Dealing with Bad Data Problems of the SAND

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

We identify dialect boundaries based on the Syntactic Atlas for Dutch Dialects (SAND) [2], a database of syntactic features observed in the language spoken by people from 267 locations in The Netherlands and Flanders. However, combining all the data in this database proved to be nontrivial because they are incomplete, imbalanced, conflicting and possibly noisy. In this paper we discuss methods for dealing with these problems.

Data

We use linguistic data from the Syntactic Atlas for Dutch Dialects (SAND) [2] for finding dialect boundaries. These data are based on interviews with people from 267 locations in The Netherlands and Flanders, the Dutch-speaking part of Belgium. The interviews have been transcribed and checked for the presence of predefined syntactic variables. We use the subset of 220 syntactic variables which are available from the digital version of the atlas: DynaSAND [1].

Although exceptional care has been put into constructing the SAND data set, it still suffers from problems common to most linguistic data sets. First, values are missing for some location-feature pairs, like for feature 2.3.1.1 *interruption of the verbal cluster: base noun* (257 rather than 267 feature values). Second, the SAND features cover a selection of syntactic phenomena (related to complementisers, fronting, negation, pronouns, quantification, subjects and verbs), and it is unknown if the number of features per phenomenon corresponds with its importance to a syntactic model. Third, on rare occasions the SAND contains conflicting data, as for location Jordwerd of the feature 2.3.1.1, which is both positive and negative. And fourth, while maps of SAND data display geographical correlation with the syntactic feature values, they also contain islands: isolated feature values which are completely surrounded by a single other value, something which could indicate noisy values.

These problems prevent a straightforward application to the data of computational models for deriving dialect boundaries. For example, because of missing values no distances can be computed between certain pairs of locations and subsequently these locations cannot be used the model that generates the dialect boundaries. A solution for this problem must be found if we want to base models on this data set.

Methods

The most important issue we need to deal with in working with the SAND data, is missing data. If there were few missing values, we could remove the problematic features or locations. Unfortunately this approach would lead to the loss of most of the data. Previous studies based on the SAND data [4, 6] have rather chosen to replace the missing values with the value *unobserved*. The other three problems mentioned in the previous sections, were ignored.

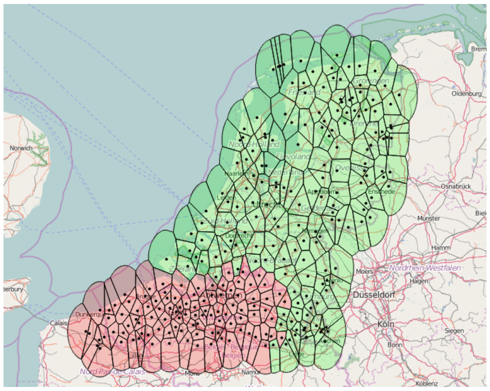

Figure 1: A geographic map with the two main Dutch-speaking regions identified by the *k*-means algorithm from the syntactic SAND data. The 267 locations were roughly divided in the geographical regions Flanders (red) and The Netherlands (green), although most of the Flemish region Limburg has been classified under The Netherlands, which contains a region with a similar dialect.

In our own work [5], we did not replace the missing values but rather adapted our models so that they could deal with them. With k-means clustering [3], we assigned a basic distance of zero to location pairs with feature values that were equal and a distance of one for values that were different. Next, we chose the distance of a half for location pairs that included a missing value. Like the other SAND studies, we ignored the other three data problems for now. This approach result in an excellent model for dividing the data in two parts (see Figure 1) although higher order divisions proved to be less satisfactory.

Discussion

We have presented one working method for dealing with missing data in clustering models: assume that distances involving missing values are half of the maximum feature value distance [5]. We believe that this approach is methodologically superior to the replacement of missing values with *unobserved*, as executed in previous studies with our data set [4, 6] because it does not create new data values. However, if the results of the alternative data analysis provides new insights to the domain experts, it is difficult to deny its usefulness.

This realization leads to an important insight: since it is not clear what the best method is for dealing with bad data, it is useful to compare different approaches and let domain experts evaluate the results. This would be the best way to deal with the problems of imbalanced and noisy data in our data set[1]: process the data with various feature weight settings and data smoothing methods. A subsequent analysis by domain experts could then determine which of the applied methods was most useful for this particular data set.

[1] The problem of conflicting data in our data set is probably caused by tagging errors which need to be corrected.

References

[1] Sjef Barbiers. Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND), 2006. http://www.meertens.knaw.nl/sand/ Accessed 26 February 2015.

[2] Sjef Barbiers, Johan van de Auwera, Hans Bennis, Eefje Boef, Gunther Vogelaer, and Margreet van der Ham. *Syntactic Atlas of the Dutch Dialects*. Amsterdam University Press, 2008.

[3] Christopher D. Manning and Hinrich Schutze. *Foundations Statistical Natural Language Processing*. MIT Press, 1999.

[4] Marco René Spruit. *Quantitative perspectives on syntactic variation in Dutch dialects*. LOT, Utrecht, The Netherlands, 2008.

[5] Erik Tjong Kim Sang. Discovering Dialect Regions in Syntactic Dialect Data. In *Workshop European Dialect Syntax VIII - Edisyn 2015*. Zurich, Switzerland, 2015.

[6] Jeroen van Craenenbroeck. The signal and noise in Dutch verb clusters – A quantitative search for parameters, 2014. Manuscript, http://jeroenvancraenenbroeck. net/s/paper_signal_noise.pdf , version 18 December 2014, Retrieved 26 February 2015.