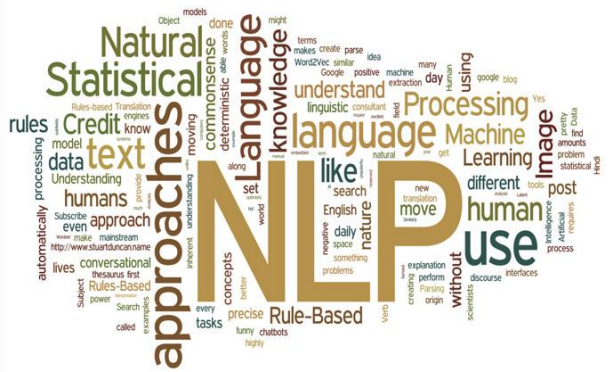


به نام خدا

گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر

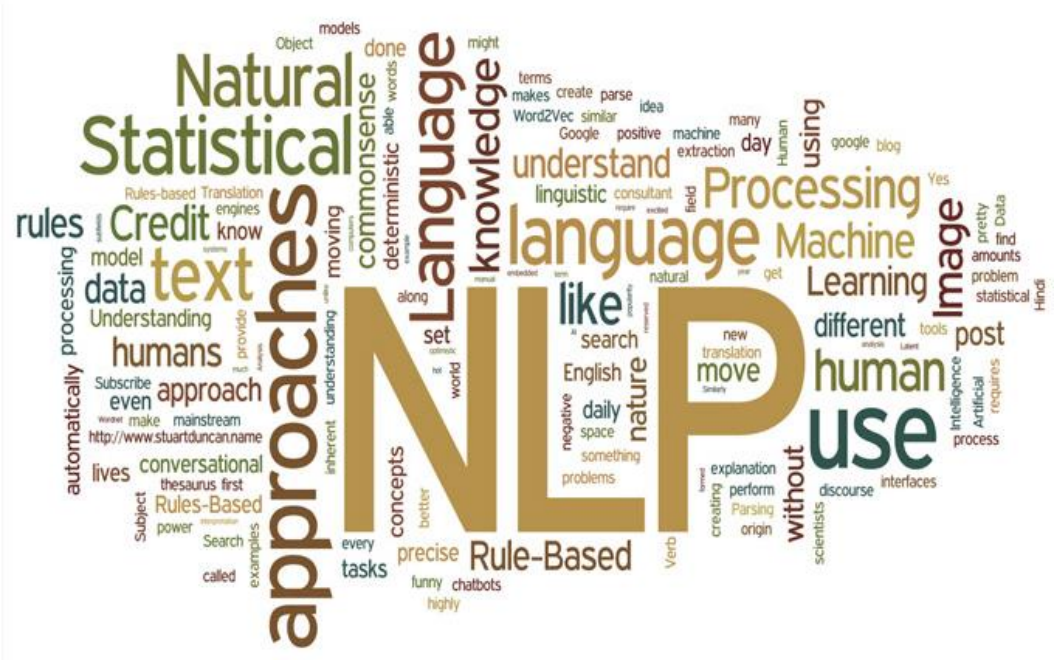
بخش سوم

مدلسازی زبانی (Language Modeling)

حمیدرضا برادران کاشانی

اهداف این بخش:

- مقدمه ای بر مدل زبانی: تعریف LM و کاربرد آن، فرضیه مارکوف
- نحوه تخمین احتمالات N-gram
- ارزیابی مدل زبانی و Perplexity
- تعمیم پذیری LM و صفرها
- هموارسازی Laplace



Hamidreza Baradaran Kashani

مقدمه ای بر مدلسازی زبانی

Hamidreza Baradaran Kashani

مدل های زبانی احتمالاتی

- هدف مدلسازی زبانی
  - تخصیص یک مقدار احتمال به یک جمله در زبان طبیعی

- کاربردهای مدل زبانی
  - ترجمه ماشینی یا MT

P(**high** winds tonite) > P(**large** winds tonite)

Hamidreza Baradaran Kashani

مدل های زبانی احتمالاتی

❖ کاربردهای مدل زبانی (ادامه)

❖ تصحیح خطا املایی (Spell Correction)

The office is about fifteen **minuets** from my house

P(about fifteen **minutes** from) > P(about fifteen **minuets** from)

❖ بازشناسی گفتار (Speech recognition)

P(I saw a van) >> P(eyes awe of an)

❖ خلاصه سازی ، QA و ...

Hamidreza Baradaran Kashani

مدل های زبانی احتمالاتی

❖ **مدلسازی زبانی احتمالاتی با هدف:**

❖ **محاسبه مقدار احتمال یک جمله یا در حالت کلی تر دنباله ای از کلمات**

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

❖ **یا محاسبه مقدار احتمال وقوع یک کلمه به شرط رخداد کلمات قبلی آن**

$$P(w_5 | w_1, w_2, w_3, w_4)$$

❖ **به مدلی که هر یک از دو مقدار فوق را محاسبه کند، گرامر (grammar) یا به اصطلاح متداولتر مدل زبانی یا LM گفته می شود.**

Hamidreza Baradaran Kashani

نحوه محاسبه P(W)

- ما متن زیر را داریم:

Its water is so transparent that …

- می خواهیم احتمال مشترک (**joint probability**) را برایش محاسبه کنیم:

P(its, water, is, so, transparent, that)

- راه حل: استفاده از قانون زنجیره احتمالات (Chain Rule of Probability)

Hamidreza Baradaran Kashani

یادآوری: قانون زنجیره احتمالات

❖ احتمال شرطی (Conditional Probability):

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

$$P(A,B) = P(A|B)\,P(B) \quad \text{or} \quad P(A,B) = P(B|A)\,P(A)$$

❖ برای بیش از دو متغیر:

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

❖ قاعده زنجیره ای در حالت کلی:

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

Hamidreza Baradaran Kashani

نحوه محاسبه P(W)

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P(“its water is so transparent”) =

P(its) ×

P(water | its) ×

P(is | its water) ×

P(so | its water is) ×

P(transparent | its water is so)

Hamidreza Baradaran Kashani

نحوه تخمین احتمالات

- سوال:
  - آیا محاسبه احتمال با استفاده از شمارش تعداد رخدادهای یک دنباله کلمات روشی منطقی است؟

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count\ (\text{its water is so transparent that the})}{Count\ (\text{its water is so transparent that})}$$

- خیر تعداد زیاد حالات برای رخداد جملات مختلف عملا این روش را غیرممکن می کند.
- برای تخمین این احتمالات به پیکره های بسیار بزرگی نیاز است.

Hamidreza Baradaran Kashani

فرضیه مارکوف (Markov Assumption)

❖ ساده سازی روابط احتمالات:

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

❖ و یا:

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

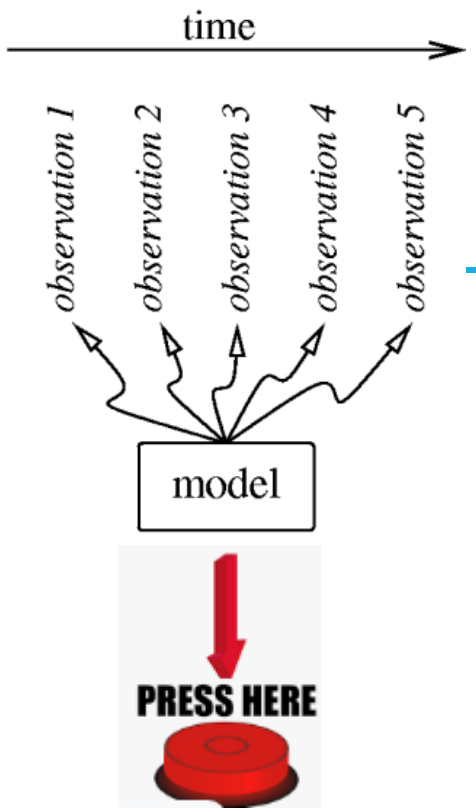
Hamidreza Baradaran Kashani

فرضیه مارکوف (Markov Assumption)

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i \mid w_1 w_2 \dots w_{i-1})$$

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-k} \dots w_{i-1})$$

Hamidreza Baradaran Kashani



فرضیه مارکوف (Markov Assumption)

❖ ساده ترین حالت مدل مارکوف: مدل Unigram

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

❖ برای مثال جملات زیر بطور خودکار از یک مدل Unigram تولید شدند:

fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, 'm, july, bullish, that, or, limited, the

Hamidreza Baradaran Kashani

فرضیه مارکوف (Markov Assumption)

❖مدل Bigram

❖ احتمال شرطی هر کلمه به کلمه قبلی آن

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

❖ برای مثال جملات زیر بطور خودکار از یک مدل Bigram تولید شدند:

texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november

Hamidreza Baradaran Kashani

فرضیه مارکوف (Markov Assumption)

❖مدل های N-gram

- ❖ می توان مدلهای **4-gram** و **5-gram** هم داشت.

- ❖ البته همچنان این مدل های زبانی نیز دقت بسیار بالایی ندارند:
  - ❖ دلیل این موضوع وابستگی های طولانی (**long-distance dependencies**) در یک زبان هست:

?

"The computer which I had just put into the machine room on the fifth floor **crashed**."

Hamidreza Baradaran Kashani

تخمین احتمالات ها N-gram

Hamidreza Baradaran Kashani

تخمین احتمالات Bi-gram

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

❖ روش تخمین حداکثر درستنمایی (Maximum Likelihood)

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

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

Hamidreza Baradaran Kashani

Sequence Notation

Corpus: This is great ... teacher drinks tea.

$w_1 \; w_2 \; w_3 \quad \ldots \quad w_{498} \; w_{499} \; w_{500}$

$m = 500$

$$w_1^m = w_1 \; w_2 \; \ldots \; w_m$$

$$w_1^3 = w_1 \; w_2 \; w_3$$

$$w_{m-2}^m = w_{m-2} \; w_{m-1} \; w_m$$

Hamidreza Baradaran Kashani

مثال: تخمین احتمال Unigram

Corpus: I am happy because I am learning

Size of corpus m = 7

$$P(I) = \frac{2}{7} \qquad P(happy) = \frac{1}{7}$$

Probability of unigram: $$P(w) = \frac{C(w)}{m}$$

Hamidreza Baradaran Kashani

مثال: تخمین احتمال Bigram

Corpus: I am happy because I am learning

$$P(am|I) = \frac{C(I\ am)}{C(I)} = \frac{2}{2} = 1$$

$$P(happy|I) = \frac{C(I\ happy)}{C(I)} = \frac{0}{2} = 0$$ ✖ I happy

$$P(learning|am) = \frac{C(am\ learning)}{C(am)} = \frac{1}{2}$$

Probability of a bigram: $$P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$$

Hamidreza Baradaran Kashani

مثال: تخمین احتمال Trigram

Corpus: I am happy because I am learning

$$P(happy|I\ am) = \frac{C(I\ am\ happy)}{C(I\ am)} = \frac{1}{2}$$

Probability of a trigram: $$P(w_3|w_1^2) = \frac{C(w_1^2\ w_3)}{C(w_1^2)}$$

$$C(w_1^2\ w_3) = C(w_1\ w_2\ w_3) = C(w_1^3)$$

Hamidreza Baradaran Kashani

مثال: تخمین احتمال N-gram

Probability of N-gram: $$P(w_N|w_1^{N-1}) = \frac{C(w_1^{N-1}\, w_N)}{C(w_1^{N-1})}$$

$$C(w_1^{N-1}\, w_N) = C(w_1^N)$$

Hamidreza Baradaran Kashani

Adding starting and end symbols

- Start of sentence symbols <s>
- End of sentence symbol </s>

Hamidreza Baradaran Kashani

Start of sentence token <s>

How to resolve the first term in the bigram approximation?

the teacher drinks tea

$$P(the\ teacher\ drinks\ tea) \approx P(the)P(teacher|the)P(drinks|teacher)P(tea|drinks)$$

<s> the teacher drinks tea

$$P(<s>\ the\ teacher\ drinks\ tea) \approx P(the|<s>)P(teacher|the)P(drinks|teacher)P(tea|drinks)$$

Hamidreza Baradaran Kashani

Start of sentence token <s> for N-grams

For trigrams, the first two words don't have enough contexts.

- Trigram:

$$P(the\ teacher\ drinks\ tea) \approx P(the)P(teacher|the)P(drinks|the\ teacher)P(tea|teacher\ drinks)$$

the teacher drinks tea => <s> <s> the teacher drinks tea

$$P(w_1^n) \approx P(w_1|<s>\ <s>)P(w_2|<s>\ w_1)...P(w_n|w_{n-2}\ w_{n-1})$$

- N-gram model: add N-1 start tokens <s>

Hamidreza Baradaran Kashani

End of sentence token </s>

$$P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$$

There is one case where the simplification does not work

When word x is the last word of the sentence

Corpus:

<s> Lyn drinks chocolate

<s> John drinks

$$\sum_w C(drinks\ w) = 1$$

$$C(drinks) = 2$$

Hamidreza Baradaran Kashani

End of sentence token </s> - solution

- Bigram

<s> the teacher drinks tea => <s> the teacher drinks tea </s>

$$P(the|<s>)P(teacher|the)P(drinks|teacher)P(tea|drinks)P(</s>|tea)$$

Corpus:

<s> Lyn drinks chocolate </s>

<s> John drinks </s>

$$\sum_{w} C(drinks\ w) = 2$$

$$C(drinks) = 2$$

Hamidreza Baradaran Kashani

End of sentence token </s> for N-grams

- N-gram => just one </s>

E.g. Trigram:

the teacher drinks tea => <s> <s> the teacher drinks tea </s>

Hamidreza Baradaran Kashani

Example: Bigram

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(sentence) = \frac{2}{3} * \frac{1}{2} * \frac{1}{2} * \frac{2}{2} = \frac{1}{6}$$

$$P(John|<s>) = \frac{1}{3}$$

$$P(chocolate|\ \text{drinks}\ ) = \frac{1}{2}$$

$$P(</s>|tea) = \frac{1}{1}$$

$$P(Lyn|<s>) =?= \frac{2}{3}$$

Hamidreza Baradaran Kashani

تخمین احتمالات Bi-gram

❖ مثال

<s> I am Sam </s>

<s> Sam I am </s>

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

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

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

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

Hamidreza Baradaran Kashani

Build a Language Model on a Corpus

1. Make count matrix
2. Make probability matrix from count matrix

Hamidreza Baradaran Kashani

Count matrix

- Rows: unique corpus (N-1)-grams
- Columns: unique corpus words

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

- Bigram count matrix

Corpus: <s>I study I learn</s>

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	1	1
study	0	0	1	0	0
learn	0	1	0	0	0

"study I" bigram →

Hamidreza Baradaran Kashani

Probability matrix

- Divide each cell by its row sum

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

$$sum(row) = \sum_{w \in V} C(w_{n-N+1}^{n-1}, w) = C(w_{n-N+1}^{n-1})$$

Corpus: <s>I study I learn</s>

Count matrix (bigram)

	<s>	</s>	I	study	learn	sum
<s>	0	0	1	0	0	1
</s>	0	0	0	0	0	0
I	0	0	0	1	1	2
study	0	0	1	0	0	1
learn	0	1	0	0	0	1

Probability matrix

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	0.5	0.5
study	0	0	1	0	0
learn	0	1	0	0	0

Hamidreza Baradaran Kashani

Language model

- probability matrix => language model
  - Sentence probability
  - Next word prediction

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	0.5	0.5
study	0	0	1	0	0
learn	0	1	0	0	0

Sentence probability:

<s> I learn </s>

$$P(sentence) =$$
$$P(I|<s>)P(learn|I)P(</s>|learn) =$$
$$1 \times 0.5 \times 1 =$$
$$0.5$$

Hamidreza Baradaran Kashani

تخمین احتمالات Bi-gram

❖ مثال دیگر: جملات مربوط به یک سیستم دیالوگ درباره رستورانی در برکلی

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i'm looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i'm looking for a good place to eat breakfast
- when is caffe venezia open during the day

Hamidreza Baradaran Kashani

تخمین احتمالات Bi-gram

❖ محاسبه تعداد Bigram ها بصورت خام (محاسبه شده از ۹۲۲۲ جمله)

	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

Hamidreza Baradaran Kashani

تخمین احتمالات Bi-gram

❖ نرمالیزه با تعداد unigram ها

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

	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

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

Hamidreza Baradaran Kashani

تخمین احتمال جمله با احتمالات Bi-gram ها

P(<s> I want english food </s>) =

P(I|<s>)

× P(want | I)

× P(english | want)

× P(food | english)

× P(</s> | food)

= .000031

Hamidreza Baradaran Kashani

عوامل تاثیرگذار در مقادیر احتمالات

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

چالش های عملی

- برای محاسبات از لگاریتم استفاده می کنیم:

- جلوگیری از underflow
- سریعتر بودن جمع نسبت به ضرب

$$\text{Log}(p_1 \acute{} \ p_2 \acute{} \ p_3 \acute{} \ p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

چالش های عملی

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1})$$

- All probabilities in calculation <=1 and multiplying them brings risk of underflow
- Use log of the probabilities in Probability matrix and calculations

$$log(P(w_1^n)) \approx \sum^{n} log(P(w_i|w_{i-1}))$$

- Converts back from log

$$P(w_1^n) = exp(log(P(w_1^n))$$

Hamidreza Baradaran Kashani

Generating text with language model

Corpus:

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

1. (<s>, Lyn) or (<s>, John)?
2. (Lyn,eats) or (Lyn,drinks) ?
3. (drinks,tea) or (drinks,chocolate)?
4. (tea,</s>) - always

Algorithm:

1. Choose sentence start
2. Choose next bigram starting with previous word
3. Continue until </s> is picked

ارزیابی و Perplexity

Hamidreza Baradaran Kashani

ارزیابی: مدل زبانی ما چقدر کارآیی دارد؟

- آیا مدل زبانی ما قادر به تشخیص جملات خوب از بد است؟

- به این معنا که آیا مدل زبانی ما احتمالات بیشتری به جملات واقعی تر و متداول تر یک زبان نتیجه می دهد؟

- و همچنین احتمال کمتر به جملاتی که متداول نبوده و مثلا از لحاظ گرامری مشکل دارند.

Hamidreza Baradaran Kashani

ارزیابی: مدل زبانی ما چقدر کارآیی دارد؟

- برای انجام این ارزیابی:
  - یک مجموعه یادگیری (Training Set) برای تخمین پارامترهای مدل یا بطور کلی یادگیری مدل استفاده می شود.
  - از یک مجموعه تست (Test Set) برای بررسی میزان کارآیی مدل استفاده می شود.
  - مجموعه تست یک مجموعه داده دیده نشده (unseen dataset) و در واقع متفاوت از مجموعه یادگیری است.
  - حال یک معیار ارزیابی (evaluation metric) به ما میگوید که چقدر مدل ما بر روی مجموعه تست کارآیی خوبی دارد.

Hamidreza Baradaran Kashani

یک روش ارزیابی: Extrinsic Evaluation

- روش ارزیابی **Extrinsic** :
  - بهترین روش برای ارزیابی و مقایسه علمکرد ۲ مدل مثلا مدلهای **A** و **B** است.

- در این روش کافی است از دو مدل در یک کاربرد خاص استفاده کرد مثلا:
  - تصحیح املایی، ترجمه ماشینی، بازشناسی گفتار و ...

- در ادامه دقت ۲ مدل را محاسبه و با یکدیگر مقایسه کرد:
  - دقت در تصحیح املایی: تعداد کلمات به درستی تصحیح شده
  - دقت در ترجمه ماشینی: تعداد جملات به درستی ترجمه شده

Hamidreza Baradaran Kashani

مشکلات روش Extrinsic Evaluation

- روش **Extrinsic** بسیار زمان بر و هزینه بر است
  - بخصوص برای کاربردی مثل **MT** یا بازشناسی گفتار

- روش پیشنهادی دیگر: **Intrinsic evaluation or Perplexity**
  - این روش به ارزیابی درونی خود مدل بدون توجه به یک کاربرد بیرونی می پردازد.
  - روش **Intrinsic** نمی تواند تقریب خوبی از روش **Extrinsic** باشد، مگر:
    - داده تست بسیار شبیه به داده یادگیری باشد.
  - روش **Intrinsic** تنها در محیطهای آزمایشی (**pilot**) مناسب است.

Hamidreza Baradaran Kashani

مفهوم Perplexity

❖ بازی شانون (**Shanon game**): بازی پیش بینی کلمه بعدی در یک جمله

I always order pizza with cheese and _____

The 33rd President of the US was _____

I saw a _____

- mushrooms 0.1
- pepperoni 0.1
- anchovies 0.01
- ….
- fried rice 0.0001
- ….
- and 1e-100

❖ واضح است که **Unigram** ها در این بازی بسیار ضعیف عمل می کنند (چرا؟)

❖ روش بهتر: استفاده از **N-gram** هایی که احتمالات معقولتری به کلمات نسبت می دهند.

Hamidreza Baradaran Kashani

محاسبه Perplexity

- برای محاسبه **Perplexity** بایستی:
  - احتمال وقوع دنباله کلمات در مجموعه تست محاسبه شود.
  - مقدار احتمال معکوس شده و سپس ریشه برابر با تعداد کلمات آن (**N**) محاسبه شود.
  - اصطلاحا گفته می شود: مقدار احتمال با مقدار "طول دنباله کلمات به صورت توان معکوس" نرمالیزه شود.

$$\begin{aligned} \mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

Hamidreza Baradaran Kashani

محاسبه Perplexity

$$\mathrm{PP}(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}$$

Chain rule:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

For bigrams:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

کمینه کردن PP معادل با بیشینه کردن احتمال وقوع کلمات است

Hamidreza Baradaran Kashani

محاسبه Perplexity

- concatenate all sentences in W

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{P(w_i|w_{i-1})}}$$

W → test set

w_i → i-th word in test set

m → number of all words in entire test set W including </s> but not including <s>

چرا سیمبل های شروع یعنی توکن <s>
را در محاسبه **m** حذف می کنیم؟

Hamidreza Baradaran Kashani

محاسبه Perplexity

E.g. m=100

$$P(W) = 0.9 => PP(W) = 0.9^{-\frac{1}{100}} = 1.00105416$$

$$P(W) = 10^{-250} => PP(W) = (10^{-250})^{-\frac{1}{100}} \approx 316$$

- Smaller perplexity = better model
- Character level models PP < word-based models PP

Hamidreza Baradaran Kashani

محاسبه Log Perplexity

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{P(w_i|w_{i-1})}}$$

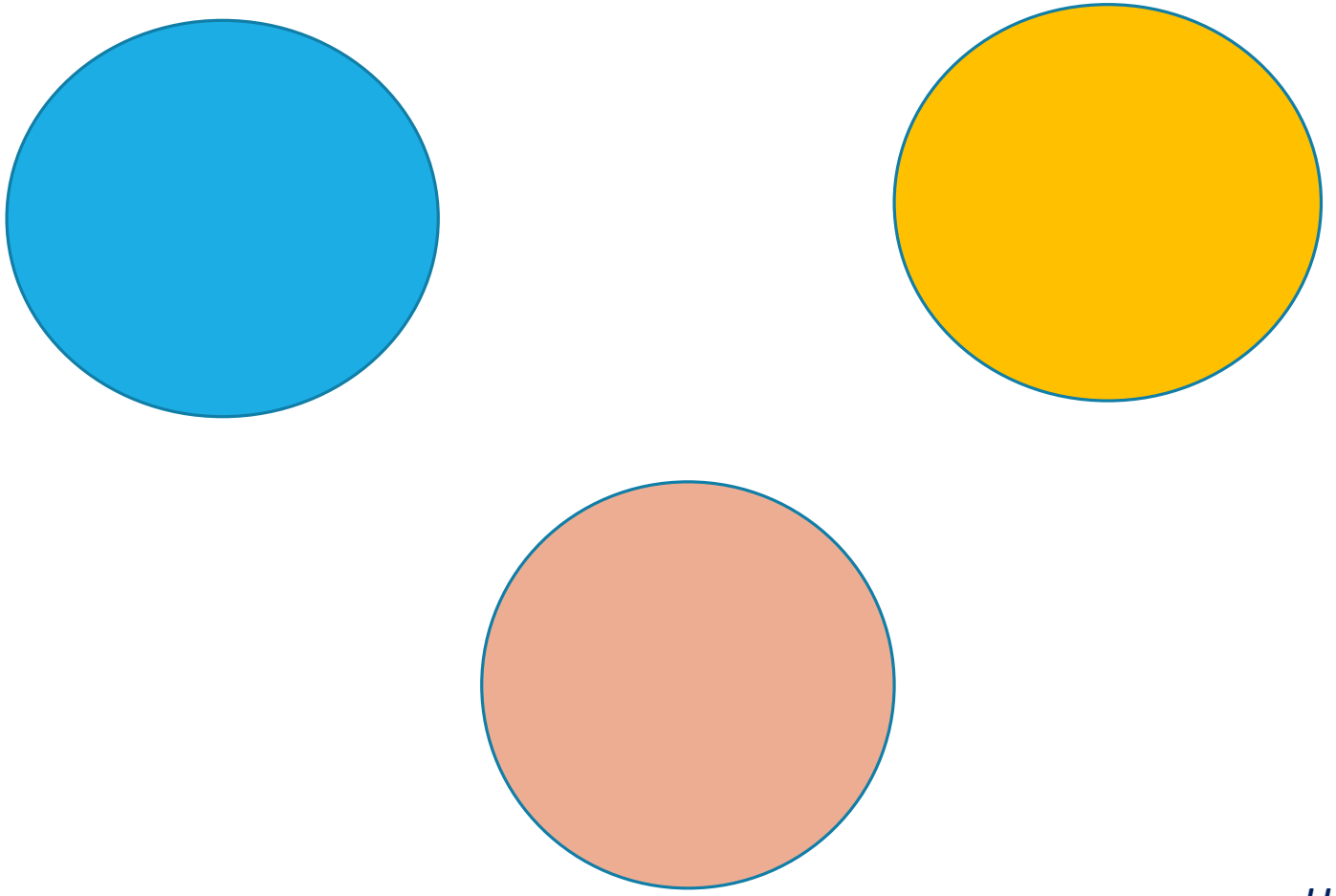
$$logPP(W) = -\frac{1}{m}\sum_{i=1}^{m} log_2(P(w_i|w_{i-1}))$$

Hamidreza Baradaran Kashani

مفهوم Perplexity بر اساس شانون

- Perplexity متوسط فاکتور شاخه بندی است (Average Branching Factor).
- مثلا در هر جای جمله که قرار داریم، چند انتخاب برای کلمه بعدی بطور متوسط وجود دارد؟ (انتخاب ما چند شاخه می شود)

- مثلا اگر ۱۰ شاخه یا انتخاب داشته باشیم و احتمال همه مساوی باشند، Perplexity برابر با ۱۰ است.



Hamidreza Baradaran Kashani

انتخاب مدل بهتر با Perplexity کمتر

- یادگیری مدل زبانی با ۳۸ میلیون کلمه
- تست روی یک مجموعه ۱.۵ میلیون کلمه ای از WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Hamidreza Baradaran Kashani

انتخاب مدل بهتر با Perplexity کمتر

Unigram
Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
Bigram
Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
Trigram
They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

[Figure from *Speech and Language Processing* by Dan Jurafsky et. al]

Hamidreza Baradaran Kashani

Out of Vocabulary Words

- Unknown word = Out of vocabulary word (OOV)
- special tag <UNK> in corpus and in input
- Choosing vocabulary

Out of Vocabulary Words

Using <UNK> in corpus

- Create vocabulary V
- Replace any word in corpus and not in V by <UNK>
- Count the probabilities with <UNK> as with any other word

Example

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

Corpus

<s> Lyn drinks chocolate </s>

<s> <UNK> drinks <UNK> </s>

<s> Lyn <UNK> chocolate </s>

Min frequency f=2

Vocabulary

Lyn, drinks, chocolate

Input query

<s>Adam drinks chocolate</s>

<s><UNK> drinks chocolate</s>

Out of Vocabulary Words

How to create vocabulary V

- Criteria:
  - Min word frequency f
  - Max |V|, include words by frequency
- Use <UNK> sparingly
- Perplexity - only compare LMs with the same V

تعمیم پذیری LM
و وجود صفرهای احتمالات

Hamidreza Baradaran Kashani

روش تولید جمله شانون

❖ روش مصورسازی شانون (Shannon visualization)

❖ در این روش برای تولید یک جمله مثلا با استفاده از Bi-gram ها

❖ ابتدا یک بایگرم (<s>, w) بصورت رندوم بر اساس احتمال آن انتخاب می شود (در واقع بایگرم هایی که مقدار احتمال بیشتری دارند، احتمال انتخاب بیشتری دارند).

❖ سپس یک بایگرم (w, x) بصورت رندوم بر اساس احتمال آن انتخاب می شود.

❖ ادامه انتخاب بایگرم ها بصورت رندوم بر اساس احتمالشان تا رسیدن به انتهای جمله یعنی </s>

```
<s> I
    I want
      want to
           to eat
              eat Chinese
                  Chinese food
                          food  </s>
I want to eat Chinese food
```

Hamidreza Baradaran Kashani

تولید جمله از پیکره شکسپیر

- یک مدل زبانی (در حالت کلی N-gram) بر روی پیکره شکسپیر یادگیری شده است، یعنی:
  - یک مدل احتمالاتی داریم که احتمال هر N-gram را به ما می دهد (محاسبه احتمال).
  - همچنین با کمک آن می توانیم جملاتی را بصورت رندوم (مشابه با روش مصورسازی شانون) تولید کنیم (تولید جمله با استخراج بایگرم های محتمل تر).

- در ادامه تعدادی جمله بر اساس Unigram، Bigram، Trigram و Quadrigram تولید شده است.

Hamidreza Baradaran Kashani

تولید جمله از پیکره شکسپیر

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first you enter
Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
Will you not tell me who I am?
It cannot be but so.
Indeed the short and the long. Marry, 'tis a noble Lepidus.

Hamidreza Baradaran Kashani

پیکره شکسپیر

- تعداد Token های آن برابر با ۸۸۴۶۴۷ است (N=884,647 tokens)
- تعداد Type ها یا دایره لغات آن برابر با ۲۹۰۶۶ است (V=29,066)
- تعداد Bigram Types (نه Bigram Tokens) برابر با ۳۰۰۰۰۰ است (یعنی جفت کلمات منحصر بفرد)
- تعداد Bigram Types ممکن در شکسپیر: V^2= 844 million
- بنابراین ۹۹.۹۶٪ از Bigram ها استخراج نشده اند، که در نتیجه آن:
  - بنابراین صفرهای بسیار زیادی در جدول محاسبه احتمالات Bigram ها ایجاد می شود.
  - این درصد برای Quadrigram بیشتر نیز است.

Hamidreza Baradaran Kashani

چالش Overfitting

- استخراج N-gram ها زمانی کارآیی بالایی خواهد داشت که پیکره یادگیری شبیه به پیکره تست باشد.
- در حالیکه در عمل و دنیای واقعی عمدتا تطبیقی میان مجموعه های یادگیری و تست وجود ندارد.
- بر این اساس نیاز به مدلهایی داریم که قابلیت تعمیم (generalization) خوبی داشته باشند.
- یکی از روشهای تعمیم پذیری مدل:
  - حل کردن مساله مقادیر صفر برای احتمالات N-gram ها بواسطه دیده نشدن کلمات مجموعه تست در مجموعه یادگیری

Hamidreza Baradaran Kashani

مثال رخداد احتمالات صفر

- Training set:
  … denied the allegations
  … denied the reports
  … denied the claims
  … denied the request

- Test set
  … denied the offer
  … denied the loan

P(“offer” | denied the) = 0

چه راه حلی برای حل این
صفرها وجود دارد؟

Hamidreza Baradaran Kashani

Smoothing

Hamidreza Baradaran Kashani

Smoothing

- Missing N-grams in corpus
- Smoothing
- Backoff and interpolation

Hamidreza Baradaran Kashani

Smoothing

Missing N-grams in training corpus

- Problem: N-grams made of known words still might be missing in the training corpus "John" , "eats" in corpus ✖ "John eats"

- Their counts cannot be used for probability estimation

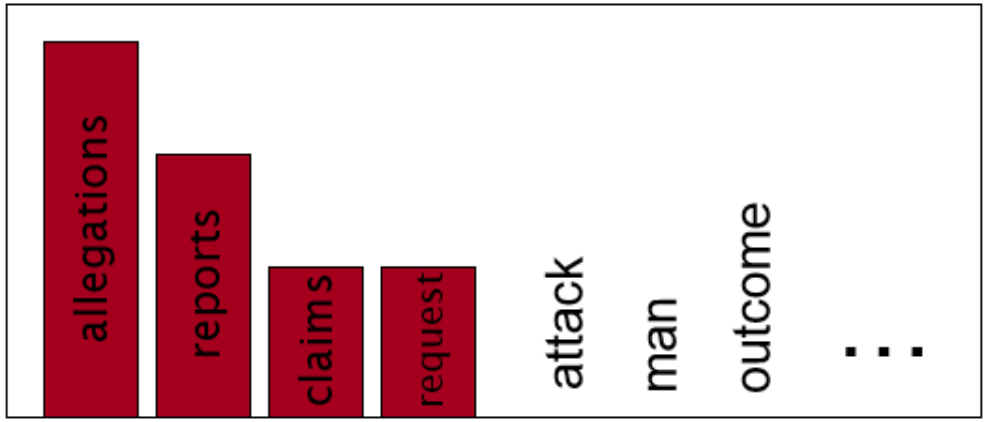
$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

← Can be 0

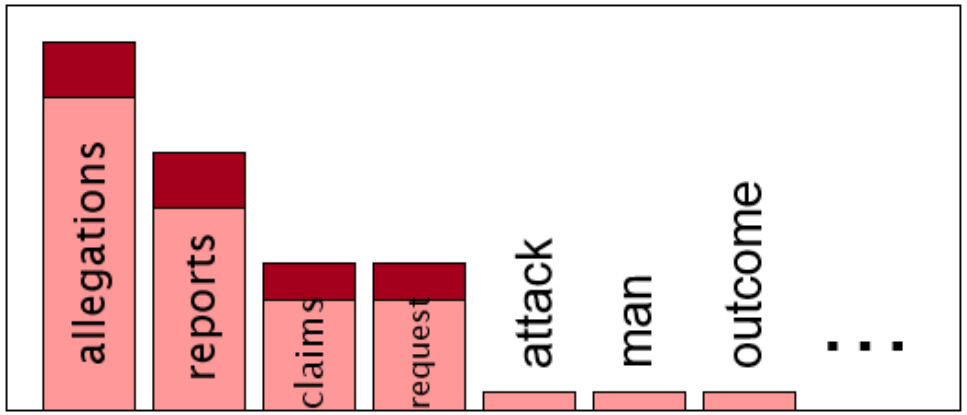
Hamidreza Baradaran Kashani

بینشی نسبت به هموارسازی

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



❖ زمانی که تعداد رخدادهای **n-gram** ها **sparse** هستند:

❖ در هموارسازی (با هدف تعمیم بهتر مدل روی دادگان دیده نشده) به نوعی از مقادیر رخداد سایر کلمات برای افزایش رخداد کلمات دیده نشده استفاده می شود.

Hamidreza Baradaran Kashani

تخمین Add-one

- نام دیگر روش: هموارسازی Laplace
- در واقع وانمود می کنیم که هر کلمه را یکبار بیشتر دیده ایم.
- بنابراین مقدار یک واحد را به تمام شمارش های هر رخداد اضافه می کنیم.
- تخمین MLE

$$P_{MLE}(w_n \mid w_{n-1}) = \frac{c(w_{n-1}, w_n)}{c(w_{n-1})}$$

- تخمین Add-one

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{\sum_{w \in V}(C(w_{n-1}, w) + 1)} = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}$$

Hamidreza Baradaran Kashani

تخمین Add-k

*** Using Add-k smoothing for larger corpus**

- Add-k smoothing

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1},w_n)+k}{\sum_{w\in V}(C(w_{n-1},w)+k)} = \frac{C(w_{n-1},w_n)+k}{C(w_{n-1})+k*V}$$

Hamidreza Baradaran Kashani

پیکره Berkeley Restaurant با اضافه کردن یک به وقوع بایگرم ها

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

Hamidreza Baradaran Kashani

احتمال رخداد بایگرم ها با روش هموارسازی Laplace

$$P^*(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Hamidreza Baradaran Kashani

محاسبه مجدد تعداد رخدادهای بایگرم ها با استفاده از احتمالات هموارشده Laplace

$$c^*(w_{n-1}w_n) = \frac{[C(w_{n-1}w_n)+1] \times C(w_{n-1})}{C(w_{n-1})+V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

Hamidreza Baradaran Kashani

مقایسه با تعداد رخدادهای بایگرم های خام

	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

	i	want	to	eat	chinese	food	lunch	spend
i	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

Backoff

- If N-gram missing => use (N-1)-gram, ...
  - With back off if N-gram information is missing, you use N-1 gram.
  - If that's also missing, you would use N-2 gram and so on until you find non zero probability.
  - Using the lower level N-grams ie N-1 gram, N-2 gram down to uni-gram.

“Stupid” Backoff

- In very large web scale corpora, a method called **stupid backoff** has been effective.
- With stupid backoff, if the higher order N-gram probability is missing the lower order N-gram probability is used just multiplied by a constant.
- A constant of about 0.4 was experimentally shown to work well.

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(chocolate|John\ drinks) = ?$$

$$0.4 \times P(chocolate|drinks)$$

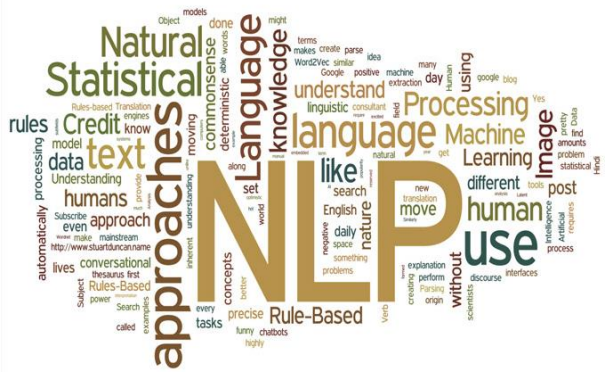
Hamidreza Baradaran Kashani

Interpolation

$$\hat{P}(chocolate|John\ drinks) = 0.7 \times P(chocolate|John\ drinks) + 0.2 \times P(chocolate|drinks) + 0.1 \times P(chocolate)$$

$$\hat{P}(w_n|w_{n-2}\ w_{n-1}) = \lambda_1 \times P(w_n|w_{n-2}\ w_{n-1}) + \lambda_2 \times P(w_n|w_{n-1}) + \lambda_3 \times P(w_n)$$

$$\sum_i \lambda_i = 1$$



با تشکر از توجه شما