

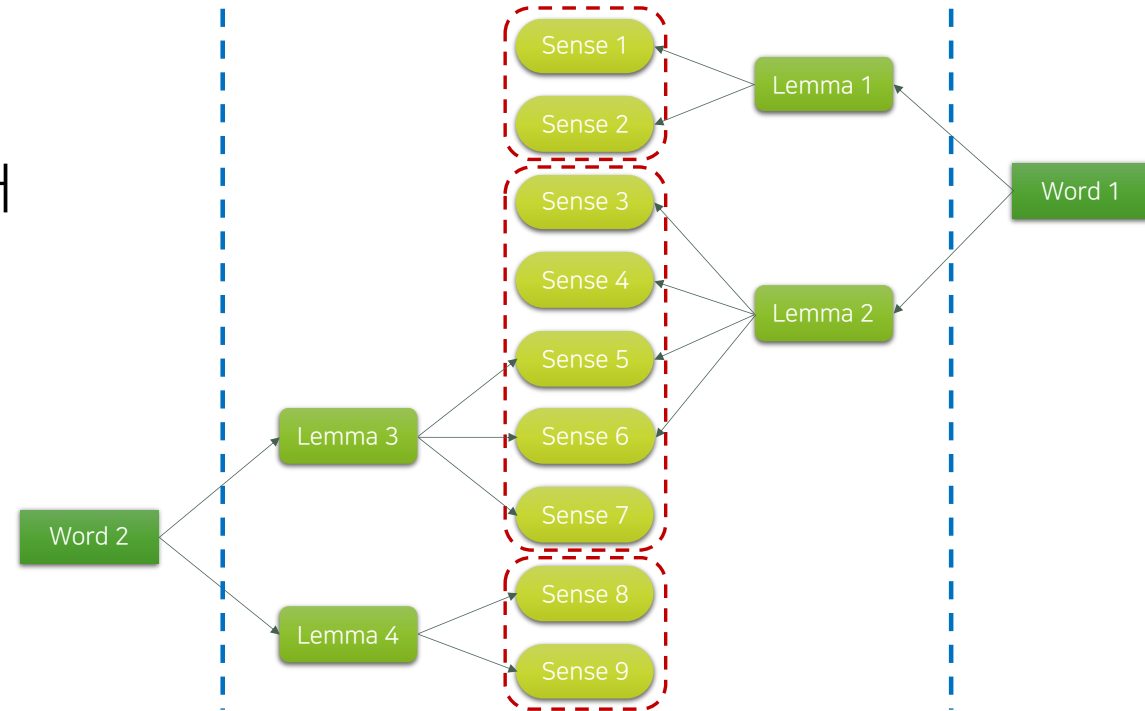
Summary

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Word Sense

- 이산적(or symbolic)인 겉 형태와 달리 내부는 연속적
- 계층적 구조를 지니고 있음
 - 상위어(hypernym)와 하위어(hyponym)의 존재
- 동의어(synonym)와 동형어(homonym), 다의어(polysemy)
- One-hot 인코딩은 이 모든 것과 배치됨



Previous Methods for Word Embedding

- TF-IDF

$$\text{TF-IDF}(w, d) = \frac{\text{TF}(w, d)}{\text{DF}(w)}$$

- TF-IDF Matrix (or Term-frequency Matrix)
 - 각 문서에서의 중요도를 feature로 삼아서 vector를 만들자
- Co-occurrence Matrix by Context Windowing
 - Window 내에서 함께 나타나는 단어들의 빈도를 count하자

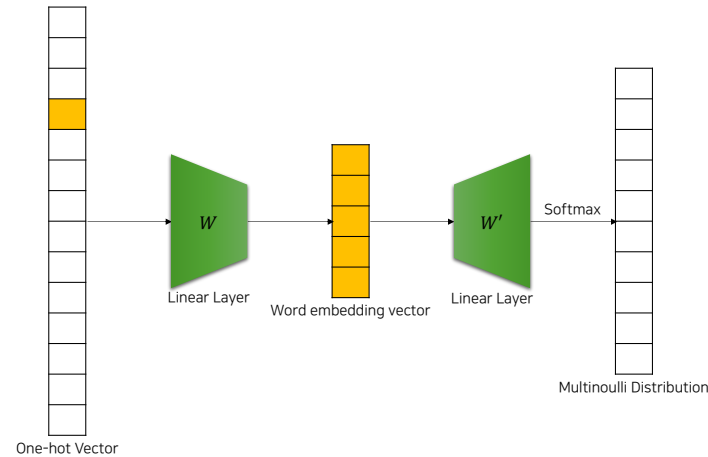


여전히 Sparse

Word Embedding Algorithms

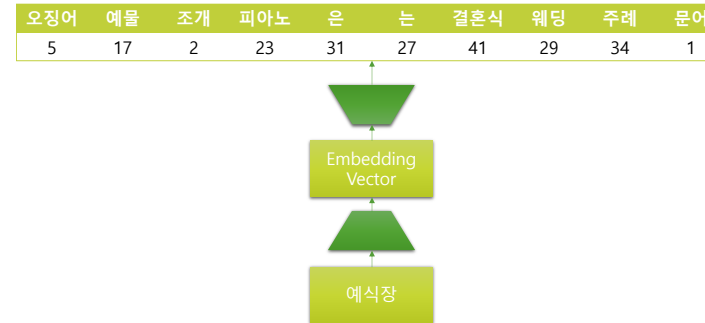
Word2Vec (Skip-gram)

- Classification



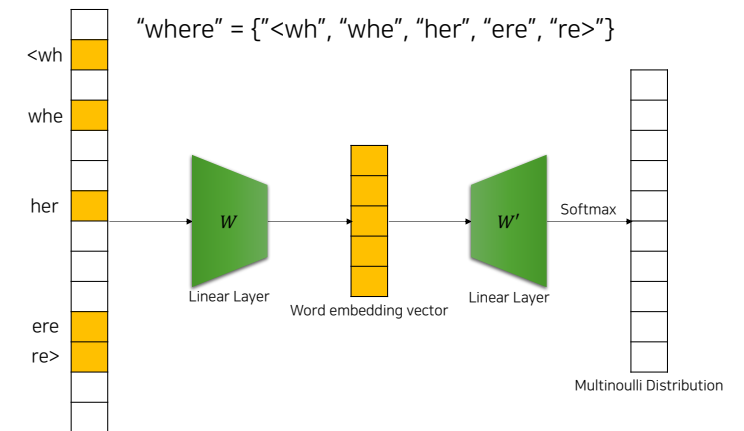
GloVe

- Regression



FastText

- Sum of subword embedding



Embedding Layer

- 신경망에는 one-hot vector를 넣는 것이 정석
- Embedding Layer는 계산의 효율성을 위해 존재

