Natural Language Processing

- A Survey Talk @ MSR UW Summer Institute 2016 -

Margaret Mitchell
Cognition Group
Microsoft Research

Yejin Choi
Computer Science & Engineering

UNIVERSITY OF WASHINGTON
Accessibility in the Past...

Reading Machine to OCR

Kurzweil Reading Machine (1976)  
$30,000-$50,000

OCR & Flatbed Scanners (Today)

Text to speech (Today)

Hi, I’m Cortana.
Accessibility in the Past…

Talking books to Audio Books

Talking Books (1930s)  Audio Books (Today)

http://www.silksoundbooks.com/history-of-audiobooks.html
Accessibility in the Past...

Telephone

Abraham Graham Bell’s experiments with hearing devices (phonautograph) for deaf individuals (1876)

Phones (Today)
What is **Novel** today will become **Standard** tomorrow
NLP Today …

- Web search
  - can handle natural language queries better
  - often presents us structured knowledge
  - Research on web search started in 90’s (or maybe 80’s)
NLP Today …

• Automatic translation
  -- not perfect, but good enough for people to use
  -- real time translation with audio
  -- first statistical model (IBM model 1) came out in 1993
  -- first MT service based on statistical model in 2007
NLP Today …

• Everyday devices with natural language interface
  -- simple question answering
  -- simple natural language commands
NLP Today

- **Sentiment Analysis**
  - Domains: twitter, product reviews, news trend...
  - Works well for some types of data
  - Used in 2012 election
  - Challenge: sarcasm, subtext, nuanced messages
  - First research on sentiment analysis started in 2002
A bit of Introspection…

• A lot of exciting results
  – After a (very) long period of hard work!

• Computers do surprisingly well on some hard tasks
  – Machine translation
  – IBM Watson winning Jeopardy

• Computers do poorly on many tasks that are trivial for us.

• Computers seem to do well when
  – Lots of data for training (where low hanging fruits are)
  – The task can be solved by primarily based on pattern matching
  – (no complex reasoning, especially not commonsense reasoning)

• What are challenges that seem just impossible today but might become possible after 10 year of hard work?
What’s in the Future?

It's hard to make predictions, especially about the future.

The best way to predict your future is to invent it.
Intelligent Communication
Intelligent Communication

Reading between the lines

- **Content**: what is said
  + what is *not* said

![Chat screen with messages](image)
Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/
Intelligent Communication

Reading between the lines
• **Content:** *what* is said
•   + *what* is **not** said
• **Intent:** *why* it is said

Language in the world
• in the physical / visual context
• in the social / cognitive context
NLP for Acting AI:
Natural Language Instructions

Smart devices and personal robots executing commands in natural language instructions
Blueberry Muffins

**Ingredients**
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

**Procedure**
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. **Bake for 20 minutes**. Serve hot.

From Kitchen to Biology Labs

DNA Precipitation

Materials
3M NaOAc pH 5.2
EtOH 95%
Glycogen (optional)

Procedure
1. Add 0.1 volumes of 3M Sodium Acetate solution to 1 volume of DNA sample.
2. Add 1ul Glycogen to the DNA sample.
3. Add 2 volumes of 95% EtOH to the DNA Sample.
4. Store the solution overnight at -20°C or for 30 minutes at -80°C.
5. Centrifuge the solution at maximum speed for least 15 minutes.
6. Decant and discard the supernatant.
7. (Optional) Add 1 ml of 70% EtOH to the pellet and let sit for 5 minutes.
8. (Optional) Centrifuge the sample at maximum speed for 5 minutes.
9. (Optional) Decant and Discard the supernatant.
10. Air-dry the pellet for 10-15 minutes at room temperature until all liquid is gone.
11. Resuspend in desired volume of water or buffer

http://openwetware.org/wiki/DNA_Precipitation
Blueberry Muffins I (http://allrecipes.com/Recipe/Blueberry-Muffins-I/)

**Ingredients**
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

**Directions**
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
What’s Next: Composing a New Recipe

Compose new recipes given a recipe title (or what’s in the fridge)!
Grounding instructions with multimodal perception
NLP for Seeing AI
Seeing AI: How Deep Learning can Empower the Blind Community

Margaret Mitchell, Anirudh Koul, Saqib Shaikh
Microsoft
Points of Contact: seeingai@microsoft.com

Margaret Mitchell
memitc

Microsoft Research

Anirudh Koul
akoul

Bing

Saqib Shaikh
saqibs

Bing
Plan

- Accessibility Accelerates Innovation
- Backend Research Technology
  - Cognitive Services
    - Object Detection and Image Captioning
Accessibility Accelerates Innovation
Our Mission

Empower every person and every organization on the planet to achieve more
Approach to Advancing AI

1. Make AI vastly capable
2. Make vastly capable AI beneficial
Approach to Advancing AI

1. Make AI beneficial
2. Make beneficial AI vastly capable
2015 OneWeek Hackathon
World’ largest hackathon
13,000+ participants
Hack Team

Silicon Valley

Redmond

London
Just saw someone at SVMT Cafe working on his hack, something with his phone strapped to his head. Can't wait to see what it's all about next Thursday at the Science Fair!
“We as developers have this tremendous opportunity and tremendous responsibility, because not only do we get to dream of the future, we get to build the future”
Video
By far the coolest thing Microsoft announced at its Build conference: Using artificial intelligence to help the blind see.
Our team cycle

Blindness scenarios

Research

Engineering

Mainstream technology
Seeing AI
A research project at Microsoft

The Connectors:
- Cornelia Carapcea,
  • Cognitive Services Vision APIs
- Margaret Mitchell
  • Microsoft Research
Seeing AI Experiences

IOT Devices like Pivothead Live Streaming glasses

Phone app
Seeing AI Experiences for Screen Readers
Backend Technology
Object Detection and Image Captioning
Object Detection and Image Captioning

- **Vision:** Deep Convolutional Neural Network (CNN)
  - Object detection accuracy has shot up in past few years
    - Due to neural networks
  - Groups all over the world beginning to connect:
    - Newly accurate vision detections
    - ... to ...
  - Language models that can use the visual representations

- **Language:** Several models with strong performance
  1. Multimodal recurrent neural network language model
  2. Multiple instance learning w/ maximum entropy language model
Vision Model

- Vision model: Multiple-layer network (16; recently 152)
  - Pre-trained on 1.2 million image, 1000-class ImageNet data set
  - Fine-tuned on the MS COCO training data
  - State-of-the-art accuracy

Image Credit: Krizhevsky et al. 2012
2012 ImageNet 1K Challenge


*the original image is from the COCO dataset*
Microsoft Captioning Model
Image Captioning

- **Vision:** Deep Convolutional Neural Network (CNN)
  - Object detection accuracy has shot up in past few years
    - Due to neural networks
  - Groups all over the world beginning to connect:
    Newly accurate vision detections
    ... to ...
    Language models that can use the visual representations

- **Language:** Several models with strong performance
  1. Multiple instance learning with maximum entropy language model
  2. Multimodal recurrent neural network language model
  3. Nearest neighbor
Vision Model

- Vision model: Multiple-layer network (16; recently 152)
  - Pre-trained on 1.2 million image, 1000-class ImageNet data set
  - Fine-tuned on the MS COCO training data
  - State-of-the-art accuracy

Image Credit: Krizhevsky et al. 2012
Captioning Model 1
1: Multimodal Recurrent Neural Network

- Use fc7 as initial state in recurrent neural network language model
- Words output in sequence

Image Credit: Karpathy and Fei-Fei 2015

- Used by: Google, Baidu, Stanford, Berkeley

Image Credit: Cho et al. 2015
Captioning Model 2:
The Microsoft System
1: Multiple Instance Learning (MIL)

- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- Noisy-Or classifier
  - Image divided into 12x12, 6x6, 3x3 label for whole image
  - fc7 overlapping regions
  - Votes from several regions used to predict vector used for image features

\[ p_i^w = 1 - \prod_{r \in i} (1 - \sigma(f_{ij} \cdot v_w)) \]

- \( i \): image id
- \( f_{ij} \): fc7 vector
- \( \sigma(x) = \text{sigmoid} \)
- \( r \): regions
- \( v_w \): learned classifier weights
2: Map features to likely image words

- Compute global precision threshold on held-out training subset
- Output all words $\tilde{V}$ with precision of $\tau$ or higher

<table>
<thead>
<tr>
<th>Metric</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Agreement</td>
<td>PHR</td>
<td>63.8</td>
<td>35.1</td>
</tr>
<tr>
<td>Classification</td>
<td>PHR</td>
<td>45.3</td>
<td>31.0</td>
</tr>
<tr>
<td>MIL NOR</td>
<td>PHR</td>
<td>51.6</td>
<td>33.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>AP</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Classification</td>
<td>AP</td>
<td>37.0</td>
<td>19.4</td>
</tr>
<tr>
<td>MIL NOR</td>
<td>AP</td>
<td>41.4</td>
<td>20.7</td>
</tr>
</tbody>
</table>
3: Maximum Entropy Language Model and Reranking

- Use MIL classifier to generate bag of detected words
- Generate many captions using 4-gram maximum entropy LM
- Then re-rank using Deep Multimodal Similarity Model within MERT
→ Turns out this works really well.

• 1st place at CVPR image captioning challenge
• Evaluated by humans

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Percentage of captions that are evaluated as better or equal to human caption.</td>
</tr>
<tr>
<td>M2</td>
<td>Percentage of captions that pass the Turing Test.</td>
</tr>
<tr>
<td>M3</td>
<td>Average correctness of the captions on a scale 1-5 (incorrect - correct).</td>
</tr>
<tr>
<td>M4</td>
<td>Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).</td>
</tr>
<tr>
<td>M5</td>
<td>Percentage of captions that are similar to human description.</td>
</tr>
</tbody>
</table>
Turns out this works really well.

- 1st place at CVPR image captioning challenge
- Evaluated by humans

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Percentage of captions that are evaluated as better or equal to human caption.</td>
</tr>
<tr>
<td>M2</td>
<td><strong>Percentage of captions that pass the Turing Test.</strong></td>
</tr>
<tr>
<td>M3</td>
<td>Average correctness of the captions on a scale 1-5 (incorrect - correct).</td>
</tr>
<tr>
<td>M4</td>
<td>Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).</td>
</tr>
<tr>
<td>M5</td>
<td><strong>Percentage of captions that are similar to human description.</strong></td>
</tr>
</tbody>
</table>
→ Turns out this works **really well.**

- Also, 2\textsuperscript{nd} place at CVPR image captioning challenge
  - When we add in GRNN forced decoding

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Percentage of captions that are evaluated as better or equal to human caption.</td>
</tr>
<tr>
<td>M2</td>
<td>Percentage of captions that pass the Turing Test.</td>
</tr>
<tr>
<td><strong>M3</strong></td>
<td><strong>Average correctness of the captions on a scale 1-5 (incorrect - correct).</strong></td>
</tr>
<tr>
<td>M4</td>
<td>Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).</td>
</tr>
<tr>
<td>M5</td>
<td>Percentage of captions that are similar to human description.</td>
</tr>
</tbody>
</table>
A baseball player throwing a ball.
Now You Try!
Image Captioning: Now You Try!

- Go to https://www.microsoft.com/cognitive-services/en-us/computer-vision-api
- Click API Reference
- Click Describe Image
- Then Open API Testing Console
Image Captioning: Now You Try!

Cognitive Services – Image Captioning

• **Request URL:**
  • [https://api.projectoxford.ai/vision/v1.0/describe?3](https://api.projectoxford.ai/vision/v1.0/describe?3) ← return 3 captions

• **Image URL:** [http://visionandlanguage.net/captioning.jpg](http://visionandlanguage.net/captioning.jpg)

• **Response:**
  
  ```json
  description: {
    tags: [man, outdoor, sitting, window, bench, building, tree],
    captions: [
      { text: A man sitting on a bench, confidence: 0.483},
      { text: A man is sitting on a bench, confidence: 0.400},
      { text: A man sitting in front of a building, confidence: 0.380 }
    ]
  }
  ```
New Challenges informed by visually impaired accessibility
Visual Question Answering
VQA: Visual Question Answering

What is the mustache made of?

AI System

bananas

visualqa.org
VQA: Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

visualqa.org
VQA: Visual Question Answering

254,721 images
764,163 questions (3 questions per image)
764,163 answers
  – 10 ground truth answers per question
  – 3 plausible (but likely incorrect) answers per question
Open-ended and multiple-choice answering tasks

visualqa.org
Photo Complexity
Pictures Taken By People Who Are Visually Impaired Are Visually Complex
Pictures Taken By People Who Are Visually Impaired Are Visually Complex

- Partnership with Jeff Bigham, VizWiz
- Visually Impaired users take picture, ask question, answered by crowdsourcing.
  - E.g., What is this? Is this a lemon or a lime? What do the instructions say?
- Annotating images for Illumination, blur, cropping, clutter
We Can Tackle This Problem!

- Need: Engineers dedicated to making this work.
Expressive Language
Abstract Conceptual Thoughts in Sequence

Microsoft Sequential Image Narrative Dataset (SIND)

What is Microsoft SIND?
We introduce the first dataset for sequential vision-to-language, and explore how this data may be used for the task of visual storytelling. The dataset includes 81,743 unique photos in 20,211 sequences, aligned to descriptive and story language.

Download Paper
Local Arxiv BibTex

sind.ai
Abstract Conceptual Thoughts in Sequence

Example Generated Story

1. The dog was ready to go.
2. He had a great time on the hike.
3. And was very happy to be in the field.
4. His mom was so proud of him.
5. It was a beautiful day for him.

Photos by kamera6chwein / CC BY-NC-ND 2.0
Looking further...

- Descriptions from video

- Video is images in sequence . . .
  . . . with serious **compute needs**. (Hololens HPU?)
“Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers”
Welcome to Amazon.com Books!

One million titles. Consistently low prices.

(If you explore just one thing, make it our personal notification service. We think it's very cool!)

**SPOTLIGHT! — AUGUST 16TH**

These are the books we love, offered at Amazon.com low prices. The spotlight moves EVERY day so please come often.

**ONE MILLION TITLES**

Search Amazon.com's [million title catalog] by author, subject, title, keyword, and more... Or take a look at the books we recommend in over 20 categories... Check out our [customer reviews] and the [award winners] from the Hugo and Nebula to the Pulitzer and Nobel... and [bestsellers] are 30% off the publishers list...

**EYES & EDITORS, A PERSONAL NOTIFICATION SERVICE**

Like to know when that book you want comes out in paperback or when your favorite author has a new book? Now you can sign up for our [notify service].
Web Today: Increasingly Visual
-- social media, news media, online shopping

- Facebook.com has over 250 billion images uploaded as of Jun 2013
- 1.15 billion users uploading 350 million images a day on average
Learning from data in the wild

Deja Image-caption corpus (NAACL 2015):
- Of 750 million image-caption pairs from Flickr
- Retain only those captions that are repeated verbatim by more than one user
- Yielding 4 million images with 180K unique captions

The sun sets for another day (12)
Sun is going to bed (21)
After the sun has set (9)
The sky looks like it is on fire (58)
Rippled sky (44)
Seeing beyond What’s in the Image

- What’s happening?
- How / why did this happen?
- What are the intent / goal of the participants?
- Sentiment: are they happy?
- Reaction: do we need to act on them (e.g., dispatching help)?
Size Graph

- dog is 83 cm tall
- dog is ~0.5 m tall
- dog is 70 - 75 cm tall

- tree is 20 m tall
- tree is about 6 m tall
- tree is 4-12 m tall
Which of objects A or B is bigger?

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>50%</td>
</tr>
<tr>
<td>Text Only Baseline</td>
<td>63.4%</td>
</tr>
<tr>
<td>Vision Only Baseline</td>
<td>72.4%</td>
</tr>
<tr>
<td>Our Model</td>
<td>83.5%</td>
</tr>
<tr>
<td>Our Model (image only)</td>
<td>78.4%</td>
</tr>
<tr>
<td>Our Model (text only)</td>
<td>75.3%</td>
</tr>
</tbody>
</table>
Segment-Phrase Table for Visual Entailment (ICCV 2015)

<table>
<thead>
<tr>
<th>Web Text Data</th>
<th>Web Image Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Segment-Phrase Table</strong></td>
<td></td>
</tr>
<tr>
<td>Horse jumping</td>
<td>![Horse jumping image]</td>
</tr>
<tr>
<td>Cat sanding up</td>
<td>![Cat sanding up image]</td>
</tr>
<tr>
<td>Bear running</td>
<td>![Bear running image]</td>
</tr>
<tr>
<td>Chimpanzee lying</td>
<td>![Chimpanzee lying image]</td>
</tr>
</tbody>
</table>

Visual Entailment

- ![Horse fighting image] (Correct)
- ![Horse rearing image] (Incorrect)
Segment-Phrase Table for Visual Entailment (ICCV 2015)

<table>
<thead>
<tr>
<th>Web Text Data</th>
<th>Web Image Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Segment-Phrase Table</strong></td>
<td></td>
</tr>
<tr>
<td>Horse jumping</td>
<td>![Horse jumping image]</td>
</tr>
<tr>
<td>Cat sanding up</td>
<td>![Cat sanding up image]</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Bear running</td>
<td>![Bear running image]</td>
</tr>
<tr>
<td>Chimpanzee lying</td>
<td>![Chimpanzee lying image]</td>
</tr>
</tbody>
</table>

**Visual Entailment**
- Horse fighting \(\models\) Horse rearing
- Horse jumping \(\models\) Horse leaping

**Visual Paraphrasing**
- Horse jumping \(\models\) Horse leaping
- Cat standing up \(\models\) Bear standing up
- Cat standing up \(\models\) Deer standing up

**Relative Similarity**
- Cat standing up \(>\) Bear standing up
- Cat standing up \(>\) Deer standing up
Thanks!!!

- Yejin Choi: yejin@cs.washington.edu

- Margaret Mitchell: memitc@Microsoft.com