Unsupervised Approaches to Sequence Tagging, Morphology Induction, and Lexical Resource Acquisition

Reza Bosaghzadeh & Nathan Schneider

LS2 ~ 1 December 2008
Unsupervised Methods

- Sequence Labeling (Part-of-Speech Tagging)
  - pronoun
  - verb
  - preposition
  - det
  - noun
  - adverb
  - She
  - ran
  - to
  - the
  - station
  - quickly

- Morphology Induction

  un-supervised learning

- Lexical Resource Acquisition
Contrastive Estimation
Smith & Eisner (2005)

• Already discussed in class

• Key idea: *exploits implicit negative evidence*
  – Mutating training examples often gives ungrammatical (negative) sentences
  – During training, shift probability mass from generated negative examples to given positive examples

• BUT: Requires a *tagging dictionary*, i.e. a list of possible tags for each word type
Prototype-driven tagging
Haghighi & Klein (2006)

Unlabeled Data + Prototype List → Annotated Data

slide courtesy Haghighi & Klein
Prototype-driven tagging
Haghighi & Klein (2006)

Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage. No dogs allowed.

Prototype List

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>president</td>
<td>IN</td>
<td>of</td>
<td></td>
</tr>
<tr>
<td>VBD</td>
<td>said</td>
<td>NNS</td>
<td>shares</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>and</td>
<td>TO</td>
<td>to</td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td>Mr.</td>
<td>PUNC</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>JJ</td>
<td>new</td>
<td>CD</td>
<td>million</td>
<td></td>
</tr>
<tr>
<td>DET</td>
<td>the</td>
<td>VBP</td>
<td>are</td>
<td></td>
</tr>
</tbody>
</table>

slide courtesy Haghighi & Klein
Prototypes

Information Extraction: Classified Ads

Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage. No dogs allowed.

Prototype List

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>kitchen, laundry</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCATION</td>
<td>near, close</td>
</tr>
<tr>
<td>TERMS</td>
<td>paid, utilities</td>
</tr>
<tr>
<td>SIZE</td>
<td>large, feet</td>
</tr>
<tr>
<td>RESTRICT</td>
<td>cat, smoking</td>
</tr>
</tbody>
</table>

slide courtesy Haghighi & Klein
Prototype-driven tagging
Haghighi & Klein (2006)

- Trigram tagger, same features as (Smith & Eisner 2005)
  - Word type, suffixes up to length 3, contains-hyphen, contains-digit, initial capitalization
- Tie each word to its most similar prototype, using context-based similarity technique (Schütze 1993)
  - SVD dimensionality reduction
  - Cosine similarity between context vectors
Prototype-driven tagging
Haghighi & Klein (2006)

Pros
• Doesn’t require tagging dictionary

Cons
• Still need a tag set
• May be hard to choose *good* prototypes
Unsupervised POS tagging
The State of the Art

Unsupervised English POS Tagging
24K tokens  48K tokens (2K sen.)

Haghighi and Klein (2006)
- Baseline (trigram), Treebank tagset: 42.4%
- Prototype-augmented, Treebank tagset: 79.1%
- Prototype-augmented, reduced tagset: 82.2%

Smith and Eisner (2005)
- CE, with 2125-entry tagging dictionary: 79.5%
- CE, with 3362-entry tagging dictionary: 88.1%
- CE, with 5406-entry tagging dictionary: 90.4%

Toutanova and Johnson (2008)
- Latent Dirichlet Allocation: 93.4%

Best supervised result (CRF): 99.5%!
Unsupervised Methods

– Sequence Labeling (Part-of-Speech Tagging)

- pronoun
- verb
- preposition
- det
- noun
- adverb

She ran to the station quickly

– Morphology Induction

un-supervised learning

– Lexical Resource Acquisition

Diagram showing relationships between words and categories.
Unsupervised Approaches to Morphology

• Morphology refers to the internal structure of words
  – A **morpheme** is a minimal meaningful linguistic unit
  – **Morpheme segmentation** is the process of dividing words into their component morphemes
  ![un-supervised-learning-of-natural-language](un-supervised-learning-of-natural-language)
  – **Word segmentation** is the process of finding word boundaries in a stream of speech or text
    ![unsupervised-learning_of_natural_language](unsupervised_learning_of_natural_language)
ParaMor: Morphological paradigms

Monson et al. (2007, 2008)

• Learns inflectional paradigms from raw text
  – Requires only a list of word types from a corpus
  – Looks at word counts of substrings, and proposes (stem, suffix) pairings based on type frequency

• 3-stage algorithm
  – *Stage 1*: Candidate paradigms based on frequencies
  – *Stages 2-3*: Refinement of paradigm set via merging and filtering

• Paradigms can be used for morpheme segmentation or stemming
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

<table>
<thead>
<tr>
<th>speak</th>
<th>dance</th>
<th>buy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
<td>comprar</td>
</tr>
<tr>
<td>hablo</td>
<td>bailo</td>
<td>compro</td>
</tr>
<tr>
<td>hablamos</td>
<td>bailamos</td>
<td>compramos</td>
</tr>
<tr>
<td>hablan</td>
<td>bailan</td>
<td>compran</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- A sampling of Spanish verb conjugations (inflections)
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

<table>
<thead>
<tr>
<th>speak</th>
<th>dance</th>
<th>buy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
<td>comprar</td>
</tr>
<tr>
<td>hablo</td>
<td>bailo</td>
<td>compra</td>
</tr>
<tr>
<td>hablamos</td>
<td>bailamos</td>
<td>compramos</td>
</tr>
<tr>
<td>hablan</td>
<td>bailan</td>
<td>compran</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• A proposed paradigm (correct): stems \{habl, bail, compr\} and suffixes \{-ar, -o, -amos, -an\}
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

• Two subsequent stages:
  – **Filtering** out spurious paradigms (e.g. with incorrect segmentations)
  – **Merging** partial paradigms to overcome sparsity: smoothing
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

\[
\begin{array}{ccc}
\text{speak} & \text{dance} \\
\text{hablar} & \text{bailar} \\
\text{hablo} & \text{bailo} \\
\text{hablamos} & \text{bailamos} \\
\text{hablan} & \text{bailan} \\
\ldots & \ldots \\
\end{array}
\]

• For certain sub-sets of verbs, the algorithm may propose paradigms with spurious segmentations, like the one at left

• The \textbf{filtering} stage of the algorithm weeds out these incorrect paradigms
ParaMor: Morphological paradigms  
Monson et al. (2007, 2008)

<table>
<thead>
<tr>
<th></th>
<th>speak</th>
<th>dance</th>
<th>buy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
<td>comprar</td>
<td></td>
</tr>
<tr>
<td>bailo</td>
<td>bailamos</td>
<td>compra</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hablan</td>
<td>compramos</td>
<td></td>
</tr>
<tr>
<td>hablan</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- What if not all conjugations were in the corpus?
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

<table>
<thead>
<tr>
<th><em>speak</em></th>
<th><em>dance</em></th>
<th><em>buy</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
<td>comprar</td>
</tr>
<tr>
<td>bailo</td>
<td>bailamos</td>
<td>compromo</td>
</tr>
<tr>
<td>hablamos</td>
<td>bailamos</td>
<td>compramos</td>
</tr>
<tr>
<td>hablan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Another stage of the algorithm **merges** these overlapping partial paradigms via clustering.
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

<table>
<thead>
<tr>
<th>speak</th>
<th>dance</th>
<th>buy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
<td>comprar</td>
</tr>
<tr>
<td>habló</td>
<td>bailó</td>
<td>compro</td>
</tr>
<tr>
<td>hablamos</td>
<td>bailamos</td>
<td>compramos</td>
</tr>
<tr>
<td>hablan</td>
<td>bailan</td>
<td>comproan</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- This amounts to smoothing, or “hallucinating” out-of-vocabulary items
ParaMor: Morphological paradigms
Monson et al. (2007, 2008)

• Heuristic-based, deterministic algorithm can learn inflectional paradigms from raw text
• Currently, ParaMor assumes suffix-based morphology
• Paradigms can be used straightforwardly to predict segmentations
  – Combining the outputs of ParaMor and Morfessor (another system) won the segmentation task at MorphoChallenge 2008 for every language: English, Arabic, Turkish, German, and Finnish
Bayesian word segmentation
Goldwater et al. (2006; in submission)

• Word segmentation results – comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>BP</th>
<th>BR</th>
<th>BF</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS-u</td>
<td>67.7</td>
<td>70.2</td>
<td>68.9</td>
<td>80.6</td>
<td>84.8</td>
<td>82.6</td>
<td>52.9</td>
<td>51.3</td>
<td>52.0</td>
</tr>
<tr>
<td>MBDP-1</td>
<td>67.0</td>
<td>69.4</td>
<td>68.2</td>
<td>80.3</td>
<td>84.3</td>
<td>82.3</td>
<td>53.6</td>
<td>51.3</td>
<td>52.4</td>
</tr>
<tr>
<td>DP</td>
<td>61.9</td>
<td>47.6</td>
<td>53.8</td>
<td>92.4</td>
<td>62.2</td>
<td>74.3</td>
<td>57.0</td>
<td>57.5</td>
<td>57.2</td>
</tr>
<tr>
<td>NGS-b</td>
<td>68.1</td>
<td>68.6</td>
<td>68.3</td>
<td>81.7</td>
<td>82.5</td>
<td>82.1</td>
<td>54.5</td>
<td>57.0</td>
<td>55.7</td>
</tr>
<tr>
<td>HDP</td>
<td>75.2</td>
<td>69.6</td>
<td>72.3</td>
<td>90.3</td>
<td>80.8</td>
<td>85.2</td>
<td>63.5</td>
<td>55.2</td>
<td>59.1</td>
</tr>
</tbody>
</table>

Goldwater et al. Unigram DP
Goldwater et al. Bigram HDP

• See Narges & Andreas’s presentation for more on this model

(table from Goldwater et al. (in submission))
Multilingual morpheme segmentation Snyder & Barzilay (2008)

- **speak** ES  **speak** FR
  - hablar  parler
  - hablo  parle
  - hablamos  parlons
  - hablan  parlent
  - ...
  - ...

- **Abstract morphemes** cross languages: (ar, er), (o, e), (amos, ons), (an, ent), (habl, parl)

- Considers **parallel phrases** and tries to find morpheme correspondences

- **Stray morphemes** don’t correspond across languages
Morphology Papers: Inputs & Outputs

<table>
<thead>
<tr>
<th>MORPHOLOGY</th>
<th>Monson et al.</th>
<th>Goldwater et al.</th>
<th>Snyder &amp; Barzilay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase/Document-Level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsegmented text</td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Parallel sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phrasal aligner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word-Level</td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Vocabulary (list of word types)</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Word-Level</td>
<td>●</td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>Paradigms</td>
<td>●</td>
<td></td>
<td>(●)</td>
</tr>
<tr>
<td>Segmentations</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonetic correspondences</td>
<td>●</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend
- ●: Training
- ○: Test
- ▽: Input
- △: Output

• What does “unsupervised” mean for each approach?
Unsupervised Methods

– Sequence Labeling (Part-of-Speech Tagging)

- pronoun
- verb
- preposition
- det
- noun
- adverb

- She
- ran
- to
- the
- station
- quickly

– Morphology Induction

un-supervise-d learn-ing

– Lexical Resource Acquisition
Bilingual lexicons from monolingual corpora

Haghighi et al. (2008)

Source Words

- state
- world
- name
- nation

Target Words

- estado
- nombre
- política
- mundo

Matching

\[ m \]

Used a variant of CCA (Canonical Correlation Analysis)

diagram courtesy Haghighi et al.
Bilingual Lexicons from Monolingual Corpora Haghighi et al. (2008)

Data Representation

Orthographic Features

<table>
<thead>
<tr>
<th></th>
<th>Source Text</th>
<th>Target Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>#st</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>tat</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>te#</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Context Features

<table>
<thead>
<tr>
<th></th>
<th>Source Text</th>
<th>Target Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>world</td>
<td>20.0</td>
<td></td>
</tr>
<tr>
<td>politics</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>society</td>
<td>10.0</td>
<td></td>
</tr>
</tbody>
</table>

Orthographic Features

<table>
<thead>
<tr>
<th></th>
<th>Source Text</th>
<th>Target Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>#es</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>sta</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>do#</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Context Features

<table>
<thead>
<tr>
<th></th>
<th>Source Text</th>
<th>Target Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>mundo</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>politica</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>sociedad</td>
<td>6.0</td>
<td></td>
</tr>
</tbody>
</table>

slide courtesy Haghighi et al.
Feature Experiments

- MCCA: Orthographic and context features

4k EN-ES Wikipedia Articles

slide courtesy Haghighi et al.
Narrative events
Chambers & Jurafsky (2008)

• Given a corpus, identifies related events that constitute a “narrative” and (when possible) predict their typical temporal ordering
  – E.g.: CRIMINAL PROSECUTION narrative, with verbs: arrest, accuse, plead, testify, acquit/convict

• Key insight: related events tend to share a participant in a document
  – The common participant may fill different syntactic/semantic roles with respect to verbs: arrest.OBJECT, accuse.OBJECT, plead.SPECIFIC
Narrative events
Chambers & Jurafsky (2008)

- A temporal classifier can reconstruct pairwise canonical event orderings, producing a directed graph for each narrative.
Statistical verb lexicon
Grenager & Manning (2006)

• From dependency parses, a generative model predicts for each verb:

  – PropBank-style semantic roles: ARG0, ARG1, etc. (do not necessarily correspond across verbs)

  – The roles’ syntactic realizations, e.g.:

    | He         | gave  | me    | a cookie       |
    | subj      | verb  | np#1  | np#2           |
    | ARG0      | give  | ARG2  | ARG1           |

    | He         | gave  | a cookie | to me          |
    | subj      | verb  | np#2   | pp_to          |
    | ARG0      | give  | ARG1   | ARG2           |

• Used for semantic role labeling
“Semanticity”: Our proposed scale of semantic richness

- text < POS < syntax/morphology/alignments < coreference/semantic roles/temporal ordering < translations/narrative event sequences
- We score each model’s inputs and outputs on this scale, and call the input-to-output increase “semantic gain”
  - Haghighi et al.’s bilingual lexicon induction wins in this respect, going from raw text to lexical translations
Semantic Gain: Comparison of Methods

<table>
<thead>
<tr>
<th></th>
<th>Sequences/POS</th>
<th>Morphology</th>
<th>Lexical Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S&amp;E</td>
<td>H&amp;K</td>
<td>M+</td>
</tr>
<tr>
<td>Input semanticity</td>
<td>.75</td>
<td>.25</td>
<td>0</td>
</tr>
<tr>
<td>Output semanticity</td>
<td>1.5</td>
<td>1.5</td>
<td>1.75</td>
</tr>
<tr>
<td>Semantic gain</td>
<td>.75</td>
<td>1.25</td>
<td>1.75</td>
</tr>
</tbody>
</table>
Robustness to language variation

• About half of the papers we examined had English-only evaluations

• We considered which techniques were most adaptable to other (esp. resource-poor) languages. Two main factors:
  – Reliance on existing tools/resources for preprocessing (parsers, coreference resolvers, ...)
  – Any linguistic specificity in the model (e.g. suffix-based morphology)
Summary

We examined three areas of unsupervised NLP:

1. **Sequence tagging**: How can we predict POS (or topic) tags for words in sequence?

2. **Morphology**: How are words put together from morphemes (and how can we break them apart)?

3. **Lexical resources**: How can we identify lexical translations, semantic roles and argument frames, or narrative event sequences from text?

In eight recent papers we found a variety of approaches, including heuristic algorithms, Bayesian methods, and EM-style techniques.
Thanks to Noah and Kevin for their feedback on the paper; Andreas and Narges for their collaboration on the presentations; and all of you for giving us your attention!

Questions?

<table>
<thead>
<tr>
<th>Target Label</th>
<th>Prototypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hablar</td>
<td>bailar</td>
</tr>
<tr>
<td>hablo</td>
<td>bailo</td>
</tr>
<tr>
<td>hablamos</td>
<td>bailamos</td>
</tr>
<tr>
<td>hablan</td>
<td>bailan</td>
</tr>
</tbody>
</table>
Improvement Ideas

• POS Tagging: Learn the tag set
• Morphology: Non-agglomerative Morphology, Also parses
• Lexical Resources: Try word classes
• All: Language variability