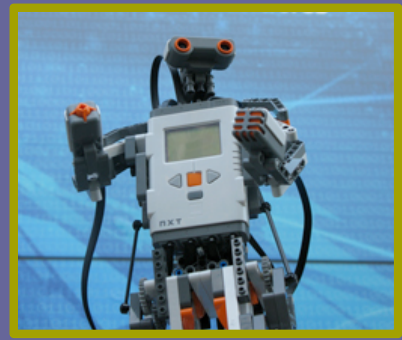
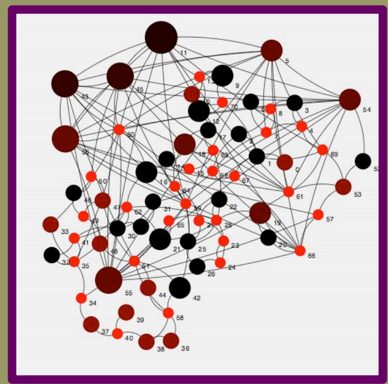
+

COMP40020 Human Language Technologies

Language Modelling

April 2019





Prof. Julie Berndsen

School of Computer Science

Julie.Berndsen@ucd.ie

HLT13

HLT13

Contents:

- Recall: competence vs. performance
- Humans, ambiguity & prediction
- Probability of a sentence
- N-Grams
- Language modelling and evaluation

Aim:

- To give an overview of probabilistic language modelling.

Acknowledgement

- For an excellent and detailed introduction to Language Modelling, I recommend the slides (and videos) of Dan Jurafsky at Stanford, building on his book (Jurafsky & Martin, 2018 – 3rd edition).

https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf

- Some of the information presented here is taken directly or adapted from Dan's slides.

Competence vs. Performance

Competence

Modelled by consistent and non-redundant systems of formal rules

Performance

Way in which humans actually use language influenced by non-linguistic factors

Competence vs. Performance

Competence

Modelled by consistent and non-redundant systems of formal rules

Human mind employs these rules to produce and understand new utterances

Humans can perceive one utterance meaning among many

Performance

Way in which humans actually use language influenced by non-linguistic factors

+ Competence vs. Performance

Competence

Modelled by consistent and non-redundant systems of formal rules

Competence grammar cannot account for all aspects of human language processing e.g. ambiguity, prediction

Performance

Way in which humans actually use language influenced by non-linguistic factors

Use of factors such as knowledge of the real world, likelihood of occurrence of words and sentences

Humans & Ambiguity

Evidence from online sentence processing experiments that the human parser is *probabilistic* with humans preferring some readings of sentences over others…

(Jurafsky & Martin, 2008)

+ Humans & Ambiguity

Evidence from online sentence processing experiments that the human parser is *probabilistic* with humans preferring some readings of sentences over others…

- **The women kept the dogs on the beach**
- The women kept the dogs which were on the beach 5%
- The women kept them (the dogs) on the beach 95%

(Jurafsky & Martin, 2008)

Humans & Ambiguity

Evidence from online sentence processing experiments that the human parser is *probabilistic* with humans preferring some readings of sentences over others…

- **The women kept the dogs on the beach**
- The women kept the dogs which were on the beach 5%
- The women kept them (the dogs) on the beach 95%

- **The women discussed the dogs on the beach**
- The women discussed the dogs which were on the beach 90%
- The women discussed them (the dogs) while on the beach 10%

(Jurafsky & Martin, 2008)

Humans & Prediction

And humans are able to predict what comes next…

- John's favourite meal is fish and __________

Humans & Prediction

And humans are able to predict what comes next…

- John’s favourite meal is fish and __________ chips

Humans & Prediction

And humans are able to predict what comes next…

- John's favourite meal is fish and __________ chips

boiled potatoes

peas

ice cream

…

he catches it himself

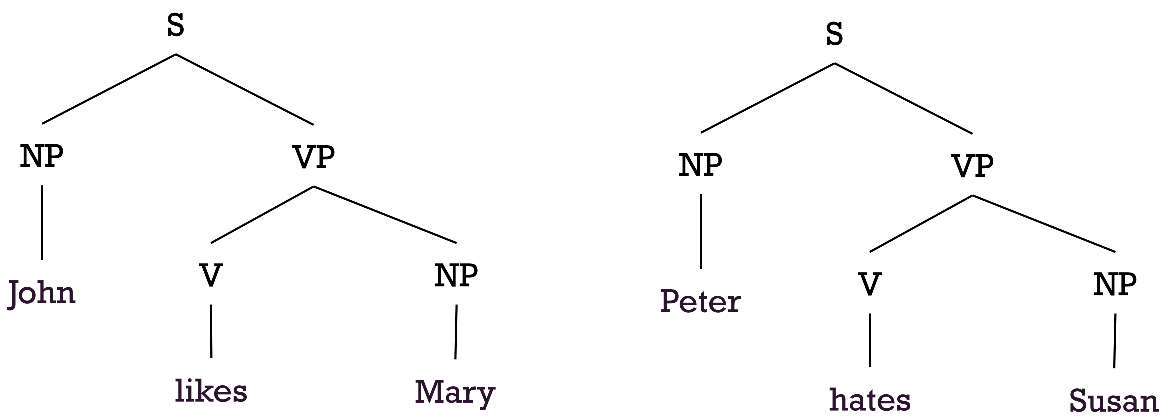
+ Adressing Ambiguity/Prediction…

- S → NP VP [.60]
- S → Aux VP PP [.35]
- S → VP [.05]
- NP → Det Noun [.20]
- NP → Proper-Noun [.35]
- NP → Noun [.05]
- NP → Pronoun [.40]
- VP → Verb [.25]
- VP → Verb NP [.40]
- VP → Verb NP PP [.35]

Augment competence grammar (grammar and lexicon) with probabilities and build a parser that takes these into account

→ stochastic grammar

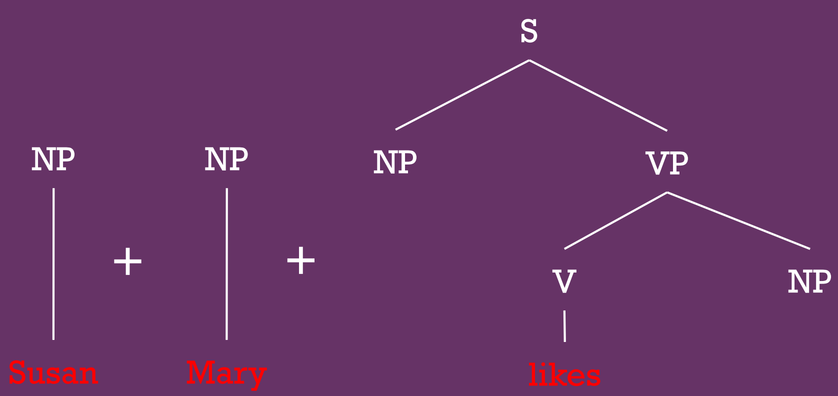
Adressing Ambiguity/Prediction…



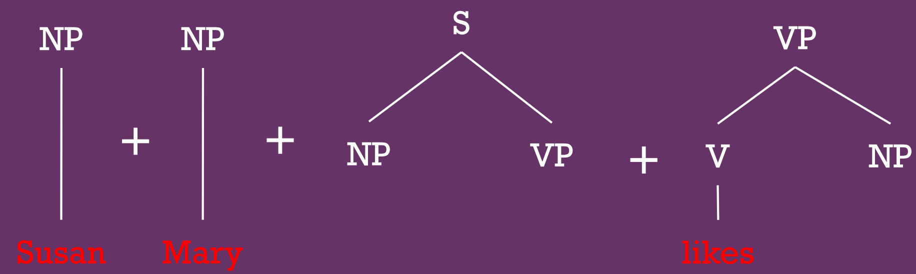
Data-Oriented Parsing (Bod, 1993)

Use experience of parse structures to find the probability of new sentence. Using e.g. a tree-bank, build the new sentence with the sub-trees of existing sentences.

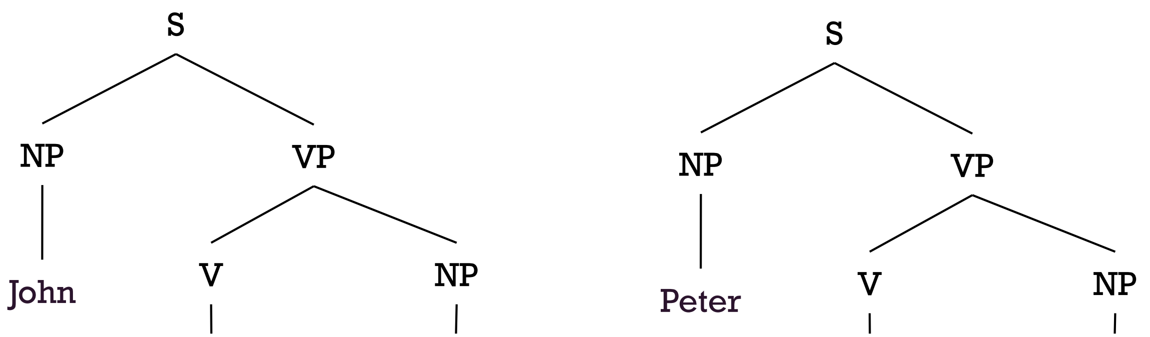
P(Susan likes Mary) = 1/4 * 1/4 * 1/20 = 0.0031



P(Susan likes Mary) = 1/4 * 1/4 * 2/20 *1/8 = 0.008



+ Adressing Ambiguity/Prediction…

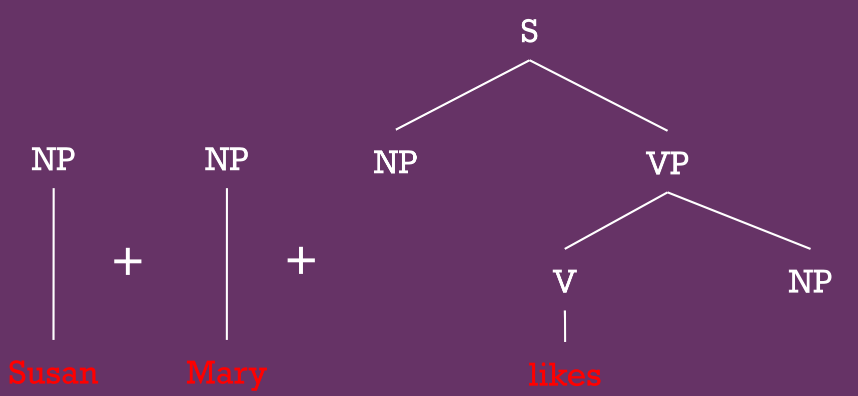


Data-Oriented Parsing
…3)

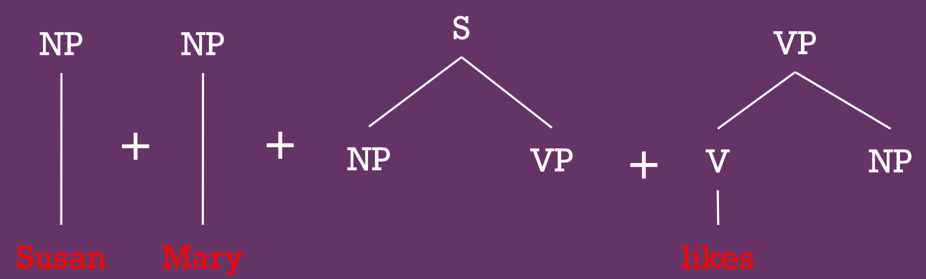
Use … g a

Both of these approaches still use knowledge about the structure of the language (i.e. knowledge-based with statistical information)

P(Su…



….008



+ Solely Data-Driven Probability…

P(the women kept the dogs on the beach)

+ Solely Data-Driven Probability…

P(the women kept the dogs on the beach)

Chain Rule

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

+ Solely Data-Driven Probability…

P(the women kept the dogs on the beach)

Chain Rule

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

P(the)P(women | the)P(kept | the women)P(the | the women kept)P(dogs | the women kept the)P(on | the women kept the dogs)P(the | the women kept the dogs on)P(beach | the women kept the dogs on the)

+ Solely Data-Driven Probability…

P(<s>the women kept the dogs on the beach</s>)

Chain Rule

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

P(<s>)P(the|<s>)P(women|<s>the)P(kept|<s>the women)P(the|<s>the women kept)P(dogs|<s>the women kept the) P(on|<s>the women kept the dogs)P(the|<s>the women kept the dogs on)P(beach|<s>the women kept the dogs on the)P(</s>|<s>the women kept the dogs on the beach)

Markov Assumption

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

Unigram

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

Bigram

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

Trigram

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx \prod_i P(w_i \mid w_{i-2} w_{i-1})$$

Solely Data-Driven Probability…

P(<s>the women kept the dogs on the beach</s>)

Unigram

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

Unigram

P(<s>)P(the)P(women)P(kept)P(the)P(dogs)P(on)
P(the)P(beach)P(</s>)

Solely Data-Driven Probability…

P(<s>the women kept the dogs on the beach</s>)

Bigram

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

Bigram

P(<s>)P(the|<s>)P(women|the)P(kept|women)
P(the|kept)P(dogs|the)P(on|dogs)P(the|on)
P(beach|the)P(</s>|beach)

Solely Data-Driven Probability…

P(<s>the women kept the dogs on the beach</s>)

Trigram

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx \prod_i P(w_i \mid w_{i-2} w_{i-1})$$

Trigram

P(<s>)P(the|<s>)P(women|<s>the)P(kept|the women)P(the|women kept)P(dogs|kept the)P(on|the dogs)P(the|dogs on)P(beach|on the)P(</s>|the beach)

+ Probabilistic Language Models

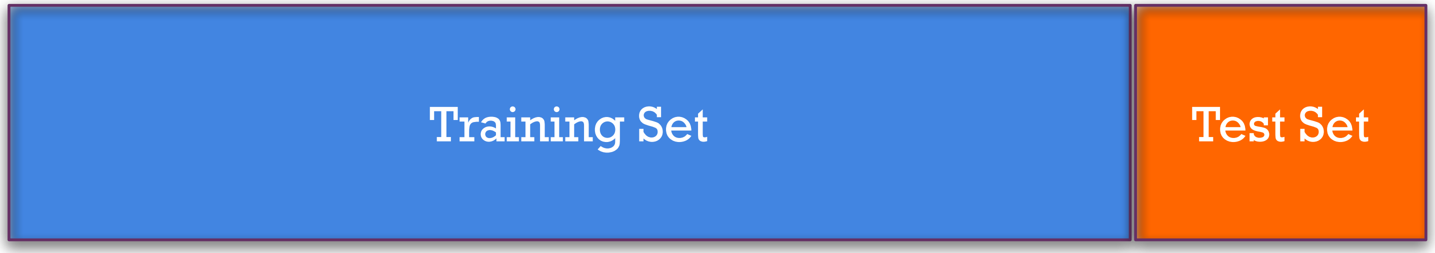
- Language Models (LM) can be used to:
  - assign a probability to a *sentence*
  - predict the next *word* in a *sentence*

- LMs can be regarded as *data-driven* grammars:
  - insufficient from linguistic perspective as they cannot capture long distance dependencies
  - but good for many language technology tasks

- Found in many language technology domains:
  - speech recognition
  - machine translation
  - spelling correction
  - information retrieval
  - …

+ Probabilistic Language Models

- The probabilities for *n-grams* of a language are estimated using a corpus or data set based on the counts (frequencies).



- E.g.
  - a *bigram* language model is built using the counts of the unigrams and the bigrams in a training set with the probabilities determined using the maximum likelihood estimate (MLE).

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts

it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts

	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts							
it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts								
	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts							
it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts								
	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts							
it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts								
	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

+ Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts

it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts

	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	$\boxed{0}$	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts							
it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts								
	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	$\boxed{0}$	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	0	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Austen's Sense and Sensibility with vocabulary size of 6833 and 141576 tokens

Unigram Counts

it	was	her	to	who	mother	declare	enough
1568	1846	2436	4063	260	258	16	103

Bigram Counts

	it	was	her	to	who	mother	declare	enough
it	0	201	1	55	0	0	0	0
was	8	0	13	52	0	0	0	14
her	2	3	1	0	0	114	0	0
to	35	0	221	0	$\boxed{0}$	0	4	0
who	1	31	0	0	0	0	0	0
mother	0	13	0	6	0	0	0	0
declare	0	0	0	0	0	0	0	0
enough	0	0	0	49	0	0	0	0

Unigram and Bigram Counts

NLTK Text: Wall Street Journal with vocabulary size of 12408 and 100676 tokens

Unigram Counts								
This	is	an	old	story	mother	to	paper	read
717	671	316	24	6	3	2164	28	11

Bigram Counts

	This	is	an	old	story	mother	paper	to	read
This	0	12	0	0	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0}$	0
is	0	0	15	0	0	0	0	0	0
an	0	0	0	2	0	0	$\boxed{0}$	0	0
old	0	0	0	0	1	0	0	0	0
story	0	1	0	0	0	0	0	0	0
mother	0	0	0	0	0	0	0	0	0
paper	0	0	0	0	0	0	0	0	0
to	0	0	19	0	0	0	0	0	3
read	0	0	0	0	0	0	0	0	0

Language Model: Small Example

Some sentences from Berkeley Restaurant Project
(cf. Jurafsky slides)

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

Maximum Likelihood Estimation

For Bigram:

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

$$P(can \mid <s>) = \frac{c(<s>\ can)}{c(<s>)} = \frac{2}{6} = \frac{1}{3}$$

$$P(looking \mid i'm) = \frac{c(i'm\ looking)}{c(i'm)} = \frac{2}{2} = 1$$

$$P(when \mid <s>) = \frac{c(<s>\ when)}{c(<s>)} = \frac{1}{6}$$

$$P(that \mid food) = \frac{c(food\ that)}{c(food)} = \frac{1}{2}$$

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

+ Unigram Counts

Example text with 6 sentences (43 words and 68 tokens)

Unigram Counts							
<s>	</s>	a	about	any	are	available	breakfast
6	6	2	2	1	1	1	1
by	caffe	can	cantonese	chez	close	day	during
1	1	2	1	1	1	1	1
eat	food	for	give	good	i'm	kinds	listing
1	2	2	1	2	2	2	1
looking	me	mid	of	open	panisse	place	priced
2	3	1	2	1	1	1	1
restaurants	tell	thai	that	the	to	venezia	what
1	2	1	1	2	1	1	1
when	you						
1	2						

+ Some Bigram Counts

Example text with 6 sentences (43 words and 68 tokens)

Some Bigram Counts							
	<s>	a	for	me	i'm	looking	good
<s>	0	0	0	0	1	0	0
a	0	0	0	0	0	0	1
for	0	1	0	0	0	0	0
me	0	0	0	0	0	0	0
i'm	0	0	0	0	0	2	0
looking	0	0	2	0	0	0	0
good	0	0	0	0	0	0	0

Some Bigram Probabilities

Example text with 6 sentences (43 words and 68 tokens)

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

Some Bigram Probabilities

	<s>	a	for	me	i'm	looking	good
<s>	0	0	0	0	0.1667	0	0
a	0	0	0	0	0	0	0.5
for	0	0.5	0	0	0	0	0
me	0	0	0	0	0	0	0
i'm	0	0	0	0	0	1	0
looking	0	0	1	0	0	0	0
good	0	0	0	0	0	0	0

Some Bigram Probabilities

Example text with 6 sentences (43 words and 68 tokens)

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

Some Bigram Probabilities

	<s>	a	fo	me	i'm	looking	good
<s>	0	0		0	0.1667	0	0
a	0			0	0	0	0.5
for							0
me							0
i'm							0
looking							0
good							0

$\log_{10}$ probabilities are typically used instead of probabilities to avoid underflow; it is also quicker to add logs rather than multiply probabilities

LM (ARPA) Format

Generated with http://www.speech.cs.cmu.edu/tools/lmtool-new.html which uses discount mass 0.5

- \data\
- ngram 1=43
- ngram 2=56
- ngram 3=53

- \1-grams:
- -1.3554 </s> -0.3010
- -1.3554 <s> -0.2747
- -1.8325 A -0.2913
- -1.8325 ABOUT -0.2946
- -2.1335 ANY -0.2946
- -2.1335 ARE -0.2978
- -2.1335 AVAILABLE -0.2814
- -2.1335 BREAKFAST -0.2814
- -2.1335 BY -0.2814
- -2.1335 CAFFE -0.2978
- -1.8325 CAN -0.2946
- …
- \2-grams:
- -0.7782 <s> CAN 0.0000
- -1.0792 <s> I'M 0.0000
- -1.0792 <s> MID 0.0000
- -1.0792 <s> TELL 0.0000
- -1.0792 <s> WHEN 0.0000
- -0.6021 A GOOD -0.1761
- -0.6021 A LISTING 0.0000
- -0.6021 ABOUT ANY 0.0000
- -0.6021 ABOUT CHEZ 0.0000
- -0.3010 ANY GOOD -0.1761
- -0.3010 ARE AVAILABLE 0.0000
- -0.3010 AVAILABLE </s> -0.3010
- -0.3010 BREAKFAST </s> -0.3010
- -0.3010 BY </s> -0.3010
- …
- \3-grams:
- -0.3010 <s> CAN YOU
- -0.3010 <s> I'M LOOKING
- -0.3010 <s> MID PRICED
- -0.3010 <s> TELL ME
- -0.3010 <s> WHEN IS
- -0.3010 A GOOD PLACE
- -0.3010 A LISTING OF
- -0.3010 ABOUT ANY GOOD
- -0.3010 ABOUT CHEZ PANISSE
- …
- \end\

LM (ARPA) Format

Generated with http://www.speech.cs.cmu.edu/tools/lmtool-new.html which uses discount mass 0.5

- \data\
- ngram 1=43
- ngram 2=56
- ngram 3=53

probability $\log_{10}$

- \1-grams:
- -1.3554 </s> -0.3010
- -1.3554 <s> -0.2747
- -1.8325 A -0.2913
- -1.8325 ABOUT -0.2946
- -2.1335 ANY -0.2946
- -2.1335 ARE -0.2978
- -2.1335 AVAILABLE -0.2814
- -2.1335 BREAKFAST -0.2814
- -2.1335 BY -0.2814
- -2.1335 CAFFE -0.2978
- -1.8325 CAN -0.2946
- …
- \2-grams:
- -0.7782 <s> CAN 0.0000
- -1.0792 <s> I'M 0.0000
- -1.0792 <s> MID 0.0000
- -1.0792 <s> TELL 0.0000
- -1.0792 <s> WHEN 0.0000
- -0.6021 A GOOD -0.1761
- -0.6021 A LISTING 0.0000
- -0.6021 ABOUT ANY 0.0000
- -0.6021 ABOUT CHEZ 0.0000
- -0.3010 ANY GOOD -0.1761
- -0.3010 ARE AVAILABLE 0.0000
- -0.3010 AVAILABLE </s> -0.3010
- -0.3010 BREAKFAST </s> -0.3010
- -0.3010 BY </s> -0.3010
- …
- \3-grams:
- -0.3010 <s> CAN YOU
- -0.3010 <s> I'M LOOKING
- -0.3010 <s> MID PRICED
- -0.3010 <s> TELL ME
- -0.3010 <s> WHEN IS
- -0.3010 A GOOD PLACE
- -0.3010 A LISTING OF
- -0.3010 ABOUT ANY GOOD
- -0.3010 ABOUT CHEZ PANISSE
- …
- \end\

LM (ARPA) Format

Generated with http://www.speech.cs.cmu.edu/tools/lmtool-new.html which uses discount mass 0.5

back-off weight $\log_{10}$

- \data\
- ngram 1=43
- ngram 2=56
- ngram 3=53

- \1-grams:
- -1.3554 </s> $\boxed{\text{-0.3010}}$
- -1.3554 <s> -0.2747
- -1.8325 A -0.2913
- -1.8325 ABOUT -0.2946
- -2.1335 ANY -0.2946
- -2.1335 ARE -0.2978
- -2.1335 AVAILABLE -0.2814
- -2.1335 BREAKFAST -0.2814
- -2.1335 BY -0.2814
- -2.1335 CAFFE -0.2978
- -1.8325 CAN -0.2946
- …
- \2-grams:
- -0.7782 <s> CAN 0.0000
- -1.0792 <s> I'M 0.0000
- -1.0792 <s> MID 0.0000
- -1.0792 <s> TELL 0.0000
- -1.0792 <s> WHEN 0.0000
- -0.6021 A GOOD -0.1761
- -0.6021 A LISTING 0.0000
- -0.6021 ABOUT ANY 0.0000
- -0.6021 ABOUT CHEZ 0.0000
- -0.3010 ANY GOOD -0.1761
- -0.3010 ARE AVAILABLE 0.0000
- -0.3010 AVAILABLE </s> -0.3010
- -0.3010 BREAKFAST </s> -0.3010
- -0.3010 BY </s> -0.3010
- …
- \3-grams:
- -0.3010 <s> CAN YOU
- -0.3010 <s> I'M LOOKING
- -0.3010 <s> MID PRICED
- -0.3010 <s> TELL ME
- -0.3010 <s> WHEN IS
- -0.3010 A GOOD PLACE
- -0.3010 A LISTING OF
- -0.3010 ABOUT ANY GOOD
- -0.3010 ABOUT CHEZ PANISSE
- …
- \end\

+ Language Model:Example

Unigram “Sentence” Generation e.g.

you me any restaurants available breakfast looking i’m during open day priced what good kinds a close mid

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

Language Model:Example

Bigram "Sentence" Generation e.g.

mid priced thai food is caffe veneizia
i'm looking for
can you give me about any

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

An Aside: Shannon Visualisation

Bigram Sentence Generation e.g.

<s>can
can you
you tell
tell me
me a
a listing
listing of
of food
food that
that are
are available
available </s>

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

+ Language Model:Example

Trigram "Sentence" Generation e.g.

tell me about…

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

+ Approximating Shakespeare

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.

Taken from: Jurafsky Language Modeling Slides
https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf

+ Approximating Shakespeare

- Shakespeare's "Corpus"
  - Number of tokens=884,647, Vocabulary=29,066
  - Only used 300,000 bigram types out of a possible 844M $\rightarrow$ 99.96% of possible bigrams never seen (i.e. have 0 in bigram count table entry)
  - Even worse for trigram and quadrigram – this is why the quadrigram text on the previous slide is so good – it is based on seen quadrigrams – i.e. it *is* Shakespeare.
  - N-grams only work if training corpus is like the test corpus – for real world applications this is generally not robust enough.

Adapted from: Jurafsky Language Modeling Slides
https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf

+ Back to Berkeley Restaurant Project

Full Berkeley Restaurant Project data with 9222 sentences
(cf. Jurafsky slides)

Unigram Counts

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

Bigram Counts

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

Back to Berkeley Restaurant Project

Full Berkeley Restaurant Project data
(cf. Jurafsky slides)

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

Bigram Probabilities

	i	want	to	eat	chinese	lunch	food	spend
i	0.002	0.3265	0	0.0036	0	0	0	0.0008
want	0.0022	0	0.6559	0.0011	0.0065	0.0054	0.0065	0.0011
to	0.0008	0	0.0017	0.2838	0.0008	0.0025	0	0.0873
eat	0	0	0.0027	0	0.0214	0.0563	0.0027	0
chinese	0.0063	0.0063	0	0	0	0.0063	0.519	0
lunch	0.0059	0	0	0	0	0	0.0029	0
food	0.0137	0	0.0137	0	0.0009	0	0.0037	0
spend	0.0036	0	0.0036	0	0	0	0	0

What should be done with <u>unseen</u> bigrams (i.e. entries with 0)?

+ Smoothing

- N-grams that are unseen in the training set may still occur in the test set $\rightarrow$ sparse training data
  - E.g. based on training set, P(food|to) = 0
  - However, the test set could contain: "to food fans everywhere"
- Laplace Smoothing:
  - Add one to each count and calculate smoothed probabilities P*
  - Only practical where small number of zeros

$$P^*(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)+1}{C(w_{i-1})+V}$$

- Advanced Smoothing Techniques:
  - Absolute Discounting
  - Good-Turing
  - Kneser-Ney
  - Witten-Bell

+ Backoff & Interpolation

- Backoff (use less context):
  - Use n-gram if the count of n-gram is greater than some threshold
  - Otherwise use (n-1)-gram
  - E.g. trigram → bigram → unigram

- Interpolation:
  - Mix trigrams, bigrams and unigrams using a weighting factor (often written as λ) selected so as to maximise the probability of held-out (or development) set.

+ Language Model Evaluation

- Extrinsic evaluation in the context of a specific application:
  - does model A perform better at the task using the test set than model B?
  - E.g. accuracy of speech recognition, machine translation, information retrieval
  - <u>but</u> it can take a long time to run these experiments

- Intrinsic evaluation based on *perplexity* (branching factor):
  - only works well if the training set and the text set are similar
  - <u>but</u> provides a general measure about the model itself

minimising perplexity ⟺ maximising probability

Some Resources & Toolkits

- See Blackboard for links
  - Google: All Our N-Gram are Belong to You
  - Google Books N-Gram Viewer
  - The SRI Language Modeling Toolkit (SRILM)
  - The CMU Statistical Language Modeling (SLM) Toolkit