“I’m sorry, Dave, I’m afraid I can’t do that”:

Linguistics, Statistics, and Natural-Language Processing in the Big Data Era

Lillian Lee
Professor, Computer Science
http://www.cs.cornell.edu/home/llee
the dream
Why is this man smiling?

ALAN TURING AT 100

Alan Turing, born a century ago this year, is best known for his wartime code-breaking and for inventing the ‘Turing machine’ – the concept at the heart of every computer today. But his legacy extends much further: he founded the field of artificial intelligence, proposed a theory of biological pattern formation and speculated about the limits of computation in physics. In this collection of features and opinion pieces, Nature celebrates the mind that, in a handful of papers over a tragically short lifetime, shaped many of the hottest fields in science today.

Image credit: Andy Potts; Turing family
The Turing test:
Intelligence ➔ human-level language use

In 1950 Alan Turing proposed that a machine could be termed "intelligent" if it could respond to queries in a manner that was completely indistinguishable from a human being.

And how are you feeling today? I think you should know I'm feeling very depressed.

Well, that's life I'm afraid.

Turing predicted we’d be close in about 50 years.
Why is this man not smiling?

Open the pod bay doors, Hal.

I’m sorry, Dave, I’m afraid I can’t do that.
from *sci-fi* to *science and engineering*
Natural-language processing (NLP)

**Goal:** create systems that use human language as input/output

- speech-based interfaces
- information retrieval / question answering
- automatic summarization of news, emails, postings, etc.
- automatic translation

... and much more!

**Interdisciplinary:** computer science; linguistics, psychology, communication; probability & statistics, information theory...
Recently deployed: Siri

“I’m in the mood for Italian food in North Beach.”

I found a number of Italian restaurants in North Beach:

- Columbus Avenue
  - The Stinking Rose
  - Kearny Street
    - Tommaso Ristorante Italiano
  - Green Street
    - Sotto Mare

State of the art: Watson

The Watson system beat human Jeopardy! champions (and didn’t have internet access; it learned by “reading” before the match)
Why are these men smiling?

The session co-organizers
But we’re not all the way there yet
Real-life error (I)

A bunch of grapes.

Hey bunch of grapes
Real-life error (2)

We can email you when we're back.

We can email you when you're fat.
Real-life error (3)

[This U.S. city’s] largest airport …

What is Toronto???
why is understanding language so hard?
Challenge: ambiguity

List all flights on Tuesday

List all flights on Tuesday = List all the flights leaving on Tuesday.

List all flights on Tuesday = Wait ‘til Tuesday, then list all flights.
More realistic example

Retrieve all the local patient files
I saw her duck with a telescope.
Baroque example

I saw her duck with a telescope.
Conversation complications

: Do you know when the train to Boston leaves?

: Yes.

: I want to know when the train to Boston leaves.

: I understand.

[Grishman 1986]
I’m sorry, Dave, I’m afraid I can’t do that.

I’m afraid you might be right.
Meeting these challenges: a brief history
1940s – 50s: From language to probability

“The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point ...

[For] the engineering problem, the significant aspect is that the actual message is one selected from a set of possible messages.”

--C. Shannon, 1948
Language, statistics, cryptography

WWII: Turing helps break the German “Enigma” code
Why is this man smiling?

I can see Alaska from my house!

Encryption process

[W. Weaver memo on translation, 1949]
Two probabilities to infer

Prob. of generating this original message?

Prob. of doing this encryption of the original?

I can see Alaska from my house!

Encryption process

[Russian]
Another use of message probs: speech recognition

(1) It’s hard to recognize speech

(2) It’s hard to wreck a nice beach

Both messages have almost the same acoustics, but different likelihoods.
1950s-1980s: Breaking with statistics

N. Chomsky (1957):

(a) Colorless green ideas sleep furiously

(b) Furiously sleep ideas green colorless

The argument: Neither sentence has ever occurred in the history of English. So any statistical model would given them the same probability (zero).

The field moved to sophisticated non-probabilistic models of language.
1990s: The empiricists strike back

- Huge amounts of data start coming online
- Advances in algorithms, models, and horsepower

2000s and beyond: integrating language insights and statistical techniques
Why is this man smiling?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision.... I do not know what the right answer is, but I think [different] approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.
computers can also (help us) understand us
Why is this man smiling?

Beyond situational effects, *phrasing* also affects memorability:

- memorable movie quotes (in aggregate) are *unusual word choices* built on a *scaffolding of common part-of-speech patterns*
  - shown via language models

- carries over to ad slogans

C. Danescu-Niculescu-Mizil et al. ACL 2012
Social interaction: who has the lead?

Communicative behaviors are “patterned and coordinated, like a dance” [Niederhoffer and Pennebaker, '02]

Those with less power tend to immediately match the function-word choices of those with more power. [C. Danescu-Niculescu-Mizil et al. WWW 2012]
Why is this man smiling?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision.... I do not know what the right answer is, but I think [different] approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.