n-Gram Dil Modelleme

Herhangi birisinin söyleyeceği sonraki birkaç kelimeyi tahmin edebilmek kolay mıdır?

Genel olarak sonraki söylenecek sözü tahmin etmek kolay değildir.

n-Gram modelleme, sonraki her kelimeye bir olasılık atayanbir modeldir.

Böylece sezgisel yapı ile modelleme tamamlanır.

n-Gram Dil Modellerine Giriş

Sonraki kelimenin Tahmini

The water of Walden Pond is beautifully ...

blue
green
clear

*refrigerator
*that

Dil Modelleri

- Sonraki sözcükler farklı sistemler tarafından tahmin edilebilir.
  - Her olası potansiyel sözcüğe bir olasılık atanır.
  - Cümlenin tümüne bir olasılık atayabilir.

Sözcük Tahmini Niçin Önemlidir?

- Bu, dil tarafından gerçekleştirilen temel ve yararlı bir iştir.
  - Dilbilgisi ya da yazım kontrolü

Their are two midterms ~~Their~~ There are two midterms

Everything has improve Everything has ~~improve~~ improved

  - Konuşma tanıma

I will be back soonish I will be bassoon dish

Sözcük Tahmini Niçin Önemlidir?

- Büyük dil modellerinin (BDM) nasıl çalıştığı sorusuna cevaptır.
- BDM, sözcükleri tahmin etmek üzere eğitilirler.
  - DM' leri soldan-sağa doğru bir sonraki sözcüğü tahmin etmek üzere öğrenme gerçekleştirirler.
    - ✓ Burada her bir gözlemin değeri, bir önceki gözlemin değerlerine bağlı olarak bir gözlemler dizisi, yani otoregresyon gerçekleştirir.
- BDM leri sözcükleri tahmin ederek metin üretirler (generate).
  - Bu süreç tekrar tekrar sonraki kelime için tekrarlanarak tamamlanır.

Dil Modelleme (DM) nin Biçimsel İfadesi

Amaç: Bir cümlenin ya da w sözcüklerinin sıralanışının olasılığı hesaplanır. :

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

Yapılacak iş: Sonraki sözcüğün olasılığını hesaplamaktır.

$$P(w_5|w_1,w_2,w_3,w_4) \text{ or } P(w_n|w_1,w_2 \ldots w_{n-1})$$

Bir DM aşağıdaki olasılıklardan birini hesaplar

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

Bu olasılıklar nasıl hesaplanır?

- Sonuca ulaşmak için sadece sözcüklerin sayılması ve daha sonra bölünmesi ile doğru hesaplama yapılmış mıdır?

$$P(\texttt{blue}|\texttt{The water of Walden Pond is so beautifully}) = \frac{C(\texttt{The water of Walden Pond is so beautifully blue})}{C(\texttt{The water of Walden Pond is so beautifully})}$$

- Bu doğru değildir.
- Sonuca ulaşmak için çok fazla cümle ile işlem yapmak gerekir.
- Bunların tahmin edilmesi için yeterli veri setini bulmak kolay değildir.

P(W) ya da $P(w_n | w_1, \ldots w_{n-1})$ hesaplanması

P(W) olasılığı nasıl hesaplanır?

P(The, water, of, Walden, Pond, is, so, beautifully, blue)

Sezgisel olarak: Zincir kuralına göre olasılığa güvenmek uygun olabilir (Chain Rule of Probability)

Zincir Kuralı / Chain Rule

- Şartlı olasılıkların tanımı

$$P(B|A) = P(A,B)/P(A) \quad \text{ya da } P(A,B) = P(A)\,P(B|A)$$

- Çoklu sayıda değişken ile:

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

- Genel Zincir Kuralı

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

Zincir Kuralı

Zincir Kuralı, cümledeki öğelerin birleşmesi / eklemlenmesi olasılığını hesaplayarak uygulanır.

$$P(w_{1:n}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2})\ldots P(w_n|w_{1:n-1})$$
$$= \prod_{k=1}^{n} P(w_k|w_{1:k-1})$$

P("The water of Walden Pond") =

P(The) × P(water|The) × P(of|The water) x

P(Walden|The water of) × P(Pond|The water of Walden)

Markov Varsayımı / Assumption


Andrei Markov

Basitleştirilmiş Varsayım

- Bir kelimenin olasılığının yalnızca bir önceki kelimeye bağlı olduğu varsayımı Markov varsayımı olarak adlandırılır.
- Markov modelleri olasılıklı modeller sınıfıdır
  - Ayrıntıyı hesaplamadan gelecek /sonraki bir birimin olasılığının tahmin edilebileceği varsayılır.

$$P(\texttt{blue}|\texttt{The water of Walden Pond is so beautifully})$$

$$\approx \quad P(\texttt{blue}|\texttt{beautifully})$$

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1})$$

Wikimedia commons

n-Gram Modellerin Kullanıldığı Problemler

- n-gram modeller birbirinden uzakta duran bağlılıkları takip edemez.
  - “**The soups** that I made from that new cookbook I bought yesterday **were** amazingly delicious."
- n-gram modeller eş anlamlılık gibi yeni /farklı dizilişleri modellemede de iyi değildir.

Bunlara çözüm: Büyük Dil Modelleridir.

- BDM, çok daha uzun içerikleri yakalayabilir.
- BDM, gömülü uzamların kullanılması nedeni ile eş anlamlılığı daha iyi modelleyebilir ve sıra dışı (alışılmamış) dizgiler üretebilir.

n-Gram Modeller Niçin Öemlidir?

- BDM' leri ile dil işlemeye ait pek çok konuya giriş yapılabilir.

Bunlardan bazıları:

- Eğitim ve test kümelerini belirlemede BDM kullanılır.
  - Perpexity, sohbete dayalı çalışma biçimi olan arama motorudur.
    - ✓ Kişinin sorusundan hareket ederek internette arama yapar ve ihtiyaç duyulan bilgiyi saniyeler içinde sağlar.
- Cümleler oluştururken (generate) örneklemede BDM 'leri kullanılır.
- Dil işlemede, interpolasyon ve geri çekilme (bask off) gibi durumlarda BDM leri kullanılır.

n-Gram Dil Modelleme

n-Gram modellere giriş

n-Gram olasılıkları tahmin etmek

Bigram Model

$P(w_n \mid w_{1:n-1}) \cong P(w_n \mid w_{n-1})$

- Klasik olarak herhangi bir sözcüğün olasılığı, önceki tüm sözcüklere göre tahmin edilir.
  - Örneğin bir sözcüğün ortaya çıkma olasılığının, önceki tüm sözcüklerinin olasılığına yaklaştığı kabul edilir.

$$P(w_n \mid w_{1:n-1})$$

- Hesaplama yapılırken, sadece önceki sözcüğün şartlı olasılığı kullanılır.

$$P(w_n \mid w_{n-1})$$

Örneğin:

P(the |Walden Pond's water is so transparent that)

yerine sonuca

P(the |that)

olasılığı ile yaklaşılır.

Bigram Olasılıkları Tahmin etmek

- bigram ya da n-gram olasılıklarını nasıl tahmin edilir?
  - Olasılıkları tahmin etmenin **sezgisel yoluna** ***Maksimum Olabilirlik Tahmini*** denir.
- **Maximum Likelihood Estimate /MLE**

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} \qquad P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

  - Soldaki eşitlik basitleştirilebilir.
  - Zira, w_{n-1} ile başlayan tüm bigram sayımlarının toplamı, o kelime için unigram sayısına w_{n-1} 'e eşit olmalıdır
- MLE tahmininde, n-gram modelin de parametreleri tahmin edilir.
  - Normalize edilmiş bir derlemden sayılar alınır ve sonra bu sayılar normalize edilir.

Olasılıkları Tahmin Etmek

- MLE tahmininde, n-gram modelin parametreleri tahmin edilir.
- Normalize edilmiş bir derlemden sayılar /counts alınır ve sonra bu sayılar da normalize edilir.
  - Değerler 0 ile 1 arasındadır.
- n-gram yaklaşımı için genel ifade bir sıralanıştaki sonraki sözcüğün şartlı olasılığının yaklaşımıdır.
- n=2 olduğunda bigram, n = 3 olduğunda trigram olarak adlandırılır.

Örnek

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

Berkeley Restaurant Project Cümleleri

- Bazı metinlerin normalleştirildiği örnek kullanıcı sorguları kullanılmıştır.
- Veri setinden alınmış 4 cümle aşağıdadır:

  can you tell me about any good cantonese restaurants close by

  tell me about chez panisse

  i’m looking for a good place to eat breakfast

  when is caffe venezia open during the day

- web sitesinde 9332 cümleden oluşan bir veri seti vardır.

Ham Bigram Saymalar

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Berkeley Restoran Projesi, bigram bir gramerin bir parçasındaki bigram sayımlar

- Değerlerin çoğu sıfırdır.
- Örnek kelimeler birbirleriyle uyumlu olacak şekilde seçilmiştir.
- Eğer matris, sekiz kelimeden oluşan rastgele bir kümeden seçilmiş olsaydı, matris buradakinden daha seyrek olacaktı.

Ham Bigram Olasılıklar

Normalize by unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Result:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

- Normalizasyondan sonra hesaplanan bigram olasılıklar gösterilmektedir.
- Şekildeki her hücre, ait olduğu satır için unigram olasılıkları setinden alınmış uygun unigram ile bölünmüştür.

Cümle Olasılıklarının Bigram Tahminleri

$$P(\langle s\rangle \text{ I want english food } \langle /s\rangle) =$$

$$P(\text{I}|\langle s\rangle) \times P(\text{want}|\text{I}) \times P(\text{english}|\text{want}) \times$$

$$P(\text{food}|\text{english}) \times P(\langle /s\rangle|\text{food})$$

$$= .25 *. 33 * .00110 *. 50 * .68 = .000031$$

What kinds of knowledge do N-grams represent?

P(english|want) = .0011

P(chinese|want) = .0065

P(to|want) = .66

P(eat | to) = .28

P(food | to) = 0

P(want | spend) = 0

P (i | <s>) = .25

n-Gram DM Araçları

SRI Dil Modeli Aracı / SRILM

https://www.sri.com/platform/srilm/

http://www.speech.sri.com/projects/srilm/

KenLM Dil Modeli Aracı

https://kheafield.com/code/kenlm/

n-Gram Language Modeling

n-Gram Olasılıkların Tahmini

Dil Modelleme

Değerlendirme and Perplexity

- DDİ araçları gibi dil modellerinin değerlendirilmesi (evaluation) gerekir.
- n-Gram modeller için olduğu gibi BDM'leri için de standart değerlendirme araçları vardır:
  - Örnek olarak eğitim setleri , test setleri ve perlexity metrikler verilebilir.

n-Gram Modellerin Değerlendirilmesi

"Gerçek / Extrinsic (in-vivo) Evaluation"

A ve B olarak iki modelin karşılaştırılması için:

- Her model gerçek bir göreve uygulanır.
  - Makine çevirisi, konuşma tanıma vs.
- Bu görev /iş çalıştırılır. A ve B için sonuçlar elde edilir.
  - Kaç sözcük doğru çevrilmiştir?
  - Kaç sözcük doğru olarak uyarlanmıştır?
- A ve B modellerinin doğruluğu karşılaştırılır.

Kendiliğinden /Yapay Değerlendirme
Intrinsic (in-vitro) Evaluation

- Gerçek /extrinsic değerlendirme her zaman mümkün olmaz
  - Pahalıdır ve fazla zaman alır.
  - Pek çok uygulamaya her zaman genelleştirilemez.
- Intrinsic evaluation: perplexity
  - Dil modelinin performansı doğrudan sözcükler tahmin edilerek ölçülür.
  - Gerçek uygulamanın performansına gerek yoktur.
  - Dil modelleri için genel bir metrik sunulur.
  - n-Grams modeller için olduğu gibi BDM ‘lerine de uygundur.