Computational Linguistics and Sociolinguistics

John Nerbonne
Rijksuniversiteit Groningen & Albert-Ludwigs-Universität, Freiburg

Copenhagen
Sept. 25, 2012
Goal

 › Seek points of potential collaboration between computational linguistics (CL) & sociolinguistics

Plan

 › Examine appl. of CL in variationist ling.
   - Own work in dialectology with CL methods
 › Critique of sociolinguistic focus on single variables (three cases)
 › Points where rapprochemen desirable
Variationist Linguistics

- Since 1980 Chambers & Trudgill have advocated fusing dialectology & sociolinguistics
- So methodological advances in one ought to accrue to the other
- Dialectology has witnessed a blossoming in methods
  - Socioling. less, w. honorable exceptions (FAVE)
- Most dialectological advances in methods due to CL
Key in Chambers-Trudgill

- Since 1980 Chambers & Trudgill have advocated fusing dialectology & sociolinguistics
- Developing tools capable of simultaneously analyzing geographic and social influences
- Later in this talk: Generalized additive mixed-effects models (Wieling)
  - First models capable of including geographic and social influences
CL & Dialectology

- Nearly 20 years “computational dialectology”
- Recognized by leading (non-computational) experts now, too
  - Focus first on problems with perspective on computational solutions
  - Many collaborations with domain experts
  - Shift in focus to larger questions
Problems needing computation

1. Variability: 87 different pronunciations of a two-segment word (at 190 sites)

Solution: “Lift” categorical comparison to numerical level using edit distance
Recipe: *Edit distance* measures degree of difference between strings

- Basics: min. number of operations needed to transform one string to another.
- The algorithm induces an alignment
- *four* pronounced in Berks vs. Philadelphia

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>^</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Refinements, too

www.gabmap.nl
...string comparison

- Used in biology, speech recognition, software engineering, ...
- Some refinements for pronunciations
- Result: Pronunciations differ to a (numeric) degree!
Problem 2

- Noise!

- Solution: Aggregate data to see broad lines first
Lots of follow-ups

- 20 languages, ...
  - Many collaborations, many projects, incl. unfunded projects
- Lots of experiments with technique
- Validation work
  - Consistency checking
  - Comparison with other work, with speakers’ judgments
- Work on aggregate visualization (wrt noise problem)
Dialectology, Phase 2

› Tackling more fundamental humanities problems

› Comparative dialectology. Toward general laws of diffusion.

› Chambers-Trudgill program, fusing dialectology and sociolinguistics
Comparative Dialectology

Challenge (Chambers & Trudgill, 1998), from Nerbonne (2010):
Realizing “Variationist” Program

> Wieling et al. (2011): one model for geographic, linguistic & social influences (age, income)
  
  - A step toward Chambers & Trudgill’s (1998) fusion of dialectology & sociolinguistics
Montemagni et al: *Gorgia Toscana*

Spirantization hierarchy: velar > dental > bilabial

- \( k \rightarrow x \)
- \( t \rightarrow \theta \)
- \( p \rightarrow \phi \)
- \( g \rightarrow \gamma \)
- \( d \rightarrow \delta \)
- \( b \rightarrow \beta \)
Swedish Dialect Leveling

- SveDia data (Eriksson, 2004)
  - > 1K speakers
  - 19 vowels, 5 rep. each
- 65-yr. olds (left), 27-yr. olds (right)
- Therese Leinonen, 2010 Royal Gustav Adolph Prize, Swedish Folk Culture
- N.B. “Leveling” good aggregate concept
Other linguistic levels

› Lexical comparison is easiest (given controlled data). Word choices are same or different.
  • Use CL lemmatization, stemming
  • *Walks, walked, walking*, ... same lexically

› Grammatical comparisons very tentative, but possible using part-of-speech tagging

› Themes, sentiment, etc. possible in CL, unused (to-date) in dialectology
Sociolinguistic studies using CL

- Modern Eur. standard-dialect dynamics
- Auer-Hinskens “cone”
- CL: reify “distance”
- Data:
  - Professional regional radio dialect broadcasters
  - Local dialects
  - Standard Dutch
Regional speakers

- Most broadcaster speech not intermediate!
- Some further from standard than basilects
- Interpretation of exaggeration
  - Broadcasters playing the “local” role
Standardization in Catalan

- Standard req. in schools in Andorra, Catalonia, **not** in Aragon
- Data: 320 speakers, 40 sites, 357 words each, all transcribed by Esteve Valls
  - 4 age cohorts, births: 1991-'96, '74-'82, '46-'60, & '17-30
- Procedure: we measure distance to standard Catalan, regression design
- Hypothesis: Aragon more distant, especially among young
Results

› Aragon less standard (right)
› Age reverses effect at border:

<table>
<thead>
<tr>
<th>Speaker year of birth</th>
<th>Coeff.</th>
<th>SE</th>
<th>p-val</th>
<th>effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Aragon)</td>
<td>0.005</td>
<td>0.004</td>
<td>0.282</td>
<td>0.014</td>
</tr>
<tr>
<td>(Catalonia and Andorra)</td>
<td>-0.012</td>
<td>0.005</td>
<td>0.028</td>
<td>-0.034</td>
</tr>
</tbody>
</table>

› (neg. coeff. more standard)
› No effect of gender, education, tourism, site pop. or income
Individual variables

- [ʌ] (stan.) vs. [j] (non-stan) (Recasens 1996) (10 items in data set)
- PRES-SUBJ [i] (stan.) vs others (Massanell 2001) (20 items)
- POSS-ADJ [β] vs [w] (Romero 2011) (6 items)

- Repeated log. regression using each of these as dependent variables (standard socioling.)
> Conform to aggregate in general, but less clearly; no significant age effect
> Note Andorra (left), north coast (middle)
Dialects in Twitter


> Data from $10^8$ tweets (2009-2012)

> Filter retweets, accounts w. > 100 followers, local toponyms, etc.

> Logistic transformation of frequencies

> EM to detect latent variables (regions, assignment of authors to regions)
Twitter diffusion (Eisenstein et al.)

- as ‘as fuck’
- ard ‘all right’
- ion ‘I don’t’
- cbvs ‘laughing but very serious’
- ctfu ‘cutting the fuck up’
- - ___ -
What’s CL add to understanding?

› Replicable techniques

› More precise definitions

› Scale!

› Research questions evolve more slowly

Spitzweg’s “lonely scholar” no longer suffices
What’s missing?

› LOTS!
  • Many more studies needed, very few have been done
› Analysis (verification) of the social meaning of variation
› Small scale studies of social dynamics (Eckert-style data collection)
› Lots of opportunity for novel work!
Goal

› Seek points of potential collaboration between computational linguistics (CL) & sociolinguistics

Plan

› Examine appl. of CL in variationist ling.
  o Own work in dialectology with CL methods
› Critique of socioling. focus on single variables:
  o Leveling, regiolects, effect of standards
› Need rapprochement, e.g., wrt social meaning
Thanks

› Thanks for your attention!

› Questions?
References


References, cont.

Extra
Extra, 2