Humans and Computers Working Together to Measure Machine Learning Interpretability

Jordan Boyd-Graber
University of Maryland
2017
Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

Information retrieval - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Information_retrieval

Introduction to Information Retrieval - Stanford University
nlp.stanford.edu/IR-book/

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to
Takeaways

- ML should be interpretable
- We should measure interpretability
- Interpretability should reflect the world we want
The Challenge of Big Data

Every second . . .

- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook
The Challenge of Big Data

Every second . . .
- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

Unstructured

No XML, no semantic web, no annotation. Often just raw text.
The Challenge of Big Data

Every second . . .
- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

**Unstructured**
No XML, no semantic web, no annotation. Often just raw text.

Common task: what’s going on in this dataset.
- Intelligence analysts
- Brand monitoring
- Journalists
- Humanists
The Challenge of Big Data

Every second . . .
- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

Unstructured
No XML, no semantic web, no annotation. Often just raw text.

Common task: what’s going on in this dataset.
- Intelligence analysts
- Brand monitoring
- Journalists
- Humanists

Common solution: unsupervised machine learning (topic models)
What does a Topic Model do?

From an input corpus and number of topics $K \rightarrow$ words to topics

Corpus
- Forget the Bootleg, Just
- Multiplex Heralded As
- The Shape of Cinema,
- A Peaceful Crew Puts
- Stock Trades: A Better Deal
- The three big Internet
- Red Light, Green Light: A
  2-Tone L.E.D. to
  Simplify Screens
What does a Topic Model do?

From an input corpus and number of topics $K \rightarrow \text{words to topics}$

**TOPIC 1**
- computer, technology, system, service, site, phone, internet, machine

**TOPIC 2**
- sell, sale, store, product, business, advertising, market, consumer

**TOPIC 3**
- play, film, movie, theater, production, star, director, stage
Evaluating Topic Models

Reading Tea Leaves: How Humans Interpret Topic Models

Evaluation

Corpus

- Forget the Bootleg, Just Download the Movie Legally
- Multiplex Heralded As Linchpin To Growth
- The Shape of Cinema, Transformed At the Click of a Mouse
- A Peaceful Crew Puts Muppets Where Its Mouth Is
- Stock Trades: A Better Deal
- The three big Internet
- Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens

Model A

Model B

Model C

Held-out Data

- Price War Brews Between Amazon and Wal-Mart
- Sony Ericsson’s Infinite Hope for a Turnaround
- For Search, Murdoch Looks to a Deal With Microsoft

-4.8

-15.16

-23.42
Evaluation

Model A → 0.10
Model B → 0.08
Model C → 0.07

Measures predictive power (likelihood)
But we don’t use topic models for prediction!
Qualitative Evaluation of the Latent Space

<table>
<thead>
<tr>
<th>“segment 1”</th>
<th>“segment 2”</th>
<th>“matrix 1”</th>
<th>“matrix 2”</th>
<th>“line 1”</th>
<th>“line 2”</th>
<th>“power 1”</th>
<th>power 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>imag</td>
<td>speech</td>
<td>robust</td>
<td>manufactur</td>
<td>constraint</td>
<td>alpha</td>
<td>POWER</td>
<td>load</td>
</tr>
<tr>
<td>SEGMENT</td>
<td>speech</td>
<td>MATRIX</td>
<td>cell</td>
<td>LINE</td>
<td>redshift</td>
<td>spectrum</td>
<td>memori</td>
</tr>
<tr>
<td>texture</td>
<td>recogni</td>
<td>eigenvalu</td>
<td>part</td>
<td>match</td>
<td>LINE</td>
<td>omega</td>
<td>vlsi</td>
</tr>
<tr>
<td>color</td>
<td>signal</td>
<td>uncertainti</td>
<td>plane</td>
<td>locat</td>
<td>galaxi</td>
<td>mpc</td>
<td>POWER</td>
</tr>
<tr>
<td>tissue</td>
<td>train</td>
<td>uncertaini</td>
<td>cellular</td>
<td>imag</td>
<td>quasar</td>
<td>hsup</td>
<td>systolic</td>
</tr>
<tr>
<td>brain</td>
<td>hmm</td>
<td>plane</td>
<td>famili</td>
<td>geometr</td>
<td>absorb</td>
<td>larg</td>
<td>input</td>
</tr>
<tr>
<td>slice</td>
<td>source</td>
<td>linear</td>
<td>design</td>
<td>impos</td>
<td>high</td>
<td>high</td>
<td>complex</td>
</tr>
<tr>
<td>cluster</td>
<td>speakerind.</td>
<td>condition</td>
<td>machinepart</td>
<td>segment</td>
<td>ssup</td>
<td>redshift</td>
<td>arrai</td>
</tr>
<tr>
<td>mri</td>
<td>SEGMENT</td>
<td>perturb</td>
<td>format</td>
<td>fundament</td>
<td>densiti</td>
<td>galaxi</td>
<td>present</td>
</tr>
<tr>
<td>volume</td>
<td>sound</td>
<td>root</td>
<td>group</td>
<td>recogn</td>
<td>veloc</td>
<td>standard</td>
<td>model</td>
</tr>
</tbody>
</table>

Figure 3: Eight selected factors from a 128 factor decomposition. The displayed word stems are the 10 most probable words in the class-conditional distribution $P(w|z)$, from top to bottom in descending order.

[Hofmann 1999]
Qualitative Evaluation of the Latent Space

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

[Blei et al. 2003]
Qualitative Evaluation of the Latent Space

Gibbs sampling involves sequentially resampling each $z_l^n$ from its conditional posterior:

$$P(z_l^n = t | w, ... children, demon-
strates the variability of everyday terminology: al-
though the four Romance languages are closely

Table 1:

<table>
<thead>
<tr>
<th>Language</th>
<th>Avg. leng.</th>
<th># docs</th>
<th>types/10k</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>178.689</td>
<td>66497</td>
<td>124.5</td>
</tr>
<tr>
<td>DA</td>
<td>160.153</td>
<td>65245</td>
<td>121.4</td>
</tr>
<tr>
<td>FI</td>
<td>161.293</td>
<td>60822</td>
<td>336.2</td>
</tr>
<tr>
<td>ES</td>
<td>170.536</td>
<td>65929</td>
<td>59.5</td>
</tr>
<tr>
<td>FR</td>
<td>186.742</td>
<td>67430</td>
<td>54.8</td>
</tr>
<tr>
<td>IT</td>
<td>187.451</td>
<td>66035</td>
<td>69.5</td>
</tr>
<tr>
<td>NL</td>
<td>176.114</td>
<td>66952</td>
<td>80.8</td>
</tr>
<tr>
<td>PT</td>
<td>183.410</td>
<td>65718</td>
<td>68.2</td>
</tr>
<tr>
<td>SV</td>
<td>154.605</td>
<td>58011</td>
<td>136.1</td>
</tr>
</tbody>
</table>

The EuroParl corpus consists of parallel texts in eleven western European languages: Danish, Ger-

4 Results on Parallel Text

|Mimno et al. 2009|
Qualitative Evaluation of the Latent Space

(a) Topic labeled as SSL

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssl</td>
<td>0.373722</td>
</tr>
<tr>
<td>expr</td>
<td>0.042501</td>
</tr>
<tr>
<td>init</td>
<td>0.033207</td>
</tr>
<tr>
<td>engine</td>
<td>0.026447</td>
</tr>
<tr>
<td>var</td>
<td>0.022222</td>
</tr>
<tr>
<td>ctx</td>
<td>0.023067</td>
</tr>
<tr>
<td>ptemp</td>
<td>0.017153</td>
</tr>
<tr>
<td>mctx</td>
<td>0.013773</td>
</tr>
<tr>
<td>lookup</td>
<td>0.012083</td>
</tr>
<tr>
<td>modssl</td>
<td>0.011238</td>
</tr>
<tr>
<td>ca</td>
<td>0.009548</td>
</tr>
</tbody>
</table>

(b) Topic labeled as Logging

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>log</td>
<td>0.141733</td>
</tr>
<tr>
<td>request</td>
<td>0.036017</td>
</tr>
<tr>
<td>mod</td>
<td>0.0311</td>
</tr>
<tr>
<td>config</td>
<td>0.029871</td>
</tr>
<tr>
<td>name</td>
<td>0.023725</td>
</tr>
<tr>
<td>headers</td>
<td>0.021266</td>
</tr>
<tr>
<td>autoindex</td>
<td>0.020037</td>
</tr>
<tr>
<td>format</td>
<td>0.017578</td>
</tr>
<tr>
<td>cmd</td>
<td>0.01512</td>
</tr>
<tr>
<td>header</td>
<td>0.013891</td>
</tr>
<tr>
<td>add</td>
<td>0.012661</td>
</tr>
</tbody>
</table>

Table 2: Sample Topics extracted from Apache source code

[Maskeri et al. 2008]
Word Intrusion

- Take the highest probability words from a topic

<table>
<thead>
<tr>
<th>Original Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
</tr>
<tr>
<td>cat</td>
</tr>
<tr>
<td>horse</td>
</tr>
<tr>
<td>pig</td>
</tr>
<tr>
<td>cow</td>
</tr>
</tbody>
</table>
**Word Intrusion**

- Take the highest probability words from a topic

<table>
<thead>
<tr>
<th>Original Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
</tr>
<tr>
<td>cat</td>
</tr>
<tr>
<td>apple</td>
</tr>
<tr>
<td>horse</td>
</tr>
<tr>
<td>pig</td>
</tr>
<tr>
<td>cow</td>
</tr>
</tbody>
</table>

- Intruder: high probability word from another topic
Interpretability and Likelihood

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM</td>
<td>50</td>
</tr>
<tr>
<td>LDA</td>
<td>100</td>
</tr>
<tr>
<td>pLSI</td>
<td>150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Topic Log Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTM</td>
<td>7.32</td>
</tr>
<tr>
<td>LDA</td>
<td>7.52</td>
</tr>
<tr>
<td>pLSI</td>
<td>7.50</td>
</tr>
</tbody>
</table>

Predictive Log Likelihood

0.65
0.70
0.75
0.80
Interpretability and Likelihood

Traditional Evaluation

Predictive Log Likelihood

0.65
0.70
0.75
0.80

New York Times

Wikipedia

Model Precision

Topic Log Odds

Model

CTM
LDA
pLSI

Number of topics

50
100
150

Traditional Evaluation
Interpretability and Likelihood

Traditional Evaluation

Interpretability

Predictive Log Likelihood

0.65

0.70

0.75

0.80

Model Precision
Topic Log Odds

Model

CTM

LDA

pLSI

Number of topics

50

100

150

Model

CTM

LDA

pLSI

Number of topics

50

100

150
Within a model, higher likelihood $\neq$ higher interpretability
What about Supervised Models?
What about Supervised Models?

Features

- viagra
- opportunity
- subscribe
- engineering
- maryland
- algorithm
What about Supervised Models?

Features

viagra
opportunity
subscribe
engineering
maryland
algorithm

Classifier

viagra
opportunity
subscribe
engineering
maryland
algorithm

It’s SPAM!
LIME: Local Interpretable Model-Agnostic Explanations
From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion. If you’d like to know more, send me a note.
What’s an Explanation

$P(\text{dog}) = 0.32$

$P(\text{guitar}) = 0.24$

$P(\text{puppy}) = 0.21$
What makes good Explanation?

- Interpretable: Humans can Understand
- Faithful: Describes Model
- Model Agnostic: Generalize to Many Models
Method

- Complicated model predicts “near” example
- Simple model explains **local variation**
- Explains what complicated model focused on
Is this a good Classifier?
Is this a good Classifier?

Didn't trust the model

"Snow insight"

<table>
<thead>
<tr>
<th>% of subjects (out of 27)</th>
<th>Before explanations</th>
<th>After explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Didn't trust the model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Snow insight&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Improving ML Algorithms

Features

viagra
opportunity
subscribe
engineering
maryland
algorithm

Classifier

It’s SPAM!
Improving ML Algorithms

Features

viagra
opportunity
subscribe
engineering
maryland
algorithm

Classifier

It's SPAM!

Interpret

REPEAT

Evaluate

Turkers don't know about this dataset

Explain

Train

20 newsgroups

test

Hidden religion dataset

predict

train
Improving ML Algorithms

Features

viagra
opportunity
subscribe
engineering
maryland
algorithm

Classifier

It's SPAM!

Human Feedback

Interpret
Improving ML Algorithms

Features

Classifier

It’s SPAM!

Interpret

Turkers don’t know about this dataset

Evaluate

Repeat

Explain

Predict

Train

Test

Hidden religion dataset

Algorithm

Algorithm

Meeting

Bishop

Subscribe

Viagra

Bitcoin

Subscribe

Algorithm

Maryland

Algorithm

Engineering

Opportunity

Subscribe

Viagra

20 newsgroups - Train

20 newsgroups - Test
Improving ML Algorithms

Example #5 of 10
True Class: Atheism

Words that the algorithm considers important.

- Host
- Posting
- NNTP
- to
- New
- Thanks
- anyone
- email
- not
- has

Bar length indicates importance, and color indicates to which topic: Christianity (green) or Atheism (Pink).

Please click on the words (right next to the bars) that you think the algorithm is using incorrectly, because they are not important to distinguish between Atheism and Christianity. They should be red and crossed off after you click them.

Document

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish. This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick
johnchad@triton.unm.edu
or
Improving ML Algorithms

<table>
<thead>
<tr>
<th>Accuracy on hidden set</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>No user input</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 round</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 rounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 rounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Train on 20 newsgroups
turkers clean data
Train on hand-cleaned
20 newsgroups
Train on 20 newsgroups
ENSLAVE HUMANITY
Sample Question

The Swiss-Italian architect Pietro Antonio Solari
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the
The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the
The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the
The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the Mongols in 1476, ending the “Tatar yoke.”
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the Mongols in 1476, ending the “Tatar yoke.” The Grand Duchy headquartered in this city came to an end in 1547 with the ascension of
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the Mongols in 1476, ending the “Tatar yoke.” The Grand Duchy headquartered in this city came to an end in 1547 with the ascension of Ivan IV, who made it his capital. For 10 points, name this city where Ivan III renovated the Kremlin,
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the Mongols in 1476, ending the “Tatar yoke.” The Grand Duchy headquartered in this city came to an end in 1547 with the ascension of Ivan IV, who made it his capital. For 10 points, name this city where Ivan III renovated the Kremlin, the capital of Russia.
Sample Question

The Swiss-Italian architect Pietro Antonio Solari built several fortified towers in this city, which often vied for power with its northern rival Tver. A ruler of this city prevailed in the Great Stand on the Ugra River. A prince from this city was nicknamed for winning a battle on the Don river. Partly because a ruler of this city married Sophia Palaiologina, the niece of the last Byzantine Emperor, this city styled itself the “Third Rome” after the fall of Constantinople. Another prince of this city stopped paying tribute to the Mongols in 1476, ending the “Tatar yoke.” The Grand Duchy headquartered in this city came to an end in 1547 with the ascension of Ivan IV, who made it his capital. For 10 points, name this city where Ivan III renovated the Kremlin, the capital of Russia.

Moscow (Moskva / Muscovy)
This is not Jeopardy

■ Jeopardy: must decide to answer once, after complete question
■ Quiz Bowl: decide after each word
This is **not** Jeopardy

- Jeopardy: must decide to answer *once*, after complete question
- Quiz Bowl: decide after each word
A Neural Network for Factoid Question Answering over Paragraphs


Deep Unordered Composition Rivals Syntactic Methods for Text Classification

Experiment 1

Colby Burnett: $375,000
Ben Ingram: $427,534
Alex Jacobs: $151,802
Kristin Sausville: $95,201

End result: 200-200 tie!
Colby Burnett: $375,000
Ben Ingram: $427,534
Alex Jacobs: $151,802
Kristin Sausville: $95,201

End result: 200-200 tie!
Humans 345-145
Boring Dot Products

\[ a \cdot u = \|a\| \cos \theta \]
Centaur Chess
Measuring Interpretability
Measuring Interpretability

Solo

6
Measuring Interpretability
Improvement through Reinforcement Learning

Visualization

Mauser Rifle
Aristophanes
Battle of Marathon
Battle of Waterloo
Quadratic Equation
Hypotenuse
Raphael

Viz 7.3
Solo 6
Improvement through Reinforcement Learning

Visualization

Exploration

Mauser Rifle
Aristophanes
Battle of Marathon
Battle of Waterloo
Quadratic Equation
Hypotenuse
Raphael

Viz
Solo
7.3
6
Improvement through Reinforcement Learning

Visualization

Exploration

Viz 7.3
Solo 6
Improvement through Reinforcement Learning

Visualization

Exploration

Viz 7.5  Solo 6
Improvement through Reinforcement Learning

Visualization

Exploration

Reward

Viz

Solo

7.5

6
Syntax-based Rewriting for Simultaneous Machine Translation

He He, Alvin Grissom II, Jordan Boyd-Graber, and Hal Daumé III. Empirical Methods in Natural Language Processing, 2015

Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation

He He, Jordan Boyd-Graber, and Hal Daumé III. North American Association for Computational Linguistics, 2016
Simultaneous Interpretation is Hard!

- Exhausting for humans
- Computers not trusted
- Differential strengths
- Same word-by-word characteristic
Takeaways

- ML should be interpretable
- We should measure interpretability
- Interpretability should reflect the world we want
Thanks

Collaborators

NAQT, Hal Daumé III (UMD), Leah Findlater (UMD), Kevin Seppi (BYU), Eric Ringger

Funders

Supporters
ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling

To rescind any unobligated discretionary appropriations returned to the F...
Active learning if time is short

Median (over participants)

Elapsed Time (min)

condition

- No Organization with Active Learning
Better than status quo
Active learning can help topic models
Topic models help users understand the collection
Moral: machines and humans together (if you let them)
After writing "The Theory of Moral..."
References

David M. Blei, Andrew Ng, and Michael Jordan.
2003.
Latent Dirichlet allocation.

David Hall, Daniel Jurafsky, and Christopher D. Manning.
2008.
Studying the history of ideas using topic models.

Thomas Hofmann.
1999.
Probabilistic latent semantic analysis.
In Proceedings of Uncertainty in Artificial Intelligence.

Girish Maskeri, Santonu Sarkar, and Kenneth Heafield.
2008.
Mining business topics in source code using latent dirichlet allocation.
In ISEC.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum.
2009.
Polylingual topic models.