3. 언어 모델 (Language Model)

1) 언어 모델 (Language Model)

언어라는 현상을 모델링하고자 단어 시퀀스(또는 문장)에 확률을 할당하는 모델

→ 가장 자연스러운 단어 시퀀스를 찾아내는 모델

→ 이전 단어들이 주어졌을 때 다음 단어를 예측

스탠포드 대학교 : 언어 모델을 문법이라고 비유하기도 한다.

(언어 모델이 단어들의 조합이 얼마나 적절하지, 또는 해당 문장이 얼마나 적합한지를 알려주는 일

언어 모델링 (language Modeling) : 주어진 단어들로부터 아직 모르는 단어를 예측하는 작업

1) 언어 모델 (Language Model)

언어라는 현상을 모델링하고자 단어 시퀀스(또는 문장)에 확률을 할당하는 모델

통계를 이용하는 방법

- 전통적 접근 방식
- 통계적 언어 모델

인공 신경망을 이용하는 방법

- 최근 핫한 자연어 처리의 신기술(GPT, BERT)
- 통계적 언어 모델

2) 단어 시퀀스의 확률 할당

기계 번역
(Machine Translation)

$P($나는 버스를 탔다$)\ vs\ P($나는 버스를 태운다.$)$

$P($나는 버스를 탔다$) > P($나는 버스를 태운다.$)$

오타 교정
(Spell Correction)

선생님이 교실로 부리나케

$P($달려갔다$)\ vs\ P($잘려갔다$)$

$P($달려갔다$) > P($잘려갔다$)$

음성 인식
(Speech Recognition)

$P($나는 메롱을 먹는다$)\ vs\ P($나는 메론을 먹는다$)$

$P($나는 메롱을 먹는다$) > P($나는 메론을 먹는다$)$

* 조건부 확률

조건부 확률의 이해

	남학생	여학생	계
중학생	100	60	160
고등학생	80	120	200
계	180	180	360

A = 학생이 남학생인 사건
B = 학생이 여학생인 사건
C = 학생이 중학생인 사건
D = 학생이 고등학생인 사건

Ex) 고등학생 중 한 명을 뽑았을 때, 남학생일 확률

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

* 조건부 확률

조건부 확률의 이해

	남학생	여학생	계
중학생	100	60	160
고등학생	80	120	200
계	180	180	360

A = 학생이 남학생인 사건
B = 학생이 여학생인 사건
C = 학생이 중학생인 사건
D = 학생이 고등학생인 사건

1. 학생을 뽑았을 때, 남학생일 확률

$$P(D) = \frac{200\,(\text{고등학생 수})}{360\,(\text{학생 전체 수})} = 0.56$$

* 조건부 확률

조건부 확률의 이해

	남학생	여학생	계
중학생	100	60	160
고등학생	80	120	200
계	180	180	360

A = 학생이 남학생인 사건
B = 학생이 여학생인 사건
C = 학생이 중학생인 사건
D = 학생이 고등학생인 사건

2. 학생을 뽑았을 때, 고등학생이면서 남학생일 확률

$$P(A \cap D) = \frac{80\,(\text{고등학생이면서 남학생 수})}{360\,(\text{학생 전체 수})} = 0.22$$

* 조건부 확률

조건부 확률의 이해

	남학생	여학생	계
중학생	100	60	160
고등학생	80	120	200
계	180	180	360

A = 학생이 남학생인 사건
B = 학생이 여학생인 사건
C = 학생이 중학생인 사건
D = 학생이 고등학생인 사건

3. 고등학생 중 한 명을 뽑았을 때, 남학생일 확률

$$P(A|D) = \frac{P(A \cap D)}{P(D)} = \frac{\frac{80}{360}}{\frac{200}{360}} = \frac{80\,(\text{고등학생이면서 남학생 수})}{200\,(\text{고등학생 수})} = 0.4$$

* 조건부 확률

조건부 확률의 이해

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

$$P(A,B) = P(A)P(B|A)$$

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(C|A,B)P(D|A,B,C)$$

조건부 확률의 연쇄 법칙 (chain rule)

$$P(x_1, x_2, x_3 \dots x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2) \dots P(x_n|x_1, \dots, x_{n-1})$$

2) 주어진 이전 단어들로부터 다음 단어 예측

단어 시퀀스의 확률

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ \ldots, w_n)$$

다음 단어 등장 확률

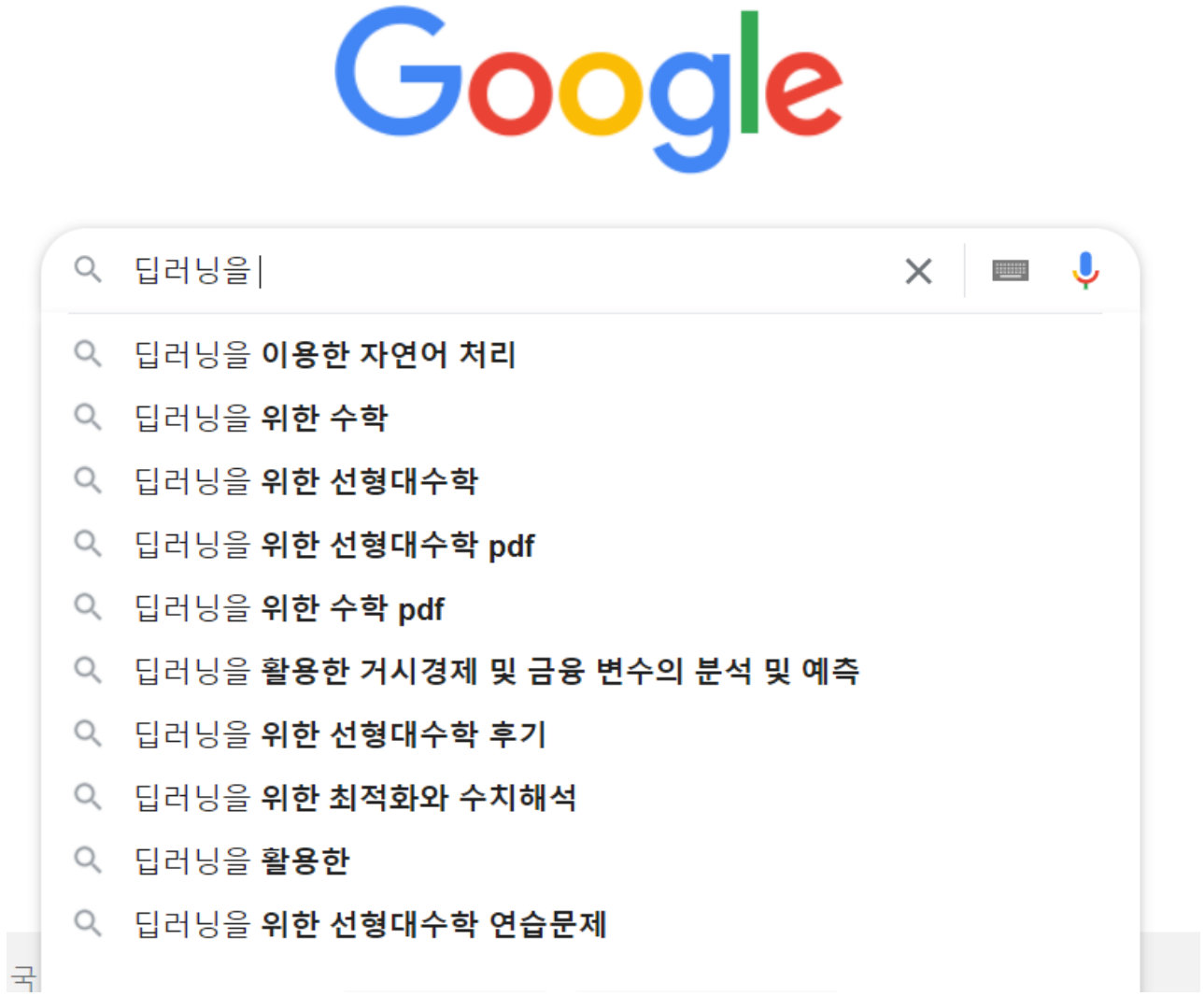
n-1개의 단어가 나열된 상태에서 n번째 단어의 확률

$$P(w_n | \, w_1, w_2, \ldots, w_{n-1})$$

전체 단어 시퀀스 W의 확률

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ \ldots, w_n) = \prod_{i=1}^{n} P(w_n | \, w_1, w_2, \ldots, w_{n-1})$$

3) 검색 엔진에서의 언어 모델의 예



4) 통계적 언어 모델 (Statistical Language Model)

언어라는 현상을 모델링하고자 단어 시퀀스(또는 문장)에 확률을 할당하는 모델

통계를 이용하는 방법

- 통계적 언어 모델 (Statistical Language Model, SLM)

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ \ldots, w_n) = \prod_{i=1}^{n} P(w_n | \ w_1, w_2, \ldots, w_{n-1})$$

$$P(An\ adorable\ little\ boy\ is\ spreading\ smiles) = P(An) \times P(adorable|An) \times P(little|An\ aborable) \times P(boy|An\ aborable\ little) \times P(is|An\ aborable\ little\ boy) \times P(spreading|An\ aborable\ little\ boy\ is) \times P(smiles|An\ aborable\ little\ boy\ is\ spreading)$$

5) 카운트 기반의 접근

전체 단어 시퀀스 W의 확률

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ \ldots, w_n) = \prod_{i=1}^{n} P(w_n | \ w_1, w_2, \ldots, w_{n-1})$$

$$P(An\ adorable\ little\ boy\ is\ spreading\ smiles) = P(An) \times P(adorable|An) \times P(little|An\ aborable) \times P(boy|An\ aborable\ little) \times P(is|An\ aborable\ little\ boy) \times P(spreading|An\ aborable\ little\ boy\ is) \times P(smiles|An\ aborable\ little\ boy\ is\ spreading)$$

이전 단어로부터 다음 단어 등장 확률

An adorable little boy가 나왔을 때, is가 나올 확률

$$P(is|An\ aborable\ little\ boy) = \frac{\text{count}(An\ aborable\ little\ boy\ is)}{count(An\ aborable\ little\ boy)}$$

An adorable little boy가 100번 등장, is가 등장한 경우 30번이라고 하면 $P(is|An\ aborable\ little\ boy) = 30\%$

6) 카운트 기반의 접근의 한계 - 희소 문제

An adorable little boy가 나왔을 때, is가 나올 확률

$$P(is|An\ aborable\ little\ boy) = \frac{\text{count}(An\ aborable\ little\ boy\ is)}{count(An\ aborable\ little\ boy)}$$

- An adorable little boy is라는 단어 시퀀스가 없었다면 이 단어 시퀀스에 대한 확률은 0이 된다.
- 또는 An adorable little boy라는 단어 시퀀스가 없었다면 분모가 0이 되어 확률은 정의되지 않는다.

- 충분한 데이터를 관측하지 못하여 언어를 정확히 모델링하지 못하는 문제를 희소문제(Sparsity Problem)

n-gram 혹은 여러가지 일반화 기법 존재

- 완화하는 방법일 뿐 근본적인 해결책은 되지 못해 통계적 언어 모델에서 인공 신경망 언어 모델로 넘어간다.

7) N-gram 언어모델

카운트에 기반한 통계적 접근을 사용하고 있으므로 SLM의 일종

이전에 등장한 모든 단어를 고려하는 것이 아니라 일부 단어만 고려하는 접근 방법

SLM의 한계는 훈련 코퍼스에 확률을 계산하고 싶은 문장이나 단어가 없을 수 있다는 점

→ $P(is|An\ aborable\ little\ boy) \approx P(is|boy)$

조금 지나친 일반화로 느껴진다면

$P(is|An\ aborable\ little\ boy) \approx P(is|little\ boy)$

단어의 확률을 구하고자 기준 단어의 앞 단어를 전부 포함해서 카운트하는 것이 아니라,

앞 단어 중 임의의 개수만 포함해서 카운트하여 근사하자는 것

8) N-gram

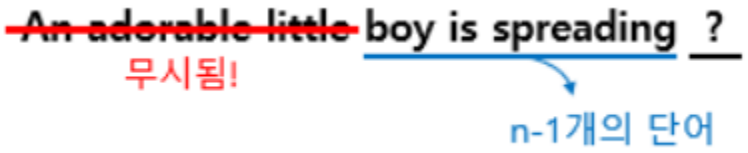
임의의 개수를 정하기 위한 기준을 위해 사용하는 것이 n-gram

n개의 연속적인 단어 나열을 의미

Ex) An adorable little boy is spreading smiles

unigrams : an, adorable, little, boy, is, spreading, smiles
bigrams : an adorable, adorable little, little boy, boy is, is spreading, spreading smiles
trigrams : an adorable little, adorable little boy, little boy is, boy is spreading, is spreading smiles
4-grams : an adorable little boy, adorable little boy is, little boy is spreading, boy is spreading smiles

n=4라고 한 4-gram을 이용한 언어 모델



$$P(w|boy\ is\ spreading) = \frac{\text{count}(boy\ is\ spreading\ w)}{count(boy\ is\ spreading)}$$

boy is spreading가 1,000번 등장

$insult$ 500번 등장

$$P(insult|boy\ is\ spreading) = 0.5$$

$smile$ 200번 등장

$$P(insult|boy\ is\ spreading) = 0.2$$

9) N-gram Language Model의 한계

1) 희소문제

문장에 존재하는 앞에 나온 단어를 모두 보는 것보다 일부 단어만을 보는 것으로 현실적으로 코퍼스에서 카운트 할 수 있는 확률을 높일 수는 있었지만,

→ n-gram 언어 모델도 여전히 n-gram에 대한 희소 문제가 존재

2) n을 선택하는 것은 trade-off 문제

몇 개의 단어를 볼지 n을 정하는 것은 trade-off가 존재

- n을 크게 선택하면 실제 훈련 코퍼스에서 해당 n-gram을 카운트할 수 있는 확률은 적어지므로 희소 문제는 점점 심각
- n이 커질수록 모델 사이즈가 커진다는 문제점

trade-off 문제로 인해 정확도를 높이려면 **n은 최대 5를 넘게 잡아서는 안 된다고 권장**

-	Unigram	Bigram	Trigram
Perplexity	962	170	109

n을 1에서 2, 2에서 3으로 올릴 때마다 성능이 올라가는 것을 보여준다.

10) 한국어에서의 언어 모델

1. 한국어는 어순이 중요하지 않다.

   이전 단어가 주어졌을 때, 다음 단어가 나타날 확률을 구한다.

   하지만, 어순이 중요하지 않다는 것은 어떤 단어이든 나타나도 된다는 의미이다.

   Ex)
   ① 나는 운동을 합니다 체육관에서.
   ② 나는 체육관에서 운동을 합니다.
   ③ 체육관에서 운동을 합니다.
   ④ 나는 운동을 체육관에서 합니다.

   예측이 어렵다.

2. 한국어는 교착어이다.

3. 한국어는 띄어쓰기가 제대로 지켜지지 않는다.

10) 한국어에서의 언어 모델

1. 한국어는 어순이 중요하지 않다.

2. 한국어는 교착어이다.

   띄어쓰기 단위인 어절 단위로 토큰화를 할 경우에는 문장에서 발생가능한 단어의 수가 굉장히 늘어난다.

   Ex) 한국어에는 조사가 존재한다.

   그녀가, 그녀를, 그녀의, 그녀와, 그녀로, 그녀께서, 그녀처럼 등

   → 토큰화를 통해 접사나 조사 등을 분리하는 것이 중요

3. 한국어는 띄어쓰기가 제대로 지켜지지 않는다.

10) 한국어에서의 언어 모델

1. 한국어는 어순이 중요하지 않다.

2. 한국어는 교착어이다.

3. 한국어는 띄어쓰기가 제대로 지켜지지 않는다.

한국어는 띄어쓰기를 제대로 하지 않아도 의미가 전달되며, 띄어쓰기 규칙 또한 상대적으로 까다로운 언어

토큰이 제대로 분리 되지 않은 상태에서 훈련 데이터로 사용된다면 언어 모델은 제대로 동작하지 않는다.