Dealing with Bad Data Problems of the SAND

Erik Tjong Kim Sang
Meertens Institute, Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

17 March 2016

SAND: Syntactic Atlas of Dutch Dialects

The SAND data set contains syntactic data of different dialects spoken in The Netherlands and the Dutch-speaking part of Belgium

The data set contains 219 linguistic features based on 671 sentences presented in 267 locations

Collected and analyzed data are available on maps in printed volumes, on a website and in a relational database

Research goal

Our goal is to use computational techniques for automatically deriving sensible dialect regions from the data

Our research question is: *Are the syntactic features of dialects sufficient for identifying dialect boundaries?*

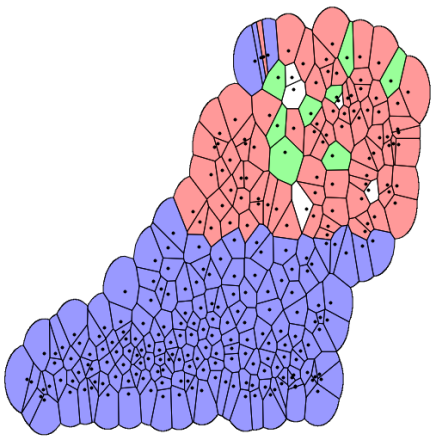
We are also interested in the kind of dialect boundaries that can be derived: sharp boundaries or transition zones

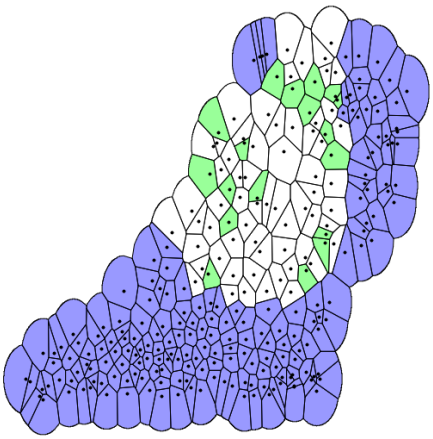
Challenges

We use the part of the data analyzed for representation on maps

We found that the data are incomplete, which presents a challenge for the analysis techniques we want to apply to them

Another challenge is presented by the values of the syntactic features: these are not atomic but they are sets of values





Dealing with missing data

Unprocessed negative data: Data annotation (expensive)

Implicit negative data: Make explicit, whenever possible

Accidental missing data: Replace with average value, or skip

volume	page	sentence	locations	values	editor	rules	missing
1	64b	356	58	58	52	+187	28
1	79a	043	96	13	26	+177	64
1	93b	321	261	60	253	-	14
1	95a	359	108	11	106	+154	7
2	32b	531	254	81	239	-	28
2	41b	248	258	61	251	-	16
2	42a	246	156	31	154	+111	2
2	42b	247	157	34	155	+110	2
2	48b	836	39	33	37	-	230
2	49a	027	262	10	241	-	26
2	50b	600	77	66	53	-	214
2	51b	474	124	33	116	-	151
2	55a	390	90	26	72	-	195
2	57a	391	87	10	66	-	201
		average	141	37	124	+61	82

Dealing with accidental missing data

We compare two techniques for dealing with accidental missing data:

Replace missing feature values with an average value:

- distance(A,A) = 0
- distance(A,B) = 1
- distance(A,missing) = 0.5
- distance(missing,missing) = 0.5

Skip missing data (Heeringa, 2004):

The distance between two locations is defined as the normalized sum of the distances between relevant features with **known** values

Comparing sets of values

Feature values are sets, for example: {A,B,C}

We assume that the distance between two feature values is zero when they share a value

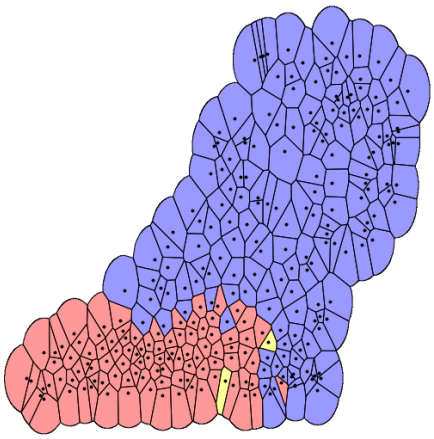
Otherwise, the distance is one:

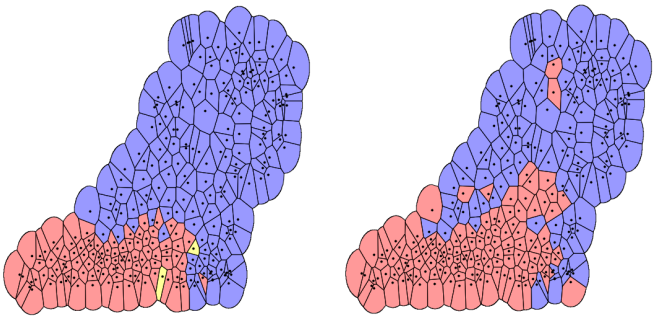
- distance({A,B,C},{A,D}) = 0 because both values contain A
- distance({A,B,C},{D,E}) = 1 becuase no feature value is shared

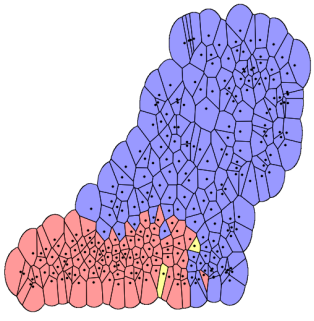
Finding dialect regions

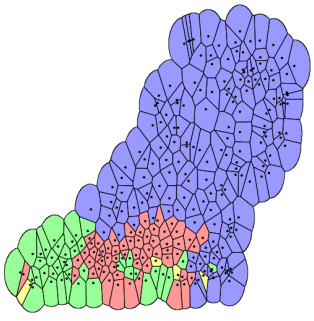
We use k-means clustering for finding dialect regions:

1. randomly choose *k* locations, the centers of *k* regions
2. assign each location to the closest center
3. compute new center locations for each region (based on distances to other points)
4. repeat steps 2-4 until the center locations remain the same









Concluding remarks

We have examined two problems related to deriving computational models from the SAND:

- **missing data**: accidental, unprocessed negatives, implicit negatives
- **values consisting of sets**

We have presented approaches for dealing with these problems: annotation, data analysis and computational methods robust for missing data

Future work

Data validation: with the goal of making explicit all implicit data values

Algorithm testing: examining other data clustering algorithms

Evaluation methods: are there apporpriate automatic evaluation methods for such clasification tasks?

THE END