Mining the Cultural Memory of Irish Industrial Schools Using Word Embedding and Text Classification

Susan Leavy
susan.leavy@ucd.ie
University College Dublin, Ireland

Emilie Pine
emilie.pine@ucd.ie
University College Dublin, Ireland

Mark T. Keane
mark.keane@ucd.ie
University College Dublin, Ireland

Introduction

The Industrial Memories project aims for new distant (i.e., text analytic) and close readings (i.e., witnessing) of the 2009 Ryan Report, the report of the Irish Government's investigation into abuse at Irish Industrial Schools. The project has digitised the Report and used techniques such as word embedding and automated text classification using machine learning to re-present the Report's key findings in novel ways that better convey its contents. The Ryan Report exposes the horrific details of systematic abuse of children in Irish industrial schools between 1920 and 1990. It contains 2,600 pages with over 500,000 words detailing evidence from the 9-year-long investigation. However, the Report's narrative form and its sheer length effectively make many of it findings quite opaque. The Industrial Memories project uses text analytics to examine the language of the Report, to identify recurring patterns and extract key findings. The project represents the Report via an exploratory web-based interface that supports further analysis of the text. The methodology outlined is scalable and suggests new approaches to such voluminous state documents.

Method

A web-based exploratory interface was designed to enable searching and analysis of the contents of the Report represented within a relational database. The relational structure detailed the categories of knowledge contained in the Report along with key information extracted from the text (Figure 1). The Ryan Report is composed of paragraphs containing an average of 87 words. These paragraphs were represented as database instances and annotations detailing semantic content were linked through the relational structure. Named entities were automatically extracted using NLTK (Looper and Bird, 2002).

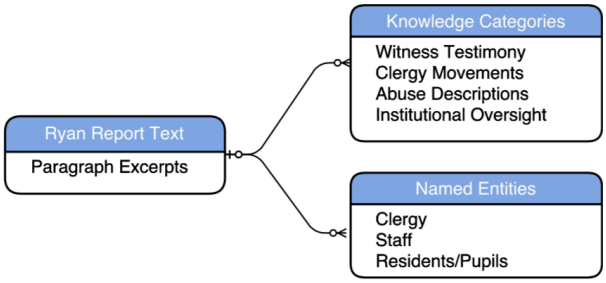

Figure 1: Knowledge Database Relational Structure

Classifying Paragraphs into Different Knowledge Categories

The Ryan Report describes key elements of an enduring system of abuse that operated in Irish industrial schools. Its paragraphs tend to focus on particular topics, allowing them to be classified and annotated. For instance, some cover the extent and nature of abuse, others present witness testimony, report on institutional oversight or on how clergy were moved from one school to another in response to allegations. By classifying paragraphs in terms of these high-level knowledge categories it becomes easier to put a shape on many of the report's findings and to analyse it to provide new readings.

Some of these paragraph-categories were identified using automated text classification. Others were extracted using a rule-based search (e.g., excerpts on institutional oversight). In building classification models, a variety feature sets were examined using a random forest classifier along with manually selected test data. A bag-of-words approach to feature selection yielded results that were over-fitted due to the small samples of training data. However, feature selection based on context-specific semantic lexicons generated from a sample of seed-words using a word embedding algorithm was found to yield accurate results. Lexicons were generated using the word2vec algorithm developed by Mikolov (2013) following an approach identifying synonyms outlined by Chanen (2016).

Movements of Staff and Clergy (Transfer Paragraphs)

An important paragraph-category covers those dealing with the Catholic Church's response to allegations of abuse. The typical response to discovered abuse was to transfer clergy from one institution to another, only for the abuse to re-occur (e.g., "*...Br Adrien was removed from Artane and transferred to another institution...*" *(CICA Vol. 1, Chapter 7, Para-*

graph 829)). Such transfers are described in many different ways in language that often obscures what was happening (e.g., transfers out of the Order, effectively sackings, are described as "dispensations to be released from vows"). We carried out a "by-hand" analysis to find transfer-paragraphs using verb-searches and then expanded this set using machine-learning classifiers.

Initial readings of the Report suggested a set of verbs frequently used to describe the transfer of staff and clergy, including 'transfer', 'dismiss', and 'sack'. The highest-ranking similar words reoccurring over five word2vec models were then identified. Features based on this lexicon, along with names of schools and clergy were extracted from 250 training examples (Table 1). A classification model then classified unseen text from the Report.

Text Category	Features Extracted
Direct Speech	Reporting verbs, personal pronouns, punctuation (colons, quotation marks, commas, question marks, contractions), newlines
Movements of Staff and Clergy	Transfer verbs, names of clergy, schools
Descriptions of Abusive Events	Clergy and staff, parts of body, abusive action, emotions and implements associated with abuse

Table 1: Optimal Features Extracted from Report Identifying Witness Testimony

Witness Testimony (Witnessing Paragraphs)

Witness testimonies in the Ryan Report are indicated through reporting verbs and structural speech markers (e.g., punctuation). Using these features, Schlör et. al. (2016) gained accuracy of 84.1 percent in automatically classifying direct speech. Reporting verbs in the Ryan Report are often specific to its context such as apology, allegation or concession. To extract these from the text, highest-ranking similar words across multiple word embedding models were identified based on seed terms generated from WordNet, 'said', 'told' and 'explained'. The resulting context-specific synonyms combined with WordNet synonyms formed a lexicon of reporting verbs tailored to the language of the Report (Table 2). A classification model was developed using these features along with punctuation information using 500 training examples.

Seed Words	Context Specific Lexicon for Reported Speech				
said	answered	alleged	warned	enounced	explained
told	learned	recounted	claimed	verbalise	believed
explained	confirmed	surmised	denied	verbalised	added
	described	relieved	asserted	assure	replied
	say	protested	witnessed	articulate	thought
	tell	stating	called	apologise	knew
	told	describes	informed	pardon	felt
	state	agreed	said	pardoned	recalled
	posit	admitted	explained	remember	saying
	posited	convinced	advised	articulated	told
	submit	presumed	assured	enounce	thinks
	submitted	screams	tells	condoned	remembered
	express	reported	requested	condone	conceded
	expressed	complained	heard	saw	realised
	narrate	asking	says	explicate	stated
	narrated	commented	confessed	explicated	insisted
	recount	questioned	remarked	apology	guarantee
	recite	accepted	alleged	concluded	asked
	recited	recollection	suggested	mentioned	

Table 2: Context Specific Synonyms Using Word Embedding

Descriptions of Abusive Events (Abuse Paragraphs)

To evaluate the scale of abuse throughout the industrial school system, excerpts from the Report detailing abusive events were extracted. The language describing abuse incorporates a broad range of linguistic features. A set of seed-words from which to base a semantic lexicon for feature extraction, was not immediately apparent on reading the Report. A support vector machine algorithm was therefore used to extract the most discriminative features based on a sample set of 200 paragraphs.

Analysis of the support vectors showed that terms distinguishing excerpts describing abuse formed five categories: abusive actions, body parts, emotions engendered in the victims, implements and names of staff and clergy. Sample words associated with each category were then used as seed-words to generate word embedding models to extract similar terms from the Report. Features based on these five lexicons, combined with names of clergy and staff were then used to generate a predictive model of abusive events.

Findings and Conclusions

This research demonstrates how word embedding can be used to compile context-specific semantic lexicons to extract features for text classification. These features allowed paragraphs for each knowledge category to be automatically classified based on manually selected training data.

Classification	No. Classified Excerpts
Clergy Movements	1,340
Direct Speech	1,920
Abusive Events	1,365

Table 3: Total Number of Classified Paragraphs

The performance of classifiers was evaluated using 10-fold cross-validation on the training data and showed high levels of accuracy in categorisations (Table 4).

Classification	Precision	Recall	F-Score	Accuracy
Clergy Movements	.91	.91	.91	91.2%
Direct Speech	.94	.93	.93	93.6%
Abusive Events	.93	.93	.93	93.3%

Table 4: Performance on Training Data: Random Forest Classifier. Weighted Average Results Using 10-fold Cross-Validation

The classification models were applied to unseen data and performance evaluated by manually inspecting classifications of 600 randomly selected

excerpts from the Report as shown in Table 5. Though overall accuracy levels remained high, precision of the classifications did fall somewhat, especially in relation to identifying speech and transfers.

Classification	Precision	Recall	F-Score	Accuracy
Clergy Movements	.58	1.0	.73	92%
Direct Speech	.84	.92	.88	94%
Abusive Events	.86	.88	.87	95%

Table 5: Performance on Unseen Text: Classification of Report Evaluated on Random Samples

Error analysis showed that incorrectly classified excerpts (false positives and negatives) were commonly those where the meaning of the language was subtle or vague. Paragraphs incorrectly classified as quoted speech for instance, were in fact quotations from letters and diary entries. Unidentified speech excerpts all consisted of short quoted phrases.

Transfers of clergy were reliably detected. However, there was a high rate of false positives due to the fact that the transfer of children throughout the school system is described using similar language (e.g. "*The witness remembered … when he was leaving Artane at nine years of age…" (CICA Vol. 1, Ch. 7, Paragraph 466))*. Classifying excerpts describing abuse yielded few false positives but it also returned the highest levels of false negatives. In these instances, references to abuse was subtle or addressed emotional abuse. As such, it was necessary to manually filter results.

This paper has demonstrated that machine learning can be used to classify text based on a limited number of examples, when used in conjunction with word embedding to generate context-specific semantic lexicons. Re-presenting the Ryan Report in the form of a relational database with a web-based exploratory interface has facilitated comprehensive analysis of the Report, and has exposed new insights about the dynamics of the system of child abuse in Irish industrial schools. In reformulating how the Ryan Report can be presented, this research presents a scalable approach to digital analysis of state reports.

Acknowledgements

This research is part of the Industrial Memories project funded by the Irish Research Council under New Horizons 2015.

Bibliography

Chanen, A. (2016). Deep learning for extracting word-level meaning from safety report narratives. In *Integrated Communications Navigation and Surveillance (ICNS), 2016* (pp. 5D2-1). IEEE.

Loper, E. and Bird, S. (2002). NLTK: e Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1.* (ETMTNLP '02). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 63–70.

Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS2013),* pp. 3111–119.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U., & Tello, J. C. (2016). Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. In *Digital Humanities 2016* (pp. 346-353).