Towards automatic geolocalisation of speakers of European French

SCHERRER, Yves, GOLDMAN, Jean-Philippe

Abstract
Starting in 2015, Avanzi et al. (2016) have launched several online surveys to inquire about regionalisms in European French (France, Belgium and Switzerland). Here, we investigate the use of data from these surveys for automatic speaker geolocalisation, both as a playful incentive to attract participants for further inquiries and as a scientific analysis method of the already collected data. Following Leemann et al. (2016), the problem of automatic speaker geolocalisation consists in predicting the dialect/regiolect of a speaker (typically, a speaker that has not participated in the survey) by asking a set of questions (typically, a small subset of the surveyed variables). Given our motivations, the success of a speaker geolocalisation method should not only be assessed by the percentage of correct answers, but also by its ability to entertain and surprise potential participants. Three parameters influence this success: - The number and type of questions to be asked. No more than 20 questions should be asked to keep the attention span short. - The number and type of the areas to predict. The areas should reflect the [...]
Towards automatic geolocalisation of speakers of European French

Yves Scherrer & Jean-Philippe Goldman
University of Geneva
Automatic speaker geolocalisation

Data

Simulation and methods:

- Clustering and shibboleth detection
- Recursive feature elimination

Crowdsourced results
Automatic speaker geolocalisation

Ask a speaker $n$ questions and predict his/her most likely area of origin (one out of $m$ areas) with $p\%$ accuracy.
Automatic speaker geolocalisation

Ask a speaker \( n \) questions and predict his/her most likely area of origin (one out of \( m \) areas) with \( p\% \) accuracy.
Automatic speaker geolocalisation

Ask a speaker $n$ questions and predict his/her most likely area of origin (one out of $m$ areas) with $p\%$ accuracy.
Automatic speaker geolocalisation

Ask a speaker $n$ questions and predict his/her most likely area of origin (one out of $m$ areas) with $p\%$ accuracy.

Goals:

- Provide a playful incentive to attract participants for further inquiries
- Collect more data
- Observation $\Rightarrow$ Prediction
- Explore scientific analysis methods of the already collected data

$\Rightarrow$ select questions and areas to maximize accuracy
Automatic speaker geolocalisation

Ask a speaker \( n \) questions and predict his/her most likely area of origin (one out of \( m \) areas) with \( p \% \) accuracy.

![Diagram showing the relationship between number and type of questions asked, expected accuracy of predictions, and number and type of predicted areas.]

- Number and type of questions asked ↓↓
- Expected accuracy of predictions ↑↑
- Number and type of predicted areas □□
Automatic speaker geolocalisation

Previous work:

- Create a geolocalisation model using data from atlases
- Select \( n \) questions on the basis of a dialectologist’s knowledge
- Use the same \( m \) areas as in the original data
- Assess accuracy post-hoc (compare model predictions with participants’ real origins)

( Leemann since 2013 )

( parlometre.ch - TSR - 2015 )
Automatic speaker geolocalisation

Previous work:
- Create a geolocalisation model using data from atlases
- Select $n$ questions on the basis of a dialectologist’s knowledge
- Use the same $m$ areas as in the original data
- Assess accuracy post-hoc (compare model predictions with participants’ real origins)

Our approach:
- ... from online inquiries
- Select optimal $n$ questions by statistics
- Select optimal $m$ areas by statistics
- Estimate accuracy (given $n$ and $m$) using the same data as for model creation and
- Assess accuracy post-hoc, compare with estimates
Data

Project *Français de nos régions* (Avanzi, Glikman et al., 2015) → online surveys to inquire about regionalisms in European French (France, Belgium, Switzerland).

<table>
<thead>
<tr>
<th>Survey 1</th>
<th>Survey 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 questions</td>
<td>90 questions</td>
</tr>
<tr>
<td>12 000 participants</td>
<td>8 000 participants</td>
</tr>
</tbody>
</table>
Comment appelez-vous cette pâtisserie ?

- Pain au raisin
- Escargot
- Cagouille
- Schnäcke
- Alsacienne
- Pain russe
- Autre (précisez):
Simulation

Simulation framework: \{questions\} + \{areas\} \rightarrow \text{prediction accuracy}

Idea: Leave-one-out method using two views of the same dataset

- Train model on aggregated data of all except one participant
- Predict origin of left-out participant, compare to ground truth

We do not leave out the test participant from the aggregated data:

- Much faster, as we don’t have to train a new model for each participant
- Since training data are aggregated and there are always > 1 participants per area, there is never an exact correspondence between training and test data
- Preliminary tests show good correlation with true leave-one-out method
Simulation

Simulation framework: \{questions\} + \{areas\} $\rightarrow$ prediction accuracy

Two preprocessing steps:

1. Settle on initial set of areas: FR départements, BE provinces, CH cantons (110)
2. Match participants from Survey 1 with participants from Survey 2 (same origin)

Two approaches to find \{questions\} and \{areas\}:

1. Clustering and shibboleth detection
2. Recursive feature elimination
Clustering and shibboleth detection

1. Determine the most relevant areal partition using Ward’s method, 5 clusters
   Ward’s method, 10 clusters
   Weighted average, 10 clusters
Clustering and shibboleth detection

1. Determine the most relevant areal partition using

- Ward’s method, 5 clusters
- Ward’s method, 10 clusters
- Weighted average, 10 clusters
2. Use the **shibboleth** detection algorithm (Prokic, Çöltekin & Nerbonne 2012) to find the most characteristic questions for each area (e.g. 5 shibboleths/cluster)
Clustering and shibboleth detection

2. Use the shibboleth detection algorithm (Prokic, Çöltekin & Nerbonne 2012) to find the most characteristic questions for each area (e.g. 5 shibboleths/cluster)
Clustering and shibboleth detection

Simulation results:

- 10 clusters, all 130 questions → 65.1% correct
  - The results are very sensitive to the cluster borders:
    - -24% between 4 and 5 clusters; -21% between 10 and 11 clusters
  - It is difficult to determine a “good” number of clusters and an optimal cluster algorithm
- 10 clusters, 14 manually defined questions → 67.0% correct
  - Few carefully selected questions are better than all questions
- 10 clusters, 20 questions determined by shibboleth detection → 61.8% correct
  - Unintuitive choice of questions (standard variants for most areas)
  - Clusters are defined on all data, not on single determining questions
Recursive feature elimination

1. The linguistic variables may have several variants with different distributions. Treat each variant separately.
2. Some variants are hardly ever used or show no geographic variation at all. Discard them first.
3. Train a classifier with the remaining variants, remove the one variant that contributes least to the classification, repeat.
4. Use the 110 atomic areas and distance between centroids throughout the process. At the end, dynamically extend the areas to their immediate and second-order neighbors.
Recursive feature elimination

1. The linguistic variables may have several variants with different distributions. Treat each variant separately.

Binarize data: 130 n-ary variables $\rightarrow$ 639 binary variables
Recursive feature elimination

2. Some variants are hardly ever used or show no geographic variation at all. Discard them first.

Single-pass feature elimination based on $\chi^2$ score

Remove variables that are least statistically dependent on area

Lowest average distance with 150 variants
Recursive feature elimination

3. Train a classifier with the remaining variants, remove the one variant that contributes least to the classification, repeat (= recursive feature elimination).

We test two classifiers:
SVM and MaxEnt

Both classifiers achieve much better simulation results than the $\chi^2$ method

MaxEnt slightly worse than SVM
Recursive feature elimination

4. At the end, dynamically extend the areas to their immediate and second-order neighbors.

Simulation results with 20 variants / 17 questions:

66.2% correct on second-order neighbors
Online speaker geolocalisation
Donnez votre français à la science!

Quiz des expressions de nos régions:
Connaissez-vous ces expressions de nos régions ?
Participez au quiz !

Localisez-moi !
Dites-nous comment vous parlez, on vous dira d'où vous venez !

Comment ça se dit chez vous ?
Comment survivent, voyagent et meurent les particularismes linguistiques ? Répondez à quelques questions sur vos usages linguistiques.

Localisez-le !
Comment sont perçus les différents accents du français ?
Essayez d'identifier la région d'origine des locuteurs que vous allez entendre.
Localisez-moi!

Question 2 sur 15

Comment appelez-vous ce fruit rouge, avec lequel on fait d’excellentes confitures ?

- Myrtilles
- Brimbelles
Localisez-moi!

Résultat: les départements en rouge représentent votre origine linguistique la plus probable.

Cliquez sur votre département d'origine

Aidez-nous à valoriser vos réponses en répondant à ce questionnaire

Où avez-vous passé la plus grande partie de votre jeunesse ?

Pays:

Code postal:

Adresse électronique (facultatif, ne sera pas diffusée à des tiers)

Année de naissance

Sexe

Sauvegarder les changements
Online speaker geolocalisation

Three versions

- Feature elimination with MaxEnt  4000  participants
- Feature elimination with SVM  4000
- Manual selection of 15 questions  200

40% of participants provided sociolinguistic info (country+zip, age, gender, email)

Social networks sharing and media coverage
## Online speaker geolocalisation

<table>
<thead>
<tr>
<th>Crowdsourced data</th>
<th>Part</th>
<th>Best</th>
<th>5-Best</th>
<th>Neighb-1</th>
<th>Neighb-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature elimination ME</td>
<td>1631</td>
<td>11 %</td>
<td>43 %</td>
<td>40 %</td>
<td>62 %</td>
</tr>
<tr>
<td>Feature elimination SVM</td>
<td>1679</td>
<td>13 %</td>
<td>47 %</td>
<td>47 %</td>
<td>64 %</td>
</tr>
<tr>
<td>Manual selection</td>
<td>54</td>
<td>5 %</td>
<td>16 %</td>
<td>12 %</td>
<td>18 %</td>
</tr>
<tr>
<td>Random</td>
<td>&lt;1 %</td>
<td>4.5%</td>
<td>~4.5%</td>
<td>~9%</td>
<td></td>
</tr>
</tbody>
</table>

(110 areas - f-score)
# Online speaker geolocalisation

## Crowdsourced data

<table>
<thead>
<tr>
<th>Feature elimination</th>
<th>Part</th>
<th>Best</th>
<th>5-Best</th>
<th>Neighb-1</th>
<th>Neighb-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>1631</td>
<td>11 %</td>
<td>43 %</td>
<td>40 %</td>
<td>62 %</td>
</tr>
<tr>
<td>SVM</td>
<td>1679</td>
<td>13 %</td>
<td>47 %</td>
<td>47 %</td>
<td>64 %</td>
</tr>
<tr>
<td>Manual selection</td>
<td>54</td>
<td>5 %</td>
<td>16 %</td>
<td>12 %</td>
<td>18 %</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>&lt;1 %</td>
<td>4.5%</td>
<td>~4.5%</td>
<td>~9%</td>
</tr>
</tbody>
</table>

## Simulated data

<table>
<thead>
<tr>
<th>Feature elimination</th>
<th>Best</th>
<th>5-Best</th>
<th>Neighb-1</th>
<th>Neighb-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>14 %</td>
<td>49 %</td>
<td>47 %</td>
<td>64 %</td>
</tr>
<tr>
<td>SVM</td>
<td>13 %</td>
<td>46 %</td>
<td>46 %</td>
<td>64 %</td>
</tr>
<tr>
<td>Manual selection</td>
<td>10 %</td>
<td>36 %</td>
<td>40 %</td>
<td>57 %</td>
</tr>
</tbody>
</table>

(110 areas - f-score)
Discussion

● Attempt to apply machine learning techniques for question (and area) selection

⇒ estimate success of crowdsourced linguistic campaign before launch

● Automatic selection better than manual? (to be confirmed)
● Crowdsourced geolocalisation also means data collection

⇒ donnezvotrefrancais.fr
Towards automatic geolocalisation of speakers of European French

Yves Scherrer & Jean-Philippe Goldman
University of Geneva
Recursive feature elimination

Retained features from the SVM classifier:

- Pain au chocolat / chocolatine / couque au chocolat / ...
- Ving[t]
- Crayon de papier / de bois / gris / ...
- Nonante / quatre-vingt-dix
- Péguer
- Gouttière / cheneau
- Il est midi vingt / et vingt / vingt
- Dîner / déjeuner
- Pain aux raisins / escargot / schnäcke
- Je vais y faire / le faire
- Faire tomber / tomber / échapper
- Séchoir / étendoir / étendage / tancarville
- Moin[s]
- Escargot / cagouille / luma
- Dégun / personne

Retained features from the MaxEnt classifier:

- Septante / soixante-dix
- Ving(t)
- Il est midi vingt / et vingt / vingt
- Pain au chocolat / chocolatine / couque au chocolat / ...
- Crayon de papier / de bois / gris / ...
- Ça joue / ça va
- Gorgée / schlouk / lichette
- Gouttière / cheneau
- Stan[d]
- Empêtrer / encoubler / achouper / ..
- Dîner / déjeuner
- Péguer
- Pain aux raisins / escargot / schnäcke
- Séchoir / étendoir / étendage / tancarville
- Papier ménage / Sopalin / essuie-tout
Si vous voulez parler d'une personne qui fumait et qui ne fume plus, il a fumé (mais il ne fume plus).