Large-Scale Language Learning

Slav Petrov

Thanks to
Ryan McDonald, Keith Hall, Alexander Rush, Dipanjan Das,
Hao Zhang, Michael Ringgaard, Terry Koo and Kuzman Ganchev
(a.k.a. the Natural Language Parsing Team at Google)
NLP: Where do we stand?
NLP: Where do we stand?

who is the president of the US

Barack Obama
United States of America, President

President of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/.../President_of_the_...
On January 20, 2009, Barack Obama became the 44th and current president. On November 6, 2012, he was re-elected ...

List of Presidents of the United ...

List of Presidents of the United States - Wikipedia, the free ...
en.wikipedia.org/.../List_of_Presidents_o...
NLP: Where do we stand?
NLP: Where do we stand?
NLP: Where do we stand?
NLP: Where do we stand?
NLP: Where do we stand?
NLP: Where do we stand?
NLP: Where do we stand?
Building a Repository

Find *all Mentions* of Entities in *Documents*

- Resolve Mentions to the *Identities* of the *Entities*
- Extract typed *Relations* between Entities
- Extract typed *Attribute* values for Entities
President Barack Obama has been re-elected to a second term, defeating Republican challenger Mitt Romney. America's first black president secured more than the 270 votes in the electoral college needed to win. In his victory speech before supporters in Chicago, Mr Obama said he would talk to Mr Romney about "where we can work together to move this country forward". Mr Obama prevailed despite lingering dissatisfaction with the economy and a hard-fought challenge by Mr Romney.

His Democrats also retained their majority in the Senate, which they have held since 2007. The Republicans kept control of the House of Representatives, which analysts say will likely result in more of the gridlock that characterized Mr Obama's first term, with the House and the president at loggerheads on most legislation.

In his address, the president challenged his opponents, asking them to work with him. With only Florida's 29 electoral votes still undecided, Mr Obama won 303 electoral votes to Mr Romney's 206. The popular vote, which is symbolically and politically important but not decisive in the race, remains very close.

Mr Obama congratulated Mr Romney and Republican vice-presidential candidate Paul Ryan on their hard-fought campaign. "We have picked ourselves up, we have fought our way back and we know in our hearts that for the United States of America the best is yet to come," he said. Mr Obama said he was returning to the White House "more determined, and more inspired than ever about the work there is to do, and the future that lies ahead". He pledged to work with Republican leaders in Congress to reduce the government's budget deficit, fix the tax code and reform the immigration system.
## Semantic Annotation of Documents

<table>
<thead>
<tr>
<th>E₁</th>
<th>Barack Obama</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>Barack Obama, Obama (6), Mr (6), President, He, His, he (3), his</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>Barack Obama</td>
<td></td>
</tr>
<tr>
<td>Freebase MID</td>
<td>/m/02mjmr</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Barack Obama</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₂</th>
<th>term</th>
<th>NON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>term</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₃</th>
<th>Mitt Romney</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>Mitt Romney, Romney (4), Mr (4), challenger</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>Mitt Romney</td>
<td></td>
</tr>
<tr>
<td>Freebase MID</td>
<td>/m/0271_s</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Mitt Romney</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₄</th>
<th>Republican</th>
<th>ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>Republican (3)</td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>Republican Party (United States)</td>
<td></td>
</tr>
<tr>
<td>Freebase MID</td>
<td>/m/07wbk</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Republican Party (United States)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₅</th>
<th>president</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>president</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₆</th>
<th>United States of America</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentions</td>
<td>America, United States of America</td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>United States</td>
<td></td>
</tr>
<tr>
<td>Freebase MID</td>
<td>/m/09c7w0</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>United States</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E₇</th>
<th>votes</th>
<th>NON</th>
</tr>
</thead>
</table>
Traditional Supervised NLP

New Task
Traditional Supervised NLP

New Task → Linguists
Traditional Supervised NLP

New Task → Linguists → Machine Learning
Traditional Supervised NLP

- Example: Syntactic analysis for English:
  - Not solved, but accuracies are high
  - 97% (or 90%?) for parts-of-speech
  - 93% (or 83%?) for parse trees

Linguists → Machine Learning

Treebank:
- 2 million words
A Note on Parallel Computing
A Note on Parallel Computing

- Most NLP models these days operate at sentence level
  - Trivially parallelizable
  - Focus on efficient single-core models
A Note on Parallel Computing

• Most NLP models these days operate at sentence level
  • Trivially parallelizable
  • Focus on efficient single-core models

• Examples:
  • Web-Scale Information Extraction
  • Machine Translation
A Note on Parallel Computing

• Most NLP models these days operate at sentence level
  • Trivially parallelizable
  • Focus on efficient single-core models

• Examples:
  • Web-Scale Information Extraction
  • Machine Translation

• Distributed systems might be needed for storage:
  • Distributed Language Models, Phrase Tables, ... in Machine Translation, Speech Recognition, ...
Universal Tagging/Parsing

- Goal: high accuracy parsing in all languages with a single universal representation of syntax
Syntactic Processing for the Web
Syntactic Processing for the Web
Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]
Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]

- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ’11, Ganchev et al. ’12]
Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]

- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ’11, Ganchev et al. ’12]

- Multilingual tagger and parser projection
  - Will never have labeled resources for all languages
  - Use parallel data to project information
    [Das & Petrov ’11, McDonald et al. ’11]
Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ‘12]

- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ‘11, Ganchev et al. ‘12]

- Multilingual tagger and parser projection
  - Will never have labeled resources for all languages
  - Use parallel data to project information [Das & Petrov ‘11, McDonald et al. ‘11]
Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

[Nivre et al. ’03–’11] [McDonald et al. ’05–’06]
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

---

Accuracy

Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

![Graph](image)

- **Accuracy**: $O(n)$
- **Time**: [Nivre et al. ’03–’11]
- **Accuracy**: greedy tr
- **Time**: [McDonald et al. ’05–’06]
Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

---

<table>
<thead>
<tr>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O(n))</td>
<td>(O(k \cdot n))</td>
</tr>
</tbody>
</table>

[Nivre et al. ‘03–’11] [McDonald et al. ‘05–’06]
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

<table>
<thead>
<tr>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[O(n)]</td>
<td>greedy tr [O(n)]</td>
</tr>
<tr>
<td>[O(k \cdot n)]</td>
<td>k-best tr</td>
</tr>
<tr>
<td>[O(n^3)]</td>
<td>1st-order gr [O(n^3)]</td>
</tr>
</tbody>
</table>

[Nivre et al. ‘03–’11] [McDonald et al. ‘05–’06]
Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

<table>
<thead>
<tr>
<th>Time</th>
<th>Accuracy</th>
<th>Greedy tr</th>
<th>O(n)</th>
<th>k-best tr</th>
<th>O(k·n)</th>
<th>1st-order gr</th>
<th>O(n³)</th>
<th>2nd-order gr</th>
<th>O(n³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Nivre et al. ’03-’11]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[McDonald et al. ’05-’06]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>greedy tr</td>
<td>$O(n)$</td>
<td></td>
</tr>
<tr>
<td>k-best tr</td>
<td>$O(k \cdot n)$</td>
<td></td>
</tr>
<tr>
<td>1st-order gr</td>
<td>$O(n^3)$</td>
<td></td>
</tr>
<tr>
<td>2nd-order gr</td>
<td>$O(n^3)$</td>
<td></td>
</tr>
<tr>
<td>3rd-order gr</td>
<td>$O(n^4)$</td>
<td></td>
</tr>
</tbody>
</table>

[Nivre et al. ’03–’11]  [McDonald et al. ’05–’06]
Dependency Parsing

- **Transition-Based (tr)**
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- **Graph-Based (gr)**
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

---

Goal

- k-best tr \(O(k \cdot n)\)
- greedy tr \(O(n)\)
- 1st-order gr \(O(n^3)\)
- 2nd-order gr \(O(n^3)\)
- 3rd-order gr \(O(n^4)\)

---

Time

- [Nivre et al. '03-'11]
- [McDonald et al. '05-'06]
As McGwire neared, fans went wild.
As McGwire neared, fans went wild.
As McGwire neared, fans went wild.
Dependency Representation

As McGwire neared, fans went wild.
As McGwire neared, fans went wild.

That’s already too much!
As McGwire neared, fans went wild.

That’s already too much!

Exploit problem structure!
Coarse-to-fine Vine Parsing

As McGwire neared, fans went wild

Vine Pruning

[Rush & Petrov ’12]
As McGwire neared, fans went wild.

Vine

Coarse-to-Fine Cascades

[Rush & Petrov ’12]
Coarse-to-Fine Cascades

As McGwire neared, fans went wild.

Vine

First-Order

[Rush & Petrov '12]
Coarse-to-Fine Cascades

Vine
First-Order
Second-Order

[McGwire neared, fans went wild]
Max-Marginals

MAP Parse
\[ y^* = \arg \max_{y \in Y} y \cdot w \]

Max-Marginal
\[ m(a) = \arg \max_{y \in Y : a \in A} y \cdot w \]
As McGwire neared, fans went wild.
As McGwire neared, fans went wild.
Structured Prediction Cascades Training

[Weiss & Taskar ’10]
Structured Prediction Cascades Training

[Weiss & Taskar ’10]

- Train to minimize pruning error (rather than 1-best)
Structured Prediction Cascades Training

[Weiss & Taskar ’10]

• Train to minimize pruning error (rather than 1-best)
• Pruning threshold:

\[ t_\alpha(w) = \alpha y^* \cdot w + (1 - \alpha) \frac{1}{|A|} \sum_{a \in A} m(a) \cdot w \]
Structured Prediction Cascades Training

- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:

\[ \alpha = 0.5 \]

\[ \alpha = 1 \]

[Weiss & Taskar ’10]
Structured Prediction Cascades Training

[Weiss & Taskar ’10]

- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:
  - Training objective:
    \[
    \min_w \lambda \|w\|^2 + \frac{1}{M} \sum_{m=1}^M \max\{0, 1 + t_\alpha(w) - y^m \cdot w\}
    \]
Structured Prediction Cascades Training

- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:

  \[
  \alpha = \text{0.5}
  \]

  \[
  \alpha = \text{1}
  \]

- Training objective:
  - Filter as many arcs as possible
  - While preserving gold arc
  - Optimize with stochastic gradient decent
    (not so different from perceptron updates)
UAS 91.0, Set pruning thresholds for no loss in accuracy
Second-Order Parsing

UAS 92.1, Set pruning thresholds for no loss in accuracy

[Relative Speed]

- No Prune
- Dictionary
- Local
- 1st-Only
- Vine
- k=16

[Bergsma & Cherry '10]
[Koo & Collins '11]
[Nivre et al. '03 - '11]
Third-Order Parsing

UAS 92.9, Set pruning thresholds for no loss in accuracy

- No Prune
- Dictionary
- Local
- 1st-Only
- Vine
- k=64

[Third-Order Parsing]

- Bergsma & Cherry '10
- Koo & Collins '11
- Nivre et al. '03 - '11

Relative Speed
Third-Order Parsing

UAS 92.9, Set pruning thresholds for no loss in accuracy

- No Prune
- Dictionary
- Local
- 1st-Only
- Vine
- k=64

Relative Speed

200 fold

[200, 200]
Google Web Treebanks

- **Google Web Treebank**
  - Funded by Google, annotated and released by LDC
  - 5 domains: Blogs, Newsgroups, Reviews, Emails, Q&A
  - ~2,000 manually annotated sentences (PTB-style)
  - >100,000 unlabeled sentences

- **Shared Task at NAACL ’12 Workshop**
  - Constituency Trees or Stanford Dependencies
  - Train on WSJ + unlabeled data
  - 2 domains released for development
  - Test on remaining 3 domains

- **New**: Google Multilingual Treebank (6 languages)
POS Accuracy (SANCL Shared Task)

<table>
<thead>
<tr>
<th>Newswire</th>
<th>Web Text</th>
</tr>
</thead>
</table>

Baseline = StanfordTagger v2.0 [Manning ’11]
POS Accuracy (SANCL Shared Task)

Newswire 

| 100 | 97.2 |

Web Text

| 90 |

80

Baseline (in-domain)
Baseline
UPenn
UMass
NAIST
IMS-2
IMS-3
IMS-1
CPH-Trento
Stanford-2
HIT-Baseline
HIT-Domain
Stanford-1
DCU-Pairs13

Baseline = StanfordTagger v2.0 [Manning ’11]
POS Accuracy (SANCL Shared Task)

Baseline = StanfordTagger v2.0 [Manning '11]
POS Accuracy (SANCL Shared Task)

Newswire

Baseline (in-domain)
Baseline
UPenn
UMass
NAIST
IMS-2
IMS-3
IMS-1
CPH-Trento
Stanford-2
HIT-Baseline
HIT-Domain
Stanford-1
DCU-Pairs13

~3%

20-30% unknown word rate

Baseline = StanfordTagger v2.0 [Manning ‘11]
Parsing Accuracy (SANCL Shared Task)

<table>
<thead>
<tr>
<th></th>
<th>Newswire</th>
<th>Web Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (in-domain)</td>
<td>90.67</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>78.77</td>
<td>80.17</td>
</tr>
<tr>
<td>UPenn</td>
<td>71.04</td>
<td>80.2</td>
</tr>
<tr>
<td>UMass</td>
<td>74.88</td>
<td>80.25</td>
</tr>
<tr>
<td>NAIST</td>
<td>76.36</td>
<td>82.54</td>
</tr>
<tr>
<td>IMS-2</td>
<td>76.87</td>
<td>82.63</td>
</tr>
<tr>
<td>IMS-3</td>
<td>77.76</td>
<td>82.63</td>
</tr>
<tr>
<td>IMS-1</td>
<td>80.25</td>
<td>83.13</td>
</tr>
<tr>
<td>CPH-Trento</td>
<td>80.17</td>
<td>83.13</td>
</tr>
<tr>
<td>Stanford-2</td>
<td>80.2</td>
<td>83.46</td>
</tr>
<tr>
<td>HIT-Baseline</td>
<td>82.54</td>
<td></td>
</tr>
<tr>
<td>HIT-Domain</td>
<td>82.63</td>
<td></td>
</tr>
<tr>
<td>Stanford-1</td>
<td>83.13</td>
<td></td>
</tr>
<tr>
<td>DCU-Pairs13</td>
<td>83.46</td>
<td></td>
</tr>
</tbody>
</table>

Baseline = MaltParser [Zhang & Nivre ’11]
Tagging Search Queries

[Tagging Search Queries

Google search results for "buy brave 2012 soundtrack"

Amazon.com: Brave: Various Artists, Patrick Doyle, James Shearmun ... www.amazon.com/Brave-Various-Artists/dp/B007XSFKAW
Audio CD, Soundtrack, 2012, $11.88, --, -- ... Songs from this album are available to purchase as MP3s. ... This item: Brave ~ Patrick Doyle Audio CD $11.88 ...

iTunes - Music - Brave (Original Score) by Various Artists
itunes.apple.com/us/album/brave-original-score/id527775518
19 Jun 2012 – Preview songs from Brave (Original Score) by Various Artists on the iTunes Store. Preview, buy, and download Brave (Original Score) for $9.99.

Brave (2012) - Soundtracks
www.imdb.com/title/tt1217209/soundtrack
IMDb > Brave (2012) > Soundtracks. Brave. Own the rights? Buy it at Amazon ... here (and in the movie credits) cannot always be found on CD soundtracks.]
Tagging Search Queries

[Tagging Search Queries - Ganchev et al. ’12]
Tagging Search Queries

ADJ or NOUN?

Google search for "buy brave 2012 soundtrack"

Amazon.com: Brave: Various Artists, Patrick Doyle, James Shearman ...
Audio CD, Soundtrack, 2012, $11.88, --, --, ... Songs from this album are available to purchase as MP3s. This item: Brave - Patrick Doyle Audio CD $11.88 ...

iTunes - Music - Brave (Original Score) by Various Artists
itunes.apple.com/us/album/brave-original-score/id527775518
19 Jun 2012 – Preview songs from Brave (Original Score) by Various Artists on the iTunes Store. Preview, buy, and download Brave (Original Score) for $9.99.

Brave (2012) - Soundtracks
www.imdb.com/title/tt1217209/soundtrack
IMDb > Brave (2012) > Soundtracks. Brave. Own the rights? Buy it at Amazon ... here (and in the movie credits) cannot always be found on CD soundtracks.
Tagging Search Queries

ADJ or NOUN?
Tagging Search Queries

[Google search results for "buy brave 2012 soundtrack"

ADJ or NOUN?]

[Ganchev et al. ’12]
Tagging Search Queries

ADJ or NOUN?

[Tagging Search Queries][Ganchev et al. ’12]
Tagging Search Queries

[Tagging Search Queries]

buy brave 2012 soundtrack

VERB NOUN NUM NOUN
Tagging Search Queries

[Tagging Search Queries, Baseline vs Retrained]

Baseline

Retrained
Query Tagging Results

et al. ’10

[Ganchev et al. ’12]
Query Tagging Results

Baseline
Bendersky et al. ’10
74.65
All
Top-1
Click
Query Tagging Results

[Query Tagging Results from Ganchev et al. '12]

- Baseline: 74.65
- Bendersky et al. '10: 77.27

Scores:
- All
- Top-1
- Click
Query Tagging Results

[Ganchev et al. ’12]

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Bendersky et al. ’10</th>
<th>All</th>
<th>Top-1</th>
<th>Click</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>74.65</td>
<td>77.27</td>
<td>77.73</td>
<td>77.73</td>
<td>77.73</td>
</tr>
<tr>
<td>Bendersky et al. ’10</td>
<td>77.27</td>
<td>77.73</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Query Tagging Results

- Baseline: 74.65
- Bendersky et al. ’10: 77.27
- All: 77.73
- Top-1: 77.6
- Click: 77.27

[Ganchev et al. ’12]
Query Tagging Results

[Ganchev et al. ’12]
The breaking bad case...
The breaking bad case...
The breaking bad case...

Before

After

Use freebase entries as features
How does a bee fly?
• Sequential Markov Models have limitations:

* How does a bee fly?
• Sequential Markov Models have limitations:

How does a bee fly?

... ‘to coach kim’
• Sequential Markov Models have limitations:

... ‘to coach kim’
• Sequential Markov Models have limitations:

How does a bee fly?

... ‘to coach kim’
• Sequential Markov Models have limitations:
• Sequential Markov Models have limitations:

How does a bee fly?

... ‘to coach kim’

• Ongoing work:
  Search jointly over POS tags and parse trees.
Manual Intervention
Manual Intervention

nc
text
NN
NOUN

ROOT
francis
NN
NOUN

dobj
that
WDT
DET

nsubj
i
PRP
PRON

aux
will
MD
VERB

rcmod
be
VB
VERB

acomp
late
JJ
ADJ
Manual Intervention

nc
text
NN
NOUN

ROOT
francis
NN
NOUN
dobj
that
WDT
DET
nsubj
i
PRP
PRON
aux
will
MD
VERB
rcmod
be
VB
VERB
acomp
late
JJ
ADJ
Manual Intervention
Machine Translation Reordering

- Source-side reordering for Japanese-English MT
  [Collins et al. ‘05]
- Dependency-based reordering for English-Japanese
  [Xu et al. ‘09]
Machine Translation Reordering

- Source-side reordering for Japanese-English MT
  [Collins et al. ’05]
- Dependency-based reordering for English-Japanese
  [Xu et al. ’09]

They solved the problem with statistics.
Reordering Score

- Source-side reordering for Japanese-English MT
  
  - Hand generated reordering data (English + Jenglish)
    - ~ 10k sentences for Augmented Loss training
    - ~ 6k evaluation sentences
  
- Score based on reordering penalty of METEOR
  
  \[
  \text{reorder-score} = \frac{\# \text{ chunks} - 1}{\# \text{ unigrams matched} - 1}
  \]

  \[
  \text{reorder-loss} = 1 - \text{reorder-score}
  \]

- Very well correlated with human eval scores.
EnJa Fuzzy Reordering Evolution

- **Uptraining [Petrov et al. ‘10]**
- **Targeted Uptraining [Katz-Brown et al. ‘11]**
- **Beam Search**
- **case-insensitive**
- **Augmented Loss [Hall et al. ‘11]**
- **More Data**

Timeline:
- Jan’10
- Mar’10
- Jun’10
- Aug’10
- Sep’10
- Oct’10
- Dec’10
- Jan’11
- May’11
- May’11
- Jun’11
EnJa Fuzzy Reordering Evolution

WSJ (LAS)
Reordering
Web (LAS)

Jan’10 Mar’10 Jun’10 Aug’10 Sep’10 Oct’10 Dec’10 Jan’11 May’11 May’11 Jun’11

70 74.5 79 83.5 88
EnJa Fuzzy Reordering Evolution

- **Uptraining** [Petrov et al. '10]
- **Targeted Uptraining** [Katz-Brown et al. '11]
- **More Data**
- **Beam Search**
- **Augmented Loss** [Hall et al. '11]

Graph showing improvements over time:
- **WSJ (LAS)**
- **Reordering**
- **Web (LAS)**

Months from Jan'10 to Jun'11 are marked on the x-axis.
Retrospective Analysis

- final parser
- greedy parser
- greedy & not retokenized
Retrospective Analysis

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>final parser</th>
<th>greedy parser</th>
<th>greedy &amp; not retokenized</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ru-en</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hu-en</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nl-en</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de-en</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zh-en</td>
<td>1.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ko-en</td>
<td>2.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Retrospective Analysis

BLEU Score Difference

- final parser
- greedy parser
- greedy & not retokenized

fr-en: 0.36
ru-en: 0.64
hu-en: 0.73
nl-en: 0.87
de-en: 0.92
zh-en: 1.85
ko-en: 2.33
Retrospective Analysis

BLEU Score Difference

final parser | greedy parser | greedy & not retokenized

- fr-en: 0.36
- ru-en: 0.64
- hu-en: 0.73
- nl-en: 0.87
- de-en: 0.92
- zh-en: 1.85
- ko-en: 2.33
Retrospective Analysis

Total: -23.4 BLEU
(i18n -2.7)
Retrospective Analysis

Total: -23.4 BLEU (i18n -2.7)
Total: -27.0 BLEU (i18n -7.7)
Standard Perceptron Training

Standard Perceptron

\[ x = \text{John likes Mary} \]

\[ y^* = \begin{array}{ccc}
\text{nsbj} & \text{ROOT} & \text{dobj} \\
\text{John} & \text{likes} & \text{Mary} \\
\text{NNP} & \text{VBZ} & \text{NNP} \\
\text{NOUN} & \text{NOUN} & \\
\end{array} \]

\[ y = \arg \max_{y \in \mathcal{Y}_x} w \cdot \phi(y) \]

\[ \theta = \theta + \phi(y^*) - \phi(y) \]
Augmented Loss (Online) Training

Intrinsic Data

$$X_1, y_1$$
$$X_2, y_2$$
$$\vdots$$
$$X_i, y_i$$
$$\vdots$$
$$X_n, y_n$$

$y \in \mathcal{Y}$

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss (Online) Training

Intrinsic Data

\[ x_n, y_n \quad y \in \mathcal{Y} \]

Similar to [Chang et al. '08, McAllester et al. '10]
Augmented Loss (Online) Training

Intrinsic Data
\[ x_1, y_1, x_2, y_2, \ldots, x_i, y_i, \ldots, x_n, y_n \]
\[ y \in \mathcal{Y} \]

Extrinsic Data
\[ x'_1, y'_1, x'_2, y'_2, \ldots, x'_i, y'_i, \ldots, x'_n, y'_n \]
\[ y' \in \mathcal{Y}' \]

Similar to [Chang et al. ’08, McAllester et al. ’10]

[Hall et al. ’11]
Augmented Loss (Online) Training

Intrinsic Data

\[
\begin{align*}
  &x_1, y_1 \\
  &x_2, y_2 \\
  &\vdots \\
  &x_i, y_i \\
  &\vdots \\
  &x_n, y_n
\end{align*}
\]

\[y \in \mathcal{Y}\]

Trainer

Extrinsic Data

\[
\begin{align*}
  &x'_1, y'_1 \\
  &x'_2, y'_2 \\
  &\vdots \\
  &x'_i, y'_i \\
  &\vdots \\
  &x'_n, y'_n
\end{align*}
\]

\[l(\hat{y}_i, y'_i) \quad \hat{y} \in \mathcal{Y}\]

Similar to [Chang et al. ’08, McAllester et al. ’10]

[Hall et al. ’11]
Augmented Loss (Online) Training

**Intrinsic Data**

\[ x_1,y_1, x_2,y_2, \ldots, x_i,y_i, x_n,y_n \]

**Extrinsic Data**

\[ x'_1,y'_1, x'_2,y'_2, \ldots, x'_i,y'_i, x'_n,y'_n \]

Similar to [Chang et al. '08, McAllester et al. '10]

[Hall et al. '11]
Augmented Loss (Online) Training

Intrinsic Data

\[ x_i, y_i \]

\[ \vdots \]

\[ x_n, y_n \quad y \in \mathcal{Y} \]

Extrinsic Data

\[ x_1', y_1' \]

\[ \vdots \]

\[ x_n', y_n' \quad y' \in \mathcal{Y}' \]

\[ l(\hat{y}_i, y_i') \quad \hat{y} \in \mathcal{Y} \]

Similar to [Chang et al. ’08, McAllester et al. ’10]

[Hall et al. ’11]
Augmented Loss (Online) Training

Intrinsic Data

\[
\begin{align*}
\vdots \\
X_n, y_n &\quad y \in \mathcal{Y} \\
\vdots
\end{align*}
\]

Extrinsic Data

\[
\begin{align*}
\hat{y} &\in \mathcal{Y} \\
x'_2, y'_2 &\quad y'_2 \in \mathcal{Y}' \\
x'_i, y'_i &\quad y'_i \in \mathcal{Y}' \\
x'_n, y'_n &\quad y' \in \mathcal{Y}'
\end{align*}
\]

 Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss (Online) Training

**Intrinsic Data**

\[ x_n, y_n \quad y \in \mathcal{Y} \]

**Extrinsic Data**

\[ x'_i, y'_i \quad y' \in \mathcal{Y}' \]

\[ l(\hat{y}_i, y'_i) \quad \hat{y} \in \mathcal{Y} \]

**Intrinsic:** supervised training data/objective

**Extrinsic:** downstream task which “uses” outputs of model

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[ x_1, y_1 \]
\[ x_2, y_2 \]
\[ \vdots \]
\[ x_i, y_i \]
\[ \vdots \]
\[ x_n, y_n \]

\[
\begin{aligned}
    y_i &= \text{nsbj} \quad \text{dobj} \\
    \hat{y}_i &= \text{nsbj} \quad \text{dobj}
\end{aligned}
\]

\[
\hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)
\]

if \( y_i \neq \hat{y}_i \) then

\[
\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)
\]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y) \]

if \( y_i \neq \hat{y}_i \) then
\[ \theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i) \]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y) \]

if \( y_i \neq \hat{y}_i \) then
\[ \theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i) \]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[
\hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)
\]

if \( y_i \neq \hat{y}_i \) then

\[
\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)
\]

Similar to [Chang et al. ’08, McAllester et al. ’10]

Extrinsic Data

\[
\begin{align*}
    x'_{1}, y'_{1} \\
    x'_{2}, y'_{2} \\
    \vdots \\
    x'_{n}, y'_{n}
\end{align*}
\]
Augmented Loss Perceptron

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y) \]

if \( y_i \neq \hat{y}_i \) then
\[ \theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i) \]

Extrinsic Data

\[ x'_1, y'_1 \]
\[ x'_2, y'_2 \]
\[ \vdots \]
\[ x'_n, y'_n \]

Similar to [Chang et al. ‘08, McAllester et al. ‘10]
Augmented Loss Perceptron

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y) \]

if \( y_i \neq \hat{y}_i \) then

\[ \theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i) \]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[
\hat{y}_i = \arg \max_{y \in \mathcal{Y}_{x_i}} \theta \cdot \phi(x_i, y)
\]

if \( y_i \neq \hat{y}_i \) then
\[
\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)
\]

\[
\hat{y}_i = \arg \max_{y \in \mathcal{Y}_{x'_i}} \theta \cdot \phi(x'_i, y)
\]

\[
y^*_i = \arg \min_{y \in \text{k-best-parse}(x'_i)} \text{loss}(y, y'_i)
\]

if \( \text{loss}(y^*_i, y'_i) < \text{loss}(\hat{y}_i, y'_i) \) then
\[
\theta^t = \theta^{t-1} + \phi(y^*_i) - \phi(\hat{y}_i)
\]

Extrinsic Data

\[
X'_1, y'_1
\]
\[
X'_2, y'_2
\]
\[
\vdots
\]
\[
X'_n, y'_n
\]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Augmented Loss Perceptron

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_{x_i}} \theta \cdot \phi(x_i, y) \]

if \( y_i \neq \hat{y}_i \) then
\[ \theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i) \]

\[ \hat{y}_i = \arg \max_{y \in \mathcal{Y}_{x'_i}} \theta \cdot \phi(x'_i, y) \]

\[ y^*_i = \arg \min_{y \in \text{k-best-parse}(x'_i)} \text{loss}(y, y'_i) \]

if \( \text{loss}(y^*_i, y'_i) < \text{loss}(\hat{y}_i, y'_i) \) then
\[ \theta^t = \theta^{t-1} + \phi(y^*_i) - \phi(\hat{y}_i) \]

Similar to [Chang et al. ’08, McAllester et al. ’10]
Example MT-Reordering

"Clean" experiment

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>1/2x extrinsic</th>
<th>1x extrinsic</th>
<th>2x extrinsic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>35.29</td>
<td>38.71</td>
<td>39.02</td>
<td>39.58</td>
</tr>
<tr>
<td>Reorder</td>
<td>76.49</td>
<td>78.19</td>
<td>78.39</td>
<td>78.67</td>
</tr>
</tbody>
</table>

Production: 0.786 -> 0.792 enja fuzzy score

On top of targeted up/self-training [Katz-Brown et al. ’11]
Syntactic Transfer

- Learn parsers for resource-poor languages from resource-rich languages
- Hwa et al. 2005 and earlier

English Treebank

- John [NOUN]
- likes [VERB]
- Mary [NOUN]

Dictionaries:

- English Treebank
- Thai
- Tagalog
- Vietnamese
Confidence Estimation
Confidence Estimation

- Sometimes it is possible to say: “I don’t know”

- Ten blue links are better than triggering an incorrect answer

- The (English) web is redundant:
  - Skip examples with low confidence predictions
  - Discount low-confidence contradicting predictions
In Summary

• Efficiency:
  • Exploit problem structure and domain knowledge
  • Train models specifically for pruning

• Adaptation:
  • Use unsupervised data
  • Use indirect signals
  • “Manual intervention”

• Confidence:
  • ?
Learning from Indirect Signals
Learning from Indirect Signals

Labeled Data
Learning from Indirect Signals

Labeled Data

Unlabeled Data
Learning from Indirect Signals

Structured Data

- The Knowledge Graph
- Wiktionary (The free dictionary)
- Ewok

Unlabeled Data

Labeled Data

Urban Dictionary
Learning from Indirect Signals

Structured Data

Labeled Data

Unlabeled Data

Learning & Modeling
Hypothesis

- Understanding arises from machine learning of relationships implicit in web content and use
  - Some expert annotation may be needed to start
  - Most evidence is not explicitly annotated: text “in the wild”
  - Aggregate information from multiple unstructured sources into a structured “knowledge base”
  - Exploit user interactions and implicit user feedback

[Based on slide from Fernando Pereira]
Thank you!

slav@google.com