물었다
나/N+는/J 나/V+는/E 날/V+ㄴ/E	수염/N+이/J	나/N+는/J 나/V+는/E 날/V+ㄴ/E	이유/N+를/J	물/V+었/PE+다/E 묻/V+었/PE+다/E

품사 태깅의 접근법

- 품사 태깅 접근법의 분류
  - 규칙 기반 접근법(rule-based approach)
    - 품사 태깅에 사용될 결정적 규칙을 이용
    - 지식기반 접근법(knowledge-based approach), 제약기반 접근법(constraint-based approach), 합리주의 접근법(rationalism)
  - 통계 기반 접근법(stochastic approach)
    - 코퍼스로부터 추출한 통계 정보를 이용
    - 경험주의 접근법(empricism), 데이타집약 접근법(data intensive approach), 코퍼스 기반 접근법(corpus-based approach)
  - 혼합 접근법(hybrid approach)

품사 태깅 접근법의 장단점

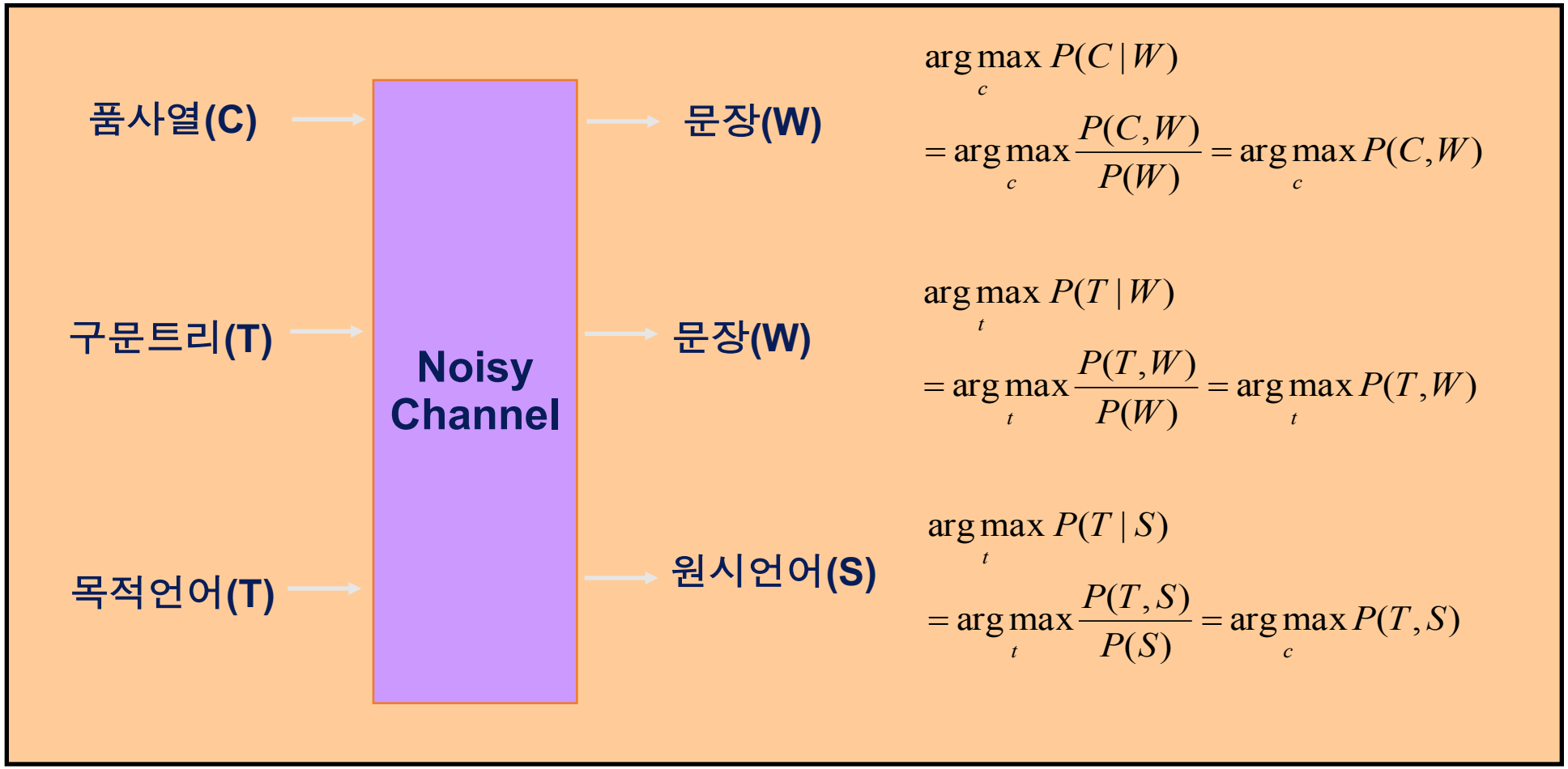
- 규칙 기반 접근법
  - 장점
    - 규칙이 적용되는 부분에 대해서 높은 정확도
    - 품사 태깅 결과를 이해/분석하기가 용이
    - 특정 코퍼스에 의존적이지 않은 정확도
  - 단점
    - 견고성이 떨어짐
    - 규칙 작성에 많은 시간과 노력 소모
    - 확장성이 떨어짐; 규칙의 개수가 증가 할수록 시소현상 발생

품사태깅 접근법의 장단점-계속

- 통계 기반 접근법
  - 장점
    - 시스템 구축이 용이; 자율 학습 또는 지도 학습
    - 넓은 적용 범위
    - 견고성, 확장성이 좋음
  - 단점
    - 규칙 기반 접근법과 비교하여 정확도가 떨어짐
      - 신뢰도가 떨어지는 확률 정보를 사용하기 때문
    - 태깅 결과의 이해/분석이 어려움
- 혼합 접근법
  - 규칙기반 접근법과 통계기반 접근법의 장단점
  - 태깅 시간/복잡도 증가

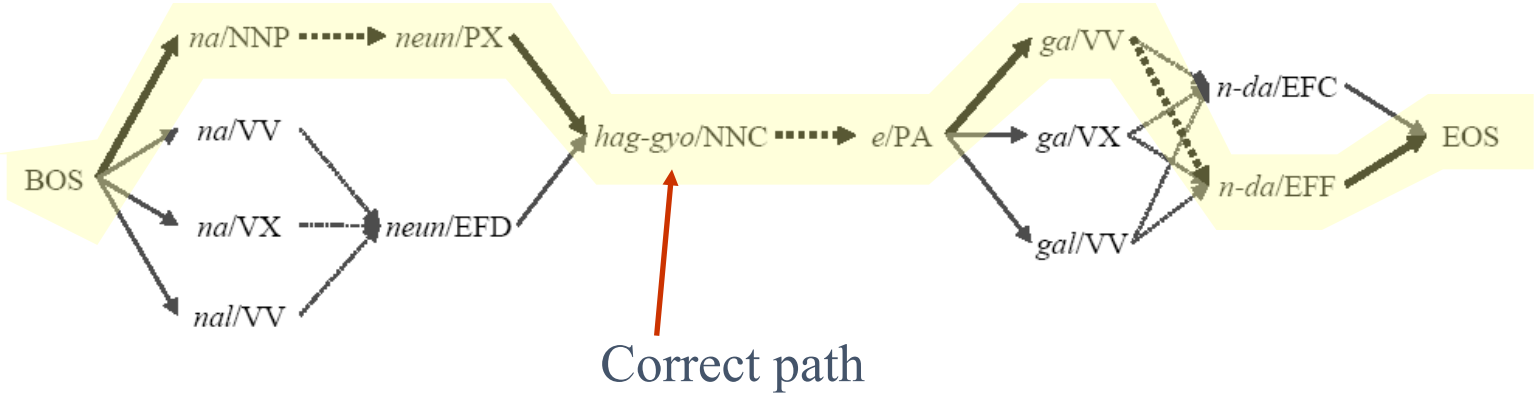
노이지 채널 모델

- 자연어처리에서의 노이지 채널 모델



POS Tagging 품사태깅 (1/9)

- Part of Speech (POS) tagging is
  - A task to assign a proper POS tag to each linguistic unit such as a word (in English), or a morpheme (in Korean) for a given sentence
    - An input of POS tagger is a morphological analysis result, and an output is a correct sequence of morpheme-POS pairs



POS Tagging 품사태깅 (2/9)

- Hidden Markov Model (HMM) based POS Tagging
  - Most popular and well-performed approach
    - Regard POS tags of morphemes in a given sentence as hidden states and find the most probable path in a lattice

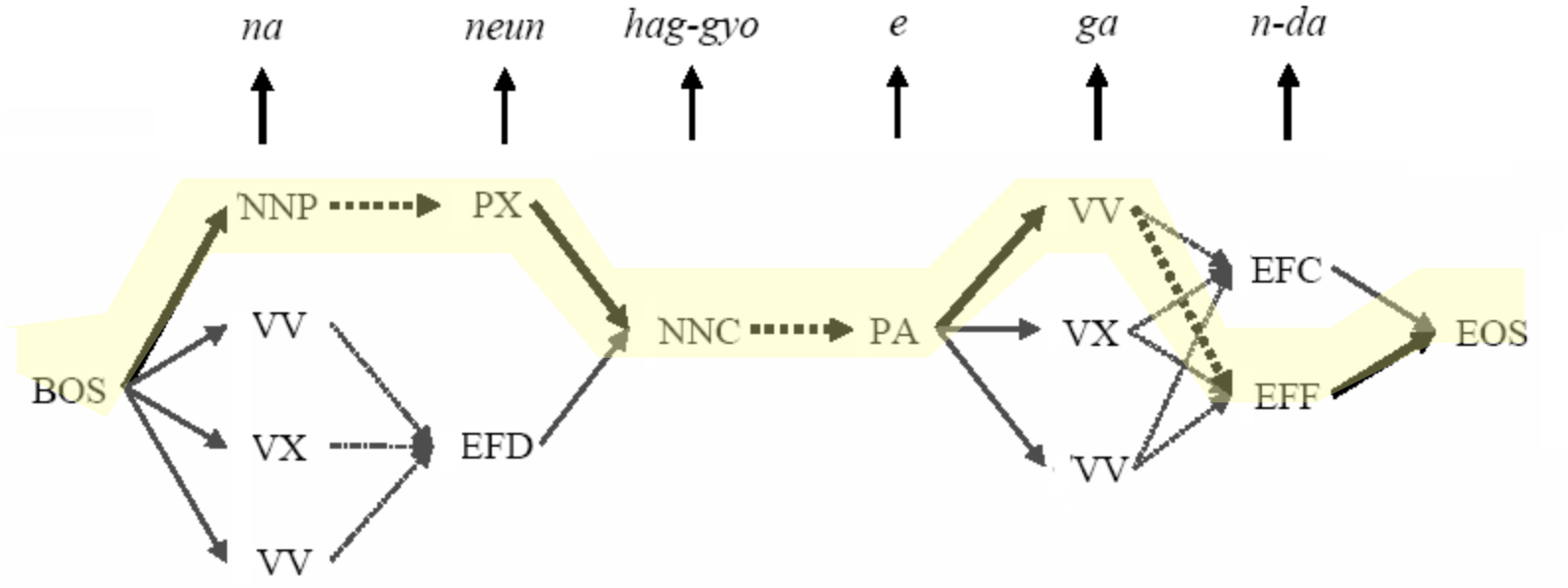


Figure: HMM representation for POS tagging problem

POS Tagging 품사태깅 (3/9)

- Example of a simple HMM method for POS tagging

$$p(W,T) = p(w_1 \ldots w_n, t_1 \ldots t_n)$$

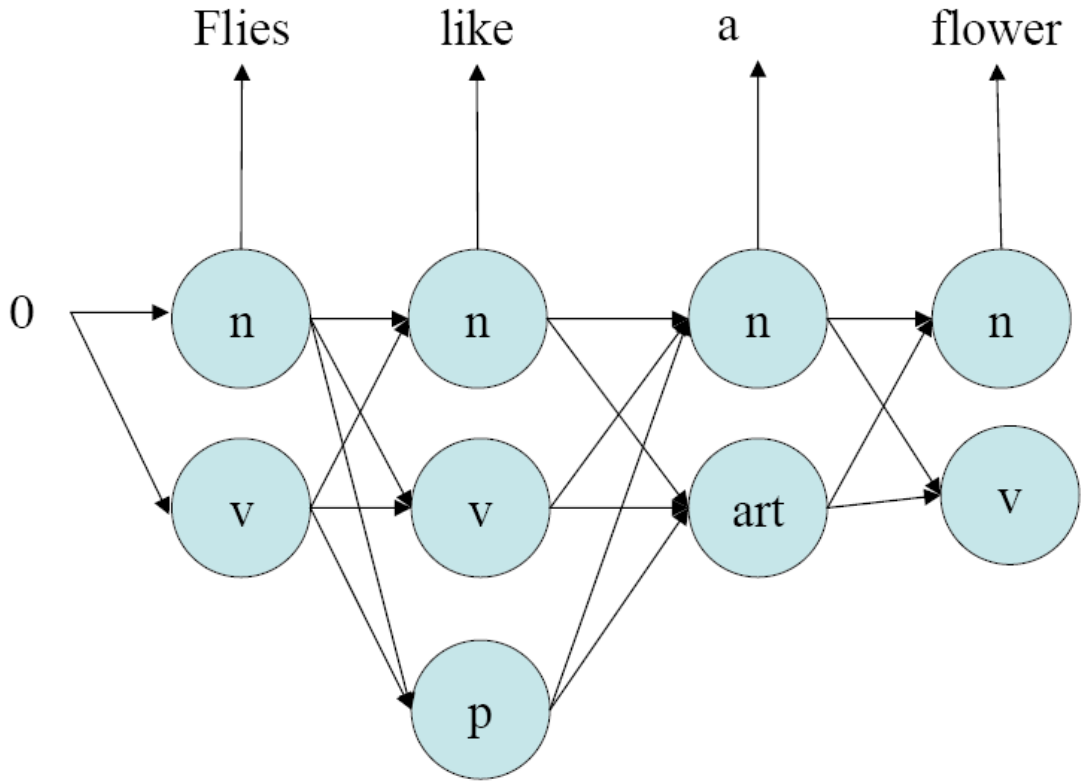
$$\approx \prod_{i=1}^{n} p(t_i \mid t_{i-1}) p(w_i \mid t_i)$$

Emission Probability

Transition Probability

POS Tagging 품사태깅 (4/9)

- Analyze the sentence "*Flies like a flower*"

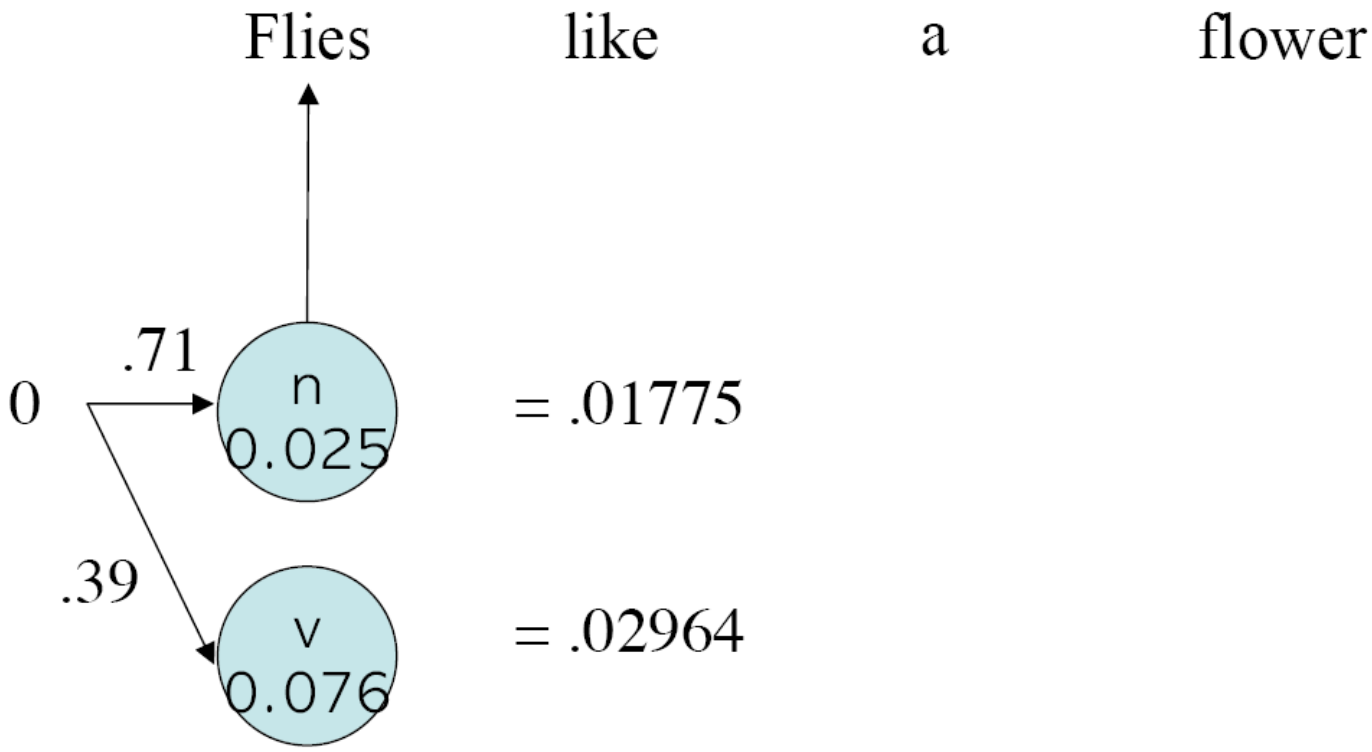


POS Tagging 품사태깅 (5/9)

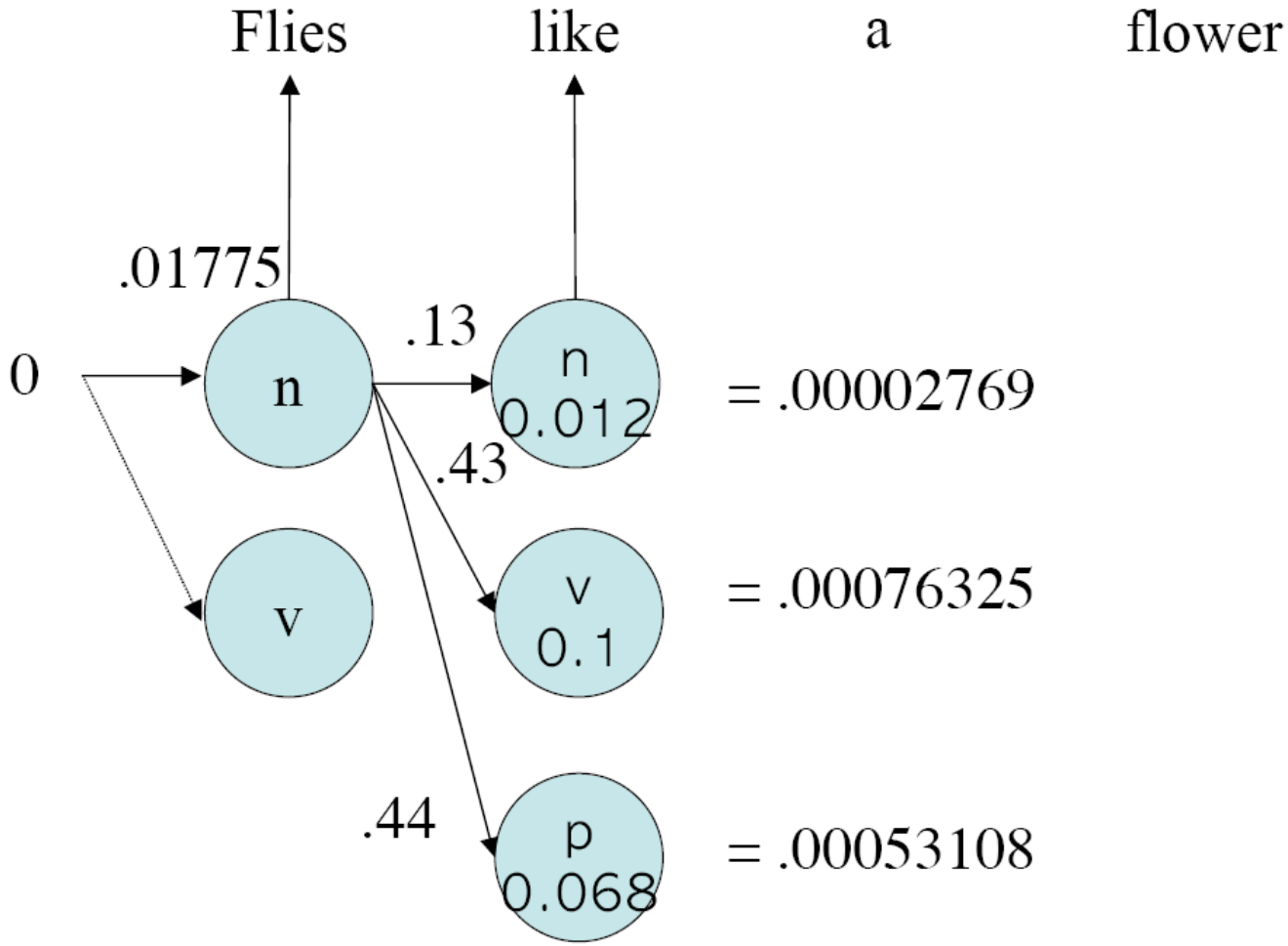
- Statistics

품사	#i	품사쌍	#i,i+1	Bigram	확률
ㅇ	300	ㅇ, ART	213	p(ART|ㅇ)	.71
ㅇ	300	ㅇ,N	87	p(N|ㅇ)	.29
ART	558	ART,N	558	p(N|ART)	1
N	833	N,V	358	p(V|N)	.43
N	833	N,N	108	p(N|N)	.13
N	833	N,P	366	p(P|N)	.44
V	300	V,N	75	p(N|V)	.35
V	300	V,ART	194	p(ART|V)	.65
P	307	P,ART	226	p(ART|P)	.74
P	307	P,N	81	p(N|P)	.26

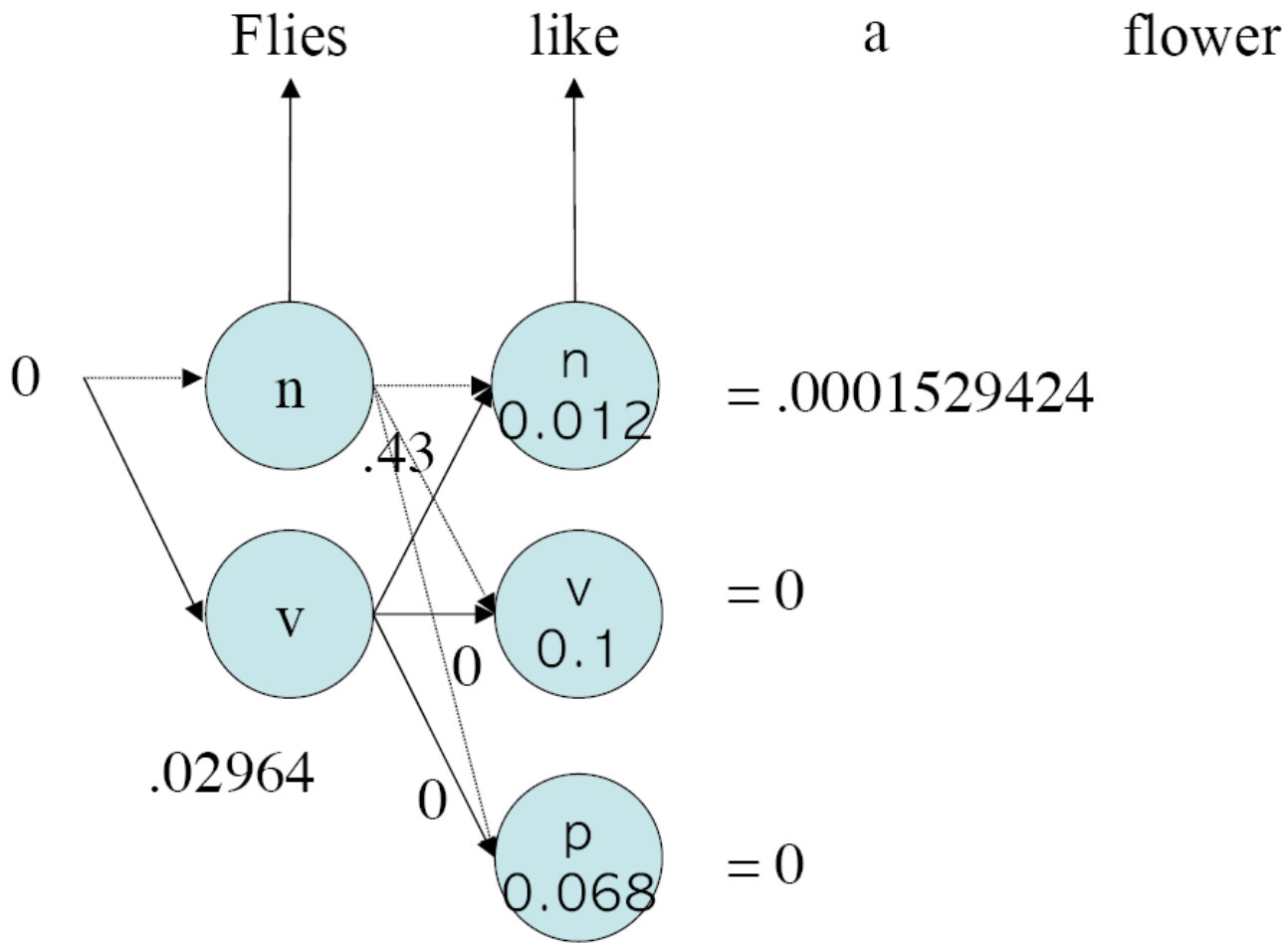
POS Tagging 품사태깅 (6/9)



POS Tagging 품사태깅 (7/9)

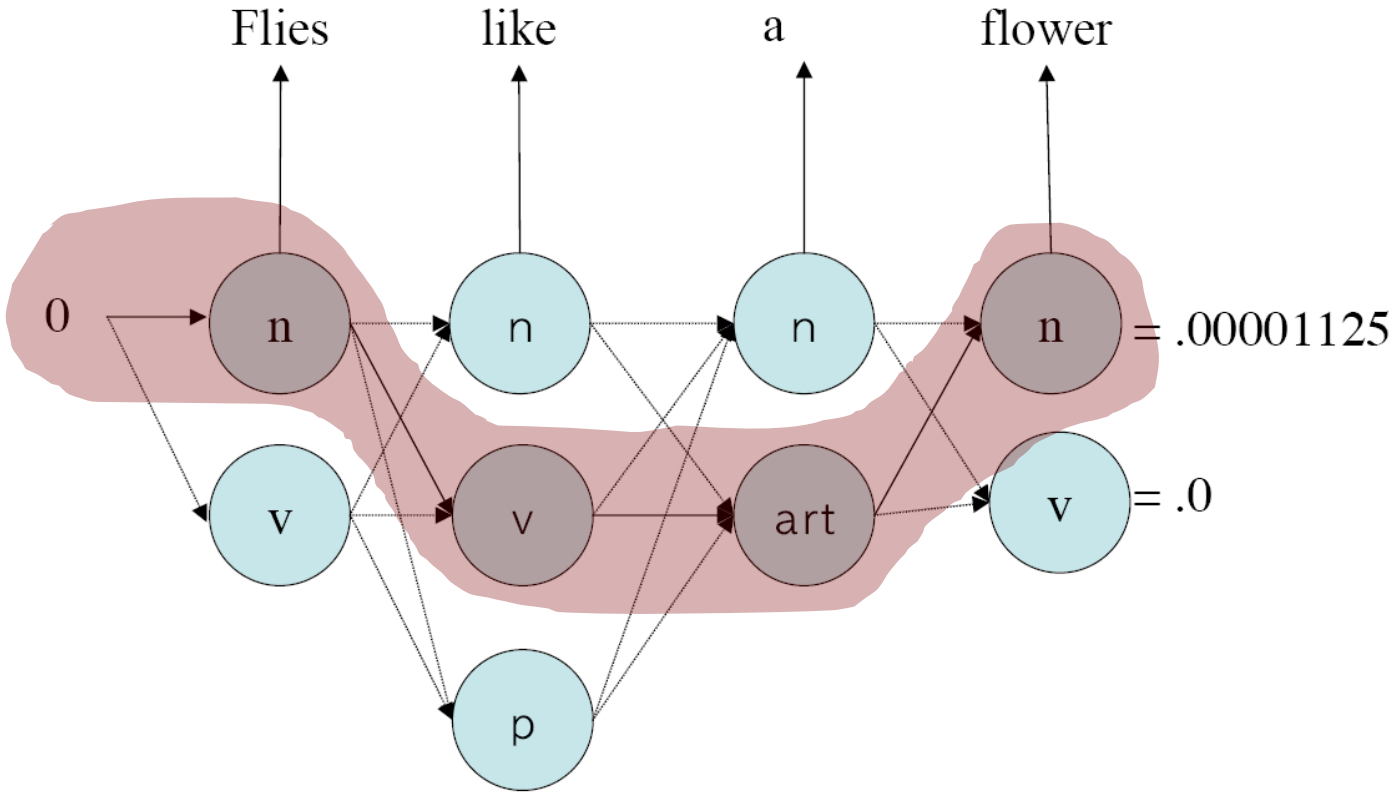


POS Tagging 품사태깅 (8/9)



POS Tagging 품사태깅 (9/9)

- Finally, we can find the POS sequence maximizing prob



품사 태깅의 확률적 모델링

- 품사 태깅 모델의 확률적 정의
  - 길이가 N인 문장(단어열) $W_N = w\ w \qquad w_N$ 이 주어졌을 때, 가장 확률이 높은 품사열 $c_{1,N} = c_1, c_2, \ldots, c_N$을 구하는 것

$$T(w_{1,N}) \overset{def}{=} \arg\max_{c_{1,N}} P(c_{1,N} | w_{1,N}) \qquad \text{식 1}$$

  - 조건부 확률의 정의에 의한 변형

모든 c1,N에 상관없이 동일하므로 제거

$$= \arg\max_{c_{1,N}} \frac{P(c_{1,N}, w_{1,N})}{P(w_{1,N})} \qquad \text{식 2}$$

$$= \arg\max_{c_{1,N}} P(c_{1,N}, w_{1,N}) \qquad \text{식 3}$$

품사 태깅의 확률적 모델링-계속

- Chain rule에 의한 변형(1) : w를 먼저 분리시킬 경우

$$
\begin{aligned}
P(c_{1,N}, w_{1,N}) &= P(w_1)P(c_1 \mid w_1) \\
&= P(w_2 \mid c_1, w_1)P(c_2 \mid c_1 w_{1,2}) \\
&\cdots \\
&= P(w_N \mid c_{1,N-1}, w_{1,N-1})P(c_N \mid c_{1,N-1} w_{1,N}) \qquad \text{식 4} \\
&= P(w_1)P(c_1 \mid w_1) \\
&\quad \prod_{i=2}^{N} P(w_i \mid c_{1,i-1,} w_{1,i-1})P(c_i \mid c_{1,i-1}, w_{1,i}) \qquad \text{식 5} \\
&= \prod_{i=1}^{N} P(w_i \mid c_{1,i-1,} w_{1,i-1})P(c_i \mid c_{1,i-1}, w_{1,i}) \qquad \text{식 6}
\end{aligned}
$$

W1,0을 문장시작기호로
c1,0을 품사열의 시작기호로 정의

품사 태깅의 확률적 모델링-계속

- Chain rule에 의한 변형(2) : c를 먼저 분리시킬 경우

$$P(c_{1,N}, w_{1,N}) = P(c_1)P(w_1 \mid c_1)$$

$$= P(c_2 \mid w_1, c_1)P(w_2 \mid w_1 c_{1,2})$$

$$\cdots$$

$$= P(c_N \mid w_{1,N-1}, c_{1,N-1})P(w_N \mid w_{1,N-1} c_{1,N}) \quad \text{식 7}$$

$$= P(c_1)P(w_1 \mid c_1)$$

$$\prod_{i=2}^{N} P(c_i \mid w_{1,i-1,} c_{1,i-1})P(w_i \mid w_{1,i-1}, c_{1,i}) \quad \text{식 8}$$

$$= \prod_{i=1}^{N} P(c_i \mid w_{1,i-1,} c_{1,i-1})P(w_i \mid w_{1,i-1}, c_{1,i}) \quad \text{식 9}$$

W1,0을 문장시작기호로
c1,0을 품사열의 시작기호로

품사 태깅의 확률적 모델링-계속

- 식 6과 식 9에 통계획득이 가능한 형태로 변형 함으로써 다양한 품사 태깅 모델을 유도

$$\prod_{i=1}^{N} P(w_i \mid c_{1,i-1}, w_{1,i-1}) P(c_i \mid c_{1,i-1}, w_{1,i}) \quad \text{식 6}$$

현실적으로 통계 정보 획득이 불가능

$$\prod_{i=1}^{N} P(c_i \mid c_{1,i-1}, w_{1,i-1}) P(w_i \mid c_{1,i}, w_{1,i-1}) \quad \text{식 9}$$

어휘 확률 기반 품사 태깅 모델

- 모델의 유도
  - [식 6]에 다음과 같은 마르코프 가정을 도입
  - [가정 1] : 현재 단어의 발생은 이전 단어에만 의존한다.
  - [가정 2] : 현재 단어의 품사는 현재 단어에만 의존한다.

$$P(w_i \mid c_{1,i-1,} w_{1,i-1}) \cong P(w_i \mid w_{i-1}) \qquad [가정1]$$

$$P(c_i \mid c_{1,i-1}, w_{1,i}) \cong P(c_i \mid w_i) \qquad [가정2]$$

어휘 확률 기반 품사 태깅 모델-계속

- [가정 1], [가정 2]를 [식 6]에 대입하고 그 결과를 [식 3]에 대입

$$\mathrm{T}(w_{1,N}) \overset{def}{=} \arg\max_{c_{1,N}} \prod_{i=1}^{N} P(w_i \mid w_{i-1}) P(c_i \mid w_i) \qquad [식10]$$

➡ [식 10]에서 $P(w_i \mid w_{i-1})$은 모든 $c_{1,N}$에 대해서 상수이므로 생략

$$\mathrm{T}(w_{1,N}) \overset{def}{=} \arg\max_{c_{1,N}} \prod_{i=1}^{N} P(c_i \mid w_i)$$

❑ **어휘 확률기반 품사 태깅 모델의 특징**

➡ 단어가 가장 빈번하게 사용된 품사를 그 단어의 품사로 결정

➡ 단어에 대한 품사 발생 정보만을 고려할 뿐 문맥 정보는 전현 고려하지 않음

HMM 기반 품사 태깅 모델

- 모델의 유도
  - [식 9]에 [가정 3]과 [가정 4]과 같은 마르코프 가정을 도입
  - [가정 3] : 현재 품사의 발생은 이전 품사에만 의존한다.
  - [가정 4] : 현재 단어의 발생은 현재 품사에만 의존한다.

$$P(c_i \mid c_{1,i-1,} w_{1,i-1}) \cong P(c_i \mid c_{i-1}) \qquad [가정3]$$

$$P(w_i \mid c_{1,i}, w_{1,i}) \cong P(w_i \mid c_i) \qquad [가정4]$$

➡ [가정 3]과 [가정 4]를 [식 9]에 대입하고, 그 결과를 [식 3]에 대입

$$\mathrm{T}(w_{1,N}) \overset{def}{=} \arg\max_{c_{1,N}} \prod_{i=1}^{N} P(c_i \mid c_{i-1}) P(w_i \mid c_i) \quad \textbf{[Bigram 모델]}$$

Named Entity Recognition

Named Entity Recognition

- What is NE?
- What isn't NE?
- Problems and solutions with NE task definitions
- Problems and solutions with NE task
- Some applications

Definition of NER

- **Named-entity recognition** (NER) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages

  - from Wikipedia

Why do NE Recognition?

- Key part of Information Extraction system
- Robust handling of entities(mostly proper names) essential for many applications
- Pre-processing for different classification levels
- Information filtering
- Information linking

NE Definition

- NE involves **identification** of *proper names* in texts, and **classification** into a set of predefined categories of interest.
- Three universally accepted categories: **person**, **location** and **organisation**
- Other common tasks: recognition of date/time expressions, measures (percent, money, weight etc), email addresses etc.
- Other domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

What NER is NOT

- NER is **not** event recognition.
- NER recognises **entities** in text, and classifies them in some way, but it does not create templates, nor does it perform co-reference or entity linking, though these processes are often implemented alongside NE as part of a larger IE system.
- NER is not just matching text strings with pre-defined lists of names. It only recognises entities which are being used as entities in a given context.
- NER is not easy!

Problems in NE Task Definition

- Category definitions are intuitively quite clear, but there are many grey areas.
- Person vs. Artefact
  - ✓ The **ham sandwich** wants his bill." vs "Bring me a **ham sandwich**."
- Organisation vs. Location
  - ✓"**England** won the World Cup" vs. "The World Cup took place in **England**".
- Company vs. Artefact
  - ✓ "shares in **MTV**" vs. "watching **MTV**"
- Location vs. Organisation
  - ✓ "she met him at **Heathrow**" vs. "the **Heathrow** authorities"

Solutions

- The task definition must be very clearly specified at the outset.
- The definitions adopted at the MUC conferences for each category listed guidelines, examples, counter-examples, and “logic” behind the intuition.
- MUC essentially adopted simplistic approach of disregarding metonymous uses of words, e.g. “England” was always identified as a location. However, this is not always useful for practical applications of NER.
- Idealistic solutions, on the other hand, are not always practical to implement, e.g. making distinctions based on world knowledge.

Basic Problems in NER

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
  - John Smith (company vs. person)
  - May (person vs. month)
  - Washington (person vs. location)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. “may”

List Lookup Approach

- System that recognises only entities stored in its lists (gazetteers).
- Advantages - Simple, fast, language independent, easy to retarget
- Disadvantages – collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Shallow Parsing Approach

- Internal evidence – names often have internal structure. These components can be either stored or guessed.

location:

CapWord + {City, Forest, Center}

e.g. Sherwood Forest

Cap Word + {Street, Boulevard, Avenue, Crescent, Road}

e.g. *Portobello Street*

Shallow Parsing Approach

- External evidence - names are often used in very predictive local contexts

Location:

"to the" COMPASS "of" CapWord

e.g. *to the south of* ***Loitokitok***

"based in" CapWord

e.g. *based in* ***Loitokitok***

CapWord "is a" (ADJ)? GeoWord

e.g. ***Loitokitok*** *is a friendly city*

Difficulties in Shallow Parsing Approach

- **Ambiguously capitalised words** (first word in sentence)

  [All American Bank] vs. All [State Police]

- **Semantic ambiguity**

  "John F. Kennedy" = airport (location)

  "Philip Morris" = organisation

- **Structural ambiguity**

  [Cable and Wireless] vs. [Microsoft] and [Dell]

  [Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith].

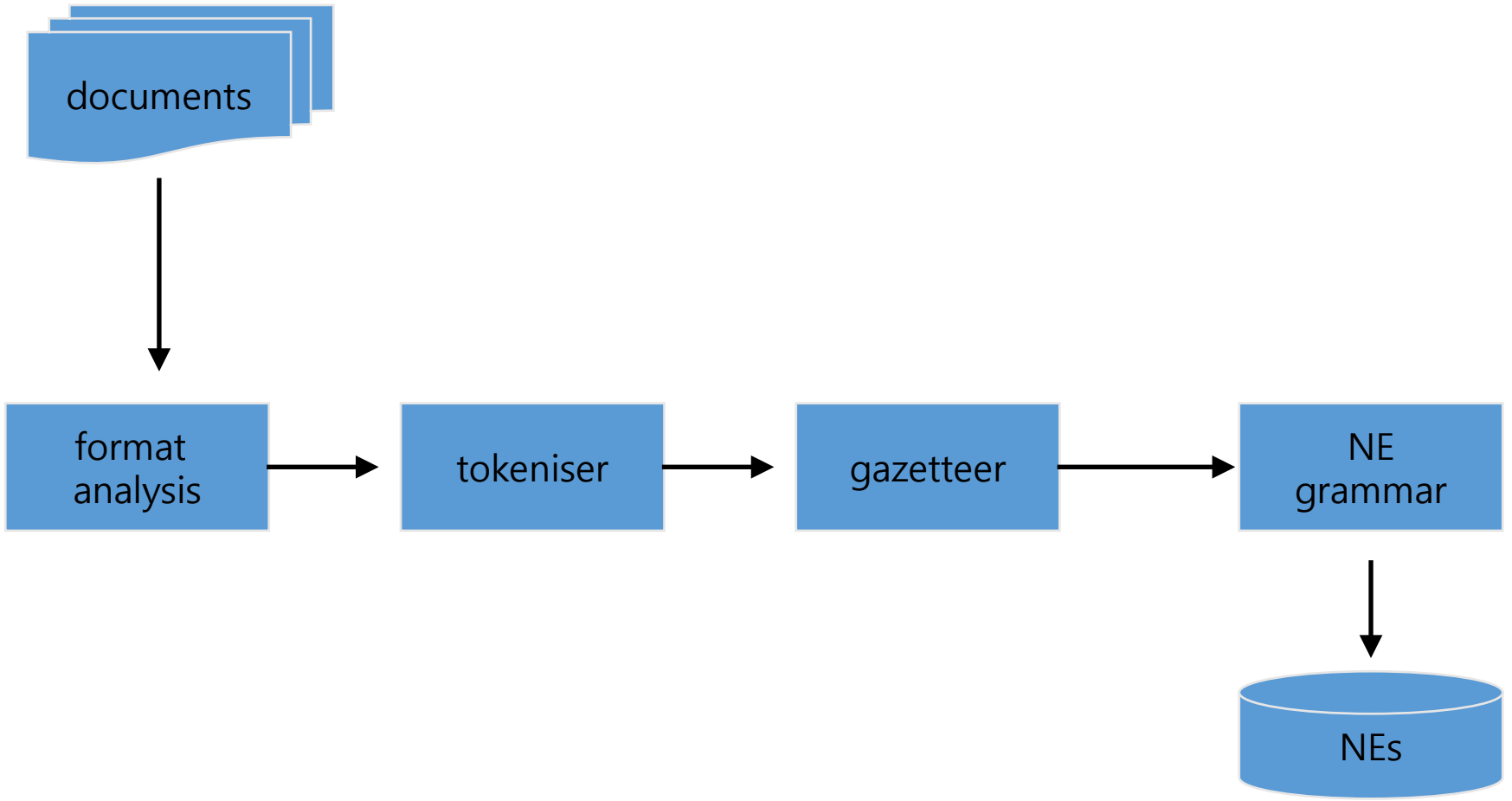
Machine Learning Approach

- Supervised approach
  - HMM
  - CRF
  - Deep Learning

Software

- JAPE – GATE
  - Graphical interface
- OpenNLP
  - Includes rule-based and statistical NER
- Stanford NER

NE System Architecture



Modules

- **Tokenizer**
  - **segments text into tokens, e.g. words, numbers, punctuation**
- **Gazetteer lists**
  - **NEs, e.g. towns, names, countries, ...**
  - **key words, e.g. company designators, titles, ...**
- **Grammar**
  - **hand-coded rules for NE recognition**

JAPE

- Set of phases consisting of pattern /action rules
- Phases run sequentially and constitute a cascade of FSTs over annotations
- LHS - annotation pattern containing regular expression operators
- RHS - annotation manipulation statements
- Annotations matched on LHS referred to on RHS using labels attached to pattern elements

Tokeniser

- Set of rules producing annotations
- LHS is regular expression matched on input
- RHS describes annotations to be added to AnnotationSet

(UPPERCASE _LETTER) (LOWERCASE_LETTER)* >

Token; orth = upperInitial; kind = word

Gazetteer

- Set of lists compiled into Finite State Machines
- Each list has attributes MajorType and MinorType (and optionally, Language)

city.lst: location: city

currency_prefix.lst: currency_unit: pre_amount

currency_unit.lst: currency_unit: post_amount

Named entity grammar

- hand-coded rules applied to annotations to identify NEs
- annotations from format analysis, tokeniser and gazetteer modules
- use of contextual information
- rule priority based on pattern length, rule status and rule ordering

Example of JAPE Grammar rule

```
Rule:      Location1
Priority: 25
( ( { Lookup.majorType == loc_key,
        Lookup.minorType == pre}
      { SpaceToken} )?
  { Lookup.majorType == location}
  ( {SpaceToken}
    { Lookup.majorType == loc_key,
      Lookup.minorType == post} ) ?
)
 : locName -->
        :locName.Location = { kind = “gazetteer”, rule = Location1
      }
```

PASTA

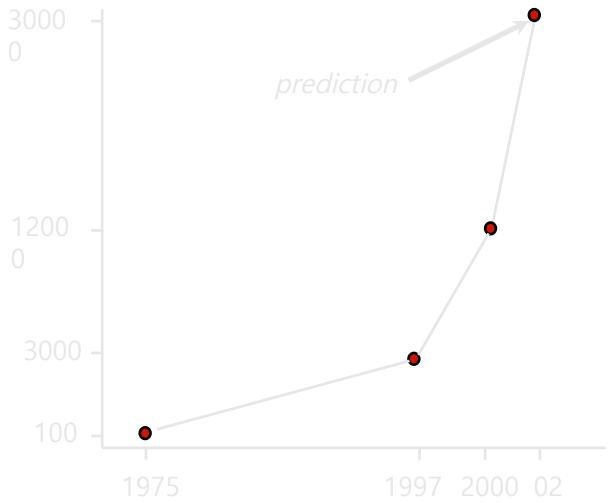
- **Protein Active Site Template Acquisition**
- **Aim: Use of IE techniques to create a database of protein active site data to support protein structure analysis**
- **Partners: Dept. of Computer Science, Information Studies, Mol. Biology and Biotechnology, Univ. of Sheffield**
- **Sponsors: BBSRC-EPSRC Bioinformatics Initiative**

Molecular Biology

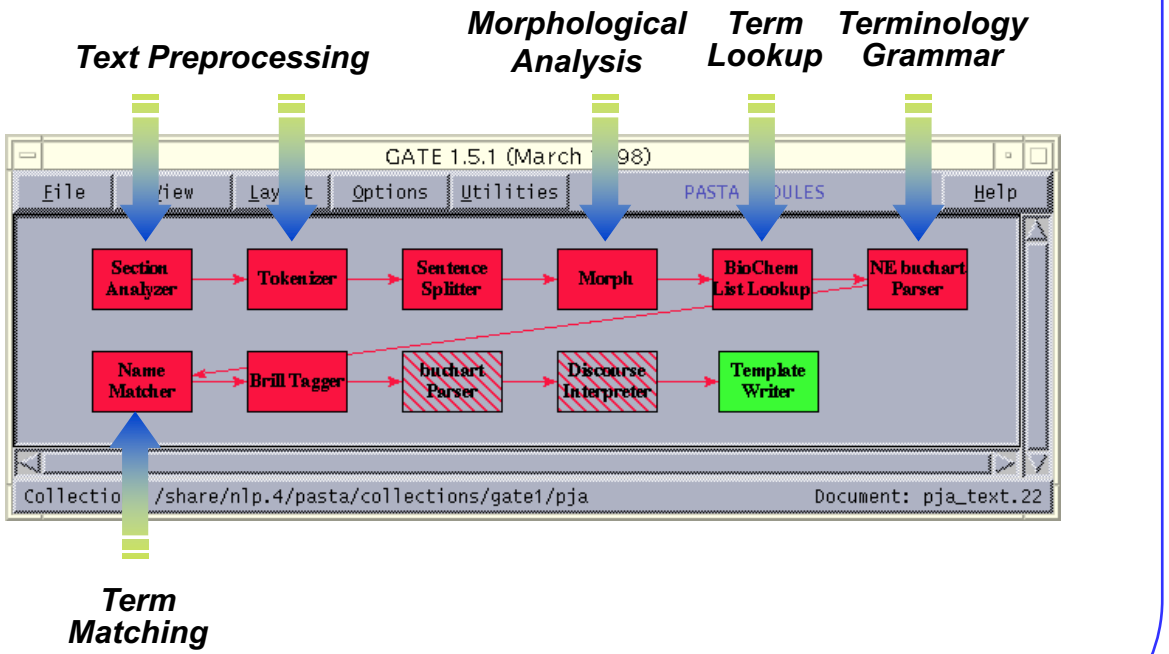
- Computer-intensive research - too many proteins are being analysed
- Too much text is being produced
- For some proteins, the literature goes back 30 years
- Wide-range research is hindered by the volume of information
- Working with more than one protein is getting increasingly difficult

The Protein Data Bank (PDB)

Currently 12,000 entries



PASTA System Architecture



Recognition of Biological Terminology

Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened 'lid', respectively.

MUMIS

- **MUltiMedia Indexing and Searching environment**
- **Application of IE technology to multimedia, multilingual video indexing in football domain**
- **2 years: June 2000 - 2002**
- **CTIT (NL), University of Sheffield (UK), DFKI (D), Max Planck Institute (D), University of Nijmegen (NL), ESTeam (SWE), VDA (NL)**

실습

한국어형태소분석 및 품사태깅 시스템

실습 및 레포트

- http://blpdemo.korea.ac.kr/MA/
- 한국어형태소분석 및 품사태깅 시스템 데모를 이용한 뉴스기사 품사태깅 및 파싱