Inferring and Revising Theories with Confidence: Data Mining the 1901 Canadian Census

Chris Drummond[1], Stan Matwin[1], and Chad Gaffield[2]

[1] School of Information Technology and Engineering,
University of Ottawa,
Ontario, Canada, K1N 6N5
{cdrummon,stan}@site.uottawa.ca
[2] Institute of Canadian Studies,
University of Ottawa,
Ontario, Canada, K1N 6N5
gaffield@uottawa.ca

Abstract. This paper shows how data mining can help historians analyze and understand important social phenomena. Using data from the Canadian census of 1901, we discover the influences on bilingualism in Canada at beginning of the last century. Our approach, based around a decision tree, not only infers theories directly from data but also evaluates existing theories and revises them to improve their consistency with the data. One novel aspect of this work is the use of confidence intervals to determine which factors are both statistically and practically significant, and thus contribute appreciably to the overall accuracy of the theory. When inducing a decision tree directly from data, confidence intervals determine when new tests should be added. If an existing theory is being evaluated, confidence intervals also determine when old tests should be replaced or deleted to improve the theory. Our aim is to minimize the changes made to an existing theory to accommodate the new data. To this end, we propose a semantic measure of similarity between trees and demonstrate how this can be used to limit the changes made.

1 Introduction

The aim of this research is to develop a data mining tool that will help historians explore the influences on the languages spoken in Canada at the beginning of the last century. At the time of Confederation in 1867, language was a secondary issue to other concerns, most notably, religion. By the turn of the century, however, language was becoming an increasingly significant concern in Canada as in other western countries, and during the following decades, it came to be seen as a principal indicator of an individual's identity. While much research has focused on the changing official views of language in Canada, little is known about the actual linguistic abilities of the Canadian population before the later twentieth century.

To address this problem, we apply a data-mining algorithm to the 1901 Canadian census. For the first time, the census asked all residents in Canada three language questions: mother-tongue, ability to speak English, and ability to speak French. Our research investigates a random five-percent sample of the 1901 enumeration that has been created by the Canadian Families Project. The sample is composed of all individuals living in households that were randomly selected from each microfilm reel of the census enumeration for that year. Households were selected to permit analysis of individuals with relevant social units. The resulting sample is a cluster sample but given the nature and large size of the sample, the design effect is not a concern is this study. For a detailed analysis of this question, see Ornstein [7]. The sample includes data on 231,909 individuals over the age of five, and it allows us to explore how factors such as ethnic origin, mother-tongue, place of birth and residence, age and sex influenced the frequency of bilingualism across Canada. We build upon research that focused on the interpretive implications of how the census questions were posed, and how the actual enumeration was undertaken [3]. We now focus on the responses to these questions written down by the census officials at the doorsteps of individuals and families across the country.

The data mining algorithm we use is the decision tree. Decision trees are easy to understand, even by non-specialists, and have been used by domain experts in many diverse applications [6]. In decision tree learning, an important issue is over-fitting avoidance. A complex tree that fits the training data well typically has unnecessary structure that does not contribute to the accuracy of the tree and may even degrade it. To make the trade-off between accuracy and tree size more principled, we use confidence intervals to prune the tree rather than one of the existing methods. Using confidence intervals allows the determination of not only a statistically significant improvement in the accuracy of the tree, but also to quantify the size of the improvement. A test then will only be added to the tree if the expected accuracy gained is sufficiently large to justify it.

We are interested not only in inferring theories directly from the data but also in testing existing theories, such as those representing the views of politicians of that era, to see if they are confirmed, or indeed contradicted, by the data. Confirmation is likely to be a matter of degree and not all parts of the theory will be affected equally. In this paper, we use a measure of the semantics of a tree to minimize the amount a theory is changed to bring it into accordance with the data. This should help historians not only evaluate an existing theory but also to identify any erroneous assumptions on which it was based.

In the following sections, we first will show how confidence intervals are used to prune a tree grown directly from the data. We then show how our semantic measure combined with confidence intervals and new data is used to evaluate and revise an existing theory on the influences on bilingualism in 1901.

2 Inducing A Decision Tree

A binary tree is used to represent the theories induced from the data. Although sometimes deeper than a tree with a greater branching factor, binary tests should help historians determine not only what are the important attributes but also the critical values of those attributes. The tree is grown in the standard greedy manner, the best test, according a splitting criterion, is selected to be added to the tree. The main difference is that a test is actually added only when there is a high confidence that a worthwhile increase in accuracy will result.

$$f(a,v) = \max_{a,v} |P(L_{a,v}|+) - P(L_{a,v}|-)| \tag{1}$$

Using the splitting criterion of equation 1, the best split has the greatest difference in the estimated probability of a positive instance going left $P(L_{a,v}|+)$ and a negative instance going left $P(L_{a,v}|-)$ [10]. The criterion is applied to each attribute and each value and the attribute-value with the greatest difference is selected. This value becomes the left branch of the split and the right branch represents the remaining values of the attribute. The difference in likelihood provides a measure of the probability that positive and negative examples come from different distributions. A large difference tends to produce branches with a large difference in class ratios and ultimately leads to better accuracy.

Our aim is to only add tests that improve the accuracy of the tree by a useful amount. But when greedily growing a decision tree adding a single test may not improve accuracy at all. This is often due to the strong imbalance in classes away from the root node. Modifying the class distribution to reduce this imbalance produces a measure that is more likely to show improvement when a single test is added but produces negative values if there is unlikely to be any advantage in adding the test. Based on the training data, the side of split where the positive likelihood is greater than the negative likelihood is labeled positive and the other side negative. Equation 2 gives the accuracy of the split if the left and right hand sides are labeled positive and negative respectively. Here, the role of the probability of each class, $P(-)$ and $P(+)$, is evident. To make a statistic less sensitive to class distribution, the values are replaced by ones closer to 0.5, by applying the squashing function $P'(a) = (P(a) + 1)/(1 + 2)$ to the class probabilities. The resultant statistic can be viewed either as accuracy with a modified class distribution or as the linear combination of two statistics, accuracy and likelihood difference, with the numbers in the squashing function controlling each statistic's influence. The statistic is divided by the fraction of instances reaching the test, and thus estimates the overall improvement in performance.

$$Acc = P(L|+)P(+) + P(R|-)P(-) \tag{2}$$

In decision tree learning, the complexity of the tree is controlled by pruning. In post-pruning, the tree is first grown until it fits the training set well and then extraneous tests, not expected to improve accuracy, are pruned away. In pre-pruning, new tests are only added if they are likely to improve accuracy. Frank [2] experimentally compared the two techniques based on significance tests and found little performance difference. Here, we use pre-pruning but based on confidence intervals rather than significance tests. To generate confidence intervals, we follow the basic procedure proposed by Margineantu and Dietterich

[4]. We apply the same bootstrapping technique [1] (but for a different purpose) as each new test is added to the tree. Rather than discuss the method in detail, we refer to their paper [4]. If two tests are being compared, a three dimensional confusion matrix is used. If we are considering adding or deleting a test, we use the confusion matrix and its row marginals. We apply our test statistic to 500 randomly generated matrices. After sorting the resulting values in ascending order, the fiftieth element will be the lower bound of a 90% one-sided confidence interval.

If this lower bound is greater than zero, we are confident that the gain is statistically significant. In Figure 1 a), H0 is the null hypothesis that the difference is less than or equal to zero, H1 is the alternative hypothesis that adding the test improves accuracy. Not only is the lower bound greater than H0, it is also greater than 0.5%. We can be confident that this test would improve the accuracy of the tree by 0.5%, so the test would be added. If the bound is smaller than the chosen percentage or smaller than H0, see Figure 1 b), the test would not be added. When starting with an existing theory, we are also interested is deleting structure. Applying the same test allows us, see Figure 1 c), to determine that we are confident that removing structure does not degrade performance.

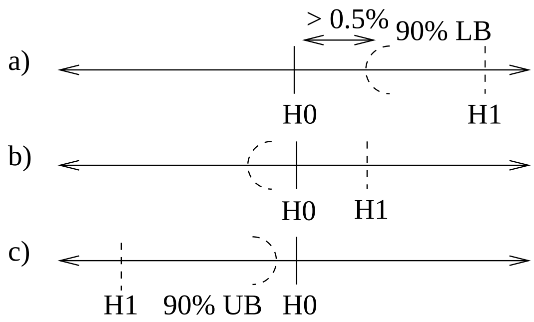

Fig. 1. Using Confidence Intervals for Pruning

The values used to decide when a test should be added where chosen by the authors to represent a reasonable confidence in a useful increase in accuracy. Future work will investigate the effect of varying these values and changing the test statistic used to estimate the increase in accuracy.

3 Theories Induced from Data

In this section, we explore theories generated directly from the data. We use eight attributes from the 1901 census data felt to be potentially relevant to the issue of bilingualism. Some of the nominal attributes have had their values combined into groups and the continuous attribute *age* has been divided into three intervals. To generate the class label *Bilingual*, we combined the attributes *Can speak English* and *Can speak French* but removed instances where one or both of the attributes were unknown. For the rest of this paper, unilingual will mean can speak French or English and bilingual will mean can speak both. To decide whether a new test should be added, an increase in performance of 0.5% is needed at a confidence level of 90%. Attributes will only be added if the number of instances on each side of the split is greater than 10. Numbers less than 10 might belong to a single family or a related group and be therefore of little interest. The instances are randomly split into a test and training set, 75% of the instances going to the training set. A pruning set is produced from a random 25% of the training set. The splits are all stratified to maintain the class ratios in each set.

Figure 2 is the tree representing the factors that affected bilingualism throughout Canada in 1901. At each leaf the classification is shown: bilingual is labeled "Y" and unilingual is labeled "N". The most important attribute, at the root of the tree, is *mother-tongue*. The split is between those that have French as their mother-tongue "MTONGUE=FR", and those that do not (divided into English, German, Gaelic and Others) "MTONGUE=oth". Notably, for this latter category the tree terminates at a leaf immediately below the root. This classifies all people that do not have French as their mother-tongue as unilingual. The former category is further divided by *birth place*, those born in urban communities "BPLACE=UR" and can write are mostly bilingual. For rural communities "BPLACE=RU", this is only true for males aged 20 to 49. The accuracy gained by adding each attribute is shown to the left of the vertical line. To the right of the line, the total accuracy (80.17%) is labeled "A", the majority classifier accuracy (72.89%) is labeled "MC", the total gain in accuracy (7.28%) is labeled "G" and its 90% lower bound (7.00%) is labeled "LB". The lower bound is generated using bootstrapping on the overall confusion matrix.

```
MTONGUE=FR                          2.91|
|  BPLACE=RU                        1.66|
|  |  AGE=20-49                     1.79|
|  |  |  SEX=F: N                   0.54|A 80.17 MC 72.89
|  |  |  SEX=M: Y                       |G  7.28 LB  7.00
|  |  AGE=oth: N                        |
|  BPLACE=UR                            |
|     CANWRITE=N: N                 0.37|
|     CANWRITE=Y: Y                     |
MTONGUE=oth: N                          |
```

Fig. 2. Decision Tree for Canada

We next explore how the factors that affected bilingualism varied across Canada. Figure 3 shows a map [3] of Canada in 1901 when the census was taken.

[3] ©2001. Government of Canada with permission from Natural Resources Canada

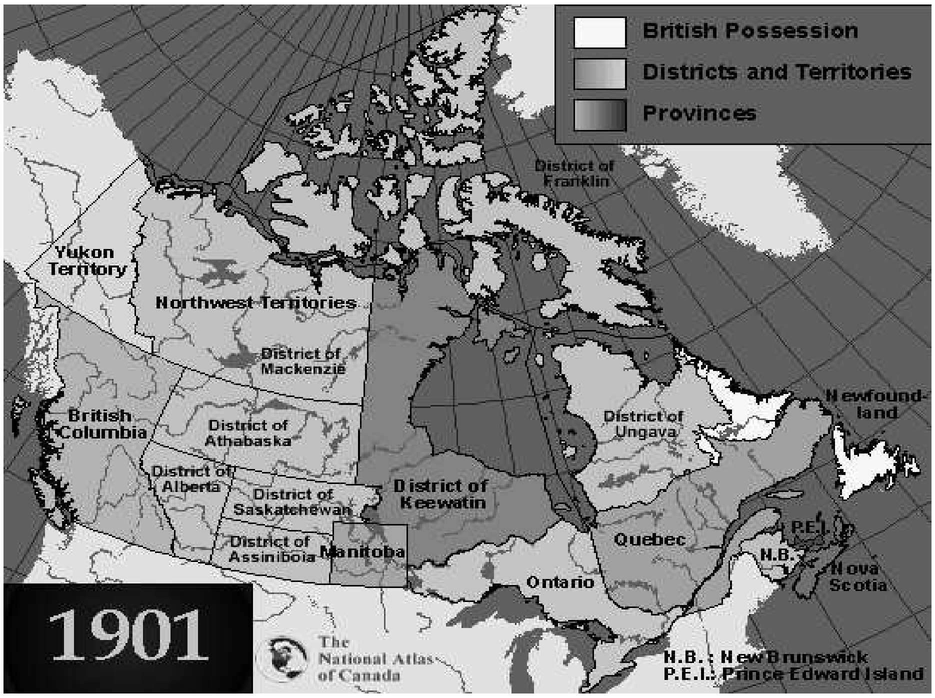

Fig. 3. Map of Canada in 1901

The territories and districts were very sparsely populated at this time. So we combine them into a single region, with a population size more in accordance with other regions. We also make a single region out of the eastern provinces; New Brunswick, Nova Scotia and PEI. We grow decision trees for each of the regions as shown in Figure 4. For British Columbia, the tree consists of the single attribute *mother-tongue* classifying all individuals with a mother-tongue of French as bilingual and all others as unilingual. The majority classifier is already quite accurate, see Figure 4, due the large preponderance of unilingual people in British Columbia. But using the attribute *mother-tongue* correctly predicts a bit over a third of the bilingual people without sacrificing much accuracy on the unilingual ones. Adding extra attributes produces no appreciable improvement. For the territories, the tree has the same root node, but an additional attribute *can read* improves accuracy when the mother-tongue is French. For Manitoba, the tree also has the same root node, but the additional attribute is now *can write*. For Ontario, as for British Columbia, only the single attribute of *mother-tongue* is used. The Eastern provinces have a tree which is similar to Manitoba. *Mother-tongue* is again the most important attribute, adding the attribute *can write* is useful, although it does not improve accuracy on its own. However with an additional attribute excluding children "AGE=5-19" , accuracy is improved.

```
British Columbia
MTONGUE=FR: Y                     4.58|A 91.88 MC 87.30
MTONGUE=oth: N                        |G  4.58 LB  3.48

Territories
MTONGUE=FR                        4.52|
|  CANREAD=N: N                   0.64|A 93.54 MC 88.38
|  CANREAD=Y: Y                       |G  5.16 LB  4.24
MTONGUE=oth: N                        |

Manitoba
MTONGUE=FR                        7.66|
|  CANWRITE=N: N                  0.63|A 89.51 MC 81.22
|  CANWRITE=Y: Y                      |G  8.29 LB  6.72
MTONGUE=oth: N                        |

Ontario
MTONGUE=FR: Y                     7.06|A 94.28 MC 87.22
MTONGUE=oth: N                        |G  7.06 LB  6.69

Eastern Provinces
MTONGUE=FR                        8.64|
|  CANWRITE=N                     0.00|
|  |  AGE=5-19: N                 1.76|A 90.66 MC 80.26
|  |  AGE=oth: Y                      |G 10.40 LB  9.55
|  CANWRITE=Y: Y                      |
MTONGUE=oth: N                        |

Quebec
BPLACE=RU                         7.15|
|  AGE=5-19: N                    0.00|
|  AGE=oth                            |
|     SEX=F: N                    1.52|
|     SEX=M                           |
|        CANWRITE=N: N            0.82|
|        CANWRITE=Y                   |
|           MTONGUE=FR: Y         0.27|
|           MTONGUE=oth: N            |A 67.05 MC 53.96
BPLACE=UR                             |G 13.09 LB 12.34
   SEX=F                          0.15|
   |  CANWRITE=N: N               1.56|
   |  CANWRITE=Y                      |
   |     MTONGUE=FR: Y            0.78|
   |     MTONGUE=oth: N               |
   SEX=M                              |
      CANWRITE=N: N               0.84|
      CANWRITE=Y: Y                   |
```

Fig. 4. Regional Decision Trees

For Quebec, a quite different tree is produced. Although the attribute *mother-tongue* is used, it appears much further down the tree, close to the leaves. The most important attribute is *birth place*, indicating if the person was born in a rural or urban community. The attributes used on both sides of this split are very similar. Although for people born in rural communities, children are immediately classified as unilingual. The overall tree is much less accurate than those of the other regions. But as there was a nearly equal number of bilingual and unilingual speakers in Quebec, it still a considerable improvement over the majority classifier.

From an algorithmic perspective, attributes seem generally to be added if, and only if, they result in an increase in accuracy at the leaves of a practically significant amount. For the larger trees this is not always the case. This might be due to using a 90% confidence limit, 10% of the time this limit will not be met. It might also be due to the test statistic not being a direct measure of accuracy. In the latter case, postpruning using accuracy might address the problem, but this remains the subject of future work. For Quebec, it was possible to increase accuracy by about 0.7%, by reducing the confidence interval to 50% and removing the requirement for any gain. But the number of tests went from 9 to 32, so is of debatable merit. With the statistic we use, it is possible to produce a split where the majority class for each branch is the same. This is makes no difference in accuracy and can be removed to make the tree smaller. In fact, for most of the trees this was unnecessary as there was no additional structure.

From a historical perspective, the decision trees are in keeping with some, though not all, of the ways in which politicians, census officials, and other observers at the time discussed the question of bilingualism. The general assumption was that English was becoming an international language of commerce, and that if Canada were to continue developing, everyone in the country should be able to speak it. In contrast, no public figure stressed the importance of learning French. In this sense, the question of bilingualism was directed to two groups: French-language residents and immigrants who did not speak either French or English. The decision trees confirm that the mother-tongue francophones accounted for much of the bilingualism in Canada. Similarly, individuals who were more likely to be involved in commerce were more bilingual. The importance of economic factors is also seen in the greater tendency of middle-aged males in rural areas in Quebec (more likely to be working in rural industries or in the forest economy) to be more bilingual. At the same time, this rural pattern shows how the decision trees diverge from the theories that underlay the contemporary public debate. Specifically, the trees reveal an extent of diversity in language patterns that is inconsistent with how observers characterized Canadian society. For the most part, for example, Quebec was assumed to be a quite homogeneous society especially in the countryside. The general picture was of a unilingual French-language rural world in Quebec that contrasted with the bilingual urban communities of Montreal and to a lesser extent Quebec City. The decision trees reveal that Quebec was indeed a quite distinct part of Canada in terms of bilingualism but that within this distinction there was still considerable diversity.

4 Revising an Existing Tree

In this section, we show how an existing tree is revised so as to minimize the change to the underlying semantics of the theory it represents. The main difference with other forms of theory revision [5, 9] lies in how we quantify changes to the theory and how we use confidence intervals to decide when those changes are worth making. Our notion of the semantics of a decision tree is based on how the tree partitions the attribute space. We capture this semantics by generating instances consistent with the tree. To limit the number of instances, we generalize the definition of an instance so that the probability of an attribute having a particular value is specified. This is similar to the treatment of unknown values in C4.5 [8]. By adding a weight to the instance we can simulate the effect of multiple examples without incurring the additional processing cost. In our approach, the user constructs a decision tree to classify a specified number of imaginary instances, say 1000. An example of what such a tree might look like is shown in Figure 5. Each leaf is marked with the number of individuals from the original thousand that are bilingual and unilingual.

```
MTONGUE=FR
|   CANWRITE=N: N Y(17) N(25)
|   CANWRITE=Y: Y Y(120) N(21)
MTONGUE=oth: N Y(52) N(765)
```

Fig. 5. Simple Domain Theory

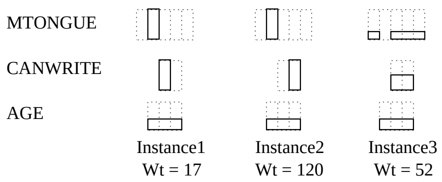

Fig. 6. The Positive Instances

To generate instances consistent with the tree, each path through the tree is represented by as many instances as there are classes at the leaf. Six instances are needed; three for the positive class, bilingual and three for the negative class, unilingual. An instance following a left hand branch has the probability of the attribute value associated with each specified test set to one. For the right hand side branch, the probability is a uniform distribution over the remaining values. Figure 6 shows the probability values for some of the attributes for the positive instances. The negative instances will be identical except for the weights shown at the bottom of Figure 6. The attribute *mother-tongue* has five possible values, indicated by the dashed rectangles. The first two instances travel down the topmost branch of the decision tree. They have the probability of the mother-tongue being French set to one, indicated by the bold continuous rectangle. The third instance, which travels down the bottommost branch, has the probability of the mother-tongue being French set to zero and all other values of mother-tongue are set to a probability of one quarter. The first two instances travel different branches of the attribute *can write*. The first instance has a one for the "N" value, the second instance a one for the "Y" value. All unused attributes on a specific path, such as *age*, have a uniform distribution across all values.

Using these instances, it is now possible to change the order of the tests, or indeed to add a new test, and produce the same partition of the attribute space into classes. Figure 7 shows the effect of changing the root node from *mother-tongue* to *can write*. The same number of instances are classified as bilingual and unilingual. The distribution on the center branch is the same, but the top and bottom most branches have changed. As these two branches are a mixture of instances where the majority class was unilingual, they still classify instances as unilingual. The topmost branch is made up of the first instance in Figure 6 plus half the third instance. The third instance had a uniform probability for *can write*. As this attribute is now the root node, this instance must be sent down both branches. This is achieved by making an additional copy of the instance. For the original instance, the probability of value "N" for *can write* is set to one, the same as the first instance. For the copy the probability for value "Y" is set to one, the same as the second instance. As there are only two values, the weight for both instances is set to half the original weight. If there were more, the weight is the original weight times the fraction of values represented by the branch. There is no longer a uniform distribution for the attribute *mother-tongue*, which was different for the first and third instances. The splitting criterion would choose this attribute as a possible additional test. This would not, however, change the classification of instances. A linear scan across the instances indicates that the classification will not change if new tests are added, so no split is made.

```
CANWRITE=N: N :- Y(43) N(407)
CANWRITE=Y
   MTONGUE=FR: Y :- Y(120) N(21)
   MTONGUE=oth: N :- Y(26) N(383)
```

Fig. 7. Changing the Root Node

To update the tree at each existing test, the splitting criterion is applied to a combination of the data generated to be consistent with the tree and the new data. If the original theory preferred certain attributes, any changes to the theory will tend to use those attributes, rather than introducing new ones, say by promoting them higher up the tree. New tests will only be introduced if the new data has a strong preference for them. To achieve this, the splitting criterion is applied separately to the old and new data. The values returned are combined linearly to form a single value. The coefficients are determined by the number of instances, or weight, of the old data versus the number of new instances.

There are four possibilities that might occur. A new test might be added where the original tree had a leaf. The original test might be replaced by a different test. The original test might deleted altogether, or the old test maintained. To determine which takes place, confidence in the new best test is determined. If the original tree had a leaf at this node, a new test will be added to the tree if the lower bound of the confidence interval is greater than 0.5. This is the same as growing the tree directly from the data. If the new test is the same as the old test nothing will change. If the new test is different and its confidence interval exceeds the threshold it is compared to the new test. If the lower bound of the confidence interval for the difference exceeds the threshold, the test will be changed. If the new test does not exceed the threshold and the upper bound of the confidence interval on the difference does not include zero, the test is deleted.

The old and new data might also differ in how an instance should be classified at a leaf. A confidence interval can be used to decide which classification should be used. Again a bootstrapping technique is used, this time based on just the binomial ratios. At the leaf we can use lower bound of accuracy directly rather than our test statistic.

5 Evaluating an Existing Theory

In this section, we present an experiment showing how the method discussed in section 4 is applied to a theory representing the views held by politicians in Canada in 1901. The theory has been developed from analyses of debate in the House of Commons and newspaper coverage of political discussion about the language questions posed in the 1901 census. For a comprehensive analysis of the political debate about language, see Gaffield [3]. The decision tree representing the theory, see Figure 8, was designed to classify an imaginary 1000 people. The design exercise began by ranking attributes according to how politicians of that era expected them to influence bilingualism. Each branch of the tree was then assigned some proportion of the 1000 people, indicated by the numbers in parentheses. Next, each attribute was considered for its effect on the proportion of bilingual speakers, and the appropriate ratio of bilingual to unilingual individuals was assigned to each branch. Politicians certainly did not all agree on the importance of various factors and their perceived influence on reported bilingualism, and therefore the experimental parameters represent a distillation of somewhat divergent views.

Ethnic origin was assessed to be the most important attribute, only those of French origin were expected to be bilingual, most other individuals were expected to be unilingual. The next most important attribute was assessed to be *birthplace*, being urban born was more strongly associated with bilingualism than being rural born. Attributes *sex*, *age* and *can write* were then added in that order. Once the tree was constructed, instances consistent with it were generated. The proportions of the classes at the leaves, indicated by the "Y()" and "N()" in the figure, were then adjusted so that the order of attributes was maintained. The tree is reasonably accurate (78.960%), only 1.2% less accurate than the tree grown directly from the data (80.169%).

Figure 9 shows the politicians' theory after revision using data for the whole of Canada. This revised theory is more accurate than the politicians' theory. It is slightly more accurate (80.204% lower bound 79.926%) than the decision tree grown directly from the data (80.169%), see Figure 2, although the base of the tree is identical. Much of the structure from the theory has been deleted, but quite a lot remains, indicated by the "o" and "+" in figure 9. The most significant

```
ORIG=FR :- (400)
|  BPLACE=RU :- (212)
|  |  SEX=F : N :- 0.235 Y(20) N(65)
|  |  SEX=M :- (127)
|  |     AGE=20-49 :- (72)
|  |     |  CANWRITE=N : N :- 0.444 Y(12) N(15)
|  |     |  CANWRITE=oth : Y :- 0.556 Y(25) N(20)
|  |     AGE=oth : N :- 0.364 Y(20) N(35)
|  BPLACE=oth :- (188)
|     SEX=F :- (78)
|     |  AGE=20-49 :- (48)
|     |  |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|     |  |  CANWRITE=Y : Y :- 0.667 Y(20) N(10)
|     |  AGE=oth :- : N :- 0.333 Y(10) N(20)
|     SEX=M :- (110)
|        AGE=>=50 :- (40)
|        |  CANWRITE=N : N :- 0.444 Y(8) N(10)
|        |  CANWRITE=Y : Y :- 0.545 Y(12) N(10)
|        AGE=oth : Y :- 0.714 Y(50) N(20)
ORIG=oth :- : N :- Y(50) N(550)
```

Fig. 8. The Politicians' Theory

change to the theory is the first test, *mother-tongue* replaces *ethnic origin* and accounts for most of the improvement in the revised theory. The additional structure, indicated by the "+'s" in figure 9, is the part of the politicians' theory which was not deleted when the tree was revised. It identifies two bilingual groups for people whose mother-tongue is not French. Urban males (labeled "+1") of French origin are predominantly bilingual, as are urban females (labeled "+2") of French origin, aged 20 to 49 who can write. This branch accounts for the slight increase in the accuracy of the tree. These groups were identified in the original theory. As the data supports this division, albeit very weakly, they have not been deleted. The additional structure, indicated by the "o's", is not supported by the data. It was not deleted, however, as the tests did not indicate a statistically significant increase in accuracy. This structure does not change the classification of the tree and so could easily be deleted.

From an algorithmic perspective, it seems that attributes were modified and deleted when there was a clear advantage in doing so. But when the data did not support such deletion, the semantics of the original theory was maintained. From a historical perspective, the Canadian politicians of 1901 used mother-tongue to help clarify ambiguities among the labels used for ethnic groups; they did not see language as being a good identifier in and of itself. These theory revision experiments suggest that mother-tongue was more important that politicians believed at the time. But they were aware that times were changing, but probably not to the extent to which the data seems to suggest, and this led to addition of language questions to the census.

```
   MTONGUE=FR :- 0.541
   |  BPLACE=RU :- 0.467
   |  |  AGE=20-49 :- 0.575
   |  |  |  SEX=F : N :- 0.451 Y(1547) N(1886)
   |  |  |  SEX=M : Y :- 0.674 Y(2892) N(1399)
   |  |  AGE=oth  : N :- 0.386 Y(3946) N(6278)
   |  BPLACE=UR :- 0.674
   |     CANWRITE=N :- 0.409
o  |     |  ORIG=FR :- 0.408
o  |     |  |  SEX=F : N :- 0.303 Y(240) N(550)
o  |     |  |  SEX=M : N :- 0.499 Y(451) N(452)
o  |     |  ORIG=oth :- : N :- 0.437 Y(35) N(45)
   |     CANWRITE=Y :- 0.732
o  |        SEX=F : Y :- 0.647 Y(2570) N(1405)
o  |        SEX=M : Y :- 0.813 Y(3399) N(782)
   MTONGUE=oth :- 0.101
+     ORIG=FR :- 0.404
+     |  BPLACE=RU : N :- 0.343 Y(348) N(666)
+     |  BPLACE=UR :- 0.506
+     |     SEX=F :- 0.448
+     |     |  AGE=20-49 :- 0.581
+     |     |  |  CANWRITE=N : N :- 0.344 Y(8) N(16)
+2    |     |  |  CANWRITE=Y : Y :- 0.621 Y(90) N(55)
+     |     |  AGE=oth : N :- 0.295 Y(43) N(103)
+1    |     SEX=M :- : Y :- 0.570 Y(162) N(122)
+     ORIG=oth :- : N :- 0.089 Y(3823) N(38942)
```

Fig. 9. The Revised Theory

6 Limitations and Future Work

From a historical perspective, the census was designed to provide evidence of the learning of English by French-language individuals. The trees, indeed, show this but they also show that a constellation of factors underlay the language patterns including age, sex, and rural-urban differences and this was not uniform across the country. It is for this reason more research is needed on specific geographic areas such as the so-called Bilingual Belt as well as on other data from the census including economic variables. From an algorithmic perspective, the test statistic and other design choices have proven effective in practice on this data set but need to be experimentally validated on other data sets. Confidence in an existing theory might not constant for all parts of the theory. The existing theory determined the old tests and influenced the choice of new tests but did not affect the confidence value. An alternative would be to take a more Bayesian approach, perhaps using credible intervals rather confidence intervals, allowing locally defined confidence values.

7 Conclusions

From a historical perspective, the most compelling conclusions concern the extent to which the Quebec patterns appear to differ from those of the other regions of Canada, and the complexity in the patterns of bilingualism at the turn of the century. From an algorithmic perspective, this work has demonstrated how confidence intervals can be used to identify factors that are both statistically and practically significant. It has also shown how combining a semantic measure of similarity between trees with confidence intervals can be used to evaluate and modify an existing theory.

References

1. B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.
2. E. Frank. *Pruning decision trees and lists*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2000.
3. C. Gaffield. Linearity, non-linearity, and the competing constructions of social hierarchy in early twentieth century canada: The question of language in 1901. *Historical Methods*, 33(4):255–260, 2000.
4. D. D. Margineantu and T. G. Dietterich. Bootstrap methods for the cost-sensitive evaluation of classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 582–590, 2000.
5. R. J. Mooney. Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning*, 10(1):79–110, 1993.
6. S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.
7. M. D. Ornstein. Analysis of household samples: The 1901 census of canada. *Historical Methods*, 33(4):195–198, 2000.
8. J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
9. G. G. Towell and J. W. Shavlik. Knowledge-based artificial neural networks. *Artificial Intelligence*, 70:119–165, 1994.
10. P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, pages 5–44, 1997.