Learning from the World

Mihai Surdeanu
msurdeanu@email.arizona.edu
Natural Language Processing (NLP): computer process that extracts meaningful information from natural language input.
NLP ≠ Neuro-linguistic Programming
What can I help you with?

“Open the pod bay doors”

We intelligent agents will never live that down, apparently.
NLP with machine learning = computers learn a task based on empirical data
Supervised Learning

- Systems duplicate correct analyses from training data

- Why?
  - Predictive accuracy by exploiting human task supervision

- Why not?
  - Time-consuming
  - Experts are expensive
  - Training data needs updating
Unsupervised Learning

- Systems take raw data and automatically detect patterns

- Why?
  - Zero cost for data annotation
  - Ability to train on big data

- Why not?
  - It’s really hard to get something for nothing
  - It’s an interesting research track, but the results are not great
The very fact that problems are of interest means that people have seen a need to produce structured data.
Barack Obama is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois.

Born in Honolulu, Hawaii, Obama is a graduate of Columbia University Law School, where he was the president of the Harvard Law Review and worked as a community organizer in Chicago before earning his law degree. He worked as an attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 40th District in the Illinois Senate from 1997 to 2004.

Following an unsuccessful bid against the Democratic incumbent for a seat in the United States House of Representatives in 2000, Obama ran for United States Senate in 2004. Several events brought him to national attention during the campaign, including his victory in the March 2004 Illinois Democratic primary for Senate election and his keynote address at the Democratic National Convention in July 2004. He won election to the U.S. Senate in Illinois in November 2004. His presidential campaign began in February 2007, and after a close campaign in the 2008 Democratic Party presidential primaries against Hillary Rodham Clinton, he won his party's nomination. In the 2008 presidential election, he defeated Republican nominee John McCain and was inaugurated as president on January 20, 2009. In October 2009, Obama was named the 2009 Nobel Peace Prize laureate.

As president, Obama signed economic stimulus legislation in the form of the American Recovery and Reinvestment Act in February 2009 and the Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act in December 2010. Other domestic policy initiatives include the Patient Protection and Affordable Care Act, the Dodd-Frank Wall Street Reform and Consumer Protection Act, the Don't Ask, Don't Tell Repeal Act, and the Budget Control Act of 2011. In foreign policy, the administration drew up the Comprehensive Nuclear Test Ban Treaty, and the New Strategic Arms Reduction Treaty (New START) in 2010.

Information from the World

Barack Obama

Team Stat Leaders

<table>
<thead>
<tr>
<th>Player</th>
<th>Team</th>
<th>Passing Yards</th>
<th>Rushing Yards</th>
<th>Receiving Yards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tarvaris Jackson</td>
<td>Sea</td>
<td>157 yds</td>
<td>Sea</td>
<td>59 yds</td>
</tr>
<tr>
<td>Alex Smith</td>
<td>SF</td>
<td>124 yds</td>
<td>SF</td>
<td>33 yds</td>
</tr>
<tr>
<td>Marshawn Lynch</td>
<td>Sea</td>
<td>124 yds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frank Gore</td>
<td>SF</td>
<td>59 yds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doug Baldwin</td>
<td>Sea</td>
<td>83 yds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vernon Davis</td>
<td>SF</td>
<td>47 yds</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scoreboard

**Thursday, Sep 8**

New Orleans 34
Green Bay 42

**Sunday, Sep 11**

Philadelphia 31
St. Louis 13
Buffalo 41
Dallas 7
The very fact that problems are of interest means that people have seen a need to produce structured data.
The very fact that problems are of interest means that people have seen a need to produce structured data.

Information from the World

- **Why?**
  - Big data + low cost + human supervision

- **Why not?**
  - Messy, low quality data
  - Data may not model exactly what you want
Information from the World

- Learning from the world outside of NLP = data science
  - Political analysis
  - Business intelligence
  - Computational advertising
  - ...

Roadmap

Learning from the world

Search

I know what I want

“How can I find a veterinary college with dorms?”

Non-factoid Question Answering

I have a topic but I don’t know what I want

“Abu Nidal”

Relation Extraction
Roadmap

Learning from the world

Search

I know what I want
I have a topic but I don’t know what I want

Non-factoid Question Answering
Relation Extraction
What is QA?

QA: Task of answering *natural language questions* with small fragments of *text*.

- **Factoid QA**: answer is a noun phrase
  - Lot of research and progress, e.g., Watson, Google

  ![Search Result](what is barack obama's date of birth)

  *Best guess for Barack Obama Date of birth is August 4, 1961
  Mentioned on at least 6 websites including wikipedia.org, imdb.com and biography.com*

- **Non-factoid QA**: answer is short text, possibly aggregated from multiple sources
  - Very little research to date
  - More common than factoid QA
Distribution of Questions for Textbook Search

- Non-factoid
- Yes/No
- Definition
- Factoid
- Aggregation
- List
- Definition (image)
Application of Non-Factoid QA: Snippet Extraction

None of these are good answers. But they should be.
Q: How do you quiet a squeaky door?
A: Spray WD-40 directly onto the hinges of the door. Open and close the door several times. Remove hinges if the door still squeaks. Remove any rust, dirt or loose paint. Apply WD-40 to removed hinges. Put the hinges back, open and close door several times again.

Q: How does a helicopter fly?
A: A helicopter gets its power from rotors or blades. So as the rotors turn, air flows more quickly over the tops of the blades than it does below. This creates enough lift for flight.
But It Is Noisy

- Correctness is subjective
  - Q: How to extract html tags from an html documents with c++?
  - A: very carefully

- Grammar and spelling are optional
  - Q: Am I really that cute? girl said that I am 10/10 on heere? :/?
  - A: boy u is so hot lol
Objective

1. Is it possible to learn an answer ranking model for non-factoid questions, in a completely automated manner, using data available in online social QA sites?

2. Which features and models are more useful in this context, i.e., ample but noisy data?
APPROACH
Architecture

Question → Answer Retrieval → Answers

Answer Collection
Architecture

- Similarity Features
- Translation Features
- Density/Frequency Features
- Web Correlation Features

Question -> Answer Retrieval -> Answer Collection

Answer Retrieval -> Answer Re-ranking

Answer Re-ranking -> Answers

- Unsupervised learning
- Generative learning
- Discriminative learning
FEATURES AND MODEL
Features (1/4)

- FG1: similarity features
  - Intuition: a better answer will reuse words from the question

<table>
<thead>
<tr>
<th>Question:</th>
<th>Answer:</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to make a bonsai plant?</td>
<td>to make a <strong>bonsai</strong> takes many many years of clipping, wiring, bending and shaping. many online places sell <strong>bonsai</strong> kits but all you will get is dirt, a pot and a couple of seeds.</td>
</tr>
</tbody>
</table>
• FG2: translation features
  • Intuition: A is a new language translated from Q. A better answer will be a more likely translation of the question.

Question: How can I cook grouse?

Answer: I’ve never cooked a grouse, but poultry is poultry... You could salt it and pepper it, put some flour on a plate and roll it in the flour to coat it lightly, then heat a few tablespoons of olive oil in a skillet and pan-fry it.
Features (3/4)

- FG3: density/frequency features
  - Intuition: a better answer will reuse patterns from the question

**Question:** How to make a bonsai plant?

**Answer:** to make a bonsai takes many many years of clipping, wiring, bending and shaping…
Features (4/4)

- FG4: web correlation features
  - Intuition: words from a good answer will appear jointly with question words in other web pages or web queries

**Question:**
how did I do for make a call and that the other dont see my number?

**Answer:**
to make a call so that the other person cant see the number... dial *67 and wait for the three beeps.. then dial the number

Google hits for “number dial” = 156M
Google hits for “call dial” = 137M
All these features can be computed for different representations of the content
Surface Lexical Representations

“A helicopter gets its power from its rotor or blades.”

- Bag of words (W)
  - A, helicopter, gets, its, power, from, its, rotor, or, blades

- Bag of n-grams (N)
  - A-helicopter, helicopter-gets, gets-its, its-power, power-from, from-its, its-rotor, rotor-or, or-blades
Syntactic Representation

- Bag of syntactic dependencies (D)
  - A \(\text{NMOD}\) helicopter
  - helicopter \(\text{SBJ}\) get
  - power \(\text{OBJ}\) get
  - from \(\text{ADV}\) get
  - its \(\text{NMOD}\) power
  - rotor \(\text{COORD}\) or
  - blade \(\text{COORD}\) or
  - or \(\text{PMOD}\) from
Semantic Representation

- Bag of semantic dependencies (R)
  - $get_{\text{Arg0}} \rightarrow \text{helicopter}$
  - $get_{\text{Arg1}} \rightarrow \text{power}$
  - $get_{\text{Arg2}} \rightarrow \text{from-rotors}$
Degree of Lexicalization

helicopter → get

PRODUCT-DESC-VEHICLE → get

WN: N-ARTIFACT
WSJ: PRODUCT-DESC-VEHICLE
16 different representations of content

All generated using open-source tools, e.g., [http://www.surdeanu.name/mihai/swirl](http://www.surdeanu.name/mihai/swirl) for SRL
Model

- **Linear models**
  - score of an answer = $w_1 \times f_1 + \ldots + w_n \times f_n$

- **Re-ranking large-margin online learner (in house)**
  - Updates model when an incorrect answer has a higher score than the correct one
  - For every question, it inspects all pairs of correct answer and one incorrect answer
  - Large margin: static margin parameter (tuned in development)

- **SVM-rank – structural SVM tailored for ranking problems**
EXPERIMENTS
Corpus

- Subset of “how to” questions from Yahoo! Answers
  - Clean questions, with a chosen/voted best answer
  - “Yahoo! Answers Manner Questions, version 1.0”. Freely available through Yahoo! Webscope.

- 142,627 (Q, best A) pairs
  - We index all answers in this set as the answer collection
  - Partitioning of questions: 60% training, 20% development, 20% testing
Evaluation Measures

- **Notations:**
  - $N$ – how many answer candidates we retrieve for each question
  - $Q$ – all queries in the collection
  - $Q^N$ – subset of queries for which the candidate answer pool contains the correct answer

- Recall @ $N = \frac{Q^N}{Q}$
- Precision at rank 1 (P@1) – percentage of questions in $Q^N$ with the correct answer on the first position
Overall Results (P@1)

Baseline – No re-ranking, IR using BM25

> 20% relative improvement
Contribution of Each Feature Group

Translation models complement well the IR baseline
Contribution of NLP

NLP features complement well the bag-of-words baseline

NLP = syntax + SRL + NER + WSD

50% of improvement comes from SRL!
ERROR ANALYSIS
### Improved Questions

**how can i cook grouse quick with normal household spices w/o going out to buy stuff?**

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Re-rank</th>
<th>Correct?</th>
<th>Answer Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>yes</td>
<td>I’ve never <em>cooked</em> a <em>grouse</em>, but poultry is poultry... You could salt it and pepper it, put some flour on a plate and roll it in the flour to coat it lightly, then heat a few tablespoons of olive oil in a skillet and pan-fry it. (If you have no olive oil, use a little vegetable oil plus a pat of butter – the oil is to keep the butter from burning.) Squeeze a few drops of lemon juice over it if you want. Or: Skip the flour. Salt and pepper the <em>grouse</em>. Pan-fry…</td>
</tr>
</tbody>
</table>

**Shared structures:**
- Arg1(*cook, grouse*)

**Many words that translate to *cook:***
- *salt, pepper, flour, tablespoons, oil, skillet,* etc.

**Well, a *grouse* is a prey animal. If there was a decline in the population of *grouse*, then the animals that usually prey on the *grouse* - coyotes, owls, etc - would probably start eating other prey animals, like the pheasants and squirrels."
**Improved Questions**

how did I do for make a call and that the other dont see my number?

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Re-rank</th>
<th>Correct?</th>
<th>Answer Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>yes</td>
<td>to <em>make</em> a <em>call</em> so that the other person cant see the <em>number</em>... dial <em>67</em> and wait for the three beeps.. then <em>dial the number</em></td>
</tr>
</tbody>
</table>

**Shared structures:**
- Arg1(make, call)
- Arg1(see, number)

**Translated structures:**
- Arg1(make, call) to Arg1(dial, number)

Oneday out of the blue *call* her. If u dont have her *number*, when u *see* her ask her if she wanted to go out oneday then get her *number*. When u talk on the phone get to know her. But dont ask her out too soon because she may not feel the same way. After a couple of days or weeks taking to her let her know how u felt about her since the first time u met her.
Questions Worsened by Re-ranking

- Complex inference
- Assumed Context
- Also good
- Redirection
- Answer quality
- Spelling
- Clarification
Questions Worsened by Re-ranking
Questions Requiring Complex Inference

<table>
<thead>
<tr>
<th>how to deal with a person in denial with M.P.D.?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

being supportive and being confrontational are a form of dealing with somebody
## Questions Worsened by Re-ranking Answers With Assumed Context

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Re-rank</th>
<th>Correct?</th>
<th>Answer Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>yes</td>
<td>Some <em>mineral ions</em> are fertilizer and will increase vegetative growth while others are poisons.</td>
</tr>
</tbody>
</table>
### Questions Worsened by Re-ranking
Model Selects Answers that Are Also Good

<table>
<thead>
<tr>
<th>How to learn the British accent?</th>
<th>Baseline</th>
<th>Re-rank</th>
<th>Correct?</th>
<th>Answer Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>yes</td>
<td>Get a dictionary where there is a pronunciation guide which gives the pronunciation in <em>British</em> English. Watch <em>british</em> movies and imitate what you can. Then just practice, practice practice. But before you go about <em>learning</em> accents, slangs or dialects, make sure you brush up on your basic grammar.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>no</td>
<td>You can do one of two things: first, go to a local bookstore, like Barnes and Noble. They sell cd’s with different accents from around the world, accompanied by a book that phonetically spells the words. This is designed for actors/actresses who need to learn different accents. Also, go rent a bunch of <em>british</em> movies, or watch <em>british</em> television. Continually pause and repeat common phrases and words.</td>
</tr>
</tbody>
</table>
Conclusion - QA

• Non-factoid QA model learned from low quality data

• Key elements:
  • Powerful multi-module system combination (similarity, translation, patterns, web correlation)
  • Data representation is key to address noise. Semantic representations built with NLP (SRL, NER, WSD)
Roadmap

Learning from the world

Search

I know what I want

Non-factoid Question Answering

I have a topic but I don’t know what I want

Relation Extraction
What is Relation Extraction?

RE: Task of answering extracting labeled relations between pairs of named/numeric entities.

“Barack Hussein Obama (born August 4, 1961) is the 44th and current President of the United States.”
Application of Relation Extraction
User-Generated Data Is Available

- Millions of infoboxes in Wikipedia
- Over 22 million entities in Freebase. Each one has multiple relations
Barack Obama is the 44th and current President of the United States.

United States President Barack Obama meets with Chinese Vice President Xi Jinping today.
Barack Obama is the 44th and current President of the United States.

United States President Barack Obama meets with Chinese Vice President Xi Jinping today.

Obama was born in the United States just as he has always said.

Obama ran for the United States Senate in 2004.
Traditional Supervised Learning

diagram with labeled examples
Multi-instance Multi-label Learning

object 1
- instance 11
- instance 12
- instance 13

label 11
label 12

... object n
- instance n1
- instance n2
- instance nm

label n1
label nk
Barack Obama is the 44th and current President of the United States.

United States President Barack Obama meets with Chinese Vice President Xi Jinping today.

Obama was born in the United States just as he has always said.

Obama ran for the United States Senate in 2004.
“This picture contains cars, one person and a bicycle.”
Objective

- Learn in the MIML scenario
  - Model multiple instances
    - Some instances are false positives
    - Some instances are false negatives
  - Model multiple labels
    - There are dependencies between labels
APPROACH
Model Intuition

- instance 1
- instance 2
- instance $m$
- object
- label 1
- label $k$
Model Intuition

- Jointly trained using discriminative EM:
  - E-step: assign latent labels using current $\theta_z$ and $\theta_y$.
  - M-step: estimate $\theta_z$ and $\theta_y$ using the current latent labels.
Plate Diagram

- Binary relation-level classifiers
- Multi-class mention-level classifier
- Relation labels
- Latent label for each mention
- Relation mentions
- Number of tuples in the DB

Diagram:

- Nodes: $w_1, \ldots, w_j, \ldots, w_k$, $y_1, \ldots, y_j, \ldots, y_k$, $z$, $x$, $M_i$, $n$
- Edges:
  - $w_1 \rightarrow y_1$
  - $w_j \rightarrow y_j$
  - $w_k \rightarrow y_k$
  - $y_1, \ldots, y_j, \ldots, y_k \rightarrow z$
  - $z \rightarrow x$
  - $x \rightarrow |M_i|$
  - $|M_i| \rightarrow n$
E-step: Assign Latent Labels

- For the entire entity tuple $i$:
  \[ z_i^* = \arg\max_z p(z|y_i, x_i, w_y, w_z) \]

- For each individual mention $m$ in tuple $i$:
  \[ p(z^{(m)}_i | y_i, x_i, w_y, w_z) \approx \frac{p(z^{(m)}_i | x^{(m)}_i, w_z)p(y_i|z_i', w_y)}{\prod_{r \in P_i \cup N_i} p(y^{(r)}_i | z_i', w^{(r)}_y)} \]

  replaced mention (m) with $z_i^{(m)}$

  probability given by the mention-level model

  probabilities that the correct relation labels are assigned and the incorrect ones are not
M-step: Maximize Log Likelihood

\[
LL(w_y, w_z) = \sum_{i=1}^{n} \log p(y_i|x_i, w_y, w_z)
\]

\[
= \sum_{i=1}^{n} \log \sum_{z_i} p(y_i, z_i|x_i, w_y, w_z)
\]

maximize this probability
M-step: Maximize Log Likelihood

\[
\log p(y_i, z_i | x_i, w_y, w_z)
= \sum_{m \in M_i} \log p(z_i^{(m)} | x_i^{(m)}, w_z) + \sum_{r \in P_i \cup N_i} \log p(y_i^{(r)} | z_i, w_y^{(r)})
\]

one multi-class logistic regression

set of binary logistic regressions
Inference

entity tuple

relation mentions

\[ z_i^{(m)*} = \arg \max_z p(z|x_i^{(m)}, w_z) \]

\[ y_i^{(r)*} = \arg \max_{y \in \{0,1\}} p(y|z_i^{*}, w_y^{(r)}) \]

relation labels
FEATURES
Features

- **Z layer:**
  - Features that model the two entities
  - Features that model syntactic context
  - Features that model surface context

- **Y layer:**
  - At least one mention with my label?
  - Co-occurrence of my label with the other K – 1 labels
EXPERIMENTS
Corpora

• Riedel:
  • DB: Freebase
  • Text: NY Times
  • Evaluation: a fragment of the DB
    • But the DB is incomplete!

• Knowledge Base Population (KBP):
  • DB: Wikipedia infoboxes
  • Text: newswire, blogs, telephone conversations, Wikipedia
  • Development: 40 queries from the 2010 and 2011 eval
  • Evaluation: 160 queries from the 2010 and 2011 eval
    • This is complete!
## Corpora

<table>
<thead>
<tr>
<th></th>
<th># of gold relations in training</th>
<th># of gold relations in testing</th>
<th># of relation labels</th>
<th>% of gold entity tuples with multiple mentions in text (training)</th>
<th>% of mentions that do not express their relation</th>
<th>% of gold entity tuples with more than one label (training)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riedel</td>
<td>4,700</td>
<td>1,950</td>
<td>51</td>
<td>46.4%</td>
<td>up to 31%</td>
<td>7.5%</td>
</tr>
<tr>
<td>KBP</td>
<td>183,062</td>
<td>3,334</td>
<td>41</td>
<td>65.1%</td>
<td>up to 39%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>
Baseline/State of the Art

- Mintz++ (ACL, 2009)
  - “Traditional” distant supervision
  - No inter-label dependencies, no modeling of incorrect labels
  - Our extensions: multi-label predictions, bagging
- Riedel (ECML, 2010)
  - Multi-instance single-label
- Hoffmann (ACL, 2011)
  - Multi-instance multi-label
  - Does not model inter-label dependencies
Example Predictions

"Mexico" is_country_of "Mexico City"

"Mexico City" contains "Mexico"

(Mexico City, Mexico)

"Mexico" is_country_of "Mexico City"
Example: Dependencies Learned

Loc2 is_country_of Loc1

×

Loc1 contains Loc2

✓

Person1 lived_in Loc2

✓

Person1 was_born_in Loc2
Conclusion - RE

• First true multi-instance multi-label approach that models distant supervision for RE
• State of the art results on two difficult domains
• Data that does not express exactly the problem of interest can be modeled using anomaly-aware ML
Summary

• Exploiting non-expert annotations in the “big data” era

• Two important issues
  • Quality of the data
  • Annotations may not model exactly the same problem
Learning from the World

Pick any three

We offer three kinds of service:
GOOD - CHEAP - FAST
You can pick any two
GOOD service CHEAP won't be FAST
GOOD service FAST won't be CHEAP
FAST service CHEAP won't be GOOD
Acknowledgements

• QA
  • Joint work with Massimiliano Ciaramita and Hugo Zaragoza (Computational Linguistics, 2011)

• RE
  • Joint work with Ramesh Nallapati, Julie Tibshirani, and Chris Manning (EMNLP, 2012)
Thank You! Questions?

Your Resolved Question

How do you respond to a question when you don't know the answer?

RunnerJo...

Annmarie

I would give a stupid, yet humorous response. How the smart community does it.

Just say...I don't know. Follow it up with a blank stare.

Best Answer - Chosen by Asker

It is simple, I do not know the answer. Doing otherwise will leave you appearing awkward. It is good to be genuine in whatever you do.
APPENDIX
Features

- FG1: similarity features
  - Intuition: a better answer will reuse words from the question

- FG2: translation features
  - Intuition: Q is a new language translated from A. A better answer will generate a more likely translation for the question.

- FG3: density/frequency features
  - Intuition: a better answer will reuse patterns from the question

- FG4: web correlation features
  - Intuition: words from a good answer will appear jointly with question words in other web pages or web queries
**FG1: Similarity Features**

*Intuition: a better answer will reuse words from the question*

- Feature value: the similarity between Q and A using IR metrics
- We used length-normalized BM25 and cosine similarity using *tf-idf* weights
FG2: Translation Features

Intuition: similarity models perform poorly because they fail to “bridge the lexical chasm” between Q and A.

Intuition: Q is a new language translated from A. A better answer will generate a more likely translation.

- Feature value: P(Q|A)

\[
P(Q|A) = \prod_{q \in Q} P(q|A)
\]

\[
P(q|A) = (1 - \lambda)P_{ml}(q|A) + \lambda P_{ml}(q|C)
\]

\[
P_{ml}(q|A) = \sum_{a \in A} (T(q|a)P_{ml}(a|A))
\]

Tuned to optimize retrieval

Adjusted to make sure T(w|w) > T(w'|w)
Intuition: a better answer will reuse patterns from the question

- **Same word sequence**: number of question words in the same order in the answer
- **Answer span**: largest distance between two question words appearing in the answer
- **Informativeness**: number of new NN/VB/JJ in the answer
Intuition: a better answer will reuse patterns from the question

- **Same sentence match**: number of question words matched in the same sentence in the answer
- **Overall match**: number of question words matched in the complete answer
- **Tree kernels**: how many dependency trees are shared between question and answer?
  - Largest value between any two sentences
  - Average of all computed kernel values
FG4: Web Correlation Features

Intuition: words from a good answer will appear jointly with question words in other web pages or web queries

- **Web correlation:**
  - Feature value = Corrected Conditional Probability (CCP)
  - \( \text{CCP}(Q, A) = \frac{\text{hits}(Q + A)}{\text{hits}(Q) \times \text{hits}(A)^{2/3}} \)
  - Needs query relaxation to work for non-factoid QA

- **Query-log correlation:**
  - Largest/average PMI and \( \chi^2 \) values between any two words in the question and answer
  - Number of \((q,a)\) word pairs that appear in the top 10, 5, and 1 percentile of PMI and \( \chi^2 \) values
Parameters
Labels of Relations

$helicopter \xrightarrow{SBJ} get$ vs. $helicopter \xrightarrow{} get$
# Feature Selection

<table>
<thead>
<tr>
<th>Iter.</th>
<th>Feature Set</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BM25(W)</td>
<td>41.1</td>
</tr>
<tr>
<td>1</td>
<td>+ translation (R)</td>
<td>46.3</td>
</tr>
<tr>
<td>2</td>
<td>+ translation (N)</td>
<td>48.0</td>
</tr>
<tr>
<td>3</td>
<td>+ overall match ($D_{WNSS}$)</td>
<td>48.9</td>
</tr>
<tr>
<td>4</td>
<td>+ translation (W)</td>
<td>49.1</td>
</tr>
<tr>
<td>5</td>
<td>+ query-log avg (PMI)</td>
<td>49.6</td>
</tr>
<tr>
<td>6</td>
<td>+ overall match (W)</td>
<td>49.7</td>
</tr>
<tr>
<td>7</td>
<td>+ overall match, normalized by Q size (W)</td>
<td>49.9</td>
</tr>
<tr>
<td>8</td>
<td>+ same word sequence, normalized by Q size (W)</td>
<td>49.9</td>
</tr>
<tr>
<td>9</td>
<td>+ BM25 (N)</td>
<td>50.0</td>
</tr>
<tr>
<td>10</td>
<td>+ informativeness: verb count</td>
<td>50.0</td>
</tr>
<tr>
<td>11</td>
<td>+ query log max (PMI)</td>
<td>50.3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Contribution of NL Structures

<table>
<thead>
<tr>
<th></th>
<th>Similarity</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>0</td>
<td>+4.2</td>
</tr>
<tr>
<td>$N$</td>
<td>-14.0</td>
<td></td>
</tr>
<tr>
<td>$N_g$</td>
<td>-18.7</td>
<td>+3.6</td>
</tr>
<tr>
<td>$D$</td>
<td>-15.2</td>
<td>+1.5</td>
</tr>
<tr>
<td>$D_g$</td>
<td>-19.3</td>
<td>+3.4</td>
</tr>
<tr>
<td>$R$</td>
<td>-27.6</td>
<td>+0.3</td>
</tr>
<tr>
<td>$R_g$</td>
<td>-28.3</td>
<td>+3.5</td>
</tr>
</tbody>
</table>

Feature classes: $W$, $N$, $N_g$, $D$, $D_g$, $R$, $R_g$.

Representations of content: $W$, $N$, $D$, $R$.
## Contribution of NL Structures

<table>
<thead>
<tr>
<th></th>
<th>Similarity</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>0</td>
<td>+4.2</td>
</tr>
<tr>
<td>$N$</td>
<td>-14.0</td>
<td>+2.5</td>
</tr>
<tr>
<td>$N_g$</td>
<td>-18.7</td>
<td>+3.6</td>
</tr>
<tr>
<td>$D$</td>
<td>-15.2</td>
<td>+1.5</td>
</tr>
<tr>
<td>$D_g$</td>
<td>-19.3</td>
<td>+3.4</td>
</tr>
<tr>
<td>$R$</td>
<td>-27.6</td>
<td>+0.3</td>
</tr>
<tr>
<td>$R_g$</td>
<td>-28.3</td>
<td>+3.5</td>
</tr>
<tr>
<td>$W + N + N_g + D + D_g + R + R_g$</td>
<td><strong>+1.65</strong></td>
<td><strong>+6.3</strong></td>
</tr>
</tbody>
</table>
M-step: Maximize Log Likelihood

one multi-class logistic regression

\[
 w_z^* = \arg \max_w \sum_{i=1}^{n} \sum_{m \in M_i} \log p(z_i^{(m)*} | x_i^{(m)}, w)
\]

set of binary logistic regressions

\[
 w_y^{(r)*} = \arg \max_w \sum_{1 \leq i \leq n \text{ s.t. } r \in P_i \cup N_i} \log p(y_i^{(r)} | z_i^*, w)
\]
### Riedel Dataset with Groups with 10+ Mentions

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoffmann (our implementation)</td>
<td>48.6</td>
<td>29.8</td>
<td>37.0</td>
</tr>
<tr>
<td>Mintz++</td>
<td>43.8</td>
<td><strong>36.8</strong></td>
<td>40.0</td>
</tr>
<tr>
<td>MIML-RE</td>
<td>56.1</td>
<td>32.5</td>
<td>41.1</td>
</tr>
<tr>
<td>MIML-RE (w/ dependencies)</td>
<td><strong>64.8</strong></td>
<td>31.6</td>
<td><strong>42.6</strong></td>
</tr>
</tbody>
</table>
Future Work

• How robust is “learning from the world” to domain transfer?
  • Learn from community QA data, apply to textbooks? Mobile?

• Can we extract complex structures from this type of data?
  • Event extraction