Automatic extraction of multiple underlying causes from textual death records

Graham Kirby[1], Masih Hajiarab Derkani[1], Alan Dearle[1], Jamie Carson[1], Fraser Dunlop[1], Chris Dibben[2], Lee Williamson[2]

[1] University of St Andrews, [2] University of Edinburgh

Overview

Data sets containing natural language strings are increasingly becoming available as outputs from various international initiatives to digitise historical population records. Analysis of such records, for example, historical causes of death, is facilitated by classification to standard systems such as ICD-10.

Most death record systems include multiple causes of death, typically a primary cause and optionally a number of contributing secondary causes. For example, the primary cause of death might be heart failure, contributed to by some underlying chronic condition.

In modern records these separate causes are clearly differentiated by being entered in different fields on the recording form. In some historical records, however, there was no imposed structure, with the person recording the death having the freedom to use any form of language.

The Digitising Scotland project is in the process of transcribing all Scottish birth, death and marriage records from 1855 to 1973. Here we describe our approach to automatic extraction of multiple causes of death from the approximately 11M death records.

Method of Classification

Our approach assumes the availability of a classifier that classifies text strings to single causes, together with some indication of the confidence in the classification, but does not make any assumptions about how this works. For the results presented here we have used an extremely simple exact-match classifier, which returns a successful classification only if it has previously seen an identical string (modulo cleaning by removing punctuation, very common words etc.) in its training data. We are currently investigating the effects of replacing the underlying exact-match classifier with one that uses approximate string matching, or machine learning, or a combination of these.

The single classifier is trained using text fragments each describing a single cause of death, and the corresponding single classification. To extract multiple causes, each death record is split into individual words, every possible sub-set of words extracted, followed by every possible ordering of these sub-sets. The rationale for considering different orderings is to enable minor variations in phrasing to be matched.

Each sub-sequence of words is classified, yielding a classification and confidence value for each one. Using the exact-match classifier, if a sub-sequence is already present in the training set then the classification specified there is returned with 100% confidence. If a different single classifier was used, some results would have lower confidence, due to the lack of identical examples in the training set.

From the set of classified sub-sequences, one or more 'valid' sets of classifications is constructed, according to the following rules:

- **non-overlapping** – the sub-sequences from which the classifications derive must not overlap, i.e. a given word in the original text cannot count towards two different classifications;
- **hierarchy** – where the classification system is hierarchical, multiple classifications must not lie in the same branch of the classification hierarchy. For example, when classifying to ICD-10 the codes J20 "Acute bronchitis" and J206 "Acute bronchitis due to rhinovirus" would not be allowed to occur together.

Finally, one of the valid classification sets is selected, using a metric that attempts to balance the number of classifications, the number of words from the text contributing to those classifications, and the confidence of the single classifier in each of its decisions.

The computational and storage costs of this scheme grow very rapidly as the number of words in the text increases. Currently we define one threshold number of words, after which the different possible orders of the word sub-sets are ignored, and another threshold beyond which the remaining words are ignored.

Evaluation

Data Sets – Since the Digitising Scotland records are not yet available, the following data sets were used to evaluate the multiple classification system.

1. **Kilmarnock** – 23,700 records (with 8,300 unique text strings) from Kilmarnock, Scotland in the period 1861-1901, derived from the 'Demography of Victorian Scotland' project. See Reid, Davies and Garrett (2002) and Reid, Garrett, Davies and Blaikie (2006) for more information on the project and for access to the related census records. These records are coded by historians into a variant of ICD-10 augmented to deal with certain historical terms, with around 800 classes occurring in the data set.
2. **Tasmania** – 93,000 records (22,000 unique text strings) from Tasmania in the period 1838-1899 (Gunn and Kippen 2008). For this data set we have only the set of unique causes; we do not know the number of occurrences of each one. These records are coded into around 1,800 distinct classes.

Classification Performance – Tables 1 and 2 show examples of the style of records found in the two data sets, and the automatically extracted codes together with the 'gold standard' historian-assigned codes. For each data set, those records that had only a single 'gold standard' code were extracted and used to train the system, and the remainder used to evaluate its accuracy.

The examples shown here do not occur in the data sets; since the data is not in the public domain, the examples shown here are fictional, and the resulting classification accuracy is somewhat worse than for real data. The performance metrics presented in Table 3 are, however, derived from the real data.

For each record, codes shown in orange are false positives, i.e. the automatic system decided that the code was present when it was not. Codes shown in red are false negatives, i.e. the system did not produce them when it should have done.

Mental and behavioural disorders due to use of alcohol: Dependence syndrome

Mental and behavioural disorders due to use of alcohol: Acute intoxication

Malaise and fatigue

Table 1. Examples from the Kilmarnock Data Set

Text String	Extracted	Gold Standard
trauma to the head followed by immersion in the sea causing finally death by drowning, this trauma being caused by accident of his having fallen from the ship rosemary then at the pier of oban verdict of jury	W74.00, Y34.07	W74.00, Y34.07
long-term hepatitis and long-standing jaundice/stupor from discharge	R17.00	K75.90, R17.00, R40.20, R58.05
short-term rheumatism/mitral illness of heart/pulmonary apoplexy/jaundice	M79.00, R09.21, R17.00	I38.01, M79.00, R09.21, R17.00
unconsciousness from fall from cliff followed by broncho pneumonia discharge into chest; heart attack	J18.00	I50.90, J18.00, Y34.01
albuminuria; anasarca; dropsy of lung	R60.11, R60.91, R80.01	J81.00, R60.11, R80.01
exposure to freezing was exhausted and had indulged in alcohol	F10.20, X31.00	F10.00, R53.07, X31.00
cold following measles	B05.90, R68.85	B05.90, R68.85
from cold after childbirth	O95.00, R68.85	O95.00, R68.85
spasm arising from teething	A09.09, R25.20	A09.09, R56.80
trauma to head from a fall down the stairs		R99.00, Y34.07
rheumatism and frailty	M79.00	M79.00, R53.03
internal wounds from being crushed between a lorry and the leg of a workbench in the workshop of the lothian district railway near livingston		X59.90, Y34.03
suffocation from having fallen face down in in bed while being intoxicated	Y20.03	F10.00, Y20.03

Exposure to excessive natural cold

Table 2. Examples from the Tasmania Data Set

Text String	Extracted	Gold Standard
Accidentally killed by a falling tree (Verdict of Jury)	39, 408, 678, 1746	39, 408, 678, 1746
Freezing and Old Age. Found in bush at Fallen Tree. Coroner decided inquest not necessary.	1200	407, 648, 1067, 1200
Broken neck and leg bone. Bed abrasions and general exhaustion.	167, 645	116, 645, 716, 717
1. Shock to the system. 2. Fractures and wounds caused by accident (leg was crushed by a threshing machine.	39, 421, 716, 960, 1492	714, 960, 1078, 1492
Suffocation by drowning in the Yellow River, Katoomba. Accidental	560, 1564	39, 97, 556
Intoxication. Disease of Liver	507, 1014	56, 355
Freezing and exposure and inhalation of sleet	648	359, 648
Cancer of Liver. Bruised wound of head	203, 960, 1056	212, 403
Spasms from Inflammation of the membrains of the brain	154, 907, 1511	405, 938
Uraemia. Paroxysms. Hemiplegia	811, 1725	405, 811
Diseased heart Inquest at Wallaby Spit	499	407, 499

Accidental death

Crushed

Fracture - leg/arm

Injury

Shock

Fracture

Machinery accident

Evaluation Metrics

Table 3 shows summary performance metrics for the two data sets. *Precision* measures the proportion of the classifications that were produced, that were correct. *Recall* measures the proportion of the classifications that should have been produced, that actually were. *F1* is a metric that combines precision and accuracy.

Macro-average summaries are produced by calculating the metric separately for each possible class, and averaging. Micro-averages are calculated over the whole data set rather then per-class. This tends to give more useful results if records are unevenly distributed across the classes.

For both data sets, macro-precision is much higher than macro-recall, indicating that most of the classifications produced are reasonable, but there are many classes that are not correctly extracted. The higher micro-precision/recall figures indicate that many of those classes include only small numbers of records.

Table 3. Performance Metrics

Data Set	Training Records	Evaluation Records	Macro-Precision	Macro-Recall	Macro-F1	Micro-Precision/Recall
Kilmarnock	18,877	3,480	84%	40%	40%	85%
Tasmania	7,768	14,249	86%	44%	44%	81%

Ongoing Work

Current and planned further development includes:

- further evaluation using varying partitions of the data into training and evaluation subsets, to verify robustness of the approach
- evaluation with respect to varying levels of the classification hierarchy, where applicable
- consultation with historians to identify types of records most in need of classifier improvement
- optimization of the algorithms to improve scalability in terms of record length
- single classification using approximate string matching and machine learning
- refinement of the approach to handle very large training sets
- synthesis of test data sets from modern data
- comparison with human-centred semi-automatic classification systems

Classification Software

The classification software is published as open source, available at:

http://digitisingscotland.cs.st-andrews.ac.uk/record_classification/

References

- Reid, A., Davies, R. and Garrett, E. (2002). Nineteenth-Century Scottish Demography From Linked Censuses and Civil Registers: a "Sets of Related Individuals" Approach. International Journal of Humanities and Arts Computing, 14(1-2), 61-86.
- Reid, A., Garrett, E., Davies, R. and Blaikie, A. (2006). Scottish Census Enumerators' Books: Skye, Kilmarnock, Rothiemay and Torthorwald, 1861-1901. Available at: *http://www.esds.ac.uk/findingData/snDescription.asp?sn=5596* (Accessed 24/8/15).
- Gunn, P. and Kippen, R. (2008). Household and Family Formation in Nineteenth-Century Tasmania, Dataset of 195 Thousand Births, 93 Thousand Deaths and 51 Thousand Marriages Registered in Tasmania, 1838-1899. Australian Data Archive. Australian National University, Canberra.

Acknowledgements

This work takes place within the Digitising Scotland project, funded by ESRC grant ES/K00574X/1. We are grateful to Eilidh Garrett and Alice Reid for access to the Kilmarnock data set, and to Rebecca Kippen for access to the Tasmania data set.

The authors can be contacted via the lead author at *graham.kirby@st-andrews.ac.uk*.