



NLP Word Sense Disambiguation



Word Sense Disambiguation

- **중의적 표현의 뜻과 특징**

- 중의적 표현이란 하나의 표현이 두 가지 이상의 의미로 해석되는 표현.
 - > 해석의 혼동으로 인해 원하는 결과에 방해를 줌.
- ex) 검색, 대화



Word Sense Disambiguation

- **중의적 표현의 종류**

- 어휘적 중의성



- > 다의어에 의한 중의성

- ex) 손 좀 봐야 되겠구나

- 1. 손이 어떻게 생겼는지 봐야겠다. (신체일부)

- 2. 기계가 고장나서 손 좀 봐야겠다. (수리)

- 3. 그 친구 손 좀 봐야겠다. (혼이 나다)

- > 동음어에 의한 중의성

- ex) 말이 많다.

- 1. 말이 많다. (언어)

- 2. 말이 많다. (동물)



Word Sense Disambiguation

- **중의적 표현의 종류**

- 구조적 중의성

- > 수식어에 의한 중의성

- ex) 게으른 토끼와 거북이가 경주를 한다.

- 1. 게으른 토끼가 거북이와 경주를 한다.

- 2. 게으른 토끼와 게으른 거북이가 경주를 한다.

- > 대칭동사에 의한 중의성

- ex) 철수와 영희는 도서관에서 공부했다.

- 1. 철수와 영희가 함께 도서관에서 공부했다.

- 2. 철수와 영희가 도서관에서 각자 공부했다.



Word Sense Disambiguation

• 중의성 해소

: 중의성 해소 과정은 문맥 정보를 보고 의미를 추론하는 방법과 이미 가지고 있는 지식을 활용해서 예측하는 방법이 있음.

- **문맥** ('배' 라는 단어의 의미를 모르는 상태)

1. 영호는 비행기를 그리고, 철수는 배를 그렸다.

(비행기를 보고 타는 배를 추측)

2. 영호는 사과를 그리고, 철수는 배를 그렸다.

(사과를 보고 먹는 배를 추측)

3. 영호는 어깨를 그리고, 철수는 배를 그렸다.

(어깨를 보고 신체 일부 배를 추측)

- **지식** ('장기'라는 단어의 의미를 아는 상태)

1. 나는 할아버지와 장기를 두었는데, 할아버지가 먼저 말을 놓으셨다.

(장기와 말이라는 사전적 정보를 알고 해석.)





Word Sense Disambiguation

- **지식 정보를 활용한 WSD(Word Sense Disambiguation) 과정**
 - 단어의 의미를 담고 있는 리소스 필요
ex) Wikipedia, WordNet
 - 문장에 등장한 단어들을 리소스 정보를 활용하여 예측.



Word Sense Disambiguation

점심을 먹고 나는
후식으로 **밤**을 먹었다.

1

2

오늘 **밤**에는 비가
많이 올 예정이니
주의하시길 바랍니다.

밤 1: 해가 저서 어두어진 때부터 다음 날 해가
떠서 밝아지기 전까지의 동안

- 유의어 = 밤중, 야간, 어둠
- 반의어 = 낮

밤 2: 밤나무의 열매, 가시가 많이 난 송이에
싸여 있고 갈색 겉껍질 안에 얇고 맛이 짧은
속껍질이 있으며, 날 것으로 먹거나 굽거나
삶아서 먹는다.



Word Sense Disambiguation

- **문맥 정보를 활용한 WSD(Word Sense Disambiguation) 추론 과정**
 - 다양한 단어 조합 정보가 포함된 학습 문장 데이터가 필요.
 - 각 단어와 같이 등장한 단어들과 관계성을 보고 어떤 의미일지 벡터에 표현.



Word Sense Disambiguation

점심을 먹고 나는
후식으로 **밤**을 먹었다.

오늘 **밤**에는 비가
많이 올 예정이니
주의하시길 바랍니다.

점심

시간 밤

아침

배

먹는 밤

사과



WordNet




WordNet

- 인간은
 - 어떻게 주변환경을 인지하는가?
 - 어떻게 지식화 하는가?
 - 어떻게 공유 하는가?
 - 새로운 지식을 어떻게 추론하는가?
 - > 다양한 답 중 하나가 어휘.
 - > 어휘가 심리학적 실재를 가진 기억의 최소 단위로 정의.





WordNet

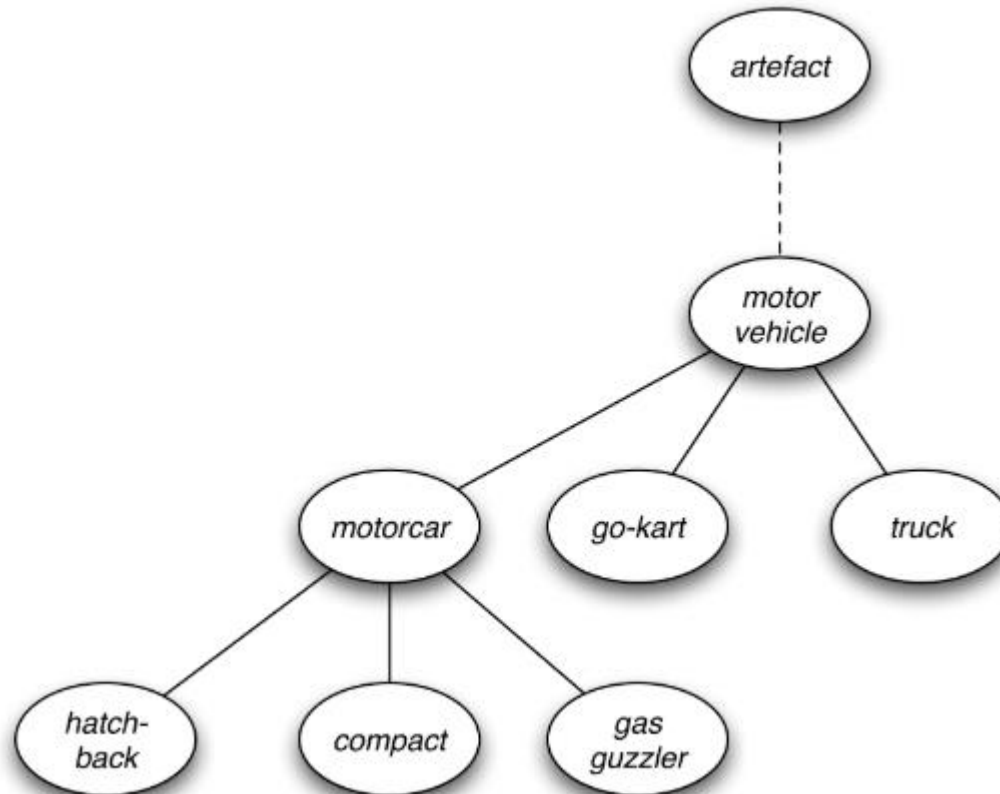
- WordNet은 영어의 의미 어휘 목록
 - 영어단어를 synset이라는 동의어 집합으로 분류. 
 - 이 집합은 간략한 정의를 제공
 - 다양한 어휘 목록 사이에서 의미 관계가 정의되어 있음.
 - 의미 관계가 상하위 관계를 이루고 있음.



WordNet

- WordNet 계층구조 예시

- 인공물을 기반으로 차량 관련 계층 구조를 나타낸 예시.





WordNet

- Synset간 의미 관계
 - Synonymy (동의어)
ex) 천사, 엔젤
 - Antonymy (반의어)
ex) 악마, 데빌
 - Hyponymy (하의어)
ex) 사랑, 평화
 - Hypernymy (상의어)
ex) 신



WordNet

- WordNet 관련 함수 코드 작성 및 이해



Knowledge Based WSD



Knowledge Based WSD

- 지식 기반 방법은 시소러스나 백과사전 등을 바탕으로 단어의 의미를 추론.
 - 사전 정의 기반 방법
 - : 사전에 정의된 문장의 단어들을 기반으로 의미 추론.
 - ex) Lesk 알고리즘
 - 그래프 기반 방법
 - : 의미관계를 가지는 사전들의 관계성을 보고 의미를 추론.



Knowledge Based WSD



• 사전 정의 기반 방법 – Lesk 알고리즘

: 중의성 단어의 사전 뜻풀이에 쓰인 단어들과 중의성 단어와 함께 주변 문맥에 나타난 단어의 사전 뜻풀이에 쓰인 단어들 사이에 중복되는 단어가 가장 많은 의미를 중의성 단어의 의미로 선택하는 것

그 **사람**은 **수술**을 통해 불편한 **다리**를 고쳤다.

단어		사전 뜻풀이에 쓰인 단어
함께 나타난 단어	사람	생각, 언어, 만들다, 쓰다, 사회, 살다, 동물 ,
	수술	피부, 점막, 조직, 기계, 병, 고치다,
	...	
<u>중의성 단어</u>	다리 01	사람, 동물 , 몸통, 신체,
	다리 02	물, 건너다, 시설물,



Knowledge Based WSD

- 사전 정의 기반 방법 – Lesk 알고리즘 한계점
 - : 단어 간의 정확한 일치가 기반.
 - : 사전 정의의 굉장히 의존적.



Knowledge Based WSD

- Lesk Algorithm 실습
> http://bit.ly/LESK_Algorithm



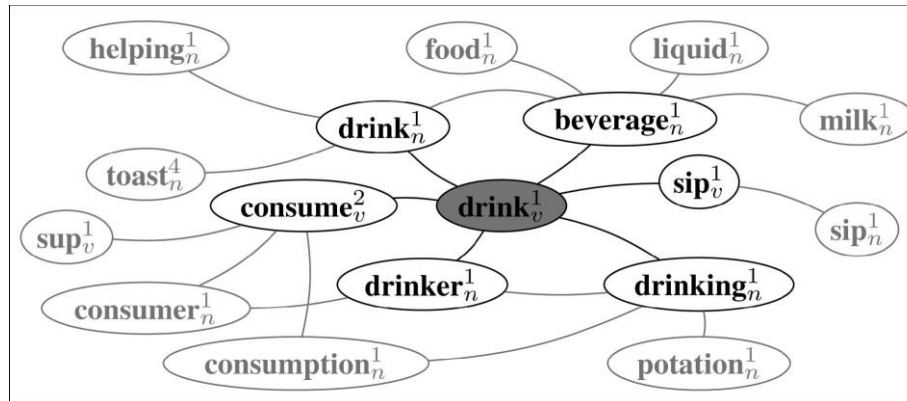
Knowledge Based WSD

- 그래프 기반 방법 – 단순 그래프 기반 방법

Ex) “She drank(l_1 =drink) some Milk(l_2 =milk)”

: $l_1 \Rightarrow$ drink : drink($v, 1$), drink($v, 2$)... $l_2 \Rightarrow$ milk : milk($n, 1$), milk($n, 2$)....

: 중의성 단어(lemma)인 drink와 milk의 Synset들을 WordNet에서 추출

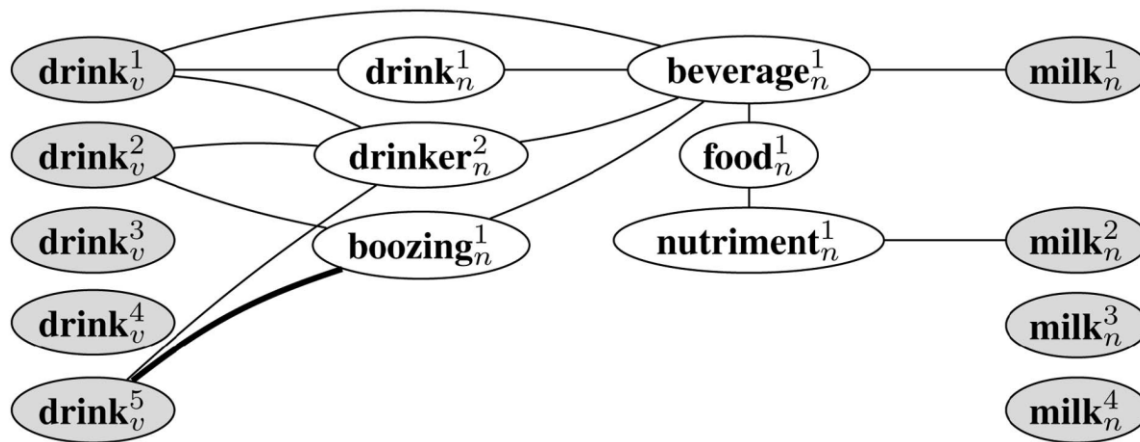




Knowledge Based WSD

• 그래프 기반 방법 – 단순 그래프 기반 방법

- DFS(Depth First Search), BFS(Breath First Search)로 검색하여 검색되는 Edge를 아래와 같이 추출하여 Subgraph 생성
- 의미 간 연결성 측정을 통해 가장 높은 의미 선택
- 같은 점수는 의미 순서가 높은 걸로 선택

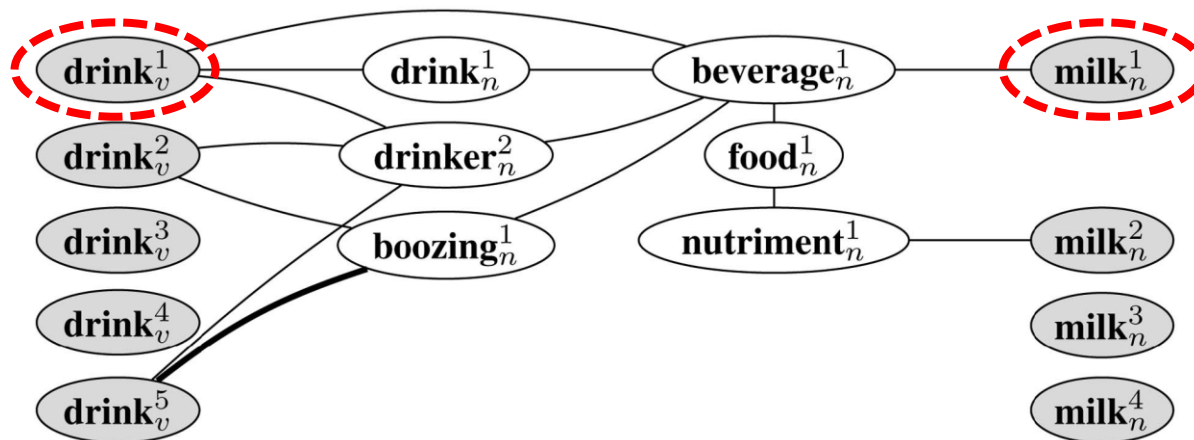




Knowledge Based WSD

• 그래프 기반 방법 – 단순 그래프 기반 방법

- DFS(Depth First Search), BFS(Breath First Search)로 검색하여 검색되는 Edge를 아래와 같이 추출하여 Subgraph 생성
- 의미 간 연결성 측정을 통해 가장 높은 의미 선택
- 같은 점수는 의미 순서가 높은 걸로 선택





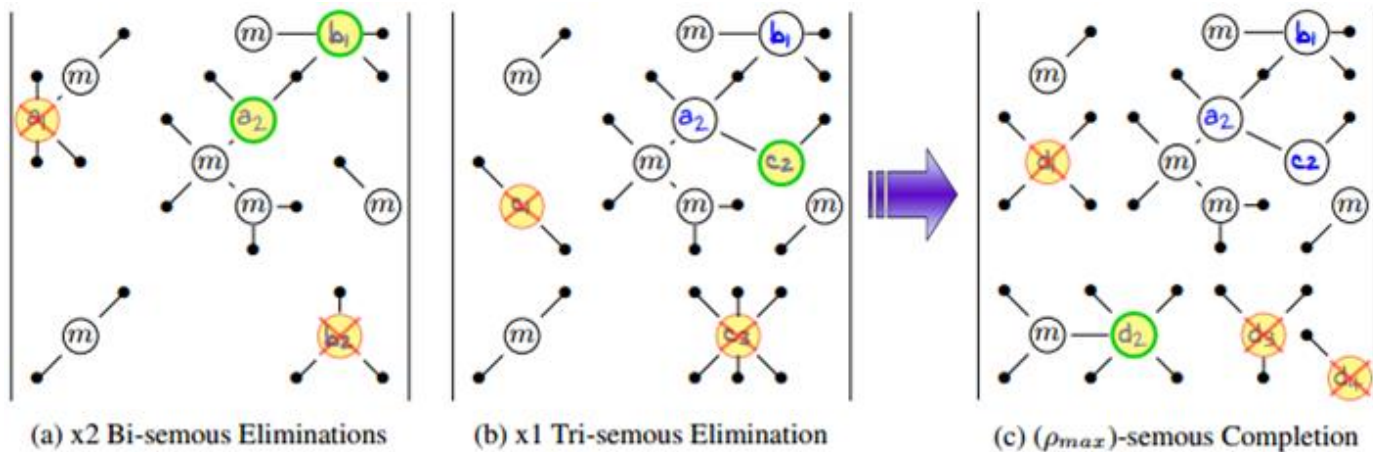
Knowledge Based WSD



- 그래프 기반 방법 – 반복적 그래프 기반 방법

- 특정 조건에 따라 그래프를 반복적으로 생성해 나가는 방법.

- Iterative 'Sense Count Score' Approach – Example



Iterative Disambiguating of Subgraphs



Knowledge Based WSD



• 그래프 기반 방법 – 반복적 그래프 기반 방법

- 반복적, 단순 그래프 방법 비교

“Spanish football players playing in the All-Star League and in powerful clubs of the Premier League of England are during the year very active in league and local cup competitions and there are high-level shocks in the European Cups and European Champions League.”

남색은 Cup의 반복 그래프, 보라색은 Cup의 단순 그래프

Cup의 의미는 Cup#8

: 단순 그래프 -> Cup#1

: 반복 그래프 -> Cup#8

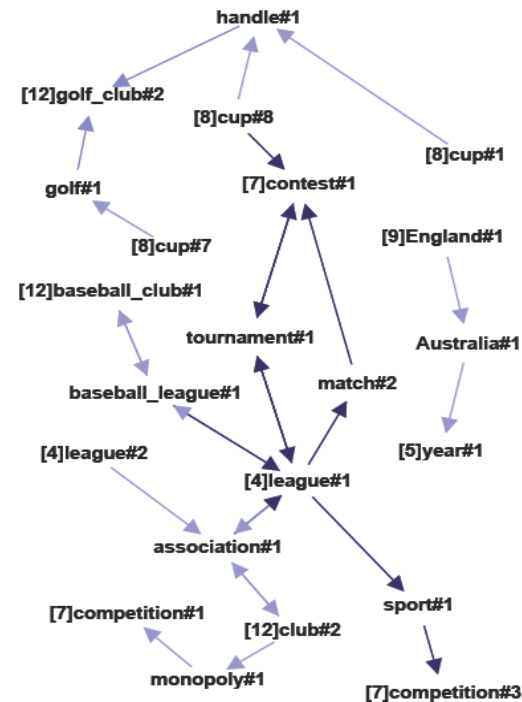


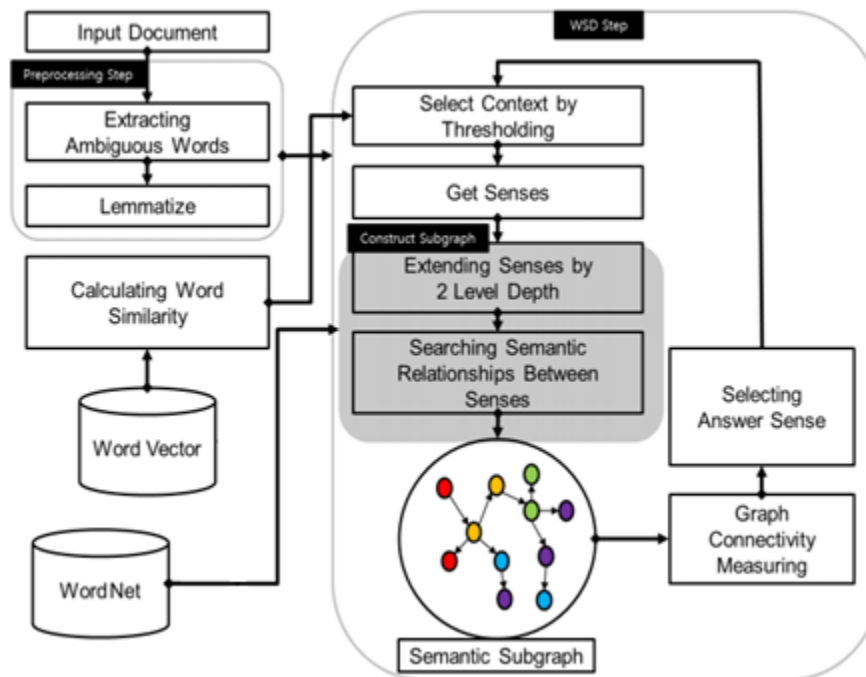
Figure 5: Conventional vs Iterative Subgraph



Knowledge Based WSD



- 그래프 기반 방법 – 반복적 그래프 기반 방법
 - Iterative 'Similarity Score' Approach – Example

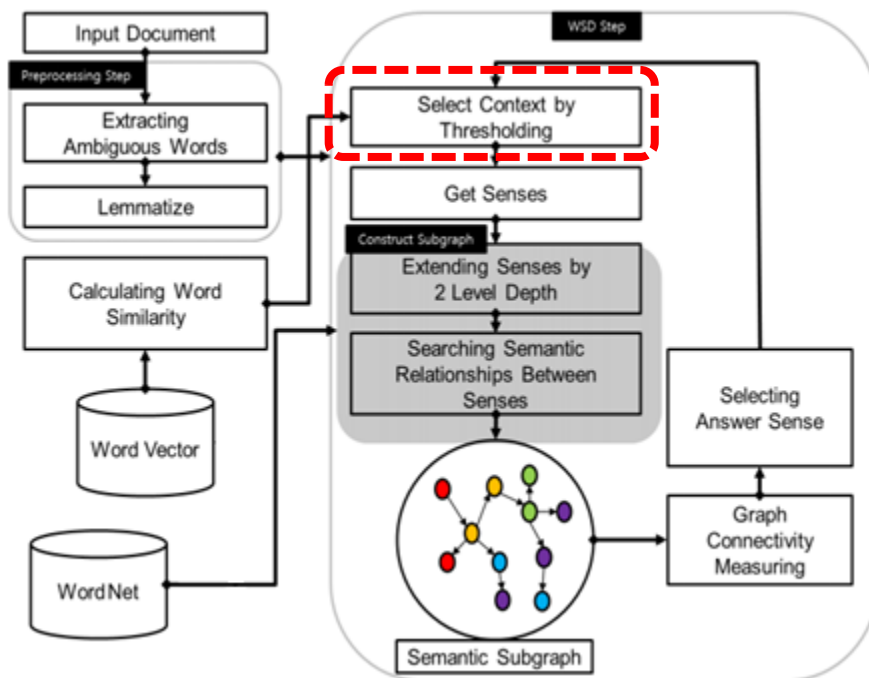




Knowledge Based WSD

- 그래프 기반 방법 – 반복적 그래프 기반 방법

- 유사도가 가장 높은 단어끼리 매칭





Knowledge Based WSD

- 그래프 기반 방법 – 반복적 그래프 기반 방법

- 유사도가 가장 높은 단어끼리 매칭

Document 1

Sen 1

U.N. group drafts plan to reduce emissions.

Sen 2

The U.N.-sponsored climate conference characterized so far by Unruly posturing and mutual recriminations gained renewed focus Friday with the release of a document outlining ambitious greenhouse-gas reductions over the next 40 years, with industrialized nations shouldering most of the burden in the near term.

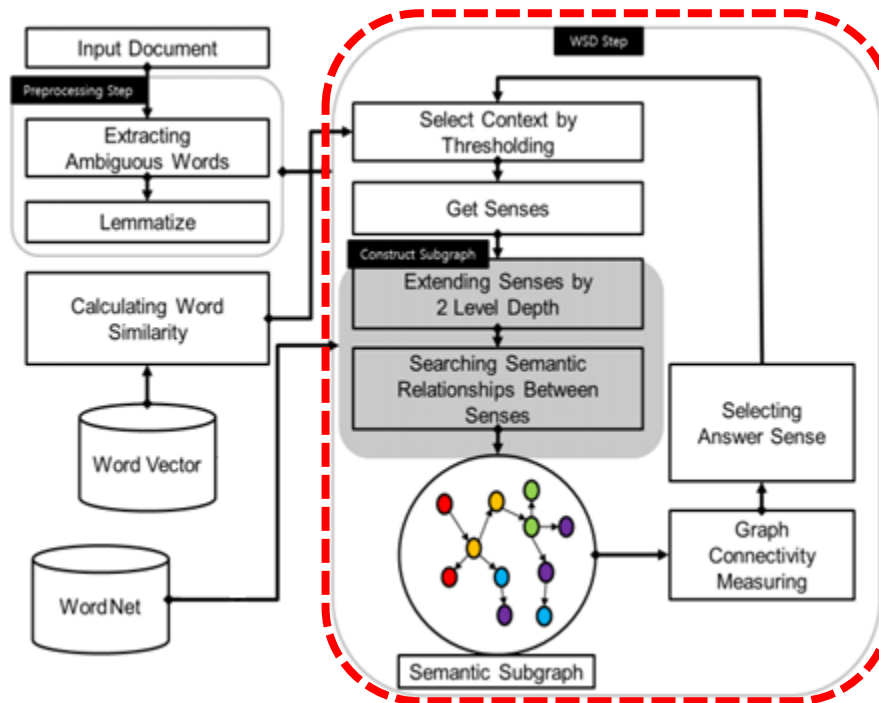
⋮



Knowledge Based WSD

• 그래프 기반 방법 – 반복적 그래프 기반 방법

- 매칭된 단어끼리 반복적 그래프 생성 및 의미 선택





Knowledge Based WSD



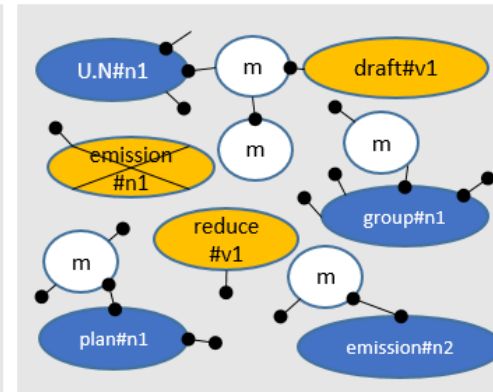
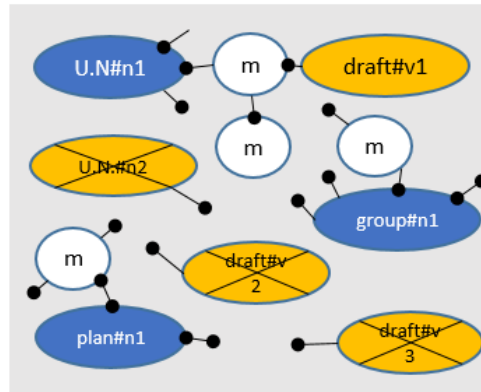
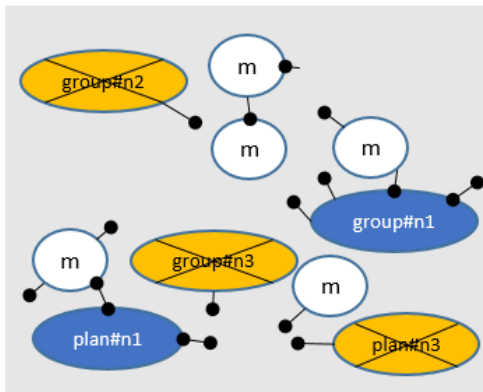
• 그래프 기반 방법 – 반복적 그래프 기반 방법

- 매칭된 단어 (U.N., draft), (group, plan), (emission, reduce)

= Group, Plan 의미 결정 - group, plan => group#n1, plan#n1

= U.N. 의미 결정 - U.N., draft, group#n1, plan#n1 => U.N.#n1

= Emission 의미 결정 - U.N.#n1, draft#v1, group#n1, plan#n1, emission, reduce => emission#n2





한계점



- 리소스의 영향을 많이 받음.
 - 정의되지 않은 단어는 풀 수 없음.
ex) 신조어
- 단어의 의미를 결정하는 핵심 단어를 고려하지 않음.
 - 문맥 단어로부터 단어의 의미들이 결정되는데 규칙에 맞춰 타겟이 정해져 있음



Knowledge Based WSD

- Knowledge Based WSD 코드이해
> http://bit.ly/KB_PPR



Supervised WSD



Supervised WSD



- 지도학습 (Supervised) WSD
 - 각종 기계 학습 알고리즘을 통해 단어 의미를 분류
 - 의미 분석에 필요한 자질을 추출하여 분류기로 분석.
ex) KNN, Deep Neural Network, SVM ...



Supervised WSD

- 지도학습 (Supervised) WSD

- 기존의 기계학습 분류기 모델은 우리가 정한 규칙에 맞춰 선택된 자질에 따라 성능을 높여옴.



- ex) 공기어, 의존 관계 정보, 형태소 정보 등.

- 최근에는 딥러닝 모델이 들어오면서 특정 규칙 기능을 담당하는 Layer를 추가하면서 성능을 개선.



Supervised WSD

- 지도학습 (Supervised) WSD

- LSTM 기반으로 다양한 기능의 Layer를 추가하여 모델 실험.

- * Refer

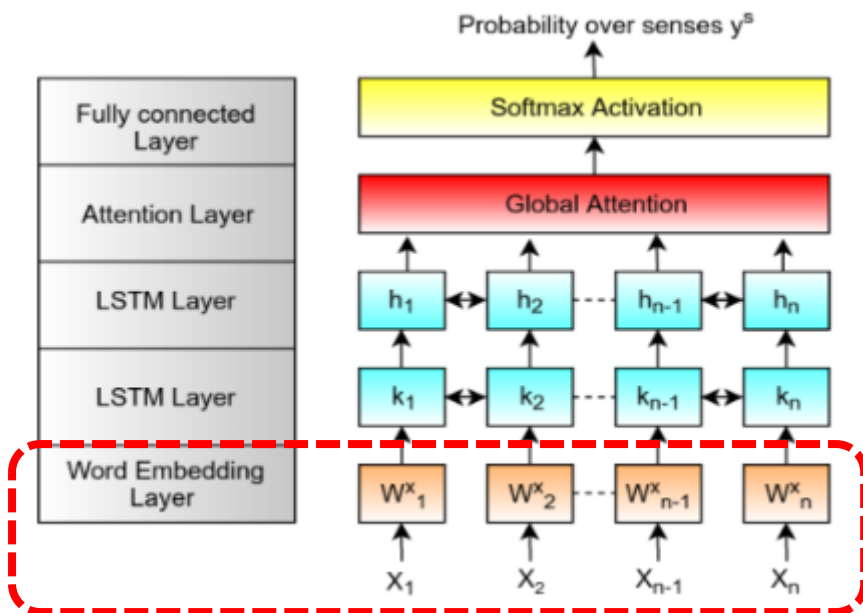
- <https://github.com/Sshanu/Hierarchical-Word-Sense-Disambiguation-using-WordNet-Senses>



Supervised WSD



- LSTM을 이용한 WSD
 - LSTM + Global Attention

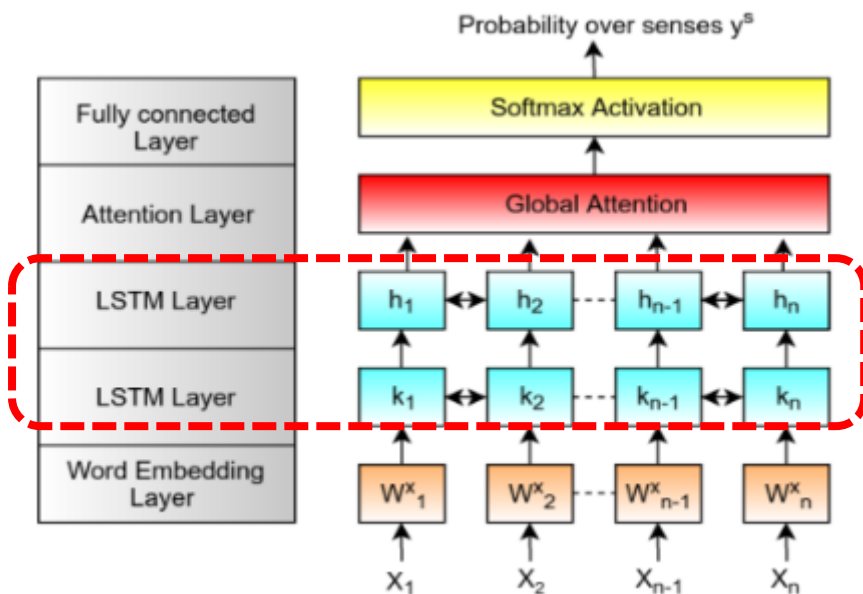


Word Embedding
기존의 Glove,
Word2vec으로 대량의
문장데이터에서 학습한
벡터값을 이용.



Supervised WSD

- LSTM을 이용한 WSD
 - LSTM + Global Attention



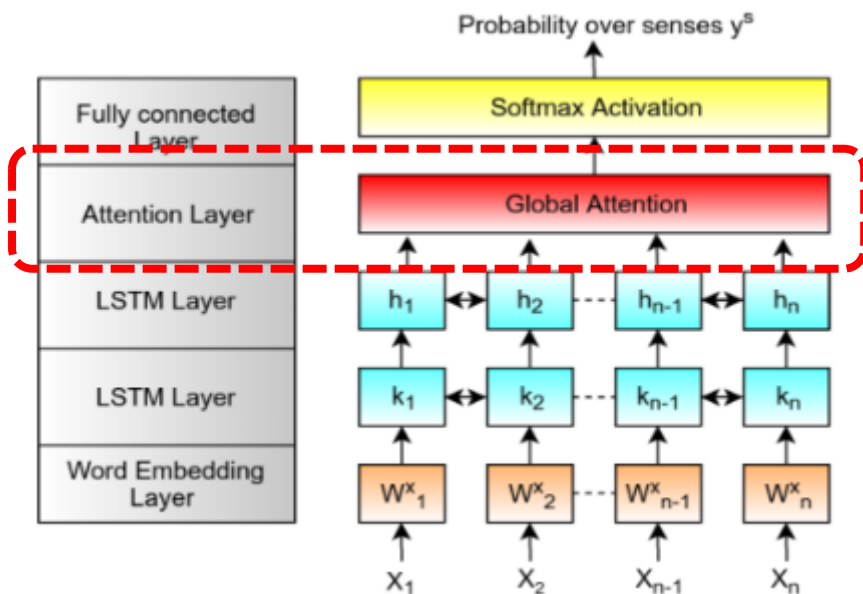
문장에 등장한 어순
정보를 모두 반영하기
위한 양방향 LSTM 구축.

RNN이 아닌 LSTM을
구축한 이유는 RNN은
의존 기간이 길어지면
정보가 손실되기 때문에
LSTM을 사용.



Supervised WSD

- LSTM을 이용한 WSD
 - LSTM + Global Attention



LSTM에서 어순 정보를 표현했다면 Global Attention에서는 문장에 등장한 단어 간의 의존 관계성을 고려함.

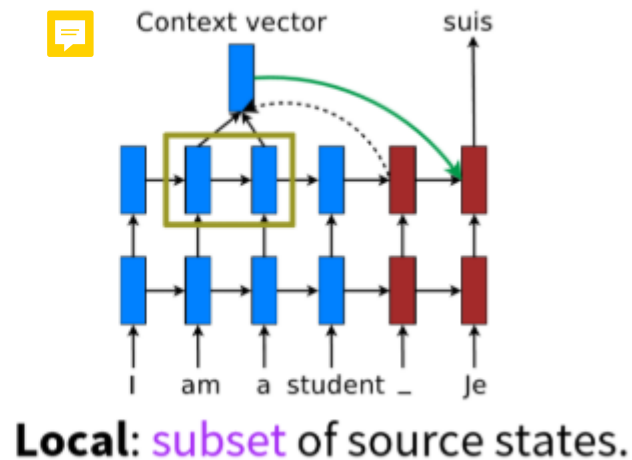
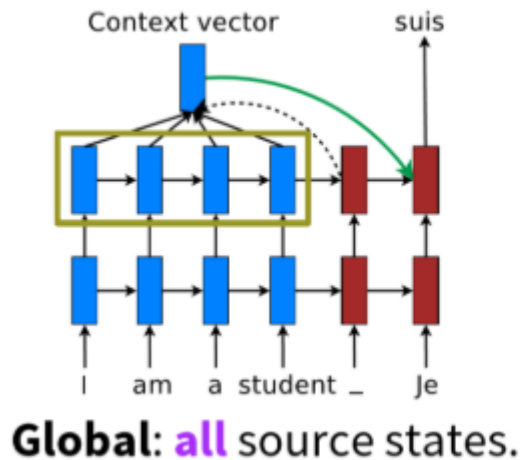
한 단어가 나머지 문맥 단어에 대해서 모든 가중치를 고려



Supervised WSD

- **Global vs Local Attention**

- Global은 Encoder 전체의 Hidden State에 대해 가중치 계산.
- Local은 Encoder 내 Window size만큼 Hidden State에 대해 가중치 계산.

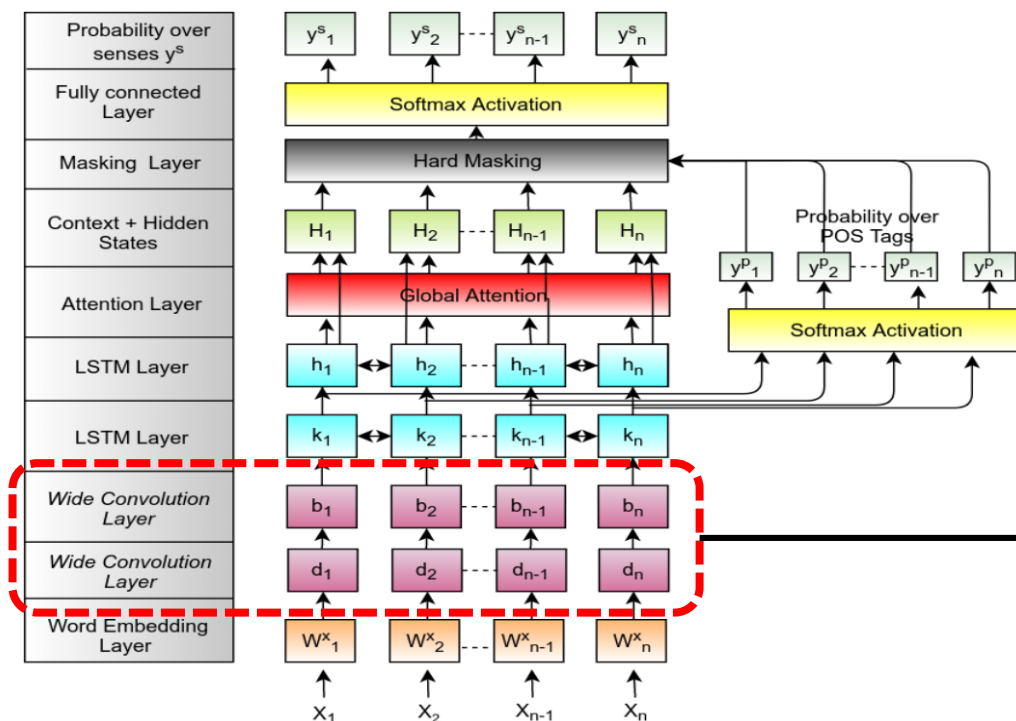




Supervised WSD



- LSTM을 이용한 WSD
 - CNN + LSTM + POS Tag Prob + Global Attention + Hard Masking



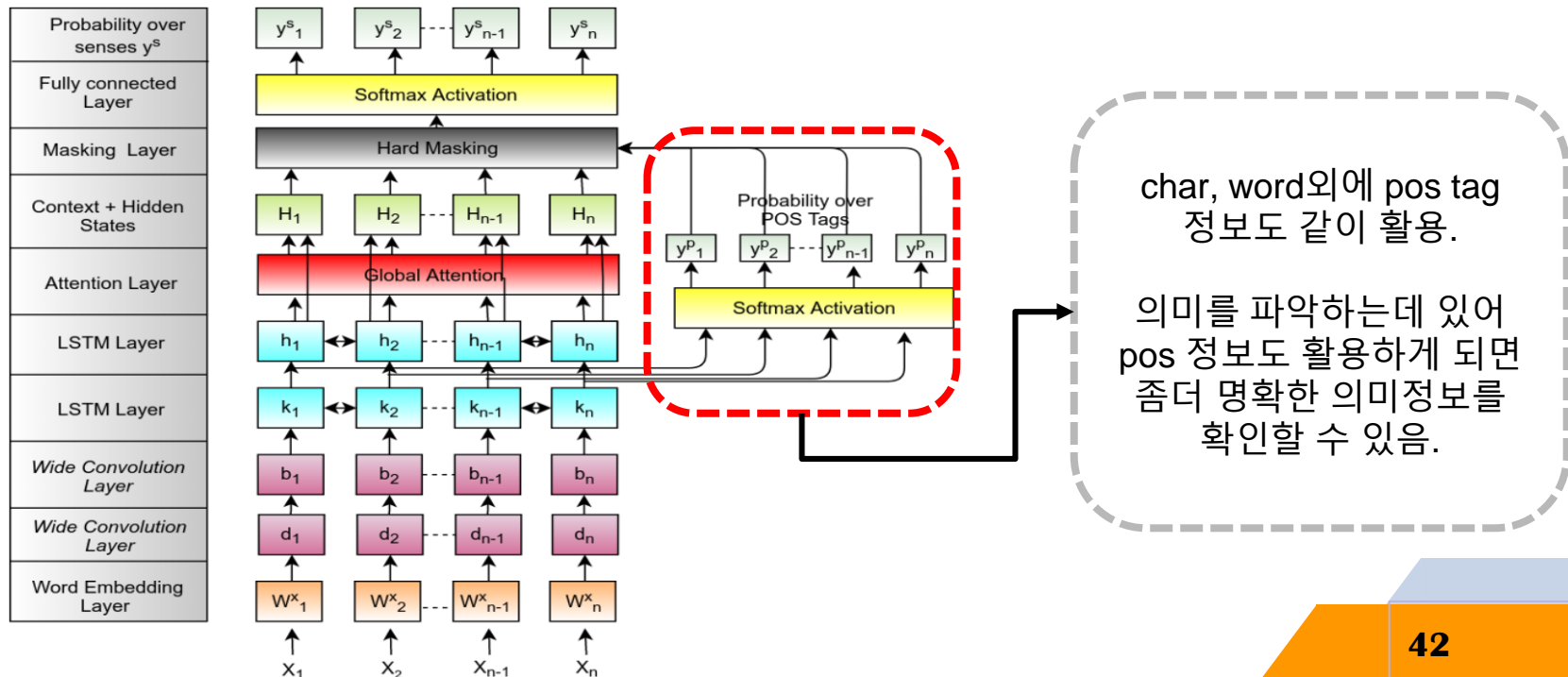
입력 단어의 음절 관계를
파악하는 정보도 추가.

입력 단어가 학습데이터에
없을 경우를 고려하여
char 관계를 고려하여
해당 단어가 어떤 단어에
의미적으로 가까운지 추론
가능.



Supervised WSD

- LSTM을 이용한 WSD
 - CNN + LSTM + POS Tag Prob + Global Attention + Hard Masking

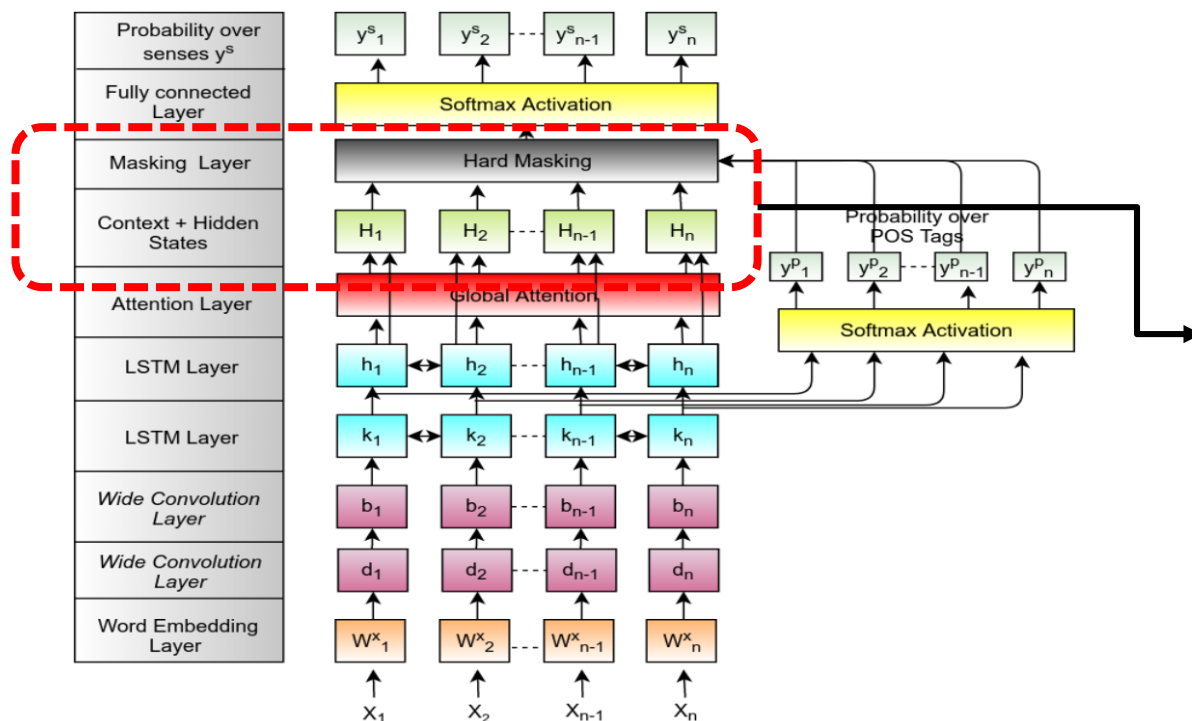




Supervised WSD

- LSTM을 이용한 WSD

- CNN + LSTM + POS Tag Prob + Global Attention + Hard Masking



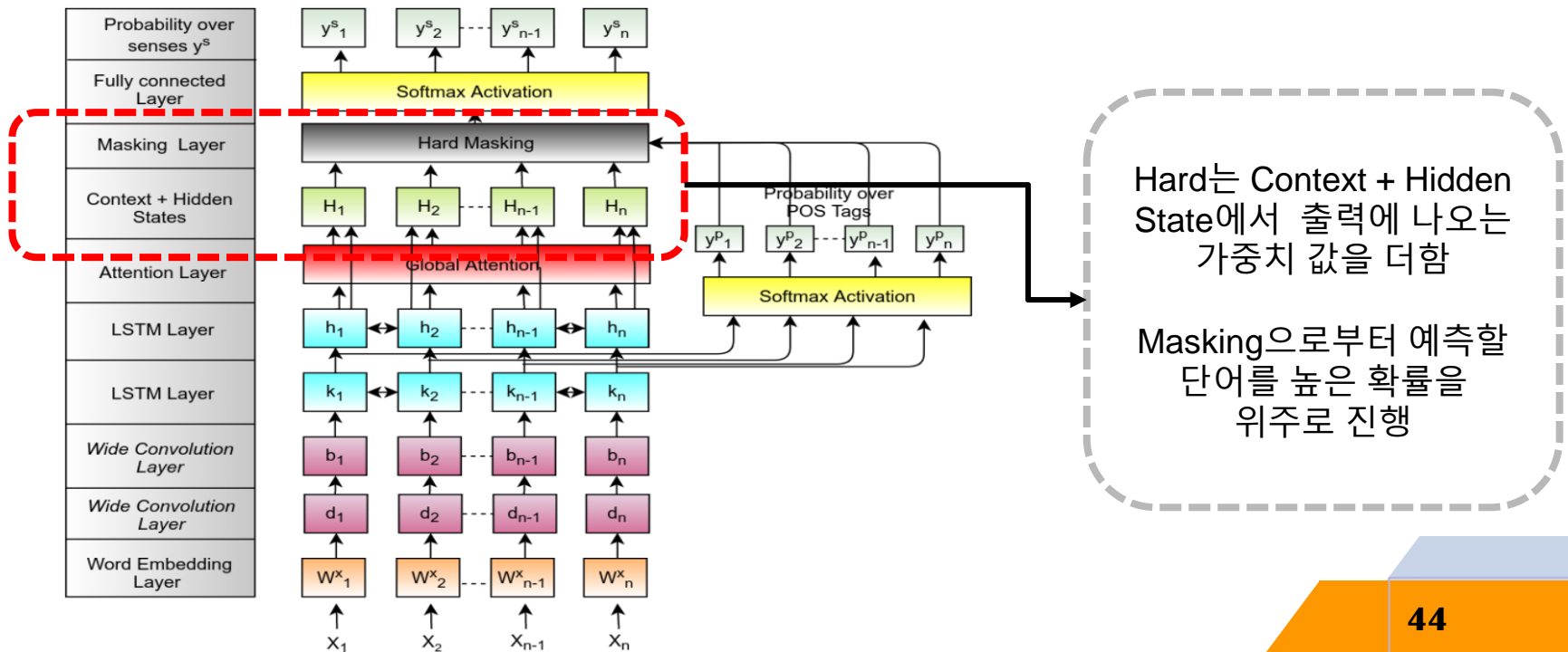
Masking은 단어에 해당하는 출력 의미를 스스로 예측하게 하여 검증하게 함.

문장의 어순과 문맥의 관계성을 보고 스스로 의미를 추론하여 예측.



Supervised WSD

- LSTM을 이용한 WSD
 - CNN + LSTM + POS Tag Prob + Global Attention + Hard Masking

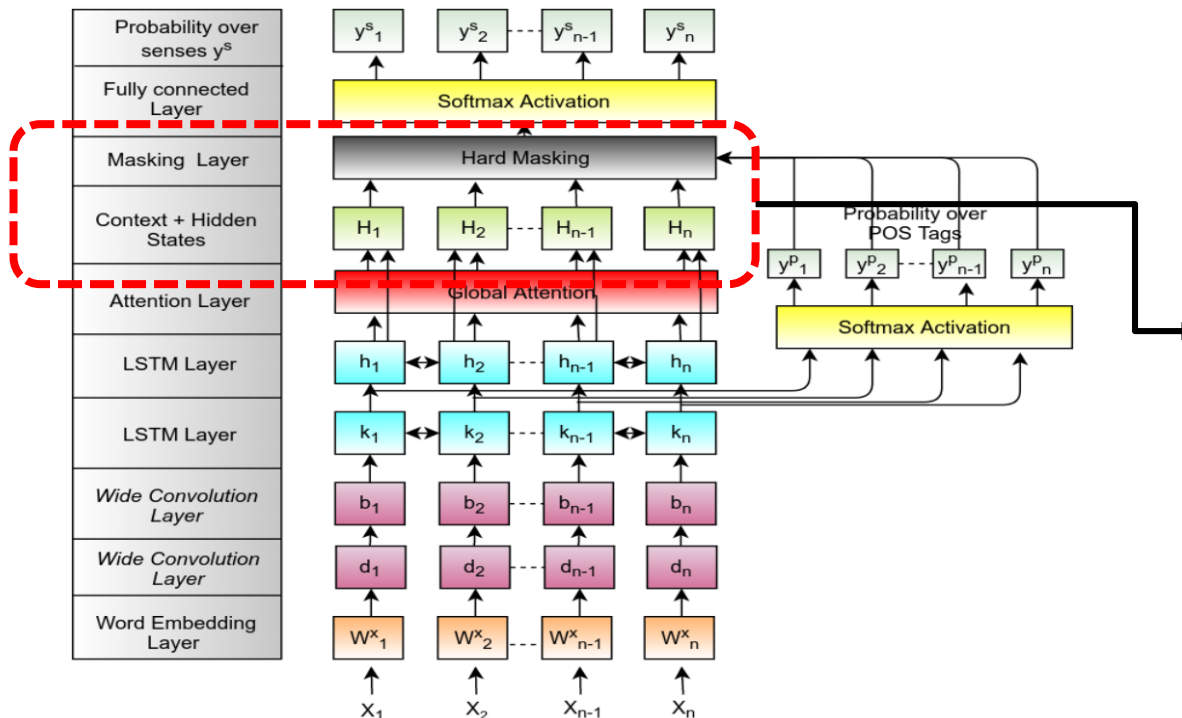




Supervised WSD

- LSTM을 이용한 WSD

- CNN + LSTM + POS Tag Prob + Global Attention + Hard Masking



곱셈으로 하게 되면 -의 값을 가중치 값이 -의 경우 +가 되어 높은 확률이 될 수 있어 soft하게 Masking이 진행될 수 있음.



Supervised WSD

- **지도학습 (Supervised) WSD 한계점**
 - 성능을 높이기 위해서는 대규모의 의미 태깅된 말뭉치가 필요.
 - 학습한 특정 중의성 단어에 대해서만 해결 가능.





문맥적 단어 의미 추론



문맥적 단어 의미 추론

- **한계점에 따른 방안**

- 대규모의 의미 태깅된 학습데이터를 작업하는데는 많은 공수가 필요.
- 단어의 상위 Entity 개념을 몇가지로 정의하여 작업량을 감소.

ex) 개체명 분석 : Person, Organization, Time, Location

- 단어의 앞뒤 정보를 파악하여 비슷한 정보를 가진 단어끼리 벡터공간에 표현.

ex) Word Embedding : Word2Vec, Glove



문맥적 단어 의미 추론

- **Word Embedding 문제**

- Word2vec, Glove와 같은 모델은 다음과 같은 문제가 생김.

- Ex) Bank Account(은행계좌), River Bank(강둑)

- > Bank라는 단어로 한 개의 벡터로 표현.

- > 이처럼 한 단어에 모든 의미를 반영하기 힘들.

- 같은 표기의 단어라도 문맥에 따라 다르게 Word Embedding을 표현해야 함.

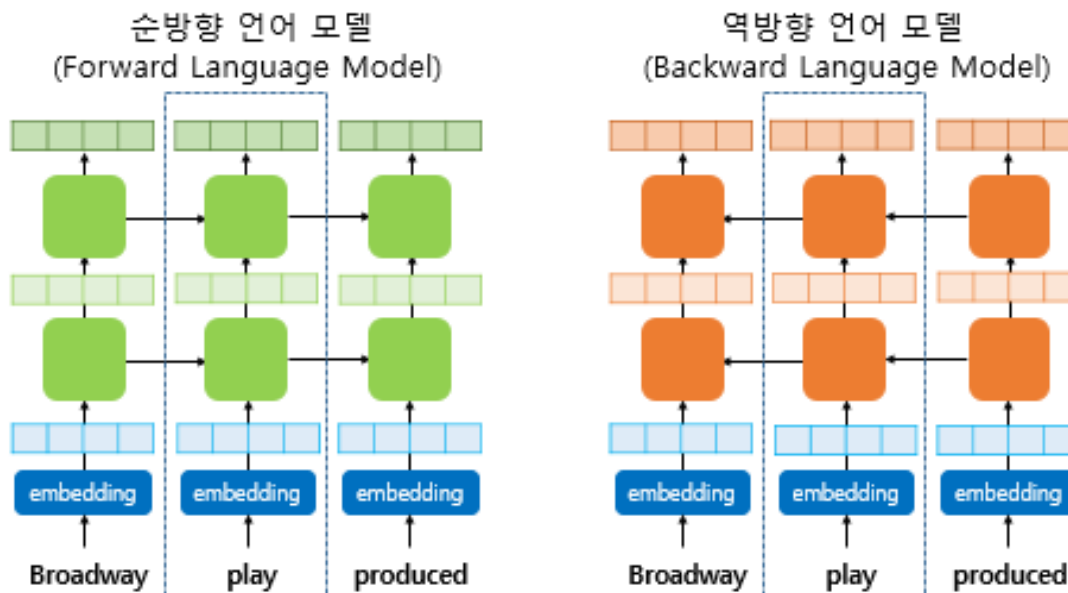


Contextualized Word Embedding



- **ELMo**

- 순방향 언어모델과 역방향 언어모델을 이용.
- RNN 모델사용.





- **ELMo**

- 

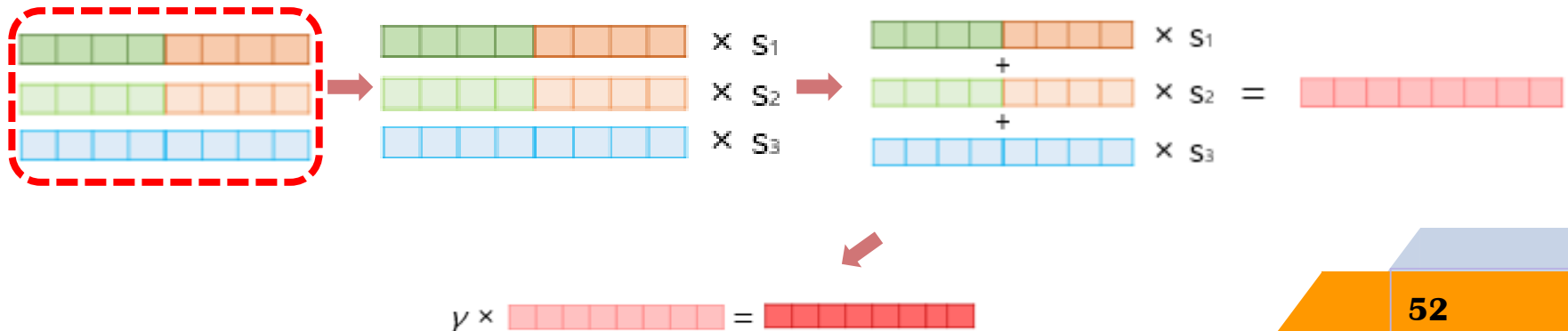




Contextualized Word Embedding

- **ELMo**

- 점선에 표시된 play의 임베딩을 표현
 - **각층의 출력 값을 연결**
 - 각층별로 가중치를 줌.
 - 각층의 출력 값을 모두 더함
 - 벡터의 크기를 결정하는 스칼라 매개변수를 곱함

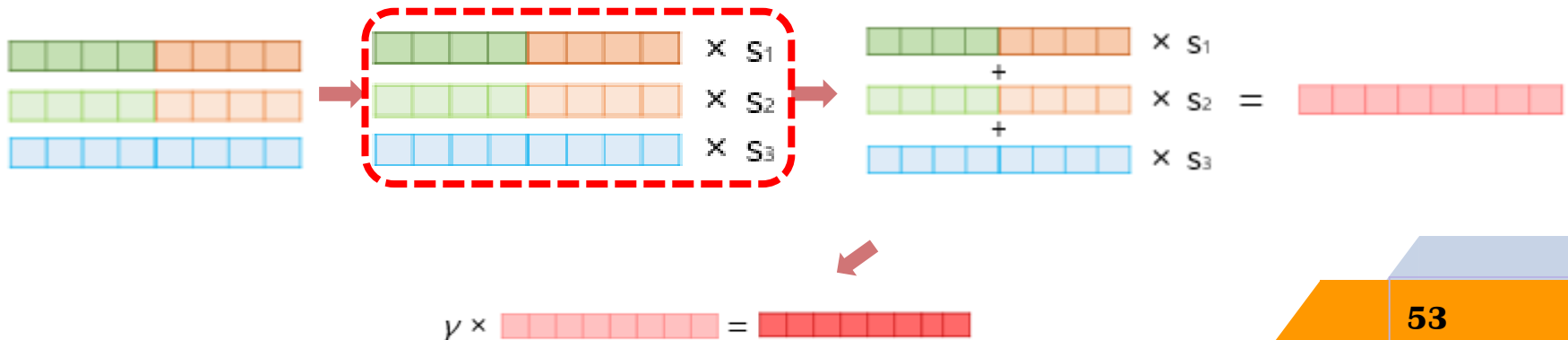




Contextualized Word Embedding

- **ELMo**

- 점선에 표시된 play의 임베딩을 표현
 - 각층의 출력 값을 연결
 - **각층별로 가중치를 줌.**
 - 각층의 출력 값을 모두 더함
 - 벡터의 크기를 결정하는 스칼라 매개변수를 곱함

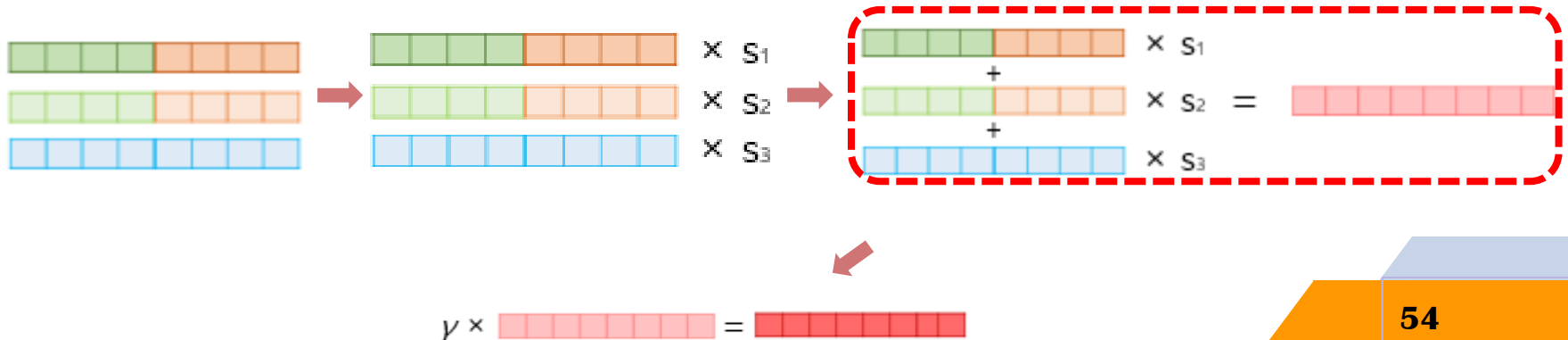




Contextualized Word Embedding

- **ELMo**

- 점선에 표시된 play의 임베딩을 표현
 - 각층의 출력 값을 연결
 - 각층별로 가중치를 줌.
 - **각층의 출력 값을 모두 더함**
 - 벡터의 크기를 결정하는 스칼라 매개변수를 곱함

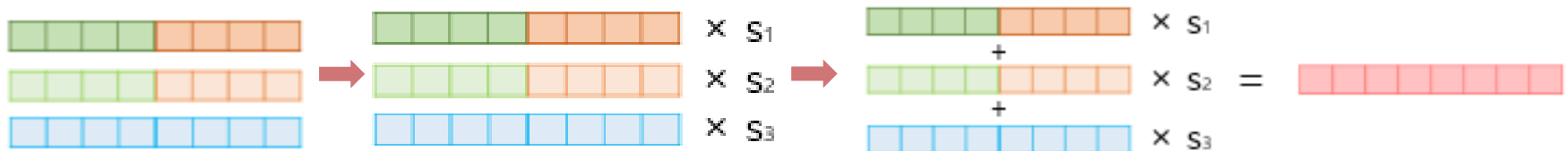




Contextualized Word Embedding

- **ELMo**

- 점선에 표시된 play의 임베딩을 표현
 - 각층의 출력 값을 연결
 - 각층별로 가중치를 줌.
 - 각층의 출력 값을 모두 더함
 - 벡터의 크기를 결정하는 스칼라 매개변수를 곱함



$\gamma \times \text{vector} = \text{scaled vector}$



Contextualized Word Embedding

- **ELMo Visaulization 실습**
> http://bit.ly/ELMO_VIS



Contextualized Word Embedding

- **ELMo Visualization**

- ELMo 모델을 이용하여 상황 별 벡터를 생성하고 시각화
- 각 Layer마다 어떤 결과를 내는지 비교





Contextualized Word Embedding

- **ELMo Visualization**

- Bank

은행관련 문장

"One can deposit money at the bank"

강둑관련 문장

"He had a nice walk along the river bank"

은행관련 문장

"I withdrew cash from the bank"

강둑관련 문장

"The river bank was not clean"

은행관련 문장

"My wife and I have a joint bank account"



Contextualized Word Embedding

- **ELMo Visaulization**

- Work

명사로써 이해

"I like this beautiful work by Andy Warhol"

동사로써 이해

"Employee works hard every day"

동사로써 이해

"My sister works at Starbucks"

명사로써 이해

"This amazing work was done in the early nineteenth century"

동사로써 이해

"Hundreds of people work in this building"



Contextualized Word Embedding

- **ELMo Visaulization**

- Plants

심는다는 의미.

"The gardener **planted** **some trees** in my yard"

심는다는 의미.

"I **plan to plant** a Joshua tree tomorrow"

심는다는 의미.

"My sister **planted** **a seed** and hopes it will grow to a tree"

식물의 의미.

"This kind of **plant** only **grows in** the subtropical region"

식물의 의미.

"Most of the **plants** **will die** without water"



Contextualized Word Embedding

- **ELMo 응용 실습**
 - ELMo Spam Classification 코드 이해
 - > http://bit.ly/ELMO_SC
 - > http://bit.ly/GLOVE_SC
 - ELMo Sentiment Analysis 실습
 - > http://bit.ly/ELMO_SA





THANKS!

Any questions?