Two Decades of Unsupervised POS Induction: How far have we come?

Christos Christodoulopoulos
Sharon Goldwater
Mark Steedman

School of Informatics
University of Edinburgh

EMNLP, 2010
Why unsupervised POS induction?

- useful pre-processing task in low-density languages 
  - e.g. our current research:
    - Bible Corpus: parallel raw text in 56 Languages
    - 14 languages have <1M speakers (3 extinct)
- use systems out of the box (no parameter tuning)
Why unsupervised POS induction?

- useful pre-processing task in low-density languages
  - e.g. our current research:
    - Bible Corpus: parallel raw text in 56 Languages
    - 14 languages have <1M speakers (3 extinct)
  - use systems out of the box (no parameter tuning)

Problem:

No comprehensive comparison of POS induction systems
Why unsupervised POS induction?

- useful pre-processing task in low-density languages
  - e.g. our current research:
    - Bible Corpus: parallel raw text in 56 Languages
    - 14 languages have <1M speakers (3 extinct)
  - use systems out of the box (no parameter tuning)

Problem:
No consensus on evaluation measures
No comprehensive comparison of POS induction systems
What evaluation measures to use?

Which system is the best?
  - especially on non-English languages
What evaluation measures to use?
- Compare 7 measures (4 in talk), argue for V-measure

Which system is the best?
- especially on non-English languages
- Older systems work as well or better than newer ones
- Systems with morphology work better
Overview

- What evaluation measures to use?
  - Compare 7 measures (4 in talk), argue for V-measure
- Which system is the best?
  - especially on non-English languages
  - Older systems work as well or better than newer ones
  - Systems with morphology work better
- How can we move forward?
  - Preliminary results using prototype-based system
Evaluation Measures

What do we want from evaluation measures?

- Intuitive (analysis and results)
- Useful across different # of clusters
- matching vs. entropy-based

Measures considered

- [many-to-1] Many-to-one accuracy
- [1-to-1] One-to-one accuracy
- [vi] Variation of Information
- [vm] V-Measure
[many-to-1]
- Match each cluster to most frequent gold-standard tag
- problem: score increases with # of clusters

[1-to-1]
- Each tag can be used only once
- problem: score decreases with # of clusters
**Entropy-based measures**

**[vi]** Variation of Information (Meilă, 2003)

- Accounts for entropy of clusters, not just matched parts.

\[
\text{VI}(C,T) = H(T|C) + H(C|T)
\]

- Problem: measured in bits → non-intuitive, non-normalized
Entropy-based Measures

[Vm] V-Measure (Rosenberg, 2007)

A

\[ V, V, V, V, N, A \]
\[ N, N, N, N, V, A \]
\[ A, A, A, A, N, V \]
\[ N, N, N, A, V \]

B

\[ V, V, V, V, N, N \]
\[ N, N, N, N, A, A \]
\[ A, A, A, A, N, V \]
\[ N, N, N, A, A \]

\[ c = 1 - \frac{H(C|T)}{H(C)} \]

Homogeneity

each cluster should contain as few tags as possible
Entropy-based Measures

[Vm] V-Measure (Rosenberg, 2007)

\[ h = 1 - \frac{H(T|C)}{H(T)} \]

\[ c = 1 - \frac{H(C|T)}{H(C)} \]

Completeness: each tag should be contained in as few clusters as possible.
POS Induction Systems

- [brown] (Brown et al., 1992)
- [clark] (Clark, 2003)
- [cw] (Biemann, 2006)
- [bhmm] (Goldwater and Griffiths, 2007)
- [vbhmm] (Johnson, 2007)
- [pr] (Graça et al., 2008)
- [feat] (Berg-Kirkpatrick et al., 2010)
Class-based n-gram models

[brown] (Brown et al., 1992)
- Similar to HMM, but type-based
- Heuristic algorithm to maximize likelihood

[clark] (Clark, 2003)
- Similar to brown, but add morphology (letter HMM)
Chinese Whispers

[cw] (Biemann, 2006)

- Graph-based clustering algorithm
- Unlike other systems, # of clusters is induced
Bayesian HMMs

[Bhmm] (Goldwater and Griffiths, 2007)
- Bigram HMM with Dirichlet priors
- Gibbs sampling for inference

[Vbhmm] (Johnson, 2007)
- Bigram HMM with Dirichlet priors
- Variational Bayes inference
Other ML methods

[pr] (Graça et al., 2009)
- Same bigram HMM
- Constraints on posteriors for sparsity

[feat] (Berg-Kirkpatrick et al., 2010)
- Log-linear feature based system
- Morphology modeling
[wsj] WSJ portion of Penn Treebank (~45k sentences)
1984 portion of Multext-East (~7k sentences)
   [bg] Bulgarian
   [cs] Czech
   [en] English
   [et] Estonian
   [hu] Hungarian
   [ro] Romanian
   [sl] Slovene
   [sr] Serbian

[wsj-s] A 7k sentence version of the WSJ corpus
System comparison - Multiple languages

V-Measure

brown clark cw bhmm vbhmm pr feat

categories: bg cs et hu ro sl sr en
Recap

- What evaluation measures to use?
- Which system is the best?
- How can we move forward?
Recap

- What evaluation measures to use?
  - V-measure: intuitive and stable over varying cluster sizes
  - Combine with matching measures for comparison with previous work
- Which system is the best?

- How can we move forward?
What evaluation measures to use?
- V-measure: intuitive and stable over varying cluster sizes
- Combine with matching measures for comparison with previous work

Which system is the best?
- Older systems are fast, as good or better than newer systems (but ML approaches are catching up...)
- Morphology helps, esp. on non-English languages
- All systems are worse for non-English

How can we move forward?
What evaluation measures to use?
- V-measure: intuitive and stable over varying cluster sizes
- Combine with matching measures for comparison with previous work

Which system is the best?
- Older systems are fast, as good or better than newer systems (but ML approaches are catching up...)
- Morphology helps, esp. on non-English languages
- All systems are worse for non-English

How can we move forward?
- Preliminary results using prototype-based system
Prototype-driven learning

Haghighi and Klein (2006) system:
- No annotated data but . . .
- Manually created lists of prototypes

JJ [new other last]
DT [The a the]
VB [sell make be]
PRP [they it he]
MD [would could will]

- Log-linear model using similarity features
- 80.5% many-to-one accuracy on WSJ
Prototype-driven learning

Haghighi and Klein (2006) system:
- No annotated data but. . .
- Manually created lists of prototypes

JJ [new other last]
DT [The a the]
VB [sell make be]
PRP [they it he]
MD [would could will]

Log-linear model using similarity features
- 80.5% many-to-one accuracy on WSJ

Can we induce prototypes automatically?
Prototype extraction

Start with POS induction system

- Algorithm chooses up to 10 prototypes that are:
  - most frequent
  - most similar to words in their clusters
  - most dissimilar to words in other clusters

- Similarity computed using SVD
- Parameters tuned on English

Use H&K system with the prototypes
Conclusions

Results from comprehensive comparison of POS induction systems using multiple systems, measures, and languages.

- For evaluation, use (at least) V-measure, and test on multiple languages
- For best results fast, use Clark (or Brown) systems

Prototype-driven POS induction:
- State-of-the-art results in WSJ
- Improvements in other languages
- Searching for prototypes instead of POS clusters
Thank you!

For all the results:
http://homepages.inf.ed.ac.uk/s0787820/pos/
