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

Motivation

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)

Motivation

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

Motivation

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

Overview

- What evaluation measures to use?
- Which system is the best?
  - especially on non-English languages

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

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

Matching-accuracy measures

[many-to-1]

- Match each cluster to most frequent gold-standard tag
- problem: score increases with # of clusters

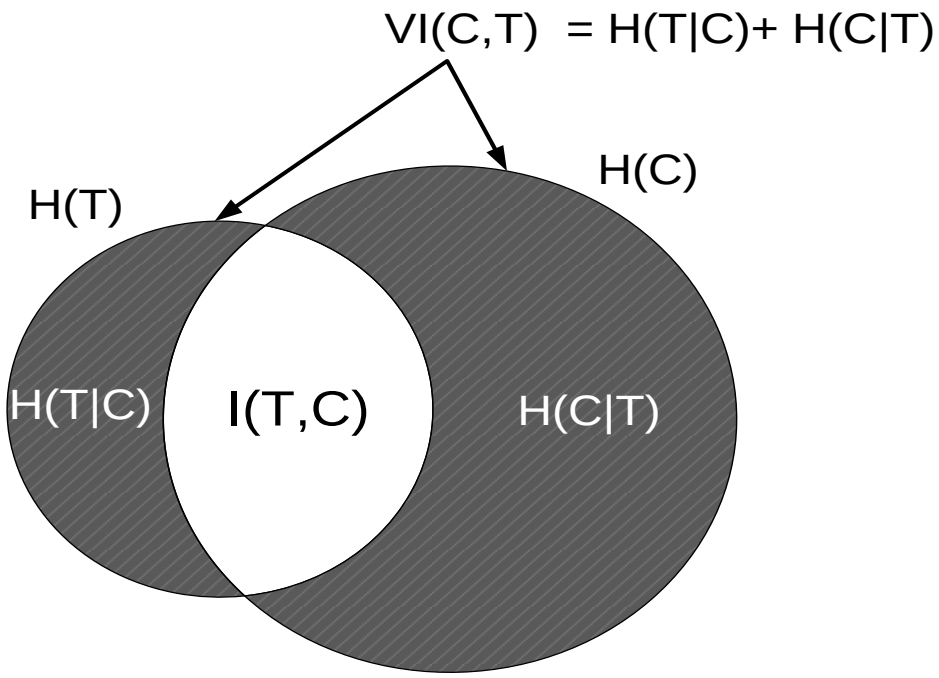
[1-to-1]

- 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.



- problem: measured in bits $\rightarrow$ 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

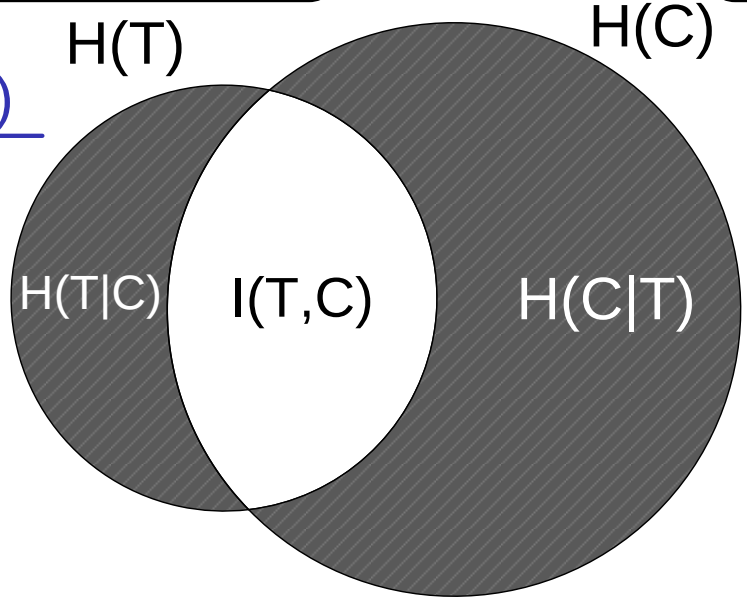
B

V, V, V, V, N, N

N, N, N, N, A, A

A, A, A, A, N, V

N, N, N, A, A



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

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

Homogeneity

each cluster should contain as few tags as possible

Entropy-based Measures

[vm] V-Measure (Rosenberg, 2007)

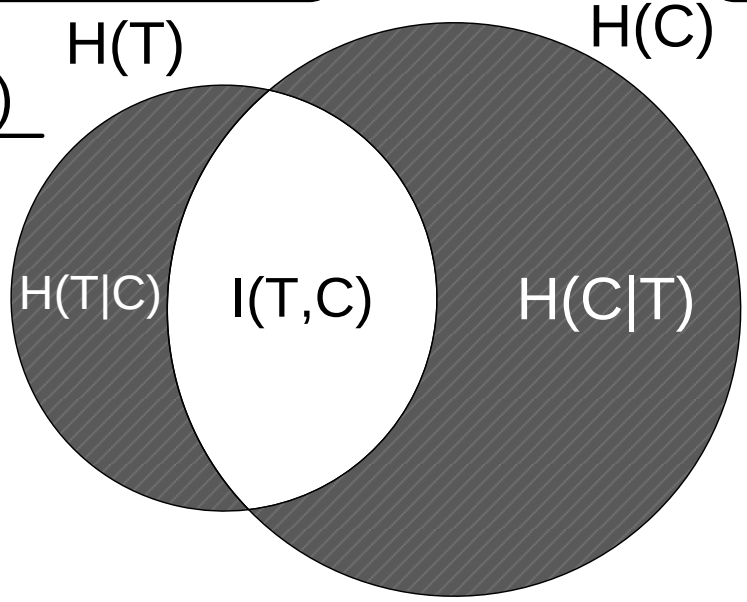
A

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

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

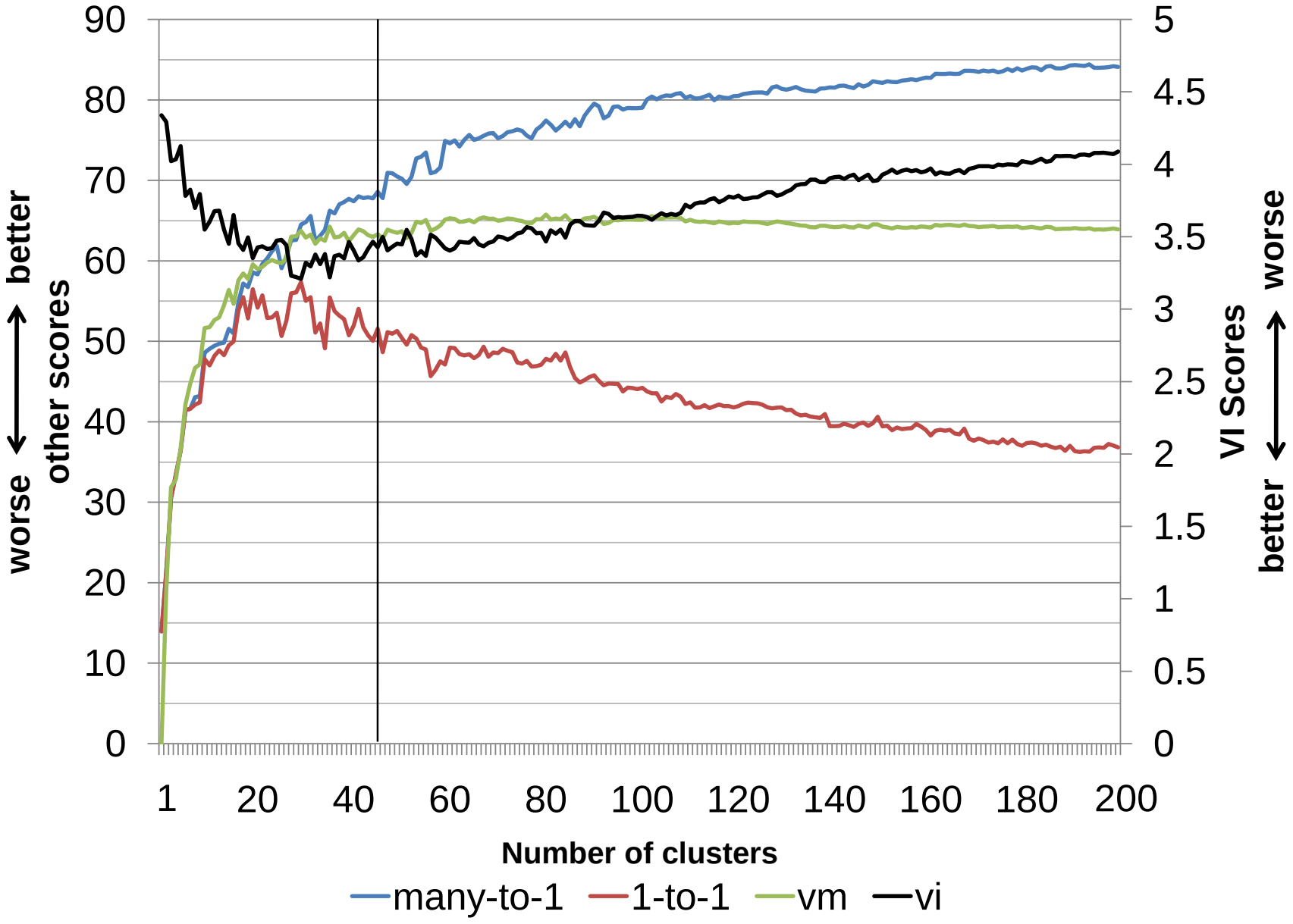


$$c = 1 - \frac{H(C|T)}{H(C)}$$

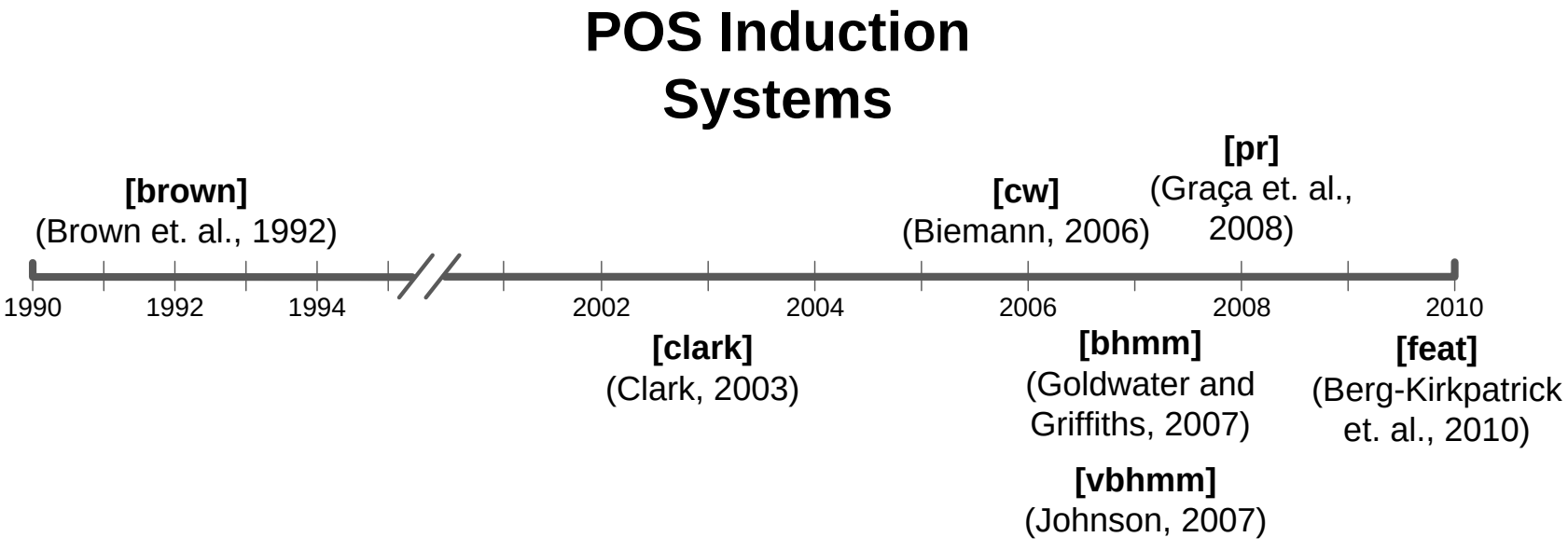
Completeness

each tag should be contained in as few clusters as possible

Evaluation of Measures



The Systems


POS Induction Systems
[brown]
(Brown et. al., 1992)
[cw]
(Biemann, 2006)
[pr]
(Graça et. al., 2008)
1990
1992
1994
2002
2004
2006
2008
2010
[clark]
(Clark, 2003)
[bhmm]
(Goldwater and Griffiths, 2007)
[vbhmm]
(Johnson, 2007)
[feat]
(Berg-Kirkpatrick et. al., 2010)

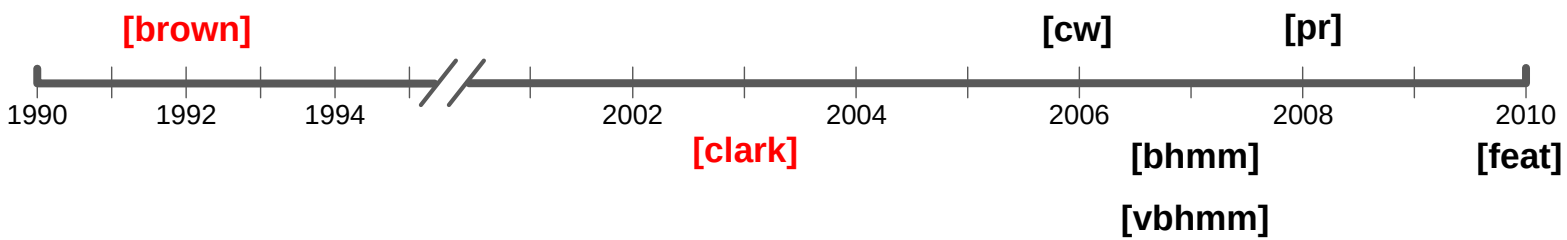
Class-based n-gram models

[brown] (Brown et al., 1992)

- Similar to HMM, but type-based
- Heuristic algorithm to maximize likelihood

[clark] (Clark, 2003)

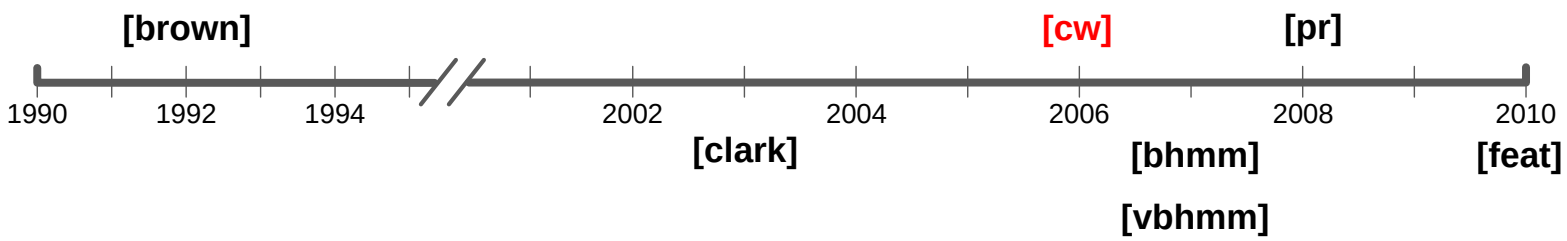
- Similar to **brown**, but add morphology (letter HMM)



Chinese Whispers

[cw] (Biemann, 2006)

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



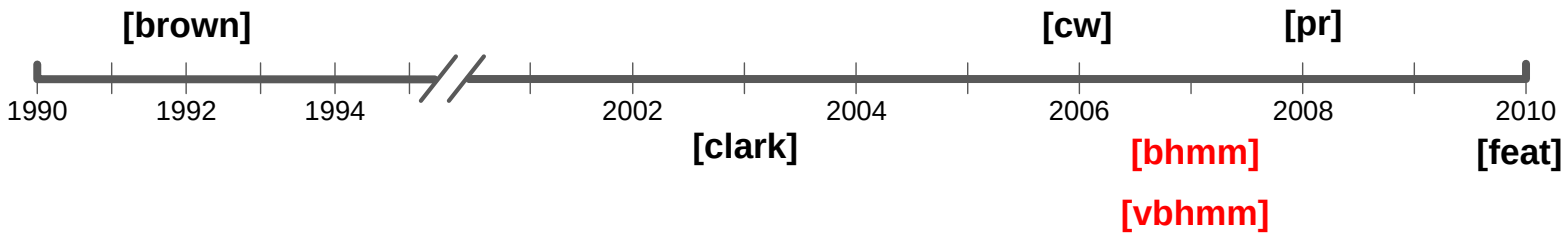
Bayesian HMMs

[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



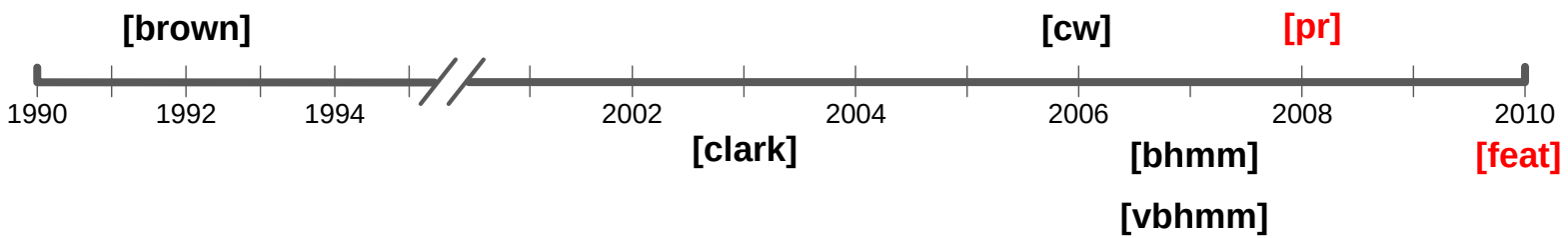
Other ML methods

[pr] (Graça et al., 2009)

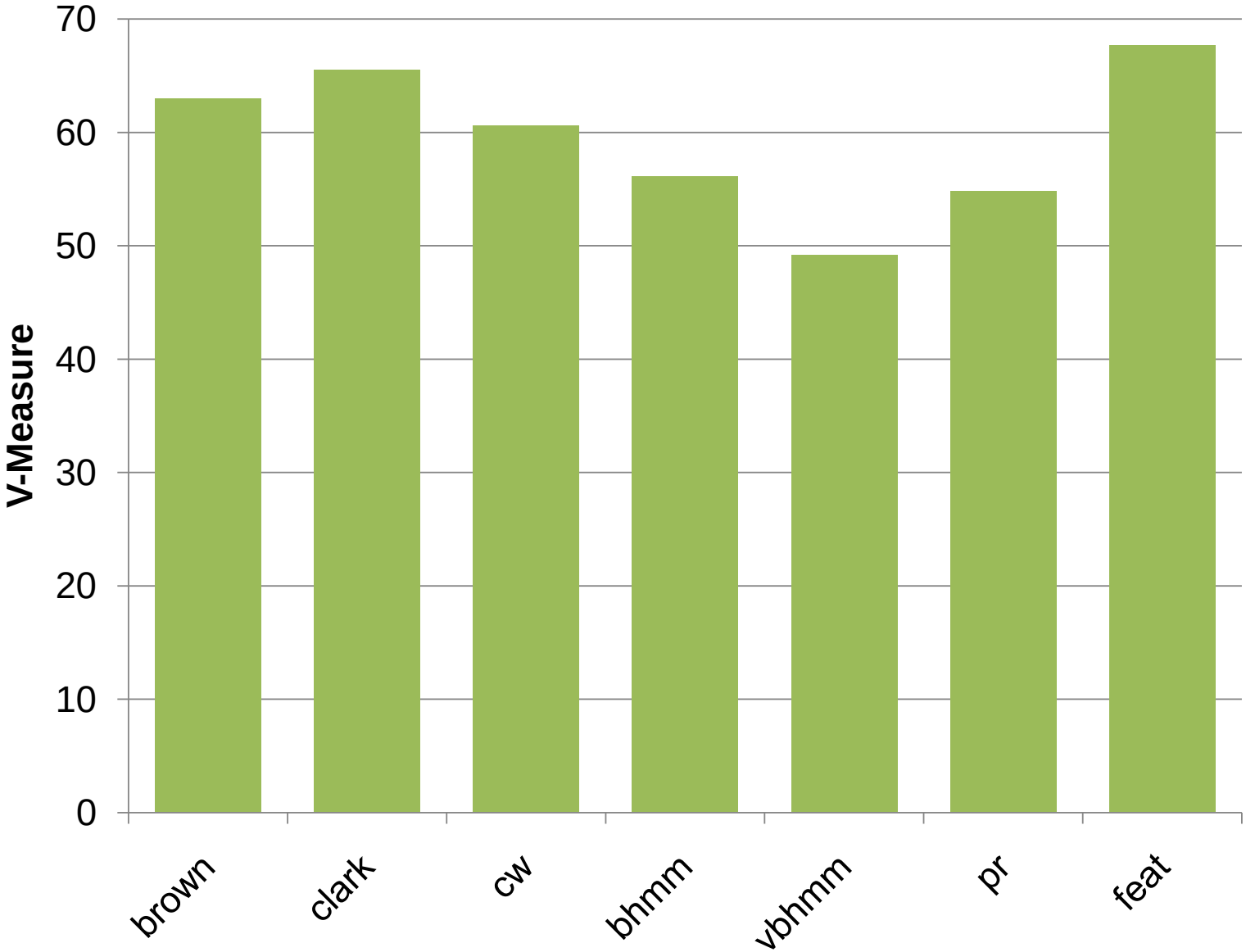
- Same bigram HMM
- Constraints on posteriors for sparsity

[feat] (Berg-Kirkpatrick et al., 2010)

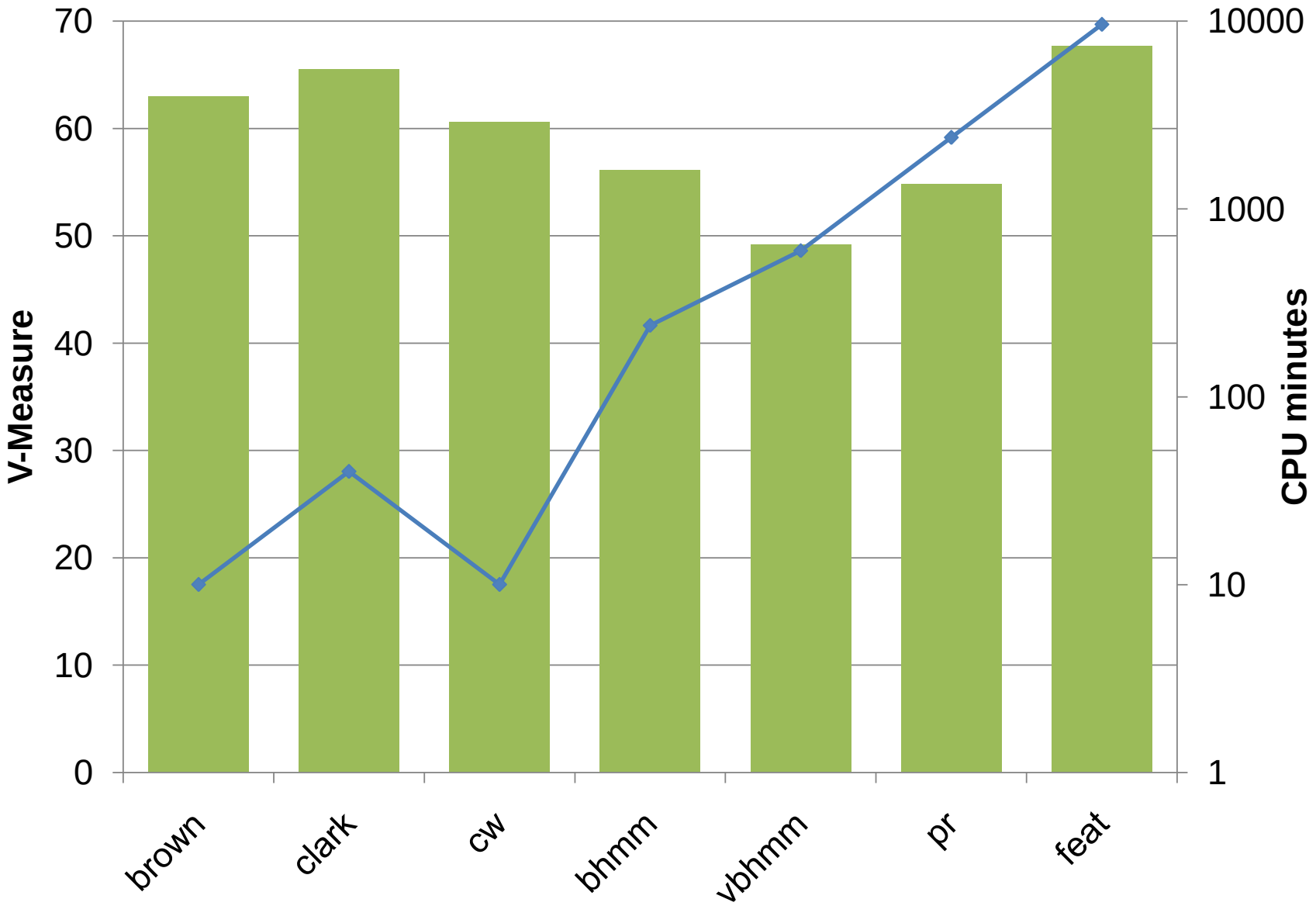
- Log-linear feature based system
- Morphology modeling



System comparison - WSJ



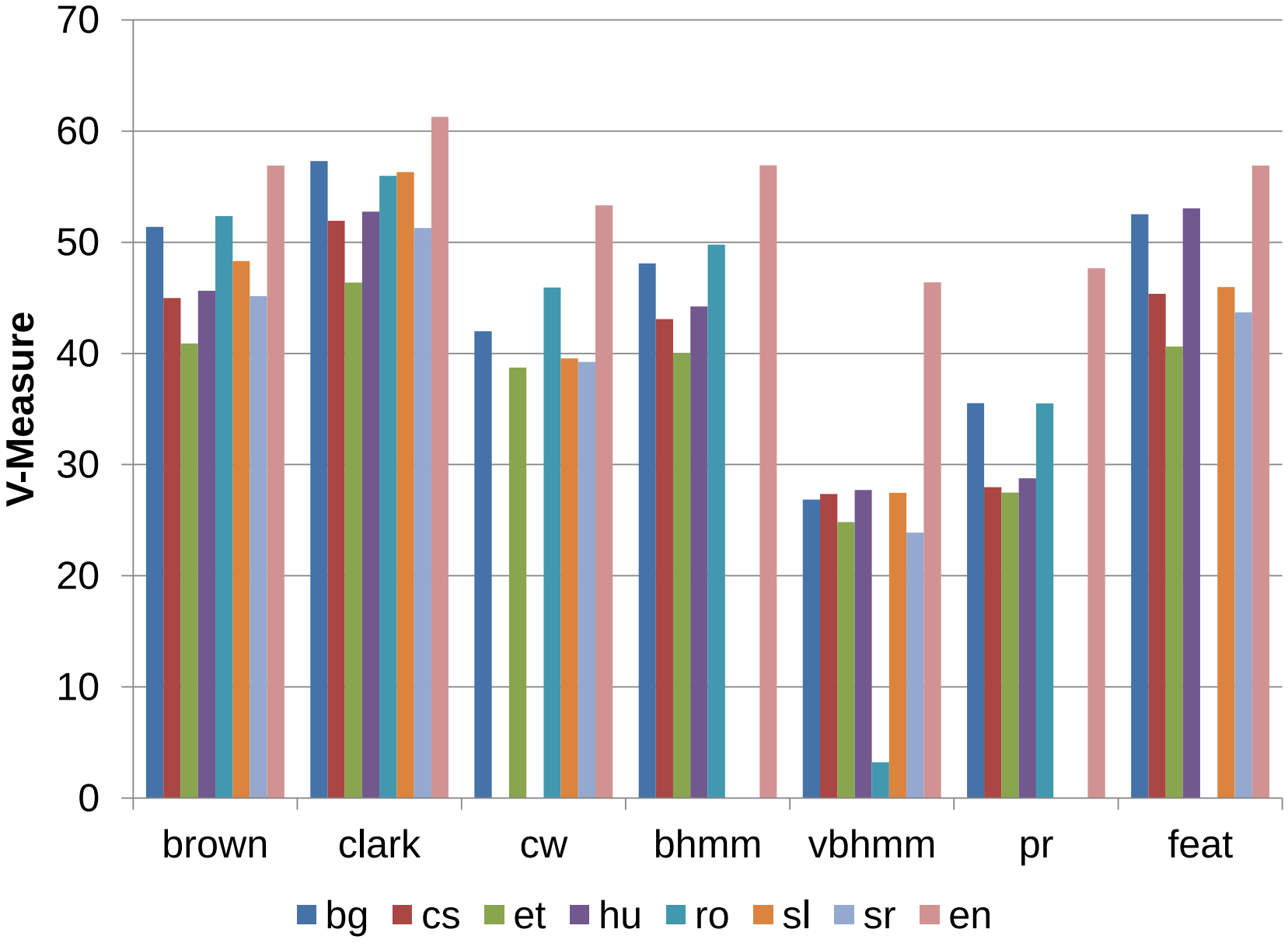
System comparison - WSJ



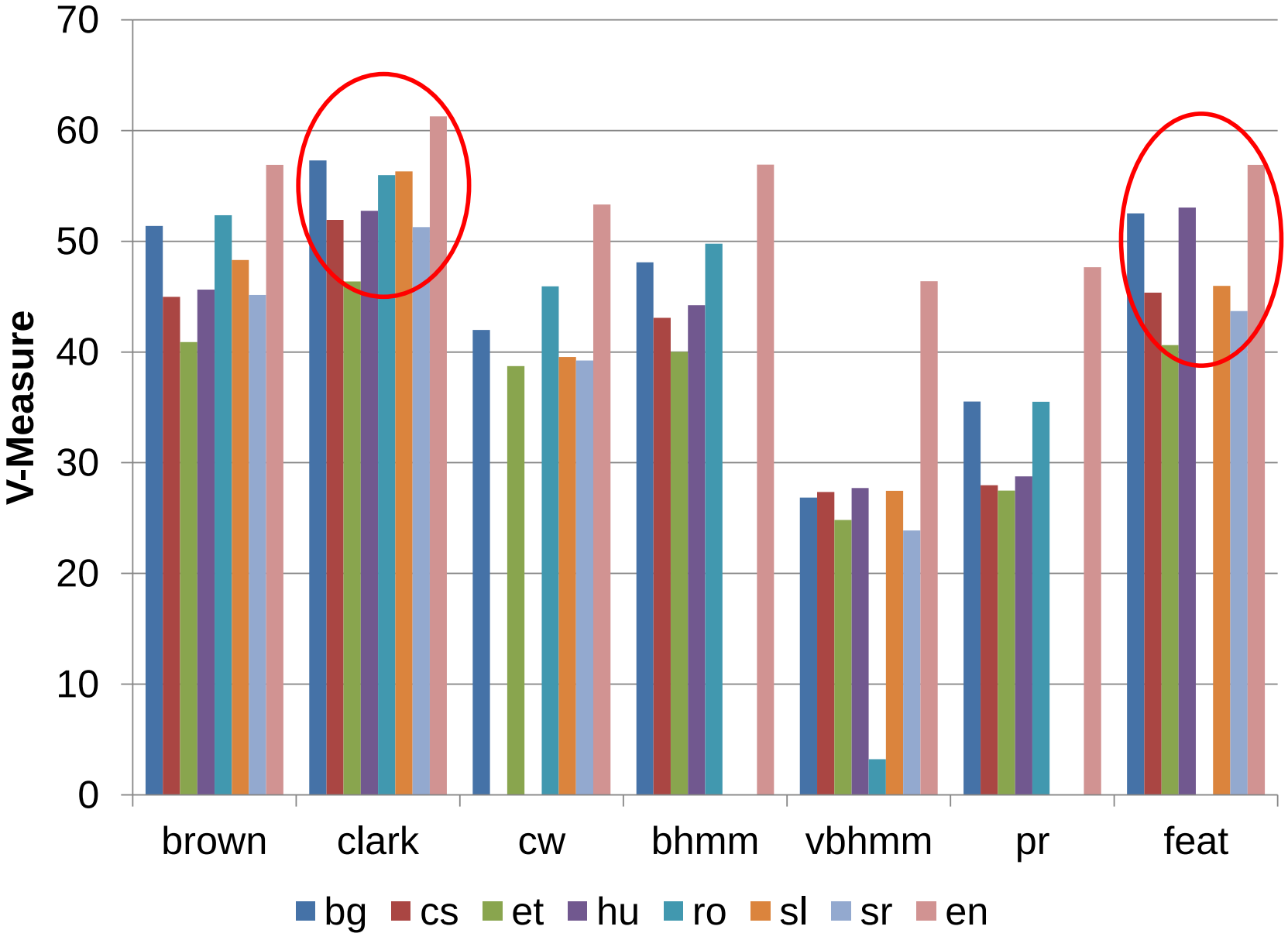
Corpora

- **[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

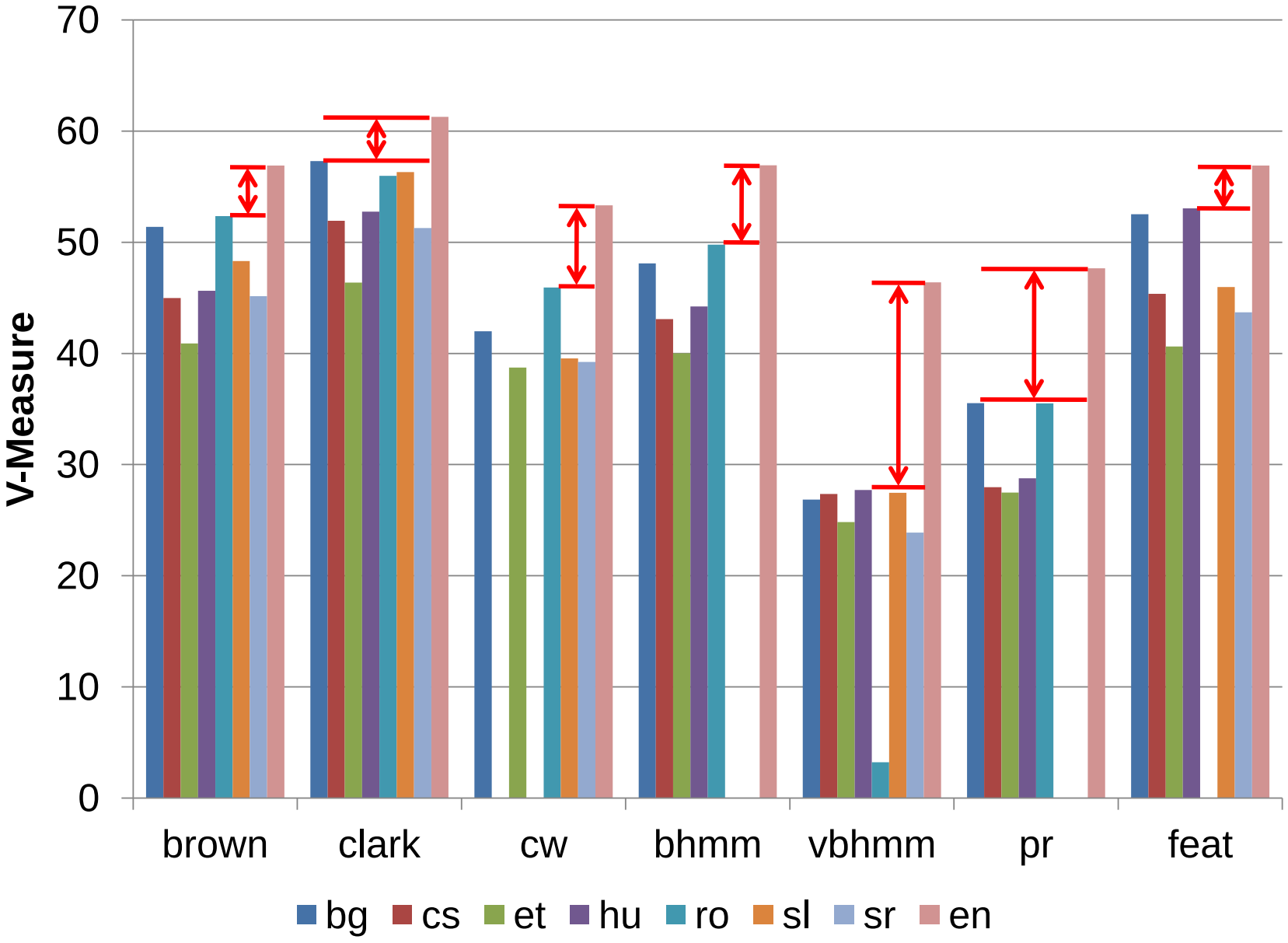
System comparison - Multiple languages



System comparison - Multiple languages



System comparison - Multiple languages



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?

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?
  - 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?

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?
  - 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?

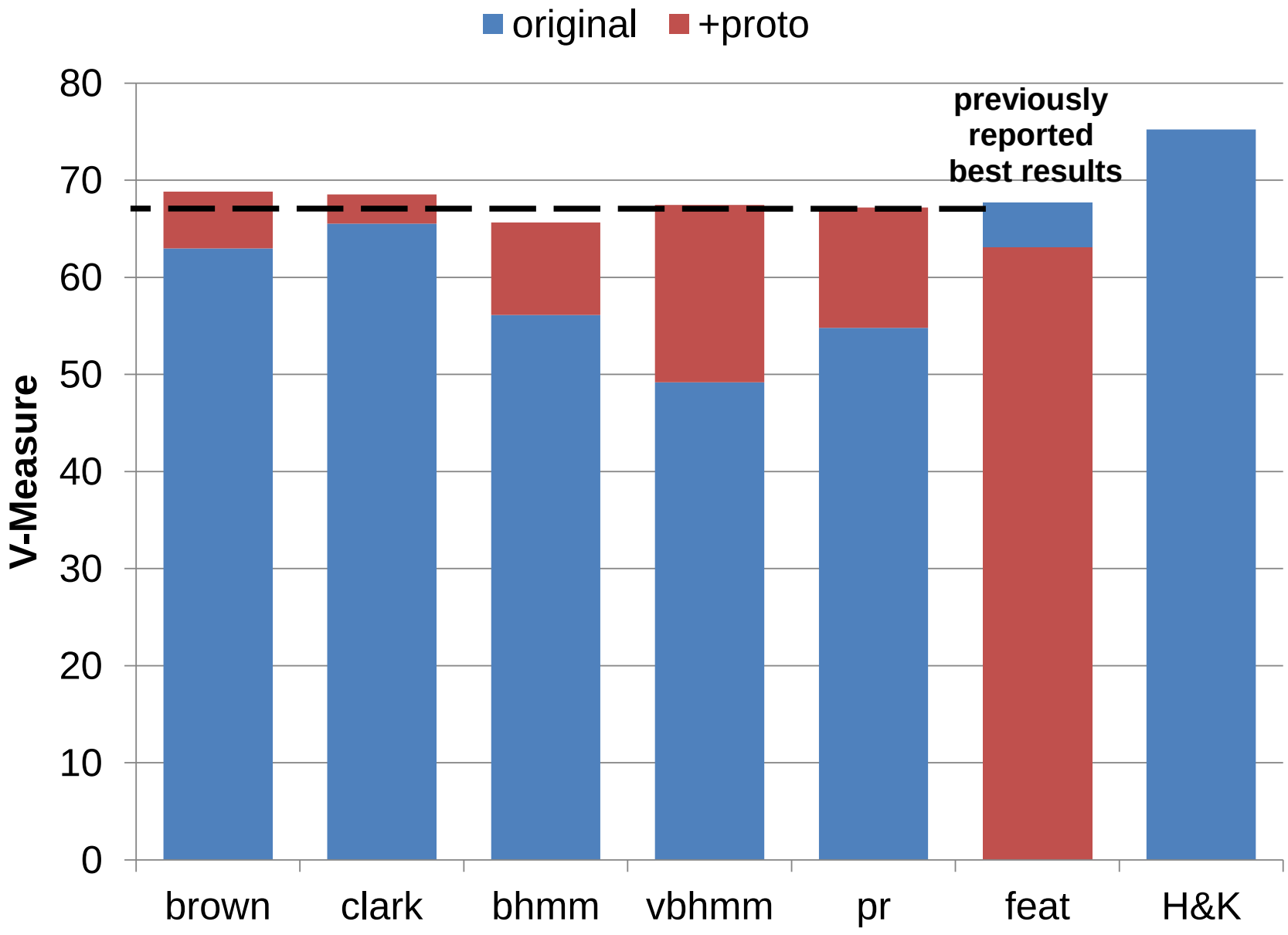
Prototype extraction

Start with POS induction system

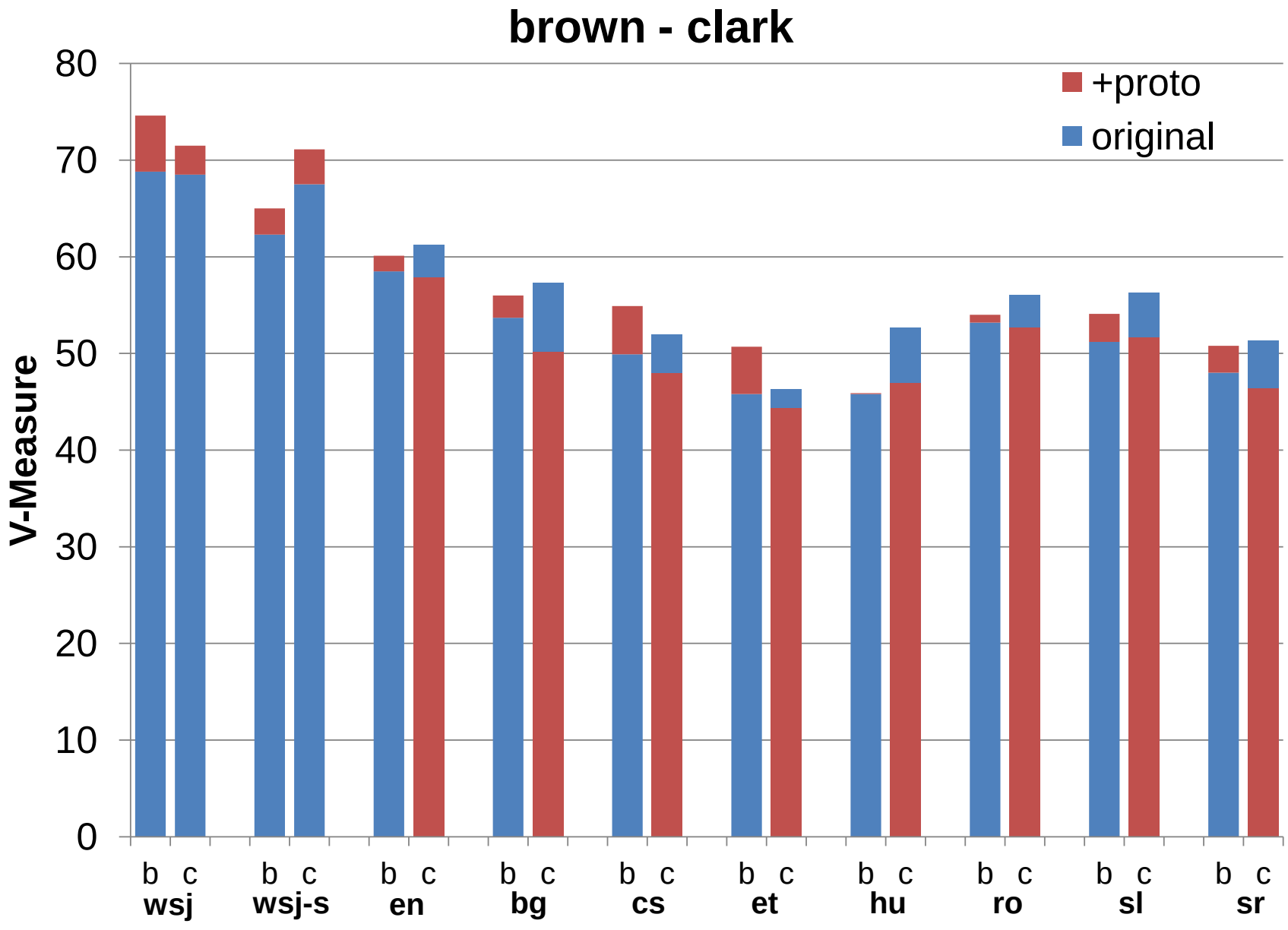
- Algorithm chooses up to 10 prototypes that are:
  - most frequent
  - most similar to words in their clusters
  - most dissimilar to words in other clusters
- Similarity computed using SVD
- Parameters tuned on English

Use H&K system with the prototypes

Results - WSJ


original
+proto
V-Measure
80
70
60
50
40
30
20
10
0
previously reported best results
brown
clark
bhmm
vbhmm
pr
feat
H&K

Results - MULTEXT



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/`

```
19      O'Brien Winston Julia   Oceania Brother
35      are     were
17      because perhaps though  but     then    yet
36      's
18      said    Big     Yes
15      that
33      had
16      when    as      if      while   than
34      was     is
13      and     or
39      .
14      what    which   how
37      must    can     will    could   did     might   would   should  don't
11      with    for     like    without after
12      ,
38      to
21      there   who
20      it      this
43      That    It      There   What
42      You     We      They    The     A       His
41      Then    As      And     To      At      All     But     Even    When    If
40      !       ?
44      She     He
22      she     he
23      I
24      you
25      we      they
26      knew    know    saw     understand      remember        think   believe wondered        thought believed
27      make    do      keep    take    have    see     say     've     'm
28      get     be
29      never   ever    not
3       up      down    off     over    out     together        round   away    ago     somewhere
2       very    better  certain no      young   quite   good   possible        different       true
1       din     piece   scrap   fragment        smell   tramp   majority        range   structure       series
10      in      into    on      at      from    by
0       room    past    book    telescreen      Party   glass   proles  street  door    voice
30      just    always  only    simply  also    probably        even    merely  still   almost
7       his     the
6       about   now     all     called  another anything        hours   nothing something       being
32      been    done    come    become  seen
5       him     himself them    her     again   me      itself
4       two     three   those   Newspeak        people  five    others  human   one     pain
31      began   walked  sat     stood   got     turned  felt    held    set     continued
9       of
8       a       an
```