Neural MT
Announcements

• HW2 directory structure penalty to be removed due to grading inconsistencies.
  • Those who lost 15 points will gain 15 points

• Dan Jurafsky will attend the beginning of class next Tuesday
  • Be prepared with questions. Your chance!!!

• Rupal Patel: Monday, Dec. 4\textsuperscript{th}, 11:30, Davis
"Does This Vehicle Belong to You?" Processing the Language of Policing for Improving Police-Community Relations

ABSTRACT

Police body-worn cameras have the potential to play an important role in understanding and improving police-community relations. In this talk I describe a series of studies conducted by our large interdisciplinary team at Stanford that use speech and natural language processing on body-camera recordings to model the interactions between police officers and community members in traffic stops. We use text and speech features to automatically measure linguistic aspects of the interaction, from discourse factors like conversational structure to social factors like respect. I describe the differences we find in the language directed toward black versus white community members, and offer suggestions for how these findings can be used to help improve the fraught relations between police officers and the communities they serve.
Today

• Multilingual Challenges for MT

• MT Approaches
  • Statistical
  • Neural net (Thursday)

• MT Evaluation
MT Evaluation

• More art than science
• Wide range of Metrics/Techniques
  • interface, ..., scalability, ..., faithfulness, ...
    space/time complexity, ... etc.
• Automatic vs. Human-based
  • *Dumb Machines vs. Slow Humans*
<table>
<thead>
<tr>
<th></th>
<th>contents of original sentence conveyed (might need minor corrections)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>contents of original sentence conveyed BUT errors in word order</td>
</tr>
<tr>
<td>4</td>
<td>contents of original sentence generally conveyed BUT errors in relationship between phrases, tense, singular/plural, etc.</td>
</tr>
<tr>
<td>3</td>
<td>contents of original sentence not adequately conveyed, portions of original sentence incorrectly translated, missing modifiers</td>
</tr>
<tr>
<td>2</td>
<td>contents of original sentence not conveyed, missing verbs, subjects, objects, phrases or clauses</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Human-based Evaluation Example
Fluency Criteria

<table>
<thead>
<tr>
<th></th>
<th>clear meaning, good grammar, terminology and sentence structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>clear meaning BUT bad grammar, bad terminology or bad sentence structure</td>
</tr>
<tr>
<td>4</td>
<td>meaning graspable BUT ambiguities due to bad grammar, bad terminology or bad sentence structure</td>
</tr>
<tr>
<td>3</td>
<td>meaning unclear BUT inferable</td>
</tr>
<tr>
<td>2</td>
<td>meaning absolutely unclear</td>
</tr>
</tbody>
</table>
Today: Crowdsourcing

- Amazon Mechanical Turk or CrowdFlower
- Create a HIT for each sentence
- Get multiple workers to rate
- Pay .01 to .10 per hit
- Complete an evaluation in hours (vs days/weeks)
- Ethics?
Automatic Evaluation Example

Bleu Metric
(Papineni et al 2001)

• Bleu
  • BiLingual Evaluation Understudy
  • Modified n-gram precision with length penalty
  • Quick, inexpensive and language independent
  • Correlates highly with human evaluation
  • Bias against synonyms and inflectional variations

Slide from Nizar Habash
Automatic Evaluation Example
Bleu Metric

Test Sentence

colorless green ideas sleep furiously

Gold Standard References

all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Slide from Nizar Habash
Automatic Evaluation Example

Bleu Metric

Test Sentence

colorless green ideas sleep furiously

Gold Standard References

all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Unigram precision = 4/5

Slide from Nizar Habash
Automatic Evaluation Example

Bleu Metric

Test Sentence

- colorless green
green ideas
- ideas sleep
sleep furiously

Gold Standard References

- all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Unigram precision = \( \frac{4}{5} = 0.8 \)
Bigram precision = \( \frac{2}{4} = 0.5 \)

Bleu Score = \( \left( a_1 a_2 \ldots a_n \right)^{1/n} \)
= \( (0.8 \times 0.5)^{\frac{1}{2}} = 0.6325 \rightarrow 63.25 \)

Slide from Nizar Habash
### BLEU scores for 110 translation systems trained on Europarl

<table>
<thead>
<tr>
<th>Source Language</th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>fi</th>
<th>it</th>
<th>nl</th>
<th>pt</th>
<th>sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>-</td>
<td>18.4</td>
<td>21.1</td>
<td>28.5</td>
<td>26.4</td>
<td>28.7</td>
<td>14.2</td>
<td>22.2</td>
<td>21.4</td>
<td>24.3</td>
<td>28.3</td>
</tr>
<tr>
<td>de</td>
<td>22.3</td>
<td>-</td>
<td>20.7</td>
<td>25.3</td>
<td>25.4</td>
<td>27.7</td>
<td>11.8</td>
<td>21.3</td>
<td>23.4</td>
<td>23.2</td>
<td>20.5</td>
</tr>
<tr>
<td>el</td>
<td>22.7</td>
<td>17.4</td>
<td>-</td>
<td>27.2</td>
<td><strong>31.2</strong></td>
<td><strong>32.1</strong></td>
<td>11.4</td>
<td>26.8</td>
<td>20.0</td>
<td>27.6</td>
<td>21.2</td>
</tr>
<tr>
<td>en</td>
<td>25.2</td>
<td>17.6</td>
<td>23.2</td>
<td>-</td>
<td><strong>30.1</strong></td>
<td><strong>31.1</strong></td>
<td>13.0</td>
<td>25.3</td>
<td>21.0</td>
<td>27.1</td>
<td>24.8</td>
</tr>
<tr>
<td>es</td>
<td>24.1</td>
<td>18.2</td>
<td>28.3</td>
<td><strong>30.5</strong></td>
<td>-</td>
<td>40.2</td>
<td>12.5</td>
<td><strong>32.3</strong></td>
<td>21.4</td>
<td><strong>35.9</strong></td>
<td>23.9</td>
</tr>
<tr>
<td>fr</td>
<td>23.7</td>
<td>18.5</td>
<td>26.1</td>
<td><strong>30.0</strong></td>
<td><strong>38.4</strong></td>
<td>-</td>
<td>12.6</td>
<td><strong>32.4</strong></td>
<td>21.1</td>
<td><strong>35.3</strong></td>
<td>22.6</td>
</tr>
<tr>
<td>fi</td>
<td>20.0</td>
<td>14.5</td>
<td>18.2</td>
<td>21.8</td>
<td>21.1</td>
<td>22.4</td>
<td>-</td>
<td>18.3</td>
<td>17.0</td>
<td>19.1</td>
<td>18.8</td>
</tr>
<tr>
<td>it</td>
<td>21.4</td>
<td>16.9</td>
<td>24.8</td>
<td>27.8</td>
<td><strong>34.0</strong></td>
<td><strong>36.0</strong></td>
<td>11.0</td>
<td>-</td>
<td>20.0</td>
<td><strong>31.2</strong></td>
<td>20.2</td>
</tr>
<tr>
<td>nl</td>
<td>20.5</td>
<td>18.3</td>
<td>17.4</td>
<td>23.0</td>
<td>22.9</td>
<td>24.6</td>
<td>10.3</td>
<td>20.0</td>
<td>-</td>
<td>20.7</td>
<td>19.0</td>
</tr>
<tr>
<td>pt</td>
<td>23.2</td>
<td>18.2</td>
<td>26.4</td>
<td><strong>30.1</strong></td>
<td><strong>37.9</strong></td>
<td><strong>39.0</strong></td>
<td>11.9</td>
<td><strong>32.0</strong></td>
<td>20.2</td>
<td>-</td>
<td>21.9</td>
</tr>
<tr>
<td>sv</td>
<td><strong>30.3</strong></td>
<td>18.9</td>
<td>22.8</td>
<td><strong>30.2</strong></td>
<td>28.6</td>
<td>29.7</td>
<td>15.3</td>
<td>23.9</td>
<td>21.9</td>
<td>25.9</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus**

Koehn, MT Summit, 2005
<table>
<thead>
<tr>
<th>Language</th>
<th>From</th>
<th>Into</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danish (da)</td>
<td>23.4</td>
<td>23.3</td>
<td>0.0</td>
</tr>
<tr>
<td>German (de)</td>
<td>22.2</td>
<td>17.7</td>
<td>-4.5</td>
</tr>
<tr>
<td>Greek (el)</td>
<td>23.8</td>
<td>22.9</td>
<td>-0.9</td>
</tr>
<tr>
<td>English (en)</td>
<td>23.8</td>
<td>27.4</td>
<td>+3.6</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>26.7</td>
<td>29.6</td>
<td>+2.9</td>
</tr>
<tr>
<td>French (fr)</td>
<td>26.1</td>
<td>31.1</td>
<td>+5.1</td>
</tr>
<tr>
<td>Finnish (fi)</td>
<td>19.1</td>
<td>12.4</td>
<td>-6.7</td>
</tr>
<tr>
<td>Italian (it)</td>
<td>24.3</td>
<td>25.4</td>
<td>+1.1</td>
</tr>
<tr>
<td>Dutch (nl)</td>
<td>19.7</td>
<td>20.7</td>
<td>+1.1</td>
</tr>
<tr>
<td>Portuguese (pt)</td>
<td>26.1</td>
<td>27.0</td>
<td>+0.9</td>
</tr>
<tr>
<td>Swedish (sv)</td>
<td>24.8</td>
<td>22.1</td>
<td>-2.6</td>
</tr>
</tbody>
</table>

Table 3: Average translation scores for systems when translating from and into a language. Note that German (de) and English (en) are similarly difficult to translate from, but English is much easier to translate into.
Automatic Evaluation Example

METEOR
(Lavie and Agrawal 2007)

- Metric for Evaluation of Translation with Explicit word Ordering
- Extended Matching between translation and reference
  - Porter stems, wordNet synsets
- Unigram Precision, Recall, parameterized F-measure
- Reordering Penalty
- Parameters can be tuned to optimize correlation with human judgments
- Not biased against “non-statistical” MT systems

Slide from Nizar Habash
Metrics MATR Workshop

- Workshop in AMTA conference 2008
  - Association for Machine Translation in the Americas
- Evaluating evaluation metrics
- Compared 39 metrics
  - 7 baselines and 32 new metrics
  - Various measures of correlation with human judgment
  - Different conditions: text genre, source language, number of references, etc.
Automatic Evaluation Example

SEPIA
(Habash and ElKholy 2008)

• A syntactically-aware evaluation metric
  • (Liu and Gildea, 2005; Owczarzak et al., 2007; Giménez and Màrquez, 2007)
• Uses dependency representation
  • MICA parser (Nasr & Rambow 2006)
  • 77% of all structural bigrams are surface n-grams of size 2,3,4
• Includes dependency surface span as a factor in score
  • long-distance dependencies should receive a greater weight than short distance dependencies
  • Higher degree of grammaticality?

[Diagram showing tree structure of a sentence]
Why do people continue to use BLEU

Start the presentation to activate live content
If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
Neural MT takes over

• WMT (Workshop on Machine Translation)

• 2015 – first neural MT, lower bleu results

• 2016: neural MT beats phrase-based and syntax-based
Results from WMT (Workshop on Machine Translation)  
German to English  
2015: Montreal  
2016 and 2017: Edinburgh
WMT 2017

• Tasks
  • News translation
  • Quality estimation
  • Automatic post-editing
  • Metrics
  • Multimodal MT and multilingual image description
  • Biomedical translation
What is being tested in the biomedical task?
News Translation Task

- 7 languages, 14 tasks (from and into English)
  - Chinese
  - Czech
  - German
  - Finnish
  - Latvian
  - Russian
  - Turkish
- Test data: 3000 sentences per language pair except Latvian: 2000 sentences
Training Data

- Europarl
- Common Crawl
- Yandex Russian-English data
- Wikipedia Headlines
- United Nations
- News Commentary V12
- EU Press Release parallel corpus for German, Finnish and Latvian
Submitted Systems

- 103 systems from 31 institutions (no companies)

- Company releases of Neural MT
  - Microsoft: February 2016
  - Systran: August 2016
  - Google: September 2016
Human Evaluation

• Assess on adequacy along a 100 point scale (Direct Assessment) (vs Relative Ranking)
  • How adequately does the translation express the meaning of the reference translation?
  • One translation per screen/hit

• 151 individual Researchers
  • 29 different groups
  • Contributed 12,693 translation scores
  • 24 days, 22 hours

• 754 AMT workers
  • Contributed 237,200 scores
  • 47 days, 23 hours
Could we control quality on an Amazon Mechanical evaluation? Is it reliable?
Some Results

Chinese → English

English → Chinese
<table>
<thead>
<tr>
<th>Czech→English</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>online-B</td>
<td>online-A</td>
<td>PJATK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English→Czech</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>online-B</td>
<td>online-A</td>
<td>PJATK</td>
</tr>
<tr>
<td>limsi-factored-norm</td>
<td>LIUM-FNMT</td>
<td>LIUM-NMT</td>
<td>CU-Chimera</td>
</tr>
<tr>
<td>online-A</td>
<td>tuning-task-ufal_8gb</td>
<td>tuning-task-afrl_4gb</td>
<td>PJATK</td>
</tr>
<tr>
<td>tuning-task-baseline_8gb</td>
<td>tuning-task-afrl_8gb</td>
<td>tuning-task-ufal_4gb</td>
<td>tuning-task-denisor_4gb</td>
</tr>
</tbody>
</table>
Today

• Multilingual Challenges for MT

• MT Approaches
  • Statistical
  • Neural net (Thursday)

• MT Evaluation
Encoder-Decoder Approach
Basic RNN Approach

ENCODER

```
<table>
<thead>
<tr>
<th>h₁</th>
<th>h₂</th>
<th>h₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td>x₂ₓ</td>
<td>x₃ₓ</td>
</tr>
<tr>
<td>das</td>
<td>ist</td>
<td>fur</td>
</tr>
</tbody>
</table>
```

DECODER

```
<table>
<thead>
<tr>
<th>Y₁</th>
<th>Y₂</th>
<th>Y₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>That</td>
<td>is</td>
<td>almost</td>
</tr>
</tbody>
</table>
```
Basic RNN Approach

**ENCODER**

```
<table>
<thead>
<tr>
<th></th>
<th>h₁</th>
<th>h₂</th>
<th>h₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x₃</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

das ist fur

**DECODER**

```
<table>
<thead>
<tr>
<th></th>
<th>Y₁</th>
<th>Y₂</th>
<th>Y₃</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

That is almost

Entire input represented here
Recurrent decoder but

ENCODER

Transition $z_t = f(z_{t-1}, y_{T-1}, h_n)$
Backpropagation = $\Sigma_t \delta z_t / \delta h$

DECODER

That is almost
Figure 1: An illustration of the proposed RNN Encoder–Decoder.

Cho et al 2014
<table>
<thead>
<tr>
<th>English Expression</th>
<th>French Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>at the end of the</td>
<td>à la fin de la [à la fin des années] [être supprimés à la fin de la]</td>
</tr>
<tr>
<td>for the first time</td>
<td>pour la première fois [été donnés pour la première fois] [été commémorée pour la première fois]</td>
</tr>
<tr>
<td>in the United States and</td>
<td>aux États-Unis et [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]</td>
</tr>
<tr>
<td>, as well as</td>
<td>que [s, ainsi que] [re aussi bien que]</td>
</tr>
<tr>
<td>one of the most</td>
<td>l’ un des plus [l’ un des plus] [être retenue comme un de ses plus]</td>
</tr>
</tbody>
</table>

### Results for Long Frequent Phrases

```
[à la fin du] [à la fin des] [à la fin de la]
[pour la première fois] [pour la première fois ,] [pour la première fois que]
[aux États-Unis et] [des États-Unis et] [des États-Unis et]
[, ainsi qu’] [, ainsi que] [, ainsi que les]
[l’ un des] [le] [un des]
```

Cho et al 2014
Other Variants: Train weights separately

ENCODER

\[
\begin{align*}
\text{das} & \xrightarrow{h_1} \xrightarrow{h_2} \xrightarrow{h_3} \\
\text{ist} & \xrightarrow{X_2} \\
\text{fur} & \xrightarrow{X_3}
\end{align*}
\]

DECODER

\[
\begin{align*}
Y_1 & \xrightarrow{X_1} \xrightarrow{X_2} \xrightarrow{X_3} \\
\text{That} & \xrightarrow{Y_1} \xrightarrow{Y_2} \xrightarrow{Y_3} \\
\text{is} & \xrightarrow{Y_2} \\
\text{almost} & \xrightarrow{Y_3}
\end{align*}
\]
Also Useful

• Train stacked RNNS using multiple layers

• Use a bidirectional encoder
  • This can help in remembering the early part of the source input sentence

• Train the input sequence in reverse order: $S_1 S_2 S_3 \rightarrow T_1 T_2 T_3$ would be trained as $S_3 S_2 S_1 \rightarrow T_1 T_2 T_3$
  • Why?
Replacing RNN with LSTM improves performance further

<table>
<thead>
<tr>
<th>Method</th>
<th>test BLEU score (ntst14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahdanau et al. [2]</td>
<td>28.45</td>
</tr>
<tr>
<td>Baseline System [29]</td>
<td>33.30</td>
</tr>
<tr>
<td>Single forward LSTM, beam size 12</td>
<td>26.17</td>
</tr>
<tr>
<td>Single reversed LSTM, beam size 12</td>
<td>30.59</td>
</tr>
<tr>
<td>Ensemble of 5 reversed LSTMs, beam size 1</td>
<td>33.00</td>
</tr>
<tr>
<td>Ensemble of 2 reversed LSTMs, beam size 12</td>
<td>33.27</td>
</tr>
<tr>
<td>Ensemble of 5 reversed LSTMs, beam size 2</td>
<td>34.50</td>
</tr>
<tr>
<td>Ensemble of 5 reversed LSTMs, beam size 12</td>
<td><strong>34.81</strong></td>
</tr>
</tbody>
</table>

Table 1: The performance of the LSTM on WMT’14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

<table>
<thead>
<tr>
<th>Method</th>
<th>test BLEU score (ntst14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline System [29]</td>
<td>33.30</td>
</tr>
<tr>
<td>Cho et al. [5]</td>
<td>34.54</td>
</tr>
<tr>
<td>State of the art [9]</td>
<td><strong>37.0</strong></td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with a single forward LSTM</td>
<td>35.61</td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with a single reversed LSTM</td>
<td>35.85</td>
</tr>
<tr>
<td>Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs</td>
<td><strong>36.5</strong></td>
</tr>
<tr>
<td>Oracle Rescoring of the Baseline 1000-best lists</td>
<td>~45</td>
</tr>
</tbody>
</table>

Table 2: Methods that use neural networks together with an SMT system on the WMT’14 English to French test set (ntst14).
Aligning and Translating

Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight $c_{ij}$ of the annotation of the $j$-th source word for the $i$-th target word (see Eq. (6), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b-d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.
Attention Mechanism - Scoring

ENCODER

```
<table>
<thead>
<tr>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>das</td>
<td>ist</td>
<td>fur</td>
</tr>
</tbody>
</table>
```

Score \((h'_{t-1}, h_s)\)

DECODER

```
<table>
<thead>
<tr>
<th>H'_1</th>
<th>H'_2</th>
<th>H'_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y_1</td>
<td>Y_2</td>
<td>Y_3</td>
</tr>
</tbody>
</table>
```

That \(?\)
Attention Mechanism - Scoring

ENCODER

```
das
```

```
ist
```

```
fur
```

Score \((h'_{t-1}, h_s)\)

DECODER

```
Y_1
```

```
Y_2
```

```
Y_3
```

That

?
Attention Mechanism - Scoring

ENCODER

h_1 → h_2 → h_3
x_1 → x_{2v} → x_{3v}

das ist fur

Score \((h'_{t-1}, h_s)\)

DECODER

H'_1 → H'_2 → H'_3
Y_1 → Y_2 → Y_3

That ?
Attention Mechanism - Scoring

\[ \alpha_t(s) = \frac{\text{escore}(s)}{\sum_{s'} \text{escore}(s')} \]

Convert into alignment weights

DECODER

That

?
Build context vector: weighted average

\[ C_t = \sum_s \alpha_t(s) h_s \]
How do you score it?

Score \((h_s, H'_t) = H'_t^T h_s\)

or \(= H'_t^T W_\alpha h_s\) (Luong et al 2015)
Performance

• Without attention, LSTM works quite well until a sentence gets longer than 30 words

• Attention does better, however, even with shorter sentences

• Other tricks in WMT 2017:
  • Improvements of 1.5 – 3 blue points (Edin)
  • Layer normalization, deeper networks (encoder depth of 5, decoder depth of 8)
  • Base Phrase Encodings (BPE)
    • Reduced vocabulary improves memory efficiency
  • Data: parallel, back-translated, duplicated monolingual
Questions?
Information Extraction

• Extraction of concrete facts from text

• Named entities, relations, events

• Often used to create a structured knowledge base of facts
Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.
Named Entities

- Kathy McKeown\textsubscript{per}, a professor from Columbia University\textsubscript{org} in New York City\textsubscript{loc}, took a train yesterday to Washington DC\textsubscript{loc}.
Named Entities, Relations

- Kathy McKeown\textsubscript{per}, a professor from Columbia University\textsubscript{org} in New York City\textsubscript{loc}, took a train yesterday to Washington DC\textsubscript{loc}.

- Kathy McKeown from Columbia
- Columbia in New York City
Named Entities, Relations, Events

• Kathy McKeown\textsubscript{per}, a professor from Columbia University\textsubscript{org} in New York City\textsubscript{loc}, took a train yesterday to Washington DC\textsubscript{loc}.

• Kathy McKeown took a train (yesterday)
Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.
State of the Art (English)

- Named Entities (news)
- Relations (slot filling)
- Events (nuggets)

F-measure
- 89%
- 59%
- 63%

Methods: Sequence labeling (MEMM, CRF), neural nets, distant learning
Features: linguistic features, similarity, popularity, gazeteers, ontologies, verb triggers
## Where Have You Been
Entity Discovery and Linking?

<table>
<thead>
<tr>
<th>Grow with DEFT</th>
<th>2006-2011</th>
<th>2012-2017</th>
<th>( \text{HENG JI, RPI} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mention Extraction</td>
<td>Human (most)</td>
<td>Automatic</td>
<td></td>
</tr>
<tr>
<td>NIL Clustering</td>
<td>None</td>
<td>64 methods</td>
<td></td>
</tr>
<tr>
<td>Foreign Languages</td>
<td>Chinese (5%-10% lower than English)</td>
<td>System for 282 languages (Chinese/Spanish comparable to/Outperform English); research toward 3,000 languages</td>
<td></td>
</tr>
<tr>
<td>Document Size</td>
<td>-</td>
<td>500 → 90,000 documents</td>
<td></td>
</tr>
<tr>
<td>Genre</td>
<td>News, web blog</td>
<td>( \text{News, Discussion Forum, Web blog, Tweets} )</td>
<td></td>
</tr>
<tr>
<td>Entity Types</td>
<td>PER, GPE, ORG</td>
<td>( \text{PER, GPE, ORG, LOC, FAC, hundreds of fine-grained types for typing} )</td>
<td></td>
</tr>
<tr>
<td>Mention Types</td>
<td>Name or all concepts (most)</td>
<td>Name, Nominal, Pronoun (for BeST)</td>
<td></td>
</tr>
<tr>
<td>KB</td>
<td>Wikipedia</td>
<td>Freebase → List only</td>
<td></td>
</tr>
<tr>
<td>Training Data</td>
<td>20,000 queries (entity mentions)</td>
<td>500 → 0 documents; unsupervised linking comparable to supervised linking</td>
<td></td>
</tr>
<tr>
<td>#(Good) Papers</td>
<td>62</td>
<td>110 (new KBP track at ACL); 6 tutorials at top conferences</td>
<td></td>
</tr>
</tbody>
</table>

\*Slide from Heng Ji\*
On the Horizon: Entity Discovery and Linking

Panel: Hoa Trang Dang, Jason Duncan, Heng Ji, Kevin Knight, Christopher Manning, Dan Roth

- Am going crazy
  - 3,000 languages
  - 10,000 entity types
  - All mention types
  - Multi-media
  - Streaming mode
  - List-only KB
  - Context-aware, living
  - No more training data
  - On-call evaluation
  - More non-traditional knowledge resources
  - Lots of dev and test sets in lots of languages

- Am staying cool
  - Success in end-to-end cold-start KBP
  - What’s still wrong with name tagging
  - Smarter collective inference
  - Resolution of true aliases
  - Resolution of handles used as entity mentions

Slide from Heng Ji