Tell me quickly, general or specific?

Junyi Jessy Li and Ani Nenkova
University of Pennsylvania
http://seas.upenn.edu/~ljunyi
Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• Goal of SPECITELLER
• First step: supervised learning
• SPECITELLER: semi-supervised learning
• Use case: sentence simplification
• SPECITELLER: hands on
Outline

• A very high level peek into Natural Language Processing
  • Sentence specificity: definition and motivation
  • Goal of SPECITELLER
  • First step: supervised learning
  • SPECITELLER: semi-supervised learning
  • Use case: sentence simplification
  • SPECITELLER: hands on
Communication

Idea & thoughts

Encode into natural language

Message

Decode from natural language

Understanding

Feedback

Encode into natural language
Natural Language Processing

- Ideas & thoughts
- Encode into natural language
- Message
- Decode from natural language
- Understanding
- Feedback
- Encode into natural language
Conversational agents

What can I help you with?

“I’m bored”

I bore you, Phil?

“No not you can you entertain me”

Perhaps not.

“Tell me a joke”

Two iPhones walk into a bar... I forget the rest.
Question answering and information extraction

Who performed at the Super Bowl 2016?

2016 Super Bowl halftime show: Beyoncé teaming up with Coldplay. Feb 6, 2016

2016 Super Bowl halftime show: Beyoncé teaming up with Coldplay. [link]
但是因为开放住宅小区牵扯了许多城市居民的切身利益，新措施引发了社会广泛关注和网络激烈讨论。激烈的网络讨论主要围绕产权、交通和居住安全等内容进行。

But because open residential area involve the vital interests of many urban residents, the new measures sparked widespread concern and heated discussion network. Intense online discussion mainly around the property, transportation, and residential security and other content.
PLATTSBURGH, N.Y. — A small favor for a killer's daughter. A stolen kiss and a furtive sexual encounter. And ultimately, sneaking tools past guards for use in an audacious escape from a maximum-security prison.

Such were the moments in Joyce E. Mitchell's evolution from a workaday prison employee to a love-struck and fully aware accomplice, as outlined in three statements Ms. Mitchell made last month to investigators.

Ms. Mitchell, 51, pleaded guilty on Tuesday to assisting Richard W. Matt and David Sweat, two convicted murderers, in their elaborate breakout early last month from the Clinton Correctional Facility in Dannemora, N.Y.

"I was already bringing stuff in to him, and didn't really feel I could stop," she said, adding that Mr. Matt assured her that "they were getting out, and we were all going to be together."

The escape set off a three-week manhunt before Mr. Matt was shot and killed on June 26 and Mr. Sweat was shot and captured on June 28.

Officials have offered ample details about the inmates' plans and the law enforcement pursuit, and Mr. Sweat himself has bragged about his role in the escape. But the central question of what drove Ms. Mitchell, a soft-spoken married woman, to help the prisoners has only now been answered.

Her statements to investigators sketch a vivid chronology of how Ms. Mitchell, a former supervisor in the prison's tailor shop, came to commit her crimes by falling for a prisoner, eventually succumbing to the seemingly irresistible momentum and dark excitement of the escape scheme.

Continue reading the main story

RELATED COVERAGE

Joyce Mitchell, Ex-New York Prison Worker, Pleads Guilty to Aiding Killers' Escape JULY 28, 2015
Broken Boys, Thieves, Killers, and Now Escapees JUNE 12, 2015
Squabbling, Hesitation and Luck Had Roles in Manhunt for New York Prison Escapees JUNE 29, 2015
David Sweat, Escaped New York Convict, Is Shot and Captured as Hunt Ends JUNE 28, 2015

Ms. Mitchell was meant to be the getaway driver. She was to meet the two men at midnight near a powerhouse outside the prison, but she had a panic attack and did not show up. That panic may have been caused by threats made by the two men directed at Ms. Mitchell's husband, Lyle, whom the inmates nicknamed "the Glitch," and whom they planned to drug and murder.

"Inmate Matt was going to kill 'the Glitch,'" Ms. Mitchell said.

Her confession was first reported by NBC, and the statements were obtained by The New York Times through a Freedom of Information request. They reveal how deeply involved she was with the men and how detailed their escape plan was.

Ms. Mitchell, for example, was meant to drive a Jeep, with four-wheel drive, "because the place we were going was in the woods." She was also meant to bring an array of supplies for the escape, including tents, sleeping bags, GPS, money, a cellphone and "fishing poles."

Mr. Matt also asked Ms. Mitchell to "bring a shotgun that he could saw off," she said.

The plan was to hide out somewhere six to seven hours from Dannemora "until it quieted down." Then, after about a week, she and Mr. Sweat would continue to hide and Mr. Matt would go on his own. "I was caught up in the fantasy," she said. "I enjoyed the attention, the feeling both of them gave me, and the thought of a different life."

But while she said she liked Mr. Sweat, her primary affections were undoubtedly for Mr. Matt. "Each time he would ask me for a tool, I would go to the store and get them," she said on June 8, two days after the escape.

The statements, which were given on June 7, 8 and 10 to investigators including federal agents and the New York State Police, show the seduction of Ms. Mitchell by Mr. Matt, who had dark, rakish good looks and a charming demeanor, according to officials and Ms. Mitchell.
Can machines understand language?

Syntactic and semantic ambiguities

I ate cherry pie with ice cream.
I ate cherry pie with a spoon.

The astronomer married the star.

Jack called John at 6am.
He woke him up.
He got quite mad at him.
Selected core tasks

• Syntax:
  • Part of speech tagging
    The astronomer married the star.
    The_DT astronomer_NN married_VBD the_DT star_NN ./

• Syntactic parsing
Selected core tasks

• Semantics:
  • Representing and computing meaning

Utterance: Compute three plus four.
Logical form: (call + 3 4)
Denotation: 7

Examples from Stanford SEMPRE parser website
Selected core tasks

- Semantics:
  - Lexical semantics
Selected core tasks

• Computational discourse
  • Coreference resolution

  Jack called John at 6am.
  He woke him up.
  He got quite mad at him.
Selected core tasks

- Computational discourse
- Discourse relation recognition

But video contents are different, as their cost of production is much higher than that of text contents, as are the costs of storage and transmission. Therefore, video searches should be classified as vertical searches, and as compared to words, their search area is easier to circle.
Selected core tasks

• Computational discourse
• Discourse relation recognition

But video contents are different, as their cost of production is much higher than that of text contents, as are the costs of storage and transmission. Therefore, video searches should be classified as vertical searches, and as compared to words, their search area is easier to circumscribe.
Outline

• A very high level peek into Natural Language Processing

• Sentence specificity: definition and motivation

• Goal of SPECITELLER

• First step: supervised learning

• SPECITELLER: semi-supervised learning

• Use case: sentence simplification

• SPECITELLER: hands on
What is sentence specificity?

Level of detail present in the sentence

Evidence of widespread cheating has surfaced in several states in the last year or so.

California’s education department suspects adult responsibility for erasures at 40 schools that changed wrong answers to right ones on a statewide test.
Motivation

Text accessibility

In consequence, as early as 1726 St. Petersburg was handling 90 percent of Russia’s foreign trade.
- Encyclopedia Britannica

The city soon started to grow rapidly as much of Russia’s foreign trade passed through the city.
- Britannica Elementary
Motivation

Text accessibility

Automatic summarization

Good summaries require generalization:

• Human summaries more general than input
• Machine summaries more specific than input
Motivation

Argument quality

- High quality arguments: stands alone
  - *If you travel to a state that does not offer civil unions, then your union is not valid there.*

- Low quality arguments: context dependent
  - *But as that’s not likely to occur, we fix what we can.*
Motivation

- Text accessibility
- Automatic summarization
- Argument quality

Sentiment analysis
- More sentiment expressions associated with sentences that are less detailed.

Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• **Goal of SPECITELLER**
  • First step: supervised learning
  • **SPECITELLER**: semi-supervised learning
  • Use case: sentence simplification
  • **SPECITELLER**: hands on
Goal of SPECITELLER

• Predict:
  • Whether a sentence is general or specific.
  • The level of detail in a sentence.
The typical supervised learning setup...

• Gather $N$ training examples
  
  \[ \{(x_1, y_1), \ldots, (x_N, y_N)\} \]
  
  $x$: sentences
  $y$: general/specific (binary), level of specificity (real)

• Transform $x$ into feature vectors

  The/DT astronomer/NN married/VBD the/DT star/NN ./.  
  DT NN VBD ADJ ADV .
  2 2 1 0 0 1

• Learn a model $M$
The typical supervised learning setup...

• Given new examples

\{x_1, \ldots, x_T\}

\(x\): sentences

• Transform \(x\) into feature vectors

The DT astronomer NN married VBD the DT star NN ./

DT NN VBD ADJ ADV .

2 2 1 0 0 1

• Apply \(M\) onto each \(f(x)\) and get output hypothesis

\(h\): general/specific (binary), level of specificity (real)
There is no dedicated, labeled examples for sentence specificity.

Repurpose discourse relation annotation (Louis and Nenkova 2011)

- Small dataset (~3K examples)
- Rich features (syntax, POS, language models, NER)

Goal: *Fast and accurate* prediction of sentence specificity.
Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• Goal of SPECITELLER
• First step: supervised learning
  • SPECITELLER: semi-supervised learning
• Use case: sentence simplification
  • SPECITELLER: hands on
Obtaining training data

• The Penn Discourse Treebank (PDTB)
  • >40K annotations of discourse relations on Wall Street Journal corpus

Explicit

The federal government suspended sales of U.S. savings bonds because Congress hasn’t lifted the ceiling on government debt.

Implicit

I never gamble too far.
I quit after one try, whether I win or lose.

Arg1 and arg2 are adjacent.

ARG2 explains in further detail the events, reasons, behaviors and attitudes mentioned in ARG1.

ARG1 He says he spent $300 million on his art business this year.

ARG2 A week ago, his gallery racked up a $23 million tab at a Sothebys auction in New York buying seven works, including a Picasso.

# instances: 1.4K
First step: supervised learning

- Training data:
  - Sentence pairs having an implicit Instantiation relation
  - ARG 1: general (+1)
  - ARG 2: specific (-1)
  - No adjacency information is preserved.
  - # training examples: 2,796
First step: supervised learning

- 2.8K training examples from Instantiation
  \[ \{(x_1, y_1), ..., (x_N, y_N)\} \]
  - \(x\): sentences
  - \(y\): +1 (general) or -1 (specific)

- Learn a **binary classifier** \(C\) (logistic regression)

- For a new sentence
  - Use \(C\) to determine whether it is general or specific
  - Sentence specificity scores: posterior probabilities of the classifier
Shallow features

- Sentence surface
  - Sentence length
- Numbers, capital letters and symbols
- Average # of characters in the words appearing in the sentence
- Stop words ("about", "either", "only"…)
- Explicit discourse connectives ("so", "because", "however"…)
Shallow features

• Word properties
  • Polar and strongly subjective words ("applaud", "surprise", "criticism"…)
  • Familiarity ("basilisk" vs. "breakfast")
  • Imageary ("convocation" vs. "telephone")
• Min, max and average Inverse Document Frequency (IDF) scores

\[
\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}
\]
Dense word representations

- Word clusters
  - Group similar words together into 100 bins
  - $C(\text{Jeep}) = C(\text{car})$, $C(\text{Jeep}) \neq C(\text{breakfast})$

- Average word embeddings
  - Each word represented by a real valued fixed dimension vector $V(w)$
  - Similar words are close to each other
    - $V(\text{man}) - V(\text{woman}) \sim V(\text{king}) - V(\text{queen})$
Lightweight features

Shallow

• Sentence surface
  • Sentence length
  • Capital letters, non-alphabeticals
  • Average word length
  • Stop words
  • Discourse connectives

• Word properties
  • Polarity, subjectivity [GI, MPQA]
  • Familiarity, imageability [MRC]
  • Min, max and mean IDF scores

Word representations

• Brown clusters
• Averaged word embeddings
First step: supervised learning

- 2.8K training examples from Instantiation
  \[ \{(x_1, y_1), \ldots, (x_N, y_N)\} \]
  - \( x \): sentences
  - \( y \): +1 (general) or -1 (specific)

- Learn a **binary classifier** \( C \) (logistic regression)

- For a new sentence
  - Use \( C \) to determine whether it is general or specific
  - Sentence specificity scores: posterior probabilities of the classifier
First step: supervised learning

- Testing data:
  - Crowdsourced annotation as in [Louis & Nenkova 2011]
  - Binary labels: general or specific
  - Out of context
  - 885 sentences, 5 annotators/sentence
  - % labeled specific: 54.58%
Evaluation

- **Accuracy:** \[
\frac{\{\text{label}==\text{prediction}\}}{\{\text{all test sentences}\}}
\]

- **Precision:** \[
\frac{\{\text{labeled general}\} \cap \{\text{predicted general}\}}{\{\text{predicted general}\}}
\]

- **Recall:** \[
\frac{\{\text{labeled general}\} \cap \{\text{predicted general}\}}{\{\text{labeled general}\}}
\]
Supervised classification results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow</td>
<td>73.56</td>
<td>69.44</td>
<td>74.63</td>
</tr>
<tr>
<td>Word representation</td>
<td>71.64</td>
<td>70.03</td>
<td>65.67</td>
</tr>
<tr>
<td>Louis &amp; Nenkova</td>
<td>77.4</td>
<td>74.4</td>
<td>76.62</td>
</tr>
</tbody>
</table>
Supervised classification results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow</td>
<td>73.56</td>
<td>69.44</td>
<td>74.63</td>
</tr>
<tr>
<td>Word representation</td>
<td>71.64</td>
<td>70.03</td>
<td>65.67</td>
</tr>
<tr>
<td>Louis &amp; Nenkova</td>
<td>77.4</td>
<td>74.4</td>
<td>76.62</td>
</tr>
</tbody>
</table>
Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• Goal of SPECITELLER
• First step: supervised learning
• SPECITELLER: semi-supervised learning
• Use case: sentence simplification
• SPECITELLER: hands on
Cotraining

- Unlabeled data: Gigaword corpus
Cotraining

Final Prediction: mean posterior probability of the classifiers

$P$: most confidently predicted examples

$U \leftarrow U \setminus P$

$L \leftarrow L \cup P$

wordrep
## Cotraining results

# unlabeled examples: 34,000

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shallow</strong></td>
<td>80.45</td>
<td>79.74</td>
<td>76.37</td>
</tr>
<tr>
<td></td>
<td>(+6.89)</td>
<td>(+10.30)</td>
<td>(+1.74)</td>
</tr>
<tr>
<td><strong>Word</strong></td>
<td>79.55</td>
<td>77.42</td>
<td>77.61</td>
</tr>
<tr>
<td><strong>representation</strong></td>
<td>(+7.91)</td>
<td>(+7.39)</td>
<td>(+11.94)</td>
</tr>
<tr>
<td><strong>Final</strong></td>
<td>81.58</td>
<td>80.56</td>
<td>78.36</td>
</tr>
<tr>
<td><strong>Louis &amp;</strong></td>
<td>77.4</td>
<td>74.4</td>
<td>76.62</td>
</tr>
<tr>
<td><strong>Nenkova</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evidence of widespread cheating has surfaced in several states in the last year or so.

Specificity: 0.048

California’s education department suspects adult responsibility for erasures at 40 schools that changed wrong answers to right ones on a statewide test.

Specificity: 0.898

J. J. Li and A. Nenkova. 2015. Fast and Accurate Prediction of Sentence Specificity. AAAI.

Download: http://goo.gl/XiDQwT
Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• Goal of SPECITELLER
• First step: supervised learning
• SPECITELLER: semi-supervised learning
• Use case: sentence simplification
• SPECITELLER: hands on
Motivation

Text accessibility

In consequence, as early as 1726 St. Petersburg was handling 90 percent of Russia’s foreign trade.

- Encyclopedia Britannica

The city soon started to grow rapidly as much of Russia’s foreign trade passed through the city.

- Britannica Elementary
Use case: sentence simplification

• Which sentences need simplification?
• Compare attributes:
  • # words in sentence
  • Automated Readability Index (ARI)
  • Specificity
Use case: simplification targets

- Simple Wikipedia / Wikipedia
  - Automatically aligned ~168K original-simplified pairs
  - ~50K sentences preserved
Use case: simplification targets

- Precision for simplification target identification

- Correlation(specificity, ARI): 0.5975
Outline

• A very high level peek into Natural Language Processing
• Sentence specificity: definition and motivation
• Goal of SPECITELLER
• First step: supervised learning
• SPECITELLER: semi-supervised learning
• Use case: sentence simplification
• SPECITELLER: hands on
• Fast, simple and accurate

Evidence of widespread cheating has surfaced in several states in the last year or so.

California’s education department suspects adult responsibility for erasures at 40 schools that changed wrong answers to right ones on a statewide test.

J. J. Li and A. Nenkova. 2015. Fast and Accurate Prediction of Sentence Specificity. AAAI.

Download: http://goo.gl/XiDQwT
Setting up SPECITELLER

- Environment: Python 2.x
- Install dependencies:
  - Numpy: `pip install numpy`
  - Liblinear:
    - [http://www.csie.ntu.edu.tw/~cjlin/liblinear/](http://www.csie.ntu.edu.tw/~cjlin/liblinear/)
    - If no `liblinear.so.*` file in its python/ directory: `make`
    - Add to `~/.bashrc`: `export PYTHONPATH=$PYTHONPATH:[path_to_liblinear_root]/python/`
  - Download SPECITELLER from [goo.gl/XiDQwT](http://goo.gl/XiDQwT)
Running SPECITELLER

• Script: speciteller.py

• Input: one sentence per line, word tokenized
  
  [California’s education department] ->
  
  [California’s education department]

• Output: specificity scores, each line corresponding to input sentences

• cd [path_to_speciteller]

• python speciteller.py --inputfile testsentences.txt --outputfile predictions.txt