Understanding and adapting statistical models
an exploration in language
Sampling bias is pervasive

- **Object detection:**
  - Objects don't come segmented in the real world
- **Translation:**
  - Can't learn to translate tense or colloquial text from news
- **Med. Diagnosis**
  - Only have data for patients you've seen

---

So my little guy had an EEG today to check for seizure activity. He was a champ. He had to be calm while they glued 26 wires on his head and lay calmly for 20 minutes for testing. I was so impressed how calm he was. Thank goodness for youtube videos. Hoping the results are negative seizures and show good brain activity. If you have a chance we could use some positive thoughts and prayers. Thank you.
Assumptions that underlie most ML

Training data is:

- independent and
- identically distributed

Test data is:

- drawn from the same
distribution as the
training data

Structured prediction:
Learning to make
consistent, interrelated
predictions, where
"success" depends on
the complete
output

NLP uses many features
Assumptions that underlie most ML

- Training data is:
  - independent and
  - identically distributed

- Test data is:
  - drawn from the same distribution as the training data

structured prediction: learning to make consistent, interrelated predictions, where “success” depends on the complete output

NLP uses many features
Language does have many flavors!

Can you guess what domain each of these sentences is drawn from?

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>Many factors contributed to the French and Dutch objections to the proposed EU constitution</td>
</tr>
<tr>
<td>Parliament</td>
<td>Please rise, then, for this minute's silence</td>
</tr>
<tr>
<td>Medical</td>
<td>Latent diabetes mellitus may become manifest during thiazide therapy</td>
</tr>
<tr>
<td>Science</td>
<td>Statistical machine translation is based on sets of text to build a translation model</td>
</tr>
<tr>
<td>Step-mother</td>
<td>I forgot to mention in yesterday's post that I also trimmed an overgrown huge hedge that spams the entire length of the front of my house and is about 3' across.</td>
</tr>
</tbody>
</table>
An example from clinical text

- Wall Street Journal:
  Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

- Clinical narrative:
  LV systolic fn normal with EF 60%.
S4 taxonomy of adaptation effects

- **Seen**: Never seen this word before
  - News to medical: “diabetes mellitus”
- **Sense**: Never seen this word used in this way
  - News to technical: “monitor”
- **Score**: The wrong output is scored higher
  - News to medical: “manifest”
- **Search**: Decoding/search erred
S4 applied to “easy” NLP problems...

Part of Speech Tagging

Inside = Old domain

Outside = New domain

Seen  Sense  Score
S4 applied to “easy” NLP problems...

Part of Speech Tagging

Shallow Parsing

Named Entity Recognition

Inside = Old domain

Outside = New domain

Seen  Sense  Score
Classic learning

Predict: $x \rightarrow y, \quad (x,y) \sim \Pr[x,y]$

Running with Scissors

Title: Horrible book, horrible.

This book was horrible. I read half, suffering from a headache the entire time, and eventually I lit it on fire. 1 less copy in the world. Don't waste your money. I wish I had the time spent reading this book back. It wasted my life.

So the topic of ah the talk today is online learning.
Domain Adaptation

Training

Source

Testing

Target

So the topic of ah the talk today is online learning.

Everything is happening online. Even the slides are produced on-line.
You could even get a Nobel prize!

- James Heckman
  - Nobel prize in economics (2000)
  - Sample selection bias as specification error. Econometrica (1979)
“MONITOR” versus “THE”

News domain:
“MONITOR” is a verb
“THE” is a determiner

Technical domain:
“MONITOR” is a noun
“THE” is a determiner

Key Idea:
Share some features (“the”)
Don't share others (“monitor”)

(and let the learner decide which are which)
Feature Augmentation

Original Features

- We: monitor
  - P: we
  - N: the
  - C: a+

- the traffic
  - W: the
  - P: monitor
  - N: traffic
  - C: a+

Augmented Features

- SW: monitor
  - SP: we
  - SN: the
  - SC: a+

- SW: the traffic
  - SP: monitor
  - SN: traffic
  - SC: a+

Why should this work?

In feature-vector lingo:

\[
\Phi(x) \rightarrow \langle \Phi(x), \Phi(x), 0 \rangle \quad (\text{for source domain})
\]

\[
\Phi(x) \rightarrow \langle \Phi(x), 0, \Phi(x) \rangle \quad (\text{for target domain})
\]
## Results – Error Rates

<table>
<thead>
<tr>
<th>Task</th>
<th>Dom</th>
<th>SrcOnly</th>
<th>TgtOnly</th>
<th>Baseline</th>
<th>Prior</th>
<th>Augment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE-</td>
<td>bn</td>
<td>4.98</td>
<td>2.37</td>
<td>2.11 (pred)</td>
<td>2.06</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>bc</td>
<td>4.54</td>
<td>4.07</td>
<td>3.53 (weight)</td>
<td>3.47</td>
<td>3.47</td>
</tr>
<tr>
<td>NER</td>
<td>nw</td>
<td>4.78</td>
<td>3.71</td>
<td>3.56 (pred)</td>
<td>3.68</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>wl</td>
<td>2.45</td>
<td>2.45</td>
<td>2.12 (all)</td>
<td>2.41</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>un</td>
<td>3.67</td>
<td>2.46</td>
<td>2.10 (linint)</td>
<td>2.03</td>
<td>1.91</td>
</tr>
<tr>
<td></td>
<td>cts</td>
<td>2.08</td>
<td>0.46</td>
<td>0.40 (all)</td>
<td>0.34</td>
<td>0.32</td>
</tr>
<tr>
<td>CoNLL</td>
<td>tgt</td>
<td>2.49</td>
<td>2.95</td>
<td>1.75 (wgt/li)</td>
<td>1.89</td>
<td>1.76</td>
</tr>
<tr>
<td>PubMed</td>
<td>tgt</td>
<td>12.02</td>
<td>4.15</td>
<td>3.95 (linint)</td>
<td>3.99</td>
<td>3.61</td>
</tr>
<tr>
<td>CNN</td>
<td>tgt</td>
<td>10.29</td>
<td>3.82</td>
<td>3.44 (linint)</td>
<td>3.35</td>
<td>3.37</td>
</tr>
<tr>
<td></td>
<td>wsj</td>
<td>6.63</td>
<td>4.35</td>
<td>4.30 (weight)</td>
<td>4.27</td>
<td>4.11</td>
</tr>
<tr>
<td></td>
<td>swbd3</td>
<td>15.90</td>
<td>4.15</td>
<td>4.09 (linint)</td>
<td>3.60</td>
<td>3.51</td>
</tr>
<tr>
<td>Tree</td>
<td>br-cf</td>
<td>5.16</td>
<td>6.27</td>
<td>4.72 (linint)</td>
<td>5.22</td>
<td>5.15</td>
</tr>
<tr>
<td>bank-</td>
<td>br-cg</td>
<td>4.32</td>
<td>5.36</td>
<td>4.15 (all)</td>
<td>4.25</td>
<td>4.90</td>
</tr>
<tr>
<td>Chunk</td>
<td>br-ck</td>
<td>5.05</td>
<td>6.32</td>
<td>5.01 (prd/li)</td>
<td>5.27</td>
<td>5.41</td>
</tr>
<tr>
<td></td>
<td>br-cl</td>
<td>5.66</td>
<td>6.60</td>
<td>5.39 (wgt/prd)</td>
<td>5.99</td>
<td>5.73</td>
</tr>
<tr>
<td></td>
<td>br-cm</td>
<td>3.57</td>
<td>6.59</td>
<td>3.11 (all)</td>
<td>4.08</td>
<td>4.89</td>
</tr>
<tr>
<td></td>
<td>br-cn</td>
<td>4.60</td>
<td>5.56</td>
<td>4.19 (prd/li)</td>
<td>4.48</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>br-cp</td>
<td>4.82</td>
<td>5.62</td>
<td>4.55 (wgt/prd/li)</td>
<td>4.87</td>
<td>4.78</td>
</tr>
<tr>
<td></td>
<td>br-cr</td>
<td>5.78</td>
<td>9.13</td>
<td>5.15 (linint)</td>
<td>6.71</td>
<td>6.30</td>
</tr>
<tr>
<td>Treebank- brown</td>
<td>6.35</td>
<td>5.75</td>
<td>4.72 (linint)</td>
<td>4.72</td>
<td>4.65</td>
<td></td>
</tr>
</tbody>
</table>
Named Entity Rec.: /bush/

- Person
- Geo-political entity
- Organization
- Location

- General
- BC-news
- Conversations
- Newswire
- Weblogs
- Usenet
- Telephone
Named Entity Rec.: p=/the/
Some Theory

Can bound expected target error:

$$\begin{align*}
\epsilon_t & \leq \frac{1}{2} \left( \hat{\epsilon}_s + \hat{\epsilon}_t \right) + O(\text{complexity}) \\
& + \left( \frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}} \right) O \left( \frac{1}{\delta} \right) + O\left( \text{disc}_H(S,T) \right)
\end{align*}$$

# source examples

# target examples
Semi-supervised Extension

For labeled data:

\[(y, x) \rightarrow (y, \langle x, x, 0 \rangle) \quad \text{(for source domain)}\]
\[(y, x) \rightarrow (y, \langle x, 0, x \rangle) \quad \text{(for target domain)}\]

What about unlabeled data?

\[(x) \rightarrow \{ (+1, \langle 0, x, -x \rangle), (-1, \langle 0, x, -x \rangle) \} \]

Encourage agreement:

\[\left[ h_t(x) = h_s(x) \right] \iff \left[ w_t \circ x - w_s \circ x = 0 \right] \]
Semi-supervised Bounds

- The complexity drops from $O(\text{tr}(B))$ to:
  $$O(\text{tr}(B) - \text{tr}(E(I + kF)^{-1}E'))$$

- Where:
  $$K = \begin{bmatrix} A & C & D \\ C' & B & E \\ D' & E' & F \end{bmatrix}$$

  - Source data
  - Target data
  - Unlab data

  $k$ depends on regularizer

- Proof based on recent results from multi-view learning
Semi-supervised Experiments

(FA++ only gets 50% labeled data!)
Is this a problem for harder tasks?
Translating across domains is hard

<table>
<thead>
<tr>
<th>Old Domain (Parliament)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
</tr>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td><strong>System</strong></td>
</tr>
</tbody>
</table>

**Key Question: What went wrong?**
Domain Shift Setting

**Old domain:** Hansard parliamentary proceedings
# New Domain Datasets

<table>
<thead>
<tr>
<th>Domain</th>
<th>News</th>
<th>Medical</th>
<th>Science</th>
<th>Subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Old domain</strong></td>
<td>8,000k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French (fr)</td>
<td>162m</td>
<td>155m</td>
<td>174m</td>
<td>162m</td>
</tr>
<tr>
<td>English (en)</td>
<td>145m</td>
<td>118m</td>
<td>145m</td>
<td>145m</td>
</tr>
<tr>
<td>French (fr)</td>
<td>3m</td>
<td>7m</td>
<td>2m</td>
<td>192k</td>
</tr>
<tr>
<td>English (en)</td>
<td>3m</td>
<td>4m</td>
<td>4m</td>
<td>187k</td>
</tr>
</tbody>
</table>

Adapting statistical models

---

Hal Daumé III, me@hal3.name
Adapting statistical models

News

OLD + Seen + Sense + Score Mixed

Medical

OLD + Seen + Sense + Score Mixed

Science

OLD + Seen + Sense + Score Mixed

Subtitles

OLD + Seen + Sense + Score Mixed
Simultaneously solving seen+sense

• Idea:
  • We have good knowledge of translations in the old domain
  • We have good knowledge of raw word frequencies in a new domain in each language individually
  • Can we “nudge” the translation probabilities to match these raw frequencies

• Assumptions:
  • Old domain parallel data
  • New domain comparable data
Marginal matching

(a) OLD-Domain Joint

(b) NEW-Domain Marginals

(c) Inferred NEW-Domain Joint

Matched Marginals
Marginal matching details

\( \Omega(p) \) : regularization term
\( f(p) \) : edit distance penalty

\[
p^{new} = \arg \min_p \| p - p^{old} \|_1 + \Omega(p) + f(p)
\]

subject to:

\[
\sum_{s,t} p(s, t) = 1, \quad p(s, t) \geq 0
\]

\[
\sum_s p(s, t) = q(t), \quad \sum_t p(s, t) = q(s)
\]
### Example Learned Translations

<table>
<thead>
<tr>
<th>French</th>
<th>Correct</th>
<th>Learned Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>cisaillement</td>
<td>shear</td>
<td>viscous crack shear</td>
</tr>
<tr>
<td>chromosomes</td>
<td>chromosomes</td>
<td>chromosomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>chromosome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>chromosomal</td>
</tr>
<tr>
<td>caractérisation</td>
<td>characterization</td>
<td>characterization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>characteristic Π</td>
</tr>
<tr>
<td>araignées</td>
<td>spiders</td>
<td>spiders ant spider</td>
</tr>
<tr>
<td>tiges</td>
<td>stems</td>
<td>usda centimeters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>flowering</td>
</tr>
</tbody>
</table>
Adapting statistical models

BLEU Scores

![Bar chart showing BLEU scores for Baseline.](chart.png)
Discussion

• Adaptation effects are real and lead to significant degradation in system performance
• Simple techniques work well in theory + practice
• Understanding source of errors helps address them
• Marginal matching addresses new S4 issues
• How can we combat sampling bias when we know the bias? When we do not?

Thanks! Questions?
### Some example “seen” errors

<table>
<thead>
<tr>
<th>Domain</th>
<th>Most frequent OOV Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>News (17%)</strong></td>
<td>behavior, favor, neighbors, fueled</td>
</tr>
<tr>
<td></td>
<td>neighboring, abe, zhao, WWII</td>
</tr>
<tr>
<td></td>
<td>favorable, phelps, favored</td>
</tr>
<tr>
<td></td>
<td>favorite, zhao, ahmedinejad</td>
</tr>
<tr>
<td></td>
<td>favorable, phelps, bernanke</td>
</tr>
<tr>
<td><strong>Medical (49%)</strong></td>
<td>renal, hepatic, subcutaneous, irbesartan</td>
</tr>
<tr>
<td></td>
<td>ribavirin, olanzapine, serum</td>
</tr>
<tr>
<td></td>
<td>dl, eine, sie, patienten</td>
</tr>
<tr>
<td></td>
<td>ritonavir, hydrochlorothiazide, pharmacokinetics</td>
</tr>
<tr>
<td></td>
<td>hydrochlorothiazide, erythropoietin</td>
</tr>
<tr>
<td><strong>Movies (44%)</strong></td>
<td>gonna, yeah, mom</td>
</tr>
<tr>
<td></td>
<td>b****, daddy, s***</td>
</tr>
<tr>
<td></td>
<td>f*****g, gotta, wanna</td>
</tr>
<tr>
<td></td>
<td>uh, namely, bye</td>
</tr>
<tr>
<td></td>
<td>hi, later</td>
</tr>
<tr>
<td></td>
<td>f***, bye</td>
</tr>
<tr>
<td></td>
<td>namely, bye</td>
</tr>
</tbody>
</table>

Intrinsic evaluation: MRR

- Ranked document pairs: learning from most science-like first
- Relative gain expected to slow, documents less and less science-y
Aspects of computational 2nd LL

- Very specific linguistic variants
  - Number, case, agreement, etc.
  - *Not enough* to get the majority case

- Focus on subtle visual differences
Aspects of computational 2ndLL

- AI-style reasoning & one-shot learning

- “It's learnable” proof of concept:
What is NLP?

- Fundamental goal: deep understanding of text
  - Not just string processing or keyword matching
- End systems that we want to build
  - Simple: Spelling correction, text categorization, etc.
  - Complex: Speech recognition, machine translation, information extraction, dialog interfaces, question answering
  - Unknown: human-level comprehension (more than just NLP?)
Macro-Analysis: TETRA

Old-Domain Phrase Table ("Old")

Both-Domains Phrase Table ("Mixed")
Macro-Analysis: TETRA

Measuring **SEEN**

![Diagram](image)

Add all phrase pairs with previously unseen F side

<table>
<thead>
<tr>
<th>Phrase Pair</th>
<th>Similarity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>voie(s) route(s)</td>
<td>0.31</td>
</tr>
<tr>
<td>mode fashion</td>
<td>0.21</td>
</tr>
<tr>
<td>mode method</td>
<td>0.41</td>
</tr>
<tr>
<td>administration directors</td>
<td>0.23</td>
</tr>
<tr>
<td>administration administration</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Macro-Analysis: TETRA

Measuring SENSE

Add all phrase pairs with previously seen F side, but unseen translation

<table>
<thead>
<tr>
<th>Phrase Pairs</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>voie(s) - route(s)</td>
<td>0.31</td>
</tr>
<tr>
<td>mode - fashion</td>
<td>0.21</td>
</tr>
<tr>
<td>mode - method</td>
<td>0.41</td>
</tr>
<tr>
<td>administration - directors</td>
<td>0.23</td>
</tr>
<tr>
<td>administration - administration</td>
<td>0.15</td>
</tr>
</tbody>
</table>
**Macro-Analysis: TETRA**

**Measuring SCORE**

Use *Mixed* scores on set of *Old* phrase pairs

<table>
<thead>
<tr>
<th>Phrase Pair</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>voie(s)</td>
<td>0.31</td>
</tr>
<tr>
<td>route(s)</td>
<td></td>
</tr>
<tr>
<td>mode</td>
<td>0.21</td>
</tr>
<tr>
<td>fashion</td>
<td></td>
</tr>
<tr>
<td>mode</td>
<td>0.41</td>
</tr>
<tr>
<td>method</td>
<td></td>
</tr>
<tr>
<td>administration</td>
<td>0.23</td>
</tr>
<tr>
<td>directors</td>
<td></td>
</tr>
<tr>
<td>administration</td>
<td>0.15</td>
</tr>
<tr>
<td>administration</td>
<td></td>
</tr>
</tbody>
</table>
Senses are domain/language specific

French
- courir
- éxécuter

English
- run
- virus
- window

Japanese
- 走る
- 病原体
- 窓
- ウィルス
- ウィンドウ
Measuring SEEN effects

Add all phrase pairs with previously unseen F side
Measuring SENSE effects

Add all phrase pairs with previously seen F side, but unseen translation
Measuring SCORE effects

Add all phrase pairs, period (and keep new domain scores)
Macro-analysis of S4 effects

- Evaluation using BLEU

<table>
<thead>
<tr>
<th></th>
<th>News</th>
<th>Medical</th>
<th>Science</th>
<th>Subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>+0.3%</td>
<td>+8.1%</td>
<td>+6.1%</td>
<td>+5.7%</td>
</tr>
<tr>
<td>Sense</td>
<td>+0.6%</td>
<td>+6.6%</td>
<td>+4.4%</td>
<td>+8.7%</td>
</tr>
<tr>
<td>Score</td>
<td>+0.6%</td>
<td>+4.5%</td>
<td>+9.9%</td>
<td>+8.4%</td>
</tr>
<tr>
<td>Search</td>
<td>+0.0%</td>
<td>+0.0%</td>
<td>+0.0%</td>
<td>+0.0%</td>
</tr>
</tbody>
</table>

- Hansard: 8m sents 161m fr-tokens
- News: 135k sents 3.9m fr-tokens
- Medical: 472k sents 6.5m fr-tokens
- Science: 139k sents 4.3m fr-tokens
- Subtitles: 19m sents 155m fr-tokens
Micro-analysis of S4 effects

Output (en): initial dose : 

Input (fr): dose initiale : 

Ref (en): starting dose : 

Correct  Seen-freebie 
Seen-error  Sense-error  Score-error
Micro-analysis of S4 effects

Output (en): initial dose :

Input (fr): dose initiale :

Ref (en): starting dose :

Correct  Seen-freebie  Sense-error  Score-error
Errors found by micro-analysis

![Bar chart showing errors in different categories: News, Medical, Science, Subtitles. The chart compares Seen %, Sense %, and Score % for each category.]

Adapting statistical models

Hal Daumé III, me@hal3.name
Senses are domain/language specific

- courir to run
- éxécuter to run
- virus to virus
- fenêtre to window
- 走る to run
- 病原体 to virus
- ウィルス to virus
- 窓 to window
- ウィンドウ to window
Case 1: No NEW domain parallel data

- Common situation
  - Lots of data in some OLD domain (e.g., government documents)
  - Need to translate many NEW domain documents
- Acquiring additional NEW domain translations is critical!
- Lots of past work in term mining
  - Distributional similarity [Rapp 1996]
  - Orthographic similarity
  - Temporal similarity
Marginal matching for “sense” errors

**Given:**
- Joint $p(x, y)$ in old domain
- Marginals $q(x)$ and $q(y)$ in the new domain

**Recover:**
- Joint $q(x, y)$ in new domain

We formulate as a L1-regularized linear program

Easier: many $q(x)$ and $q(y)$s
Additional features

• Sparsity: # of non-zero entries should be small

• Distributional: document co-occurrence \( \geq \) translation pair

• Spelling: Low edit dist \( \geq \) translation pair

• Frequency: Rare words align to rare words; common words align to common words

c-caractérisation
characterization

<table>
<thead>
<tr>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>le</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>spiders</td>
<td>araignées</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Intrinsic evaluation: Mean Reciprocal Rank (MRR)

- Ranked Wikipedia document pairs: learning from most science-like first
- Decreasing benefits after ~50,000 document pairs
- Relative gain expected to slow, as documents are less and less science-y
# Example learned translations (Science)

<table>
<thead>
<tr>
<th>French</th>
<th>Correct English</th>
<th>Learned Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>cisaillement</td>
<td>shear</td>
<td>viscous</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crack</td>
</tr>
<tr>
<td></td>
<td></td>
<td>shear</td>
</tr>
<tr>
<td>chromosomes</td>
<td>chromosomes</td>
<td>chromosomes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>chromosome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>chromosomal</td>
</tr>
<tr>
<td>caractérisation</td>
<td>characterization</td>
<td>characterization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>characteristic</td>
</tr>
<tr>
<td>araignées</td>
<td>spiders</td>
<td>spiders</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>spider</td>
</tr>
<tr>
<td>tiges</td>
<td>stems</td>
<td>usda</td>
</tr>
<tr>
<td></td>
<td></td>
<td>centimeters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>flowering</td>
</tr>
</tbody>
</table>
Case 2: Add NEW domain parallel data

• Say we have a NEW domain translation memory
• How can we leverage our OLD domain to achieve the greatest benefit?
Initial adaptation baselines

1. Do nothing
   - OLD

2. Ignore old data
   - OLD
   - NEW

3. Concatenate the two
   - OLD
   - NEW
Use both models (log-linear mixture)

Baseline:
\[ \alpha_1 \log p(f|e) + \alpha_2 \log p(e) + ... \]

New:
\[ \alpha_{1OLD} \log p_{OLD}(f|e) + \alpha_{1NEW} \log p_{NEW}(f|e) + \alpha_2 \log p(e) + ... \]
Combine models (linear mixture)

- **Baseline:**
  \[ p(f|e) = \frac{c(f,e)}{c(e)} \]

- **New** — mix with \( \lambda \) picked on dev set:
  \[ p(f|e) = \lambda \frac{c_{old}(f,e)}{c_{old}(e)} + (1 - \lambda) \frac{c_{old}(f,e)}{c_{old}(e)} \]
### BLEU results

<table>
<thead>
<tr>
<th></th>
<th>OLD</th>
<th>NEW</th>
<th>OLD+ NEW</th>
<th>Use both models</th>
<th>Combine models</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>23.8</td>
<td>21.7</td>
<td>22.0</td>
<td>16.4</td>
<td>21.4</td>
</tr>
<tr>
<td>EMEA</td>
<td>28.7</td>
<td>34.8</td>
<td>34.8</td>
<td>32.9</td>
<td>36.6</td>
</tr>
<tr>
<td>Science</td>
<td>26.1</td>
<td>32.3</td>
<td>27.5</td>
<td>30.9</td>
<td>32.2</td>
</tr>
<tr>
<td>Subtitles</td>
<td>15.1</td>
<td>20.6</td>
<td>20.5</td>
<td>18.4</td>
<td>18.5</td>
</tr>
</tbody>
</table>
Next steps

• These mixtures are simple but coarse

• More fine-grained approaches:
  – Data selection: pick OLD data most like NEW
  – Data reweighting: use fractional counts on OLD data; greater weight to sentence pairs more like NEW
  – Can reweight at the word or phrase level rather than sentence pair [Foster et al., 2010]

• Similar in spirit to statistical domain adaptation
  – but existing machine learning algorithms can’t be applied
  – because SMT is not a classification task
Phrase Sense Disambiguation (PSD)

Proposed solution: **Phrase Sense Disambiguation**

[Carpuat & Wu 2007]

- Incorporate **context** in lexical choice
  - Yields $P(e|f, \text{context})$ features for phrase pairs
  - Unlike usual $P(e|f)$ relative frequencies

- Turns phrase translation into **discriminative classification**
  - Just like standard machine learning tasks

Why PSD for domain adaptation?

- Disambiguating English senses of *rapport*
  - report: Il a rédigé un *rapport*.
  - relationship: Quel est le *rapport*?
  - ratio: le *rapport* longueur / largeur
  - balance: le *rapport* bénéfique / risque

Source context can prevent translation errors when shifting domain

- P(elf) in Hansard
- Occurs in new domains but not as often as in Hansard!
- Highest P(elf) in Science!
- New sense in medical domain!
Phrase Sense Disambiguation

- PSD = phrase translation as classification
- PSD at test time
  - use context to predict correct English translation of French phrase
  - local lexical and POS context, global sentence and document context
- PSD at train time
  - extract French phrases with English translations from word alignment
  - throw into off-the-shelf classifier + adaptation techniques

[Blitzer & Daumé 2010]
Domain adaptation in PSD

- Train a classifier over OLD and NEW data

Baseline:

\[ \alpha \log p(f|e) + \beta \log p(e) + \ldots \]

New:

\[ \alpha_{OLD} \log p_{OLD}(f|e) + \alpha_{NEW} \log p_{NEW}(f|e) + \beta \log p(e) + \ldots \]
Feature augmentation

Original features

OLD

NEW

\[
\varphi_{e,f} \mapsto \langle \varphi_{e,f}, 0, \varphi_{e,f} \rangle
\]

\{rédigé …\} rapport

\{rédigé …\} rapport \rightarrow report

\{aucun …\} rapport

\{rédigé …\} rapport \rightarrow report

rapport \{… valeurs\} \rightarrow ratio
Domain adaptation results: Science

<table>
<thead>
<tr>
<th>Domain</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>75.6</td>
</tr>
<tr>
<td>OLD+Science</td>
<td>75.9</td>
</tr>
<tr>
<td>OLD+Science: FEDA</td>
<td>76.0</td>
</tr>
</tbody>
</table>
PSD in Moses: VW-Moses integration

• First general purpose classifier in Moses

• Tight integration
  • Can be built and run out-of-the-box, extended with new features, etc
  • Fast!
    • 180% run time of standard Moses, fully parallelized in training (multiple processes) and decoding (multithreading)
Other areas of investigation

PSD for Hierarchical phrase-based translation

Discovering latent topics from parallel data

Spotting new senses: determining when a source word gains a new sense (needs a new translation)
Spotting New Senses

- **Binary classification problem:**
  - +ve: French token has previously unseen sense
  - -ve: French token is used in a known way
- **Gold standard as byproduct of S4 analysis**
- **Many features considered**
  - Frequency of words/translations in each domain
  - Language model perplexities across domains
  - Topic model “mismatches”
  - Marginal matching features
  - Translation “flow” impedance
Adapting statistical models

Experimental Results

Selected features:
- **EMEA**: ppl || matchm flow || matchm topics flow
- **Science**: ppl || matchm ppl || matchm topics ppl
- **Subs**: topics || matchm topics || matchm topics flow
Discussion

• Introduced taxonomy and measurement tools for adaptation effects in MT

• “Score” errors – target of prior work – only a part of what goes wrong

• Marginal matching introduced as a model for addressing all S4 issues simultaneously: +2.4 BLEU

• Data and outputs released for you to use (both in MT and as a stand-alone lexical selection task)

• Feature-rich approaches integrated into Moses via VW library, applied to adaptation

• Range of other problems to work on: identifying new senses, cross-domain topic models, etc.)
Domain Shift Setting

**Old domain:** Hansard parliamentary proceedings
# New Domain Datasets

<table>
<thead>
<tr>
<th>Domain</th>
<th>News</th>
<th>Medical</th>
<th>Science</th>
<th>Subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong># Phrases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tokens</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Types</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>en</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>fr</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **News**: 135k tokens, 4m types, 63k phrases
- **Medical**: 472k tokens, 7m types, 35k phrases
- **Science**: 139k tokens, 4m types, 118k phrases
- **Subtitles**: 19m tokens, 155m types, 362k phrases

---

Adapting statistical models

Hal Daumé III, me@hal3.name
Two methods for measuring adaptation effects

Macro Analysis: corpus-level analysis using BLEU

TETRA: table enhancement for translation analysis

Micro Analysis: word-level analysis using word alignments

WADE: word alignment driven evaluation

[Irvine et al. TACL 2013]
SenseSpotting

**Why?** MT performance across domains degrades due to lexical choice errors

**What?** New task to identify word occurrences (tokens) that gain a new sense in new domains

**How?** Automatic annotation from parallel text + supervised learning
SenseSpotting Task Definition

Old domain translation lexicon

| rapport || report || 0.8  
| rapport || connection || 0.1 
| rapport || study || 0.05  
| rapport || relationship || 0.05 

New domain sentences

ces données sont basées sur le rapport d’ étude clinique

le rapport cholestérol total / hdlc est resté stable
Key aspects of SenseSpotting

Sense inventory is defined by the MT lexicon
[Chan et al. 2007, Carpuat & Wu, 2007, inter alia]

New Senses are detected at the token-level
SenseSpotting is related to...

**novel sense detection**, but SenseSpotting...

- operates at the token-level
- is specific to domain-adaptation

**most frequent sense shift detection**, but SenseSpotting

- considers *all* previously seen senses


[McCarthy et al. 2004, 2007; Erk 2006; Chan & Ng, 2007]
SenseSpotting is related to...

word sense disambiguation, but SenseSpotting expects sense inventory to grow...

[e.g., Sinha et al. 2010, Lefever & Hoste 2010]

... similar to word sense induction, but we have old senses defined

[e.g., Agirre and Soroa 2007]
Data Requirements

Hansard

| F | E |

Medical

| F | E |

Extract candidate terms and useful statistics

Train model parameters
Generating Annotated Data

Hansard

F  E

Medical

F  E
Generating Annotated Data

Hansard

Medical

F | E

F | E

F | E

Deduce the ratio of h/m.

No connection!

Il a rédigé un rapport.

He wrote a report.

Aucun rapport!
Data Requirements

Hansard

F | E

Medica

F | E

Medical

F | E

Adapting statistical models

Extract candidate terms and statistics

Extract useful statistics

Train model parameters
Adapting statistical models

Data Requirements

Hansard

F | E

Medicai

F | E

Medical

F | E

we can remove this requirement using a surrogate new domain
Classification set-up

Logistic regression model trained with vw

- L1 or L2 regularized based on tuning data

16-fold cross validation at the type level

- Never test on type seen in training!
- E.g., train on “mode”, “administration”; test on “rapport”

Evaluation metric: AUC

- area under the ROC curve
<table>
<thead>
<tr>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>ramenez</td>
<td>report</td>
</tr>
<tr>
<td>recevez</td>
<td>researched</td>
</tr>
<tr>
<td>recouvrement</td>
<td>refuge</td>
</tr>
<tr>
<td>rendez</td>
<td>rends</td>
</tr>
<tr>
<td>rigidité</td>
<td>rigueur</td>
</tr>
<tr>
<td>scalaire</td>
<td>sebastian</td>
</tr>
<tr>
<td>sorties</td>
<td>souches</td>
</tr>
<tr>
<td>stériles</td>
<td>subissant</td>
</tr>
<tr>
<td>ramification</td>
<td>report</td>
</tr>
<tr>
<td>recherché</td>
<td>rechercher</td>
</tr>
<tr>
<td>réflexes</td>
<td>refuge</td>
</tr>
<tr>
<td>rends</td>
<td>rendu</td>
</tr>
<tr>
<td>rompre</td>
<td>rond</td>
</tr>
<tr>
<td>service</td>
<td>signalant</td>
</tr>
<tr>
<td>souhaitée</td>
<td>soulèvement</td>
</tr>
<tr>
<td>substituant</td>
<td>sui</td>
</tr>
</tbody>
</table>
New Sense Indicators

New senses alter corpus-level word frequency

New senses alter document-level context
  - topic distribution

New senses alter local context
  - n-gram language model
  - distributional similarity
  - context-dependent translation model
SenseSpotting Results

Area Under the ROC Curve (cross-validation)

Medical
(52% positive, 35k tokens)

Subtitles
(43% positive, 23k tokens)

Science
(24% positive, 8k tokens)

All Features
Only Token
Only Type
Random
Constant

(43% positive, 23k tokens)
Adapting statistical models

Indicators to reach peak performance

- New senses alter corpus-level word frequency
- New senses alter document-level context
  - topic distribution
- New senses alter local context
  - n-gram language model
  - distributional similarity
  - context-dependent translation model

Computed at both type and token levels
SenseSpotting summary

new task motivated by cross-domain machine translation errors
free *token*-level annotation from parallel text
minimal new domain parallel text required
AUC as high as 80%
on word types *never seen* during training
requires both type and token level indicators
Adapting statistical models

Hal Daumé III, me@hal3.name

News
Medical
Subtitles
Science

OLD + Seen + Sense + Score Mixed

BLEU

35 30 25 20 15

OLD + Seen + Sense + Score Mixed
Adapting statistical models

Hal Daumé III, me@hal3.name

BLEU

OLD + Seen + Sense + Score
OLD + Seen + Sense + Score Mixed

+0% +2% +8% +28%
+6% +23% +6% +22%
+1% +1% +7%
+2% +4% +9% +5%
Adapting statistical models

BLEU

OLD + Seen + Sense + Score

+0% +1% +2%

+6% +4% +23%

+6% +9% +22%

+8% +7% +28%
Adapting statistical models

News
Medical
Subtitles
Science

OLD + Seen + Sense + Score Mixed

BLEU
15
20
25
30

+0% +1% +1% +2%
+8% +7% +5%
+6% +4% +10%
+6% +9% +8%
+28%
Adapting statistical models

Hal Daumé III, me@hal3.name
Adapting statistical models

Hal Daumé III, me@hal3.name

BLEU

OLD +Seen +Sense +Score OLD +Seen +Sense +Score Mixed

+0% +1% +2%
+8% +7%
+23%
+6% +9%
+28%
+22%
Gadovist
S⁴ ontology of adaptation effects

- **Seen:** Never seen this word before
  - News to medical: “diabetes mellitus”
- **Sense:** Never seen this word used in this way
  - News to technical: “monitor”
- **Score:** The wrong output is scored higher
  - News to medical: “manifest”
- **Search:** Decoding/search erred (ignored)

![Diagram showing the ontology of adaptation effects with categories for seen, sense, score, and search.](image)
Adaptation effects in MT

- **Quick observations:**
  - New D language model helps (10%-63% improvement)
  - Tuning on new D data helps (10%-90% improvement)
  - Weighting new D data helps (4%-150% improvement)

- **Identifying errors in MT (w/o parallel new D data):**
  - **Seen:** old-only model + unseen input word pairs
  - **Sense:** old-only model + seen input/unseen output pairs
  - **Score:** intersect old and mixed model, score from old

<table>
<thead>
<tr>
<th></th>
<th>News</th>
<th>Medical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>Little effect</td>
<td>~ 40% of error</td>
</tr>
<tr>
<td>Sense</td>
<td>Little effect</td>
<td>~ 40% of error</td>
</tr>
<tr>
<td>Score</td>
<td>~ 90% of error</td>
<td>~ 20% of error</td>
</tr>
</tbody>
</table>

(As measured by Bleu score)

Consistent in:
- movie subtitles
- scientific pubs
- PHP tech docs
Translating across domains is hard

<table>
<thead>
<tr>
<th>Dom</th>
<th>Most frequent OOV Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>behavior, favor, neighbors, centered, favorite</td>
</tr>
<tr>
<td></td>
<td>fueled, preferred,BERNANKE, AHMEDINEJAD</td>
</tr>
<tr>
<td>Medical</td>
<td>renal, hepatic, subcutaneous, IRBESARTAN</td>
</tr>
<tr>
<td></td>
<td>ribavirin, olanzapine, serum, PATIENTEN</td>
</tr>
<tr>
<td></td>
<td>dl, eine, hydrochlorothiazide, PHARMACOKINETICS</td>
</tr>
<tr>
<td>Movies</td>
<td>gonna, yeah, mom, hi</td>
</tr>
<tr>
<td></td>
<td>b***, daddy, s***, later</td>
</tr>
<tr>
<td></td>
<td>f**<em><strong>g, f</strong></em>, gotta, wanna</td>
</tr>
<tr>
<td></td>
<td>uh, namely, bye, bye</td>
</tr>
<tr>
<td></td>
<td>f***, bye, bye, bye</td>
</tr>
<tr>
<td></td>
<td>uw, uw, uw, uw</td>
</tr>
</tbody>
</table>

Adapting statistical models

Hal Daumé III, me@hal3.name
Dictionary mining for “seen” errors

- Find frequent terms in new domain
- Use those that exist in old domain as “training data”
- Extract context and orthographic features
- Find low-dimensional subspace on training data (CCA)

Pair input words with <=5 output words
Add four features to SMT model
Rerun parameter tuning

<table>
<thead>
<tr>
<th></th>
<th>DE</th>
<th>FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>+0.80</td>
<td>+0.36</td>
</tr>
<tr>
<td>Emea</td>
<td>+1.44</td>
<td>+1.51</td>
</tr>
<tr>
<td>Subs</td>
<td>+0.13</td>
<td>+0.61</td>
</tr>
<tr>
<td>PHP</td>
<td>+0.28</td>
<td>+0.68</td>
</tr>
</tbody>
</table>

(Bleu score improvements)
Marginal matching for “sense” errors

**Given:**
- A joint $p(x,y)$ in the old domain
- Marginals $q(x)$ and $q(y)$ in the new domain

**Recover:**
- Joint $q(x,y)$ in the new domain

We formulate as a $L_1$-regularized linear program

Easier: many $q(x)$ and $q(y)$s
Intrinsic evaluation: MRR

• Ranked document pairs: learning from most science-like first

• Work in progress: increasing document pairs

• Relative gain expected to slow, documents less and less science-y
## Example Learned Translations

<table>
<thead>
<tr>
<th>French</th>
<th>Correct</th>
<th>Learned Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>cisaillement</td>
<td>shear</td>
<td>viscous crack shear</td>
</tr>
<tr>
<td>chromosomes</td>
<td>chromosomes</td>
<td>chromosomes</td>
</tr>
<tr>
<td>caractérisation</td>
<td>characterization</td>
<td>characterization</td>
</tr>
<tr>
<td>araignées</td>
<td>spiders</td>
<td>spiders ant spider</td>
</tr>
<tr>
<td>tiges</td>
<td>stems</td>
<td>usda centimeters flowering</td>
</tr>
</tbody>
</table>
## Bleu Scores

<table>
<thead>
<tr>
<th>Variation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.91</td>
</tr>
<tr>
<td>Baseline + Strip Accents</td>
<td>22.20</td>
</tr>
<tr>
<td>Append Top-1 translation for OOVs</td>
<td>23.25</td>
</tr>
<tr>
<td>Append Top-1 translation for $\text{freq}(fr)&lt;11$</td>
<td>23.86</td>
</tr>
<tr>
<td>Append Oracle OOV translations</td>
<td>26.38</td>
</tr>
</tbody>
</table>
Automatically identifying new senses

- Context + existence of translations in comparable data

is a **window** of opportunity
have a **window** of opportunity
in the **run** up to
, we **run** the risk

via une **fenêtre** insérée .
vers ma **fenêtre** ou vers
voulons pas **courir** le risque
, sans **courir** le risque

dans la **fenêtre** . cet
dans la **fenêtre** . </s>

courir not found

the browser **window** ' s
in the **window** to give
time to **run** when applied
or have **run** vcvars.bat ,

ne pouvez **exécuter** que les
pour l' **exécuter** elle va
Spotting New Senses

• **Binary classification problem:**
  - +ve: French token has previously unseen sense
  - -ve: French token is used in a known way

• **Lots of features considered...**
  - Frequency of words/translations in each domain
  - Language model perplexities across domains
  - Topic model “mismatches”
  - Marginal matching features
  - Translation “flow” impedance
Experimental Results

Selected features:
- **EMEA**: ppl || matchm flow || matchm topics flow
- **Science**: ppl || matchm ppl || matchm topics ppl
- **Subs**: topics || matchm topics || matchm topics flow

Colors:
- **Constant**
- **One Feature**
- **Two Features**
- **Three Features**
- **All Features**
Discussion

- Introduced taxonomy and measurement tools for adaptation effects in MT
- “Score” errors – target of prior work – only a part of what goes wrong
- Marginal matching introduced as a model for addressing all S4 issues simultaneously: +2.4 BLEU
- Data and outputs released for you to use (both in MT and as a stand-alone lexical selection task)
- Feature-rich approaches integrated into Moses via VW library, applied to adaptation
- Range of other problems to work on: identifying new senses, cross-domain topic models, etc.)

Thanks! Questions?
Discussion and Future Work

- With labels, or labeler, for target, simple methods go a long way
- Active sampling can use adaptation knowledge
- Adaptive co-regularization ties domains together
- Similar ideas can be applied to multitask learning

- What happens with tons of domains?
  - Or non-discrete domains?
  - Or only domain features?
- Can we make better use of unlabeled data (ala Blitzer)

Thanks! Care to actively query?

Happy Thanksgiving!
Some experimental results

<table>
<thead>
<tr>
<th>Task</th>
<th>Dom</th>
<th>SrcOnly</th>
<th>TgtOnly</th>
<th>Baseline</th>
<th>Prior</th>
<th>Augment</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td></td>
<td>2.45</td>
<td>2.45</td>
<td>2.12 (all)</td>
<td>2.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D1</td>
<td>3.67</td>
<td>2.46</td>
<td>2.10 (linint)</td>
<td>2.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.32</td>
<td>4.98</td>
<td>2.37</td>
<td>2.11 (pred)</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.76</td>
<td>4.54</td>
<td>4.07</td>
<td>3.53 (weight)</td>
<td>3.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.61</td>
<td>3.47</td>
<td>1.98</td>
<td>1.98</td>
<td>1.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.41</td>
<td>5.15</td>
<td>5.15</td>
<td>5.15</td>
<td>5.15</td>
<td></td>
</tr>
</tbody>
</table>

| CoNLL | tgt | 2.49 | 2.95 | 1.75 (wgt/li) | 1.89  |
|       |     | 12.02 | 4.15 | 3.95 (linint) | 3.99  |
|       | CNN | tgt | 10.29 | 3.82 | 3.44 (linint) | 3.35  |
|       |     | wsj | 6.63  | 4.35 | 4.30 (weight) | 4.27  |
|       |     | swbd3 | 15.90 | 4.15 | 4.09 (linint) | 3.60  |
|       |     | br-cf | 5.16  | 6.27 | 4.72 (linint) | 5.22  |
|       | CNN | tgt | 10.29 | 3.82 | 3.44 (linint) | 3.35  |
|       |     | wsj | 6.63  | 4.35 | 4.30 (weight) | 4.27  |
|       |     | swbd3 | 15.90 | 4.15 | 4.09 (linint) | 3.60  |
|       |     | br-cf | 5.16  | 6.27 | 4.72 (linint) | 5.22  |
|       |     | br-cg | 4.32  | 5.36 | 4.15 (all) | 4.25  |
|       |     | br-ck | 5.05  | 6.32 | 5.01 (prd/li) | 5.27  |
|       | Chunk | br-cl | 5.66  | 6.60 | 5.39 (wgt/prd) | 5.99  |
Some Theory

- Can bound expected target error:

\[
\hat{\text{Number of}} \quad \hat{\text{Number of}} \quad \hat{\text{Number of}}
\]
Feature Hashing

- Feature augmentation creates \((K+1)D\) parameters
- Too big if \(K \geq 20\), but very sparse!
Hash Kernels

- How much contamination between domains?

\[
\frac{(\Pr \in \mathcal{P} \cap \mathcal{E} \cap \mathcal{O} \cup \mathcal{W} \circ \mathcal{F})}{\Pr \in \mathcal{E} \cap \mathcal{O} \cup \mathcal{W} \circ \mathcal{F}}
\]

Hash vector for domain \( u \)

Weights excluding domain \( u \)

Target (low) dimensionality

![Graph showing the relationship between spam miss-rate (relative to baseline) and b bits in hash-table.](image-url)
What is a domain anyway?

- Time?
  - News the day I was born vs news today?
  - News yesterday vs news today?

- Space?
  - News back home vs news in Haifa?
  - News in Tel Aviv vs news in Haifa?

- Do my data even come with a domain specified?
  
  Stream of \(<x,y,d>\) data with \(y\) and \(d\) sometimes hidden?
We’re all domains: personalization

- adapt learn across millions of “domains”?
- share enough information to be useful?
- share little enough information to be safe?
- avoid negative transfer?
- avoid DAAM (domain adaptation spam)?
Discussion