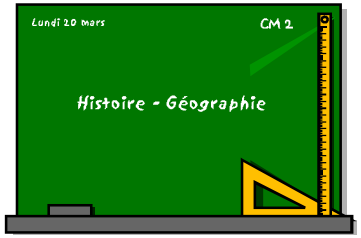
무지개
동산

Dil Modelleri

Prof..Dr. Banu Diri



Dilin modellenmesinin amacı

- Konuşma tanıma (Speech recognition)
- El yazısı tanıma (Handwriting recognition)
- İmla hatalarının düzeltilmesi (Spelling correction)
- Makine çeviri sistemleri (Machine translation systems)
- Optik karakter tanıma (Optical character recognizers)

El yazısı tanıma (Handwriting recognition)

Bankadaki veznedara bir not verildiğini düşünün,
ve veznedar notu "I have a gub" olarak okusun.
(cf. Woody Allen)

NLP burada yardımcı olur ....

gub ingilizcede anlamlı bir kelime değildir.

gun, gum, Gus, ve gull olabilir, fakat gun kelimesinin banka ile ilişki olasılığı daha fazla olduğundan "gub", "gun" olarak alınır.

İmla hatalarının kontrolünde

Birbirinin yerine sıklıkla geçebilen kelimeler

piece/peace, whether/weather, their/there …

Örnek:

"On Tuesday, the whether …"

"On Tuesday, the weather …"

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

W

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

Wh

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

Wha

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What d

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Letter-based Language Models

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do you

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do you think

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do you think the

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do you think the next

Harf-tabanlı (Letter-based) dil modelleri

Shannon's Game

Guess the next letter:

What do you think the next letter is?

Guess the next word:

What do you think the next word is?

- zero-order approximation: harflerin sıraları birbirinden bağımsız
  - xfoml rxkhrjffjuj zlpwcwkcy ffeyvkcqsghyd

- first-order approximation: harfler birbirinden bağımsızdır, fakat dildeki (Ingilizce) harflerin dağılımlarına göre meydana gelir
  - ocro hli rgwr nmielwis eu ll nbnesebya th eei alhentppa oobttva nah

- second-order approximation: bir harfin görülme olasılığı bir önceki harfe bağlıdır
  - On ie antsoutinys are t inctore st bes deamy achin dilonasive tucoowe at teasonare fuzo tizin andy tobe seace ctisbe
- third-order approximation: bir harfin görülme olasılığı kendisinden önce gelen iki harfe bağlıdır
  - in no ist lat whey cratict froure birs grocid pondenome of demonstures of the reptagin is regoactiona of cre

Farklı diller için yüksek frekanslı trigram'lar:

İngilizce: THE, ING, ENT, ION
Almanca: EIN, ICH, DEN, DER
Fransızca: ENT, QUE, LES, ION
İtalyanca: CHE, ERE, ZIO, DEL
İspanyolca:QUE, EST, ARA, ADO

Dillerdeki hece benzerlikleri

Aynı aile içerisinde bulunan diller birbirlerine diğer dillere göre daha fazla benzer

Aynı aile içerisinde yer alan diller birbirlerine nasıl benzerler ?

- Hece tabanlı benzerlik

Aile içerisinde yer alan her bir dildeki en fazla kullanılan kelimeler çıkarılır;

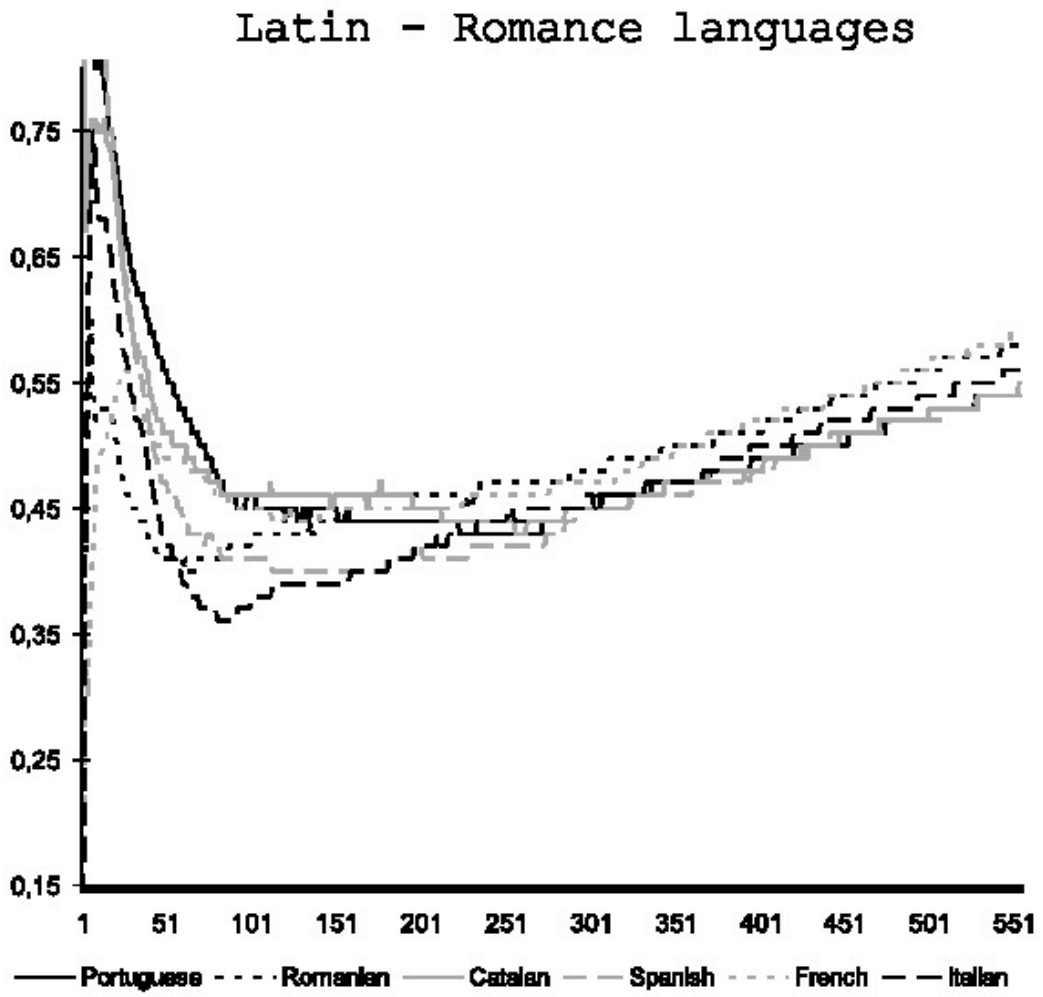
- Kelimeler hecelerine ayrılır;
- Hecelerin frekansları hesaplanır;
- Heceye dayalı dildeki benzerlik hesaplanır

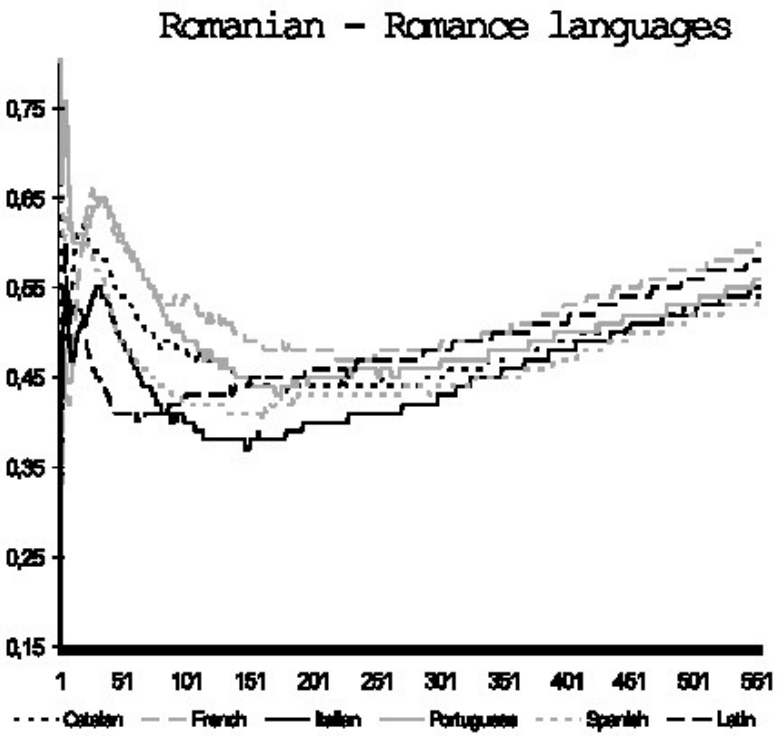
Örnek: Romance dilleri ailesi

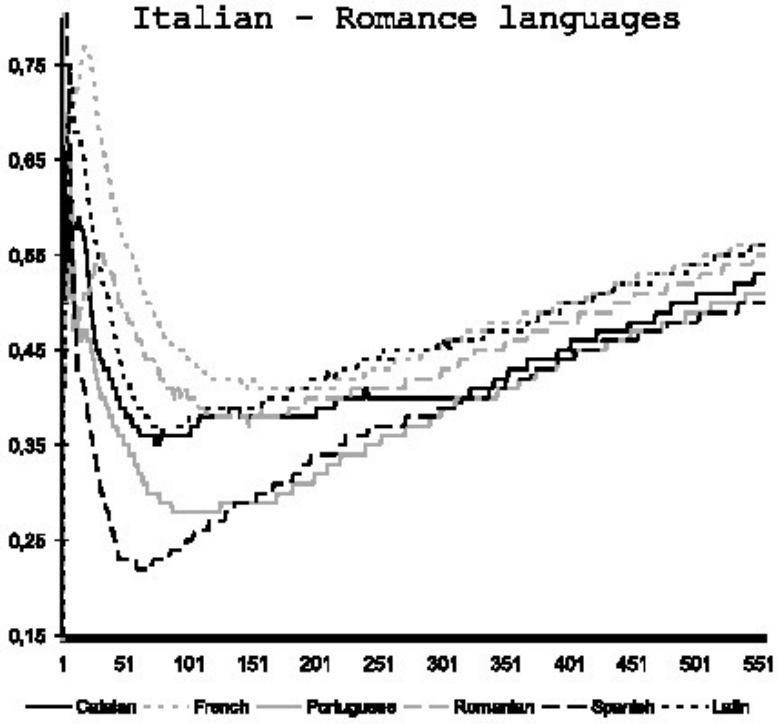
Romance dillerindeki heceler

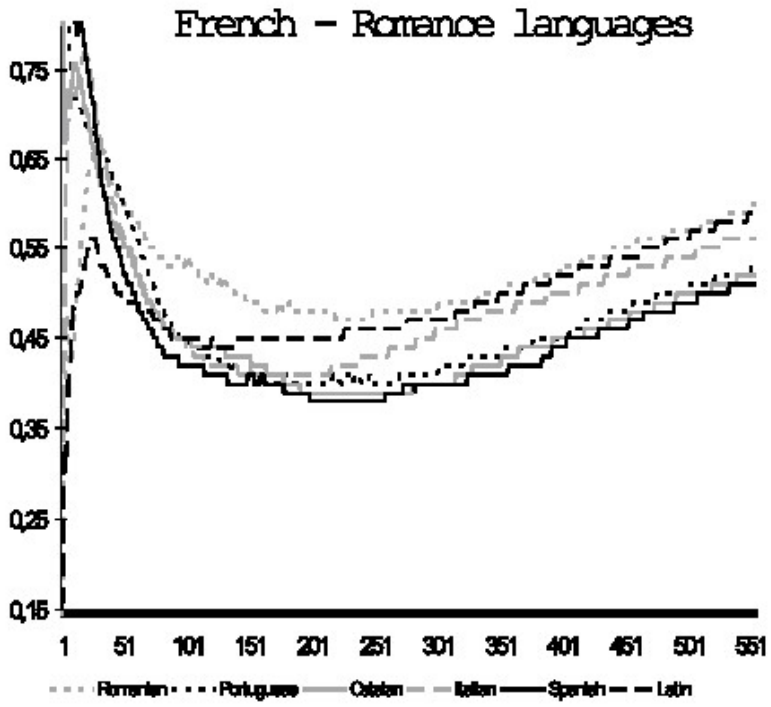
Language	The percentage covered by the first ··· syllables						No. syllables	
	100	200	300	400	500	561	type	token
Latin	72%	86%	92%	95%	98%	100%	561	3922
Romanian	63%	74%	80%	84%	87%	90%	1243	6591
Italian	75%	85%	91%	94%	96%	97%	803	7937
Portuguese	69%	84%	91%	95%	97%	98%	693	6152
Spanish	73%	87%	93%	96%	98%	99%	672	7477
Catalan	62%	77%	84%	88%	92%	93%	967	5624
French	48%	61%	67%	72%	76%	78%	1738	5691

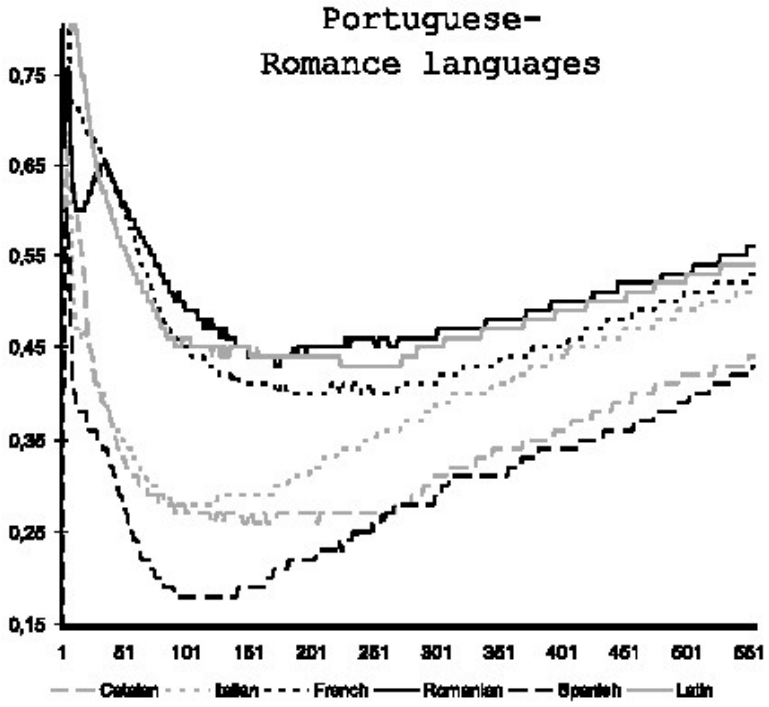
Latin-Romance Dillerinin Benzerliği

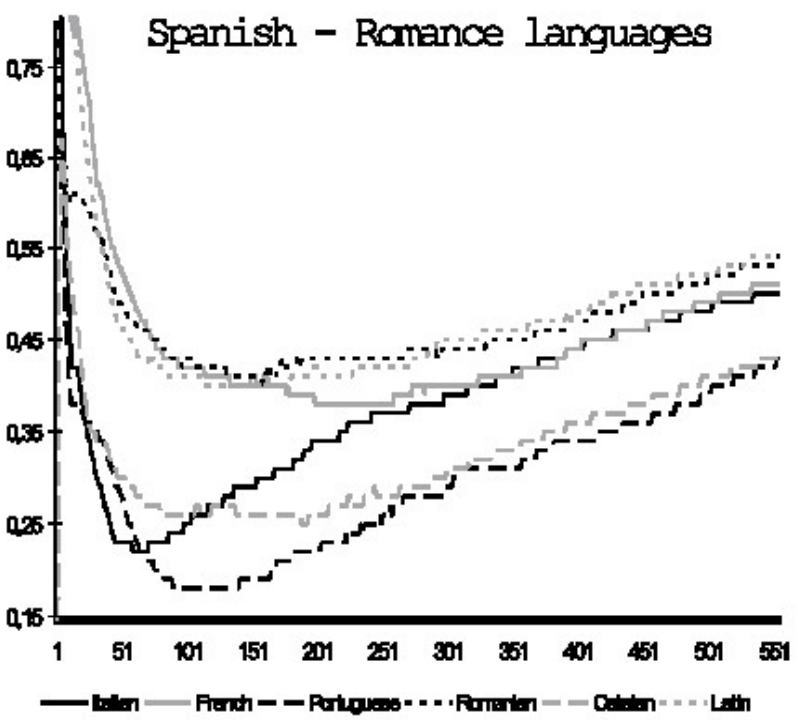


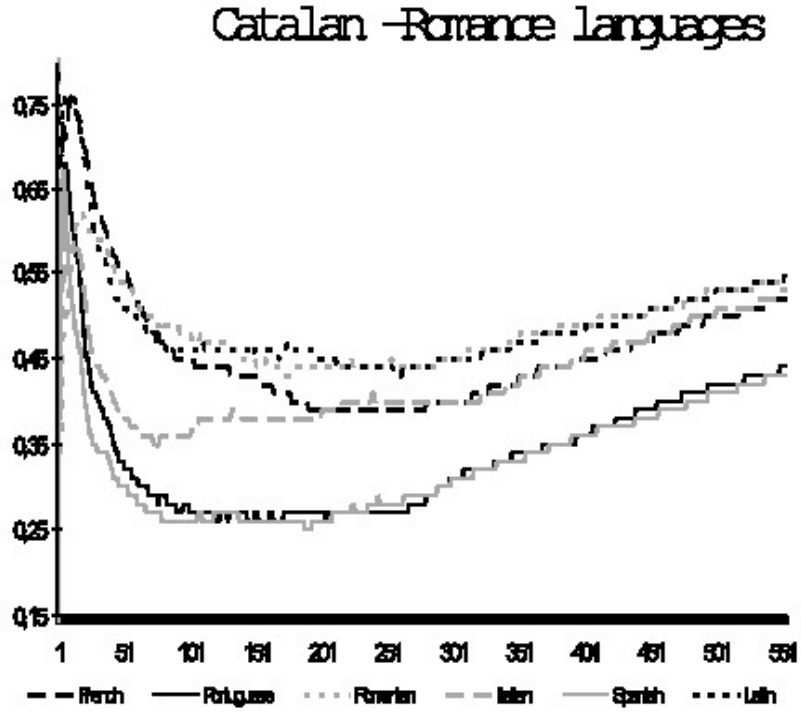

Romanian – Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
Catalan
French
Italian
Portuguese
Spanish
Latin


Italian – Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
Catalan
French
Portuguese
Romanian
Spanish
Latin


French – Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
Romanian
Portuguese
Catalan
Italian
Spanish
Latin


Portuguese-
Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
Catalan
Italian
French
Romanian
Spanish
Latin


Spanish - Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
Italian
French
Portuguese
Romanian
Catalan
Latin


Catalan -Romance languages
0,75
0,65
0,55
0,45
0,35
0,25
0,15
1
51
101
151
201
251
301
351
401
451
501
551
French
Portuguese
Romanian
Italian
Spanish
Latin

Kelime Tabanlı Dil Modelleri

Dil modeli, S cümlesinin olasılığını (likelihood/probability) hesaplamaya yardımcı olur, P(S).

En basit haliyle, herbir kelime bir sonraki w kelimesini eşit olasılıkla izler (0-gram).

V sözlüğünün boyunun |V| olduğunu farzedelim. Buna göre n uzunluğundaki S cümlesinin olasılığı (likelihood) = 1/|V| × 1/|V| … × 1/|V| olarak hesaplanır.

Eğer bir dilde 100,000 kelime varsa, gelecek olan herbir kelimenin olasılığı

1/100000 = .00001 dır.

- Kesin: gelecek olan her kelimenin olasılığı kelimenin frekansı ile ilişkilidir.
  - cümlenin olasılığı $S = P(w_1) \times P(w_2) \times \ldots \times P(w_n)$
  - herbir kelimenin olasılığı diğer kelimelerin olasılıklarından bağımsızdır.
- En kesin: daha önce verilmiş olan kelimenin olasılığına bakılır (n-gram).
  - S cümlesinin olasılığı = $P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_{n-1})$
  - her kelimenin olasılığının diğer kelimelerin olasılıklarına bağlı olduğu farzedilir.

Zincir Kuralı (Chain Rule)

Şartlı olasılık

$P(A_1,A_2) = P(A_1) \cdot P(A_2|A_1)$

Zincir kuralı, çoklu olaylar ile genelleştirildiğinde

$(A_3|A_1,A_2)...P(A_n|A_1...A_{n-1})$

yazılır.

Örnekler:

P(the dog) = P(the) P(dog | the)

P(the dog bites) = P(the) P(dog | the) P(bites| the dog)

For a Word String

Bir string $w_1^n = w_1 \ldots w_n$ oluşsun.
Bu stringin olasılığı

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1..w_2)\ldots P(w_n|w_1\ldots w_{n-1})$$

$$= \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

Fakat bu yaklaşım genelde, bir kelime sırasının olasılığını belirlemek için çok yararlı değildir. Hesaplama maliyeti çok yüksektir.

Markov Yaklaşımı

$P(w_n|w_1^{n-1})$ nasıl hesaplanır ?

P(rabbit|I saw a) yerine, P(rabbit|a) kullanılabilir.

Bigram model:
P(the barking dog) = P(the|<start>)P(barking|the)P(dog|barking)

Markov modeller olasılık modeli olup, gelecek bir birimin olasılığını çok uzak geçmişine bakmaksızın yakın geçmişe bakarak olasılıklarını tahmin etmedir.

N=2 (bigram): $P(w_1^n) \approx \Pi_{k=1}^{n} P(w_k|w_{k-1})$; w_0 = <start>

Terminoloji

Cümle/Sentence: yazım dilinin birimi

Söz-laf/Utterance: konuşma dilinin birimi

Kelime formu/Word Form: Kullanıma göre kelimenin formunun değiştirilmesi

Types (V): corpus içerisinde yer alan ayrık kelime sayısı (vocabulary size)

Token (N_T): corpus içerisindeki toplam kelime sayısı

Şimdiye kadar gözüken kelime sayısı (T): corpus da görülen ayrık kelime sayısı (V ve N_T küçüktür)

Basit N-Grams

N-gram model, gelecek kelimeyi tahmin edebilmek için önceki N-1 adet kelimeyi kullanır.

$$P(w_n \mid w_{n-N+1}\, w_{n-N+2\ldots}\, w_{n-1})$$

unigrams: P(dog)
bigrams: P(dog | big)
trigrams: P(dog | the big)
quadrigrams: P(dog | chasing the big)

N-Grams kullanımı

Hatırla

N-gram: $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$

Bigram: $P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1})$

Bigram grameri:

Cümlenin olasılığı P(cümle), cümle içerisinde yer alan bütün bigram'ların olasılıklarının çarpına yakın bir değerdir.

Örnek:

P(I want to eat Chinese food) =

P(I | <start>) P(want | I) P(to | want) P(eat | to)

P(Chinese | eat) P(food | Chinese)

Bigram Gramer Parçaları

Eat on	.16	Eat Thai	.03
Eat some	.06	Eat breakfast	.03
Eat lunch	.06	Eat in	.02
Eat dinner	.05	Eat Chinese	.02
Eat at	.04	Eat Mexican	.02
Eat a	.04	Eat tomorrow	.01
Eat Indian	.04	Eat dessert	.007
Eat today	.03	Eat British	.001

Ek gramer

<start> I	.25	Want some	.04
<start> I'd	.06	Want Thai	.01
<start> Tell	.04	To eat	.26
<start> I'm	.02	To have	.14
I want	.32	To spend	.09
I would	.29	To be	.02
I don't	.08	British food	.60
I have	.04	British restaurant	.15
Want to	.65	British cuisine	.01
Want a	.05	British lunch	.01

Cümlenin olasılığının hesaplanması

P(I want to eat British food) =
P(I|<start>) P(want|I) P(to|want) P(eat|to)
P(British|eat) P(food|British) =
.25×.32×.65×.26×.001×.60 = .000080

P(I want to eat Chinese food) = .00015

Olasılıklar dünyanın bildiği bir gerçeği göstermektedir.

N-grams sonuçları

Sparse data

Eğitim seti içerisinde bütün N-gram'lar yer almayabilir ve bu n-gram'ların frekansı sıfır olarak alınır, bu yüzden yumuşatma (smoothing) tekniklerine ihtiyaç duyulur.

P("And nothing but the truth") ≈ 0.001

P("And nuts sing on the roof") ≈ 0

Bigram Sayıları

	I	Want	To	Eat	Chinese	Food	lunch
I	8	1087	0	13	0	0	0
Want	3	0	786	0	6	8	6
To	3	0	10	860	3	0	12
Eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
Food	19	0	17	0	0	0	0
Lunch	4	0	0	0	0	1	0

Bigram Olasılıkları: Unigram sayısını kullanarak

Unigram değerleri

I	Want	To	Eat	Chinese	Food	Lunch
3437	1215	3256	938	213	1506	459

Computing the probability of I I

$P(I|I) = C(I\ I)/C(I) = 8 / 3437 = .0023$

Bigram grameri VxV boyutunda bir olasılıklar matrisidir. V, sözlük boyutu

Bigram Grameri

Bigram için kullanılan formül (parameter estimation)

$P(w_n|w_{n-1}) = C(w_{n-1}w_n)/C(w_{n-1})$

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Yumuşatma (Smoothing) Teknikleri

Çok geniş bir corpora' ya sahip olsak bile N-gram eğitim matrisi sparse bir matrisdir (Zipf's law).

Çözüm: gözükmeyen n-gram olasılıklarını tahmin etmek

Add-one Smoothing

Her n-gram değerine 1 eklenir.

$N_T/(N_T+V)$ katsayısı ile normalize edilir

Yumuşatılmış toplam: $c_i' = (c_i+1)\frac{N_T}{N_T+V}$
(smoothed count)

Yumuşatılmış olasılık
(smoothed probability):

$P'(w_i) = c_i'/N_T$

Add-one Smoothed Bigrams

$$P(w_n|w_{n-1}) = C(w_{n-1}w_n)/C(w_{n-1})$$

	I	want	to	eat	Chinese	food	lunch
I	8	1087	0	13	0	0	0
want	3	0	786	0	6	8	6
to	3	0	10	860	3	0	12
eat	0	0	2	0	19	2	52
Chinese	2	0	0	0	0	120	1
food	19	0	17	0	0	0	0
lunch	4	0	0	0	0	1	0

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

$$P'(w_n|w_{n-1}) = [C(w_{n-1}w_n)+1]/[C(w_{n-1})+V]$$

	I	want	to	eat	Chinese	food	lunch
I	9	1088	1	14	1	1	1
want	4	1	787	1	7	9	7
to	4	1	11	861	4	1	13
eat	1	1	3	1	20	3	53
Chinese	3	1	1	1	1	121	2
food	20	1	18	1	1	1	1
lunch	5	1	1	1	1	2	1

	I	want	to	eat	Chinese	food	lunch
I	.0018	.22	.00020	.0028	.00020	.00020	.00020
want	.0014	.00035	.28	.00035	.0025	.0032	.0025
to	.00082	.00021	.0023	.18	.00082	.00021	.0027
eat	.00039	.00039	.0012	.00039	.0078	.0012	.021
Chinese	.0016	.00055	.00055	.00055	.00055	.066	.0011
food	.0064	.00032	.0058	.00032	.00032	.00032	.00032
lunch	.0024	.00048	.00048	.00048	.00048	.00096	.00048

	Bugün	de	her	zamanki	gibi	eve	gidiyorum
Bugün	0	1	0	0	0	0	0
de	0	0	1	0	0	0	0
her	0	0	0	1	0	0	0
zamanki	0	0	0	0	1	0	0
gibi	0	0	0	0	0	1	0
eve	0	0	0	0	0	0	1
gidiyorum	0	0	0	0	0	0	0

$$P_{Bugün\ \text{de}} = \frac{1}{5},\ P_{de\ \text{her}} = \frac{1}{5},\ P_{her\ \text{zamanki}} = \frac{1}{5},\ P_{\text{zamanki gibi}} = \frac{1}{5},\ P_{gibi\ \text{gidiyorum}} = \frac{1}{5}$$

	Bugün	de	her	zamanki	gibi	eve	gidiyorum
Bugün	1	2	1	1	1	1	1
de	1	1	2	1	1	1	1
her	1	1	1	2	1	1	1
zamanki	1	1	1	1	2	1	1
gibi	1	1	1	1	1	2	1
eve	1	1	1	1	1	1	2
gidiyorum	1	1	1	1	1	1	1

$$P_{Bugün\ de} = \frac{2}{5+49} = \frac{2}{54},\ P_{de\ her} = \frac{2}{54},\ P_{her\ zamanki} = \frac{2}{54},\ P_{zamanki\ gibi} = \frac{2}{54},\ P_{gibi\ gidiyorum} = \frac{2}{54}$$

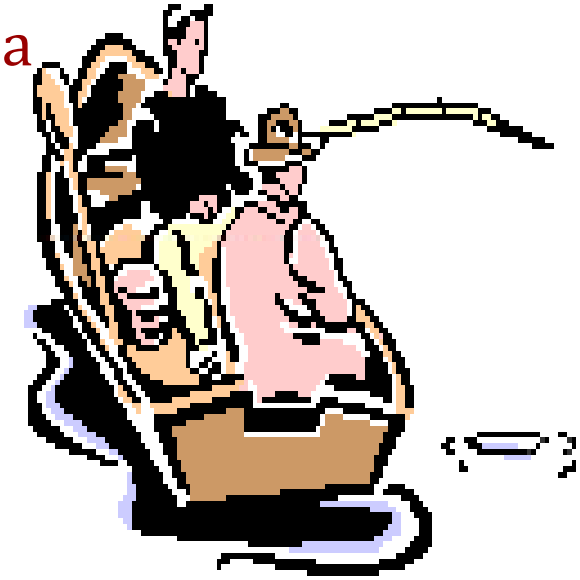
.

$$P_i^* = \frac{C_i + 1}{N + V},\ \mathrm{i} = 1, 2, \ldots, \mathrm{t}$$

Diğer yumuşatma teknikleri: Good-Turing

Balık tutmaya çıktığınızı hayal edin...

Ve 10 tane aynalı sazan (carp), 3 tane morina (cod), 2 tane tuna, 1 tane alabalık (trout), 1 tane som balığı (salmon), 1 tane de yılan balığı (eel) yakalamış olun.



Bir sonra ki yakalanacak olan balığın yeni bir tür olma olasılığı nedir ?

3/18

Back-off Yöntemi

Hatırlatma :N-gram'lar her zaman (N-1) gram'a göre daha duyarlıdır.

Fakat, N-gram'lar (N-1) gram'a göre daha fazla sparse 'tır.

Bu ikisi nasıl birleştirilir ?

N-gram'ın frekans değeri uygun değilse (sıfır ise) vazgeçilir (back-off) ve (n-1) gram'a dönülür. Monogram'a kadar devam edilir. Recursive bir yapı sözkonusudur.

$$\hat{P}(w_i|w_{i-2}w_{i-1}) = \begin{cases} \tilde{P}(w_i|w_{i-2}w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 \tilde{P}(w_i|w_{i-1}), & \text{if } C(w_{i-2}w_{i-1}w_i) = 0 \text{ and } C(w_{i-1}w_i) > 0 \\ \alpha_2 \tilde{P}(w_i), & \text{otherwise} \end{cases}$$