Transfer learning in language

Hal Daumé III

Computer Science
University of Maryland

me@hal3.name

IWSML
Kyoto, Japan

31 Mar 2012
Linguistic ambiguity

- Teacher Strikes Idle Kids
- Enraged Cow Injures Farmer With Ax
- I saw the Grand Canyon flying to New York
- Dog collar vs. flea collar
- Plastic cat food can cover
- The BUG in the room …
  - … flew out the window
  - … was planted by spies
- Everyone on the island speaks two languages
Typical NLP pipeline

Source Words

The man ate a sandwich

Target Words

The man eat+ a sandwich

Interlingua

Morphology
Tagging
Parsing
Role labeling
Interpretation

Source Semantics

∃ a ∃ t ∃ e
man(a) & sandwich(t) & eat(e,a,t) & past(e)

Target Semantics

Shallowmantics

Source Syntax

VP

Theme

Source Morphology

Agent

Target Morphology

Analysis

Generation

Source Words

Target Words
Typical NLP pipeline

These tasks are **highly related!**
Pipeline models break down (sorta)

- Tagging + Parsing  
  + 0% / +3%

- Parsing + Named Entities  
  +0.5% / +4%

- Parsing + Role Identification  
  +0% / -0.3%  
  (upper bound: +13%)

- Named Entities + Coreference  
  +0.3% / +1.3%  
  (upper bound: +8%)

Why? Maybe simpler model already has a lot of the fancier information

[Finkel & Manning; ACL 09  
Sutton & McCallum; NAACL 07  
Daumé III & Marcu; EMNLP 06  
Many others...]

Hal Daumé III (me@hal3.name)  
Transfer learning in language
This talk is about...

1. Joint Parsing and Entity Recognition

Mark Gales spoke at IWSML

2. Transfer via Multilinguality

3. Transfer from unlabeled data
This talk is about...

1. Joint Parsing and Entity Recognition

2. Transfer via Multilinguality

3. Transfer from unlabeled data
Agreement-based transfer

- Entities are subsequences of NPs
- NNPs are subsequences of entities
Lots of approaches in 2008

- Semi-supervised learning with constraints
  - Force outputs to obey constraints and do self-training
  - [Chang, Ratinov, Rizzolo, Roth; AAAI 2008]

- Co-regularization
  - Encourage learned models to have similar structure
  - [Ganchev, Graca, Blitzer, Taskar; UAI 2008]

- Cross-task Co-training
  - Do self-training only on outputs that obey constraints
  - [Daumé III; EMNLP 2008]
Simple black-box algorithm

- Learn a parser on labeled data
- Learn an entity recognizer on labeled data
- Run both on unlabeled data
- Assume outputs are both correct for any data point that obeys the constraints in outputs space
- Retrain models on original data plus new data
- Rinse and repeat

If the constraints are:
- Correct (true outputs always agree)
- Discriminating (the probability of agreement is at most $1/[4 \(|Y| - 1|^2$ ])

Then this algorithm “works” (in a PAC sense)

[Daumé III; EMNLP 2008]
Black-box results

[Daumé III; EMNLP 2008]
This talk is about...

1. Joint Parsing and Entity Recognition

2. Transfer via Multilinguality

3. Transfer from unlabeled data
Multilinguality as a source of x-fer

The man ate a tasty sandwich

+ 21% on average over 8 languages

English, Dutch, Danish, Swedish, Spanish, Portuguese, Slovene, Chinese

See also: Snyder, Barzilay et al....

[Berg-Kirkpatrick & Klein; ACL10]
[Iwata, Mochihashi & Sawada; ACL10]
Implicational Universals

**English:**
I eat dinner in restaurants.

**French:**
je mange le diner dans les restaurants
I eat the dinner in the restaurants

**Japanese:**
boku-wa bangohan-o resutoran -ni taberu
I -topic dinner -obj restaurants -in eat

**Hindi:**
main raat ka khaana restra mein khaata hoon
I night-of-meal restaurants in eat am

[Daumé III & Campbell; ACL 2007]
Typological Map: VO

[Daumé III & Campbell; ACL 2007]
Typological Map: PreP

[Daumé III & Campbell; ACL 2007]
Unsupervised part of speech tagging

- Seeds (frequent words for each tag)
  - N: membro, milhoes, obras
  - D: as [the,2f] o [the,1m] os [the,2m]
  - V: afector, gasta, juntar
  - P: com, como, de, em

- Typological rules:
  - Art ← Noun
  - Prp → Noun

- Tag knowledge:
  - Open class
  - Closed class

[Teichert & Daumé III; NIPSWS 2009]
Does typology help?

Can also transfer across languages *Even for typologically distinct ones!*

**Graph:**
- **X-axis:** No Rules, Prp->N, Both
- **Y-axis:** 20, 25, 30, 35, 40, 45, 50, 55

- **Legend:**
  - **Blue:** No Rules, Prp->N, Both
  - **Red:** Art<-N, Prp->N, Both

**References:**
- [Teichert & Daumé III; NIPSWS 2009; Sanders & Daume III, EMNLP 2012 sub]
This talk is about...

1. Joint Parsing and Entity Recognition

Mark Gales spoke at IWSML

2. Transfer via Multilinguality

3. Transfer from unlabeled data
Spectral Clustering

- Represent datapoints as the vertices $V$ of a graph $G$.
- All pairs of vertices are connected by an edge $E$.
- Edges have weights $W$.
- Large weights mean that the adjacent vertices are very similar; small weights imply dissimilarity.
Graph partitioning

- Clustering on a graph is equivalent to partitioning the vertices of the graph.
- A loss function for a partition of $V$ into sets $A$ and $B$
  \[ \text{cut}(A, B) = \sum_{u \in A, v \in B} W_{u,v} \]
- In a good partition, vertices in different partitions will be dissimilar.
- Mincut criterion: Find partition $A, B$ that minimizes $\text{cut}(A, B)$
Graph partitioning

- Mincut criterion ignores the size of the subgraphs formed.

- Normalized cut criterion favors balanced partitions.

\[ N\text{cut}(A, B) = \frac{\text{cut}(A, B)}{\sum_{u \in A, v \in V} W_{u,v}} + \frac{\text{cut}(A, B)}{\sum_{u \in B, v \in V} W_{u,v}} \]

- Minimizing the normalized cut criterion exactly is NP-hard.
Spectral Clustering

- One way of approximately optimizing the normalized cut criterion leads to spectral clustering.

- Spectral clustering
  - Find a new representation of the original data points.
  - Cluster the points in this representation using any clustering scheme (say 2-means).

- The representation involves forming the row-normalized matrix $Y$ using the largest 2 eigenvectors of the matrix $L$

$$D = \text{diag}(W1) \quad \text{and} \quad L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$W_{uv} = \exp(- \| s_u - s_j \|^2 / (2\sigma^2))$$
Example: 2-means
Example: Spectral clustering
Multiview spectral clustering
Multiview spectral clustering

Algorithm
1. Run SVD on each view
2. Project each view onto subspace spanned by other's top-left sv's
3. Goto 1 unless converged

Look ma: no hyperparameters!
Multiview spectral clustering

Algorithm
1. Run SVD on each view
2. Project each view onto subspace spanned by other’s top sv’s
3. Goto 1 unless converged

Look ma: no hyperparameters!

Results (Reuters)

<table>
<thead>
<tr>
<th></th>
<th>F-score</th>
<th>Norm. MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best View</td>
<td>0.342</td>
<td>0.287</td>
</tr>
<tr>
<td>Concat</td>
<td>0.368</td>
<td>0.298</td>
</tr>
<tr>
<td>SofA</td>
<td>0.381</td>
<td>0.342</td>
</tr>
<tr>
<td>Co-Spec</td>
<td>0.412</td>
<td>0.388</td>
</tr>
</tbody>
</table>
This talk is about...

1. Joint Parsing and Entity Recognition

2. Transfer via Multilinguality

3. Transfer from unlabeled data
This talk is about...

Simple algorithms can achieve great transfer

Mark Gales spoke at IWSML

Person VP PP Event

2. Transfer via Multilinguality

3. Transfer from unlabeled data
This talk is about...

Simple algorithms can achieve great transfer

Plentiful multilingual data + knowledge = strong models

1. Joint Parsing and Entity Recognition
2. Transfer via Multilinguality
3. Transfer from unlabeled data

Mark Gales spoke at IWSML

3. Transfer from unlabeled data
This talk is about...

Simple algorithms can achieve great transfer

Mark Gales spoke at IWSML

Plentiful multilingual data + knowledge = strong models

3. Transfer

Unlabeled (paired) data can be exploited efficiently
This talk is about...

1. Joint Parsing and Entity Recognition
2. Transfer via Multilinguality
3. Transfer from unlabeled data

Simple algorithms can achieve great transfer

Plentiful multilingual data + knowledge = strong models

When will transfer help?
Has transfer helped?
How to incorporate knowledge?
Scaling to billions of examples?

Unlabeled (paired) data can be exploited efficiently
This talk is about...

Simple algorithms can achieve great transfer

Plentiful multilingual data + knowledge = strong models

When will transfer will help?
Has transfer helped?
How to incorporate knowledge?
Scaling to billions of examples?

Unlabeled (paired) data can be exploited efficiently

ありがとうございます！質問は？