Interpolation and Back-off

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Interpolation

- 다른 Language Model을 linear하게 일정 비율(λ)로 섞는 것

- general domain LM + domain specific LM
  = general domain에서 잘 동작하는 domain adapted LM

- Examples:
  - 의료 domain ASR, MT system
  - 법률 domain ASR, MT system
  - 특허 domain MT system

$$\tilde{P}(w_n|w_{n-k},\cdots,w_{n-1}) = \lambda P_1(w_n|w_{n-k},\cdots,w_{n-1}) + (1-\lambda)P_2(w_n|w_{n-k},\cdots,w_{n-1})$$

Interpolation

- 그냥 domain specific corpus로 LM을 만들면 장땡 아닌가?
  - 그럼 unseen word sequence가 너무 많을 것 같은데?

- 그냥 전체 corpus를 합쳐서 LM을 만들면 장땡 아닌가?
  - Domain specific corpus의 양이 너무 적어서 반영이 안될 수도?

- Interpolation에서 ratio(λ)를 조절하여 중요도(weight)를 조절
  - 명시적(explicit)으로 섞을 수 있다.
  - General domain test set, Domain specific test set 모두에서 좋은 성능을 찾는 hyper-parameter λ를 찾아야 한다.

Interpolation Example

- "준비 된 진정제 를 투여 합 시다"

- General domain
  - $P(\text{진정제} \mid \text{준비}, \text{된}) = 0.00001$
  - $P(\text{사나이} \mid \text{준비}, \text{된}) = 0.01$
- Domain specialized
  - $P(\text{진정제} \mid \text{준비}, \text{된}) = 0.09$
  - $P(\text{약} \mid \text{준비}, \text{된}) = 0.04$

- $P(\text{진정제} \mid \text{준비}, \text{된}) = 0.5 * 0.09 + (1 - 0.5) * 0.00001 = 0.045005$

Back-off

- 희소성에 대처하는 방법
  - Markov assumption 처럼 n을 점점 줄여가면?
    - 조건부 확률에서 조건부 word sequence를 줄여가면,
    - Unknown(UNK) word가 없다면 언젠가는 확률을 구할 수 있다!

$$\begin{aligned}\tilde{P}(w_n|w_{n-k},\cdots,w_{n-1}) =& \lambda_1 P(w_n|w_{n-k},\cdots,w_{n-1}) \\ &+ \lambda_2 P(w_n|w_{n-k+1},\cdots,w_{n-1}) \\ &+ \cdots \\ &+ \lambda_k P(w_n),\end{aligned}$$

$$where \sum_i \lambda_i = 1.$$

Back-off Example

- P(분석했다 | 비핵화, 선언과는, 거리가, 멀다고)
  - C(비핵화, 선언과는, 거리가, 멀다고, 분석했다) $>$ 0?

- P(분석했다 | 거리가, 멀다고)
  - C(거리가, 멀다고, 분석했다) $>$ 0?

- P(분석했다 | 멀다고)

- P(분석했다)

Summary

- Back-off를 통해 확률 값이 0이 되는 현상은 방지할 수 있음 – OoV 제외
  - 하지만 unseen word sequence를 위해 back-off를 거치는 순간 확률 값이 매우 낮아져 버림
  - 여전히 음성인식(ASR) 등의 활용에서 어려움이 남음

- 전통적인 방식의 NLP에서는 단어를 discrete symbol로 보기 때문에 문제 발생
  - Exact matching에 대해서만 count를 하여, 확률 값을 approximation
  - 다양한 방법을 통해 문제를 완화하려 하지만 근본적인 해결책은 아님
    - Markov Assumption
    - Smoothing and Discounting
    - Interpolation and Back-off