HOW STATISTICAL LANGUAGE MODELS CHANGED THE GAME

David Thomson, VP Speech Research
MAD GAB: HOW MANY CAN YOU READ?

1. eic tchu notp parbt uv the s solution dhe eup ahrt of s a per soup a tet

2. haum itz have nope year prescher

3. hi dr whey to fas tou were ia bout clest roll

4. tea early burdg ets the wai erm but the s aee can mouse kaets a cheese

5. eagle 's me 's or but weasels don't gets suck tin to jad anjun xiz

6. peep a oo eathe urr luff me orh eatm eae orr the synch i moak aa
MAD GAB: HOW MANY CAN YOU READ?

1. eic tchu notp parbt uv the s solution dhe eup ahrt of s a per soup a tet
   If you're not part of the solution, you're part of the precipitate

2. haum itz have nope year prescher
   Hermits have no peer pressure

3. hi dr whey to fas tou were ia bout clest roll
   I drive way too fast to worry about cholesterol

4. tea early burdg ets the wai erm but the s aee can mouse kaets a cheese
   The early bird gets the worm, but the second mouse gets the cheese.

5. eagle 's me 's or but weasels don't gets suck tin to jad anjun xiz
   Eagles may soar, but weasels don't get sucked into jet engines.

6. peep a oo eathe urr luff me orh eatm eae orr the synch i moak aa
   People either love me or hate me or they think I'm OK.
EVERY INTERACTION SHOULD BE PRODUCTIVE

For over a decade, we’ve been focused on building technology that allows people to communicate naturally—technology that responds to the person, not the other way around. Ultimately, we help turn frustrating experiences into productive interactions.
interaction n. (ĭnˈtər-ăkˈshən)
1. The act or process of interacting.
2. The state of undergoing interaction.
“The state governors met with their respective legislatures convening in the capital city.

- The first “s” in “legislatures” is completely obscured.
- Subjects were told there would be a cough and would be asked to indicate if all speech sounds were present.
- 19 of 20 subjects reported no phonemes missing.
- 11 subjects placed the cough early, 9 late.
REPLACING “WHEEL” WITH “HEEL”

It was found that the wheel was on the axle.

It was found that the heel was on the shoe.

It was found that the wheel was on the shoe.

It was found that the (cough)heel was on the shoe.
PHONEMIC RESTORATION EFFECT (1970)

- Participants heard a sentence with a phoneme removed
- People claim to hear the missing phoneme
- The perceived phoneme was influenced by context

“It was found that the #eel …” in the axle.”

“It was found that the #eel was on the shoe.”

“It was found that the #eel was on the orange.”

“It was found that the #eel was on the table.”
• Hey Janice, i thought i'd give you a quick call. Uh
• Claire, honey, it's John. give me a call when you get ✔
• Hey Greg, this is Allison. i didn't know your name, but
• Hey buddy, i just uh called to wish you a happy ✔
• Hey , it's Jorge at ABC, Inc. i tried to fax
• Hey, it's Audrey. um i'm done unpacking, cleaning, and getting the
• Did you just do that, just so you could get the ✔
• Hey Ben, it's Mark. um real quick questions. some of these
• Hey Antonio, it's about one forty six, and this is my
• They said you're on the phone. call me back and let ✔
that
the
whel
pee
l

Select first half of “wheel”

Mute 125 ms of speech

Replace silence w/cough
HUMAN-MACHINE CONVERSATION

- Speech recognition
- Text
- Natural language processing
- Intent
- Application
- Response
- Generation

Audio flow:
1. Speech recognition
2. Natural language processing
3. Application
4. Generation

© 2016 Interactions LLC Proprietary
SPEECH RECOGNITION OVERVIEW

1. Feature extraction
2. Feature transform
3. Compute probabilities
4. Decode (search)
5. Estimate score
6. Text

Confidence
SPEECH RECOGNITION OVERVIEW

Models

- HLDA matrix
- Acoustic model
- Language model
- Confidence model

Feature extraction

Feature transform

Compute probabilities

Decode (search)

Estimate score

Confidence

Text

40 years of development

© 2016 Interactions LLC Proprietary
Data is a strategic asset
SPEECH RECOGNITION OVERVIEW

1. Feature extraction
2. Feature transform
3. Compute probabilities
4. Decode (search)
5. Estimate score
6. Confidence

Text
I LOVE TOMATOES

Speech waveform

Spectrogram
FEATURE EXTRACTION

- Compress information
- Eliminate irrelevant information
- Preserve important information

~320 audio samples

~60 features

$\alpha_1$
$\alpha_2$
$\alpha_2$
$\alpha_n$
• Higher discriminative power
• Compact representation
• Orthogonal (uncorrelated elements)
SPEECH RECOGNITION OVERVIEW

1. Feature extraction
2. Feature transform
3. Compute probabilities
4. Decode (search)
5. Estimate score
6. Confidence
7. Text
Compute phoneme probabilities

\[ \theta = \left( \begin{array}{c} b_1 \\ b_2 \\ \vdots \\ b_n \end{array} \right) \]

Transformed features

Compute phoneme probabilities

\[ P(\, \ddot{a} \mid \theta \, ) \]
\[ P(\, \dot{a} \mid \theta \, ) \]
\[ P(\, \ddot{a} \mid \theta \, ) \]
\[ P(\, b \mid \theta \, ) \]
\[ P(\, \ddot{e} \mid \theta \, ) \]
\[ P(\, \dot{e} \mid \theta \, ) \]
\[ P(\, f \mid \theta \, ) \]
\[ \vdots \]

Acoustic model
NEURAL NETWORK – ONE HIDDEN LAYER

Inputs (band energy)

Weights

Hidden Layer

Weights

Outputs

\( P(\ddot{u}) \)

\( P(\ddot{O}) \)

\( P(\ddot{a}) \)

\( P(\ddot{e}) \)

\( P(s) \)

\( P(m) \)
SPEECH RECOGNITION OVERVIEW

Acoustic model \( \rightarrow P(\hat{a} | \theta) \)

Feature extraction

Feature transform

Compute probabilities

Decode (search)

Estimate score

Confidence

Text
SPEECH RECOGNITION OVERVIEW

Feature extraction
Feature transform
Compute probabilities
Decode (search)
Estimate score
Confidence
Text

\[ P(\, \bar{a} \mid \theta) \]
I love tomatoes
Acoustic model

Probability(phoneme=ũ | features_{frame f=2})

Features extracted from each frame

ũv təmət ź

© 2016 Interactions LLC Proprietary
VAST NUMBER OF POSSIBLE PATHS TO SEARCH

Search for highest probability path
$P(\text{phoneme}=\ddot{u} \mid \text{features}) = \text{local score} = \bigcirc$

Pick lowest cumulative score, then add local score
BACKWARDS PATH

P(phoneme=ũ | features) = local score =

Rule for each node:
If this node is in the best path, which is the previous node? Remember the answer.
SCORE “I LOVE TOMATOES” VS. ACOUSTIC EVIDENCE

Acoustic model

ǐ ū v t ĕ m ā t ź
BACKTRACK TO FIND BEST PATH

Acoustic model

Ilvtematōz
Uh, please tell me what is the balance in my checking account.
Please give me what is the balance in my checking account. For my savings account, check my balance.
Uh, please tell me what is the balance for my savings account. My savings account balance is 0.8. Checking account for my savings account balance is 0.9. Checking 0.33 for 0.5. Uh, give me what is my check account balance. My checking account balance is 0.7. Checking 0.33 for 0.5.
INCLUDING PRIOR PROBABILITIES

Prior probabilities (example):

\[ P(\text{savings}) = 0.8 \]
\[ P(\text{checking}) = 0.2 \]

Probability based on audio:

\[ P(\text{savings} \mid \text{audio}) \]
\[ \equiv P(\text{savings} \mid \text{features}) \]
\[ \equiv \prod_{f \in \text{frames}} P(\text{savings}_f \mid \text{features}_f) \]
\[ \log[P(\text{savings} \mid \text{features})] \]
\[ \equiv \sum_f \log[P(\text{savings}_f \mid \text{features}_f)] + \psi \log[P(\text{savings})] \]
WHAT WE CAN DO WITH GRAMMARS

• Menus
• IVR
• Name dialing
• Home automation
• Car dashboards
• Video games
SEVERAL PROBLEMS REMAIN

• How to compute probabilities
• Difficult to anticipate all possible combinations
• Bad grammar
• Repeated phrases
• Scaling for unequal path lengths
• Probabilities are not independent
WHAT REQUIRES AN SLM?
(STATISTICAL LANGUAGE MODEL)

• High accuracy w/freeform input
• Call transcription and analytics
• Medical/legal transcription
• Voicemail-to-text
• Video captioning
• Prompts robust to out-of-vocabulary speech
• “How may I help you?”
• Personal assistants
• Comfortable language translation
• Q&A, navigation, relay for the deaf
• Human-like conversations
Please tell me what my checking savings account balance is for.
PROBABILITY OF BREAD *

P(bread) = 0.0026%

* From Google Books Ngram Viewer
Unigram:
Prior probability = \( P(\text{savings}) \)
PROBABILITY OF “BREAD” GIVEN “LOAF OF”

Conditional probability:

\[ P(A, B) = P(A) \ P(B \mid A) \]

Rearrange to get:

\[ P(B \mid A) = \frac{P(A, B)}{P(A)} \]

Let B = "bread," A="loaf of":

\[ P(bread \mid loaf of) = \frac{P(loaf of, bread)}{P(loaf of)} \]

\[ = \frac{0.0000590833\%}{0.0000703839\%} \]

\[ = 0.839 \]

Remember: \( P(bread) = 0.000026 \)
4-gram:
Prior probability $= P(savings|tell+me+my)$
THE EFFECT OF LM SCALE ($\Psi$)

$$\log[P(savings)] = \sum_{f} \log[P(savings \mid audio_f)] + \psi \log[P(savings)]$$

[Acoustic model] [Language model]

$\Psi = 0$  high dr -way to fas tou were ia bout kless –trol

$\Psi = 1$  hi drive way to fas to worry about kless troll

$\Psi = 5$  hi drive way too fast to worry about quest roll

$\Psi = 20$  i drive way too fast to worry about cholesterol

$\Psi = 70$  i drive way too fast too worried about cholesterol

$\Psi = 100$  driveway to fast worry about question-mark
oh i see said pooh
do you know what i mean pooh
and pooh said oh i see
pooh sat down on the name
of the end of the room
and he had gone out of the hole
and at the end of the wood
REMAINING PROBLEMS

• Data sparseness

  From 520M-word Corpus of Contemporary American English:

<table>
<thead>
<tr>
<th>N-gram</th>
<th>Actual number</th>
<th>Possible number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram</td>
<td>1.0 million</td>
<td>$10^6$</td>
</tr>
<tr>
<td>2-gram</td>
<td>6.2 million</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>3-gram</td>
<td>11.9 million</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>4-gram</td>
<td>8.3 million</td>
<td>$10^{24}$</td>
</tr>
</tbody>
</table>

• Impossible to get good statistics on all n-grams.

• Difficult to generalize, cluster, or smooth estimates
  – “I think David is a terrific guy.”
  – “I think Jerry is a slovenly fellow.”

• Distance limits
  – “The car/flavor, said the bartender, was clearly a lemon.”

• Cost of SLM development
UNLIKE ANY OTHER SPEECH SOLUTION

Interactions virtual assistant applications are based on our proprietary Adaptive Understanding platform.

CURO SPEECH AND LANGUAGE PLATFORM

CONTINUOUS TRAINING AND TUNING

HUMAN ASSISTED UNDERSTANDING

VOICE CHAT OUTBOUND TEXTING MOBILE

APPLICATION

iProxy

ADAPTIVE UNDERSTANDING™ TECHNOLOGY

© 2016 Interactions LLC Proprietary
DNN-BASED LANGUAGE MODEL

\[ \text{word}_{n-1} = \text{“and”} \]
\[ \text{word}_{n-2} = \text{“to”} \]
\[ \text{word}_{n-3} = \text{“the”} \]

\[ \begin{align*}
\text{“the”} &= 0 \\
\text{“and”} &= 1 \\
\text{“of”} &= 0 \\
\text{“to”} &= 0
\end{align*} \]

\[ \begin{align*}
\text{“the”} &= 1 \\
\text{“and”} &= 0 \\
\text{“of”} &= 0 \\
\text{“to”} &= 0
\end{align*} \]

\[ \begin{align*}
\text{“the”} &= 0 \\
\text{“and”} &= 0 \\
\text{“of”} &= 0 \\
\text{“to”} &= 1
\end{align*} \]

\[ \cdots \]

\[ \begin{align*}
\text{P(} \text{word}_n = \text{“finish”}) &= \text{P(} \text{word}_n = \text{“rest”}) \\
\text{P(} \text{word}_n = \text{“the”}) &= \text{\ldots}
\end{align*} \]
SPEECH RECOGNITION OVERVIEW

1. Feature extraction
2. Feature transform
3. Compute probabilities
4. Decode (search)
5. Estimate score
6. Confidence

\[ P(\bar{a} | \theta) \]
INTERPOLATION EXAMPLE

TV remote
Telephone number
Name dialing
How may I help you?
Street address
SMS
Voicemail-to-text

Q&A

\[ \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \end{bmatrix} \]

Small dev. set

Generic Model

\( \lambda_0 \)
\( \lambda_1 \)
\( \lambda_2 \)
\( \lambda_3 \)
\( \lambda_4 \)
\( \lambda_5 \)
\( \lambda_6 \)
\( \lambda_7 \)
NEED FOR SUBGRAMMARS
**Hierarchical Language Models**

Top Grammar:

- Please lookup
- Information
- Address
- Phone
- Email
- Skype

Sub Grammar:

- (Device)
DATA INGESTION DEMO

Dynamic hierarchical statistical language models
- or -
Class-based hierarchal statistical language models
THE FUTURE

- DNNs
- More data
- Unsupervised learning
- Integrate decoder, DNN, natural language processing
- Domain transplanting
- Focus shift to natural language, dialogs, content
MORE INFORMATION:
(or to send your resume)

David Thomson <davidt@interactions.com>
www.interactions.com

David Thomson, VP Speech Research