Can Character Embeddings Improve Cause-of-Death Classification for Verbal Autopsy Narratives?

Zhaodong Yan
Dept of Electrical and Computer Engineering
University of Toronto
Toronto, Ontario, Canada
zhaodong.yan@mail.utoronto.ca

Serena Jeblee
Dept of Computer Science
University of Toronto
Toronto, Ontario, Canada
sjeblee@cs.toronto.edu

Graeme Hirst
Dept of Computer Science
University of Toronto
Toronto, Ontario, Canada
gh@cs.toronto.edu

Abstract

We present two models for combining word and character embeddings for cause-of-death classification of verbal autopsy reports using the text of the narratives. We find that for smaller datasets (500 to 1000 records), adding character information to the model improves classification, making character-based CNNs a promising method for automated verbal autopsy coding.

1 Introduction

1.1 Verbal autopsies

Each year, two-thirds of the 60 million deaths in low-and-middle-income countries do not have a known cause of death (CoD), usually because they occurred outside of health facilities and no physical autopsy was performed (United Nations, 2013). Verbal autopsy (VA) surveys are one method of assessing the true distribution of CoDs in these regions. These surveys are conducted by lay interviewers and typically include demographic data, multiple-choice questions, and a free-text narrative, which details the events leading up to the person's death. These records are later coded by physicians for cause of death.

Although several attempts have been made to automate this coding process, including systems such as InterVA (Byass et al., 2012), InSilicoVA (McCormick et al., 2016), the Tariff method (Serina et al., 2015), and others (Miasnikof et al., 2015), the results have not been adequate, in part because they have focused only on the multiple-choice questions and not at all, or only to a limited extent, on the narrative text. However, using the narrative is more convenient because it does not require a specific questionnaire format, and also because it takes less time to collect a short questionnaire and narrative than a long, very detailed survey. Although the narratives present some text processing problems, they allow for more detail and explanation than the structured data alone.

Only a few methods have used the full text of the narrative for CoD classification. Danso et al. (2013) used term frequency and TF-IDF (term frequency–inverse document frequency) features to classify CoD from VA narratives of neonatal deaths. The Tariff method (Serina et al., 2015) uses a small set of word occurrence features from the narrative, but both of these methods ignore word order. Jeblee et al. (2018) used VA narrative text to jointly predict CoD and a list of keywords for each record using a neural network model with word embeddings.

In our work, we therefore focus on the narrative text. However, the models that have been developed to date for VA classification using the narrative, including SVMs (Danso et al., 2013) and neural networks (Jeblee et al., 2018), have used only word-level information. However, recent research has shown that character-level information can improve text classification models, especially in cases where there are many spelling errors and variations, which is the case with VA narratives. Therefore, we investigate here the use of character embeddings for the VA CoD classification task.

1.2 Character embedding models

Instead of representing each word as a vector, as is typically done with word embeddings, we can represent each character in the text as a vector. With traditional word embeddings, any word that is not found in the vocabulary is represented as a vector of zeros, essentially losing all the information from that word. The character-based model does not have this limitation, and therefore can represent unseen words as well as misspelled words.

Another benefit to character-based models is that because of the much smaller vocabulary size, they result in less variation in the input representa-

Proceedings of the BioNLP 2019 workshop, pages 234–239
Florence, Italy, August 1, 2019.

tion, which can be especially useful for very small datasets such as our verbal autopsy records.

Zhang et al. (2015) used a character-level convolutional neural network (CNN) for text classification tasks on a dataset of news articles and internet reviews, demonstrating that the character-level model could outperform word-level models. Verwimp et al. (2017) combined character-level and word-level embeddings by concatenation with padding, and used them with a Long Short-Term Memory (LSTM) language model, achieving better perplexity than similar word-based models.

Si and Roberts (2018) used an LSTM model to learn character embeddings, which were then concatenated with pre-trained word embeddings to extract cancer-related information such as diagnosis, showing that combined character and word-based models can be used successfully for tasks in the medical domain.

2 Data: The Million Death Study

Our dataset of informal medical narratives consists of verbal autopsy reports from the Million Death Study (MDS) (Westly, 2013), a program that collects VAs in India that cover adult, child, and neonatal deaths. We currently have a dataset of 12,045 adult records, 1851 child records, and 572 neonatal records with English narratives (transcribed from handwritten forms). The records are classified into several broad CoD categories: 18 for adult deaths, 9 for child deaths, and 5 for neonatal deaths. (See Table 4 in the Appendix for the list of CoD categories.)

The process of translating the local languages and converting handwritten texts to digital format creates many errors. Many narratives have frequent spelling and grammar errors, such as inconsistent pronouns, sentence fragments, incorrect punctuation, and transcription errors, in addition to many local terms. See Table 1 for an example narrative. The nature of the text means that purely word-based models, especially ones trained on other corpora, are likely to miss key information. In order to address this issue, we add character embedding representations to the classification model to see whether it will improve the results. We also compare this model to the word-only model.

Narrative
Heart failure. The patient death due to breathlessness. The person sufering paralysis and stroke lost on year with chest pain very pressure after then person was head.
CoD category: Other cardiovascular diseases

Table 1: A verbal autopsy narrative with spelling and grammar errors, and the associated CoD category.

3 Models

3.1 Pre-processing

All text is lowercased before being passed to the model, and punctuation is separated from words. Spelling is corrected using PyEnchant's English dictionary (Kelly, 2015) and a 5-gram language model for scoring candidate replacements, using KenLM (Heafield et al., 2013). However, many instances remain where misspellings result in another valid word (such as *dead* being mistyped as *head*) or are too badly misspelled to be corrected. Moreover, many local terms are not handled properly by our automated spelling correction, so while the spelling correction model fixes some of the more apparent errors, many misspellings persist even after this step.

3.2 Word–based model

For the word-based model, we represent each word in the narrative as a 100-dimensional word embedding. The embeddings are trained using the word2vec CBOW algorithm (Mikolov et al., 2013) on the training set of the VA narratives, as well as data from the ICE corpus of Indian English[1] and about 1M posts from MedHelp, an online medical advice forum for patients[2]. The maximum length of input is 200 words, and shorter narratives are padded with zeros.

The classification model is a convolutional neural network (CNN) implemented in PyTorch (Paszke et al., 2017), with windows of 1 to 5 words, max-pooling, and 0.1 dropout.

3.3 Character–based model

For the character-based model we use publicly available pre-trained character embeddings[3] de-

[1] http://ice-corpora.net/ice/avail.htm
[2] http://www.medhelp.org
[3] https://github.com/minimaxir/char-embeddings

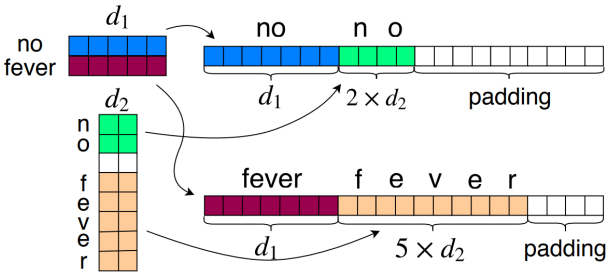

Figure 1: Embedding concatenation model architecture. d_1 is the dimensionality of the word embedding (100), and d_2 is the dimensionality of the character embedding (24).

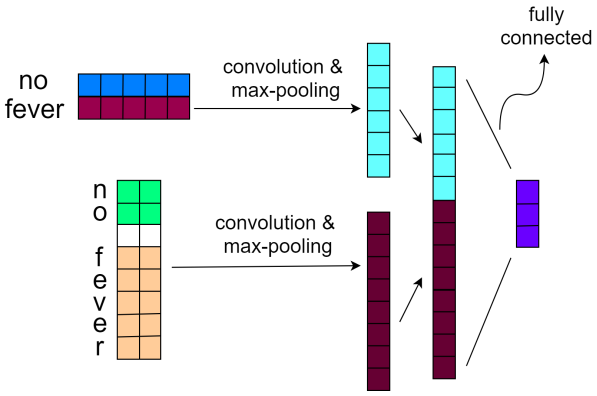

Figure 2: Model combination architecture.

rived from GloVe vectors (Pennington et al., 2014) trained on Common Crawl. The dimensionality of the character embeddings is reduced from 300 to 24 with principal component analysis (PCA).

We also tried learning the embeddings directly as a first layer in the model, but the model was unable to learn useful embeddings, likely because our training set is too small.

The character-based classification model is also a CNN, with a maximum of 1000 characters for each narrative. We also remove punctuation for the character-based model.

3.4 Combined models

We use two different methods of combining the word and character embeddings: embedding concatenation and model combination.

For embedding concatenation, we simply concatenate the word embedding for each word with the ordered character embeddings for the characters in the word. Since words have different numbers of characters, we keep only the first 7 characters of the word, and if the word is shorter than 7 characters we pad the embedding with zeros. In the dataset, 87% of words have 7 characters or fewer, and no improvement was seen by using thresholds of 5, 6, 8, 9, or 10 characters. See Figure 1 for a diagram of the embedding concatenation.

For the model combination, we use all but the final layer of both the word-based CNN and the character-based CNN in parallel, which each produce a feature vector. Before the final classification layer, we concatenate the output vectors from these two networks, and use the combined vector as input to the final feed-forward layer that produces the classification probabilities. See Figure 2 for the diagram of the model architecture[4]. This model allows us to combine the full information from both the word-level and character-level models. However, it also requires the model to learn almost twice as many parameters.

4 Results

We evaluate the four different models using precision, recall, and F_1 score. We also report *cause-specific mortality fraction accuracy* (Murray et al., 2011, 2014), which measures how similar the predicted CoD distribution is to the true distribution. A *cause-specific mortality fraction* (CSMF) is the fraction of a population whose death is attributable to a specific cause. *CSMF accuracy* (CSMFA) is then defined in terms of the difference between the true and predicted fraction for each of k causes:

$$CSMFA = 1 - \frac{\sum_{j=1}^{k} |CSMF_j^{true} - CSMF_j^{pred}|}{2(1 - min(CSMF_j^{true})}$$

The results of CoD classification using 10-fold cross-validation are presented in Table 2.

Since we hypothesized that the character information would improve results particularly for smaller datasets, we also evaluated the models on a subset of the adult data, which consists of 10% of the original adult dataset, evaluated with 10-fold cross-validation (about 137 records in each test set). We call this dataset "Adult small".

5 Discussion

Overall, the embedding concatenation model performs the best across all individual-level metrics, except on the full adult dataset, where the word

[4]The model code is available at: https://github.com/sjeblee/verbal-autopsy

Model	Precision	Recall	F_1	CSMFA
Adult (18 categories)				
Word embedding	**.759**	**.755**	**.751**	**.933**
Char. embedding	.690	.684	.680	.922
Emb. concatenation	.716	.699	.699	.912
Model combination	.629	.620	.609	.872
Adult small (18 categories)				
Word embedding	.453	.500	.456	.773
Char. embedding	.609	.603	.589	.837
Emb. concatenation	**.691**	**.669**	**.660**	**.861**
Model combination	.590	.596	.571	.827
Child (11 categories)				
Word embedding	.713	.707	.697	**.902**
Char. embedding	.655	.638	.623	.851
Emb. concatenation	**.740**	**.718**	**.712**	.890
Model combination	.640	.638	.627	.890
Neonate (5 categories)				
Word embedding	.515	.556	.515	.795
Char. embedding	.504	.502	.482	.795
Emb. concatenation	**.562**	**.585**	**.556**	**.819**
Model combination	.502	.530	.495	.807

Table 2: Results from 10-fold cross-validation for each age group in the MDS dataset.

Cat	1	2	3	4	5	*n*
1	0.870	0.043	0.000	0.043	0.043	23
2	0.294	0.588	0.118	0	0	17
3	0.818	0.091	0.091	0	0	11
4	0.500	0.250	0	0.250	0	4
5	0.500	0.333	0.167	0	0	6

Cat	1	2	3	4	5	*n*
1	0.826	0.043	0.043	0	0.0869	23
2	0.235	0.588	0.176	0	0	17
3	0.545	0.182	0.273	0	0	11
4	0.500	0	0.250	0	0	4
5	0.500	0	0.333	0	0.167	6

Table 3: Confusion matrices for the neonatal test set (iteration 0). **Top:** results from the word embedding model. **Bottom:** results from the embedding concatenation model. Rows are the correct CoD categories and columns are the predicted categories. *n* is the number of records belonging to that category in the test set.

embedding model performs the best. For the child dataset, the word-based model performs the best in terms of CSMF accuracy, which means that it best captures the distribution of CoD categories, but the character-based model achieves better accuracy on classifying individual records.

For the adult data, reducing the dataset size to 10% of the original size causes a sharp decrease in accuracy for the word-based model, but only a smaller decrease for the character-based and combined models, showing that the character embeddings are more robust to data size.

Table 3 shows the confusion matrix for the five classes of the neonatal test set from the word embedding model versus the embedding concatenation model. We can see that both models have a heavy preference for the most frequent class (1 *Prematurity and low birthweight*). The embedding concatenation model achieves better accuracy on class 3 (*Birth asphyxia and birth trauma*) and class 5 (*Ill-defined*), but performs worse on class 4 (*Congenital anomalies*), which is the smallest class.

For the child data, the embedding concatenation performs much better on class 1 (*Pneumonia*) (68% accuracy vs. 44%) and class 6 (*Non-communicable diseases*) (83% vs.78%), and class 10 (*Ill-defined*) (33% vs. 11%), while the word-based model performs better on class 4 (*Other infections*) (76% with the embedding concatenation model vs. 84% with the word model).

The best performing classes for the adult dataset are class 5 (*Maternal*), 15 (*Road traffic incidents*), and 16 (*Suicide*), which are also the categories which have the highest physician agreement. For the small adult dataset, the embedding concatenation model performs noticeably better on classes 4 (*Unspecified infection*), 8 (*Neoplasms*), 16 (*Suicide*), and 18 (*Ill-defined*).

Overall the character information seems to improve accuracy with the smaller datasets, due to its much smaller vocabulary size and its ability to handle spelling variations and unknown words. The combined model performs the best on all of the small datasets, regardless of the number of categories, and especially seems to perform better on more ambiguous categories like *Ill-defined* and *Unspecified infections*.

6 Conclusion and future work

We have shown that character information can improve classification of CoD for verbal autopsies, for smaller datasets, which are very common in the case of VAs. To our knowledge, this is the first application of character-based neural network models to VA narratives.

Due to differences in the datasets, we cannot make direct comparisons to other automated methods. However, since they typically have recall

scores around 0.6, our method is competitive. In addition, this method can be applied to any VA dataset with narratives, regardless of the country of origin or the specific survey form.

Future work may include using a language model with character information, such as ELMo (Peters et al., 2018), but we would have to rely on out-of-domain data since the VA dataset is too small to effectively train ELMo or a similar model. The paucity of VA data is one of the biggest obstacles to automated coding.

In the future we also plan to expand these models to other languages, as there are larger VA datasets in languages such as Portuguese and Hindi. We will also investigate using the structured data in addition to the narrative to improve performance.

Acknowledgments

We thank Prabhat Jha of the Centre for Global Health Research for providing the dataset. Our work is supported by funding from the Natural Sciences and Engineering Research Council of Canada and by a Google Faculty Research Award.

References

Peter Byass, Daniel Chandramohan, Samuel Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy Graham, Abraham Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, Anand Krishnan, Jordana Leitao, Frank Odhiambo, Osman Sankoh, and Stephen Tollman. 2012. Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Global Health Action*, 5:19281.

Samuel Danso, Eric Atwell, and Owen Johnson. 2013. A comparative study of machine learning methods for verbal autopsy text classification. *International Journal of Computer Science Issues*, 10(6).

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Serena Jeblee, Mireille Gomes, and Graeme Hirst. 2018. Multi-task learning for interpretable cause of death classification using key phrase prediction. In *Proceedings of the BioNLP 2018 Workshop*, pages 12–17, Melbourne, Australia. Association for Computational Linguistics.

Ryan Kelly. 2015. Pyenchant. `http://pythonhosted.org/pyenchant/`.

Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel Clark. 2016. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(15):1036–1049.

Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha. 2015. Naïve Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine*, 13(1):286–294.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. 2011. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population Health Metrics*, 9:28. Erratum (Murray et al., 2014).

Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. 2014. Erratum to: Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population Health Metrics*, 12:7.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS 2017 Autodiff Workshop*, pages 1–4.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237.

Peter Serina et al. 2015. Improving performance of the Tariff method for assigning causes of death to verbal autopsies. *BMC Medicine*, 13(1):291.

Yuqi Si and Kirk Roberts. 2018. A frame-based NLP system for cancer-related information extraction. *AMIA Annual Symposium Proceedings*, 2018:1524–1533.

Department of Economic and Social Affairs, Population Division, United Nations. 2013. *World Population Prospects: The 2012 revision*. ST/ESA/SER.A/334.

Lyan Verwimp, Joris Pelemans, Hugo Van hamme, and Patrick Wambacq. 2017. Character-word LSTM language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 417–427, Valencia, Spain. Association for Computational Linguistics.

Erica Westly. 2013. Global health: One million deaths. *Nature*, 504(7478):22–23.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS '15, pages 649–657.

A Appendix

Num	Category
	Adult
1	Acute respiratory infections
2	Tuberculosis
3	Diarrhoeal
4	Unspecified infections
5	Maternal
6	Nutrition
7	Chronic respiratory diseases
8	Neoplasms
9	Ischemic heart disease
10	Stroke
11	Diabetes
12	Other cardiovascular diseases
13	Liver and alcohol
14	Other non-communicable diseases
15	Road traffic incidents
16	Suicide
17	Other injuries
18	Ill-defined
	Child
1	Pneumonia
2	Diarrhoea
3	Malaria
4	Other infections
5	Congenital anomalies
6	Non-communicable diseases
7	Injuries
8	Nutritional
9	Other
10	Ill-defined
11	Cancer
	Neonate
1	Prematurity and low birthweight
2	Neonatal infections
3	Birth asphyxia and birth trauma
4	Congenital anomalies
5	Ill-defined

Table 4: Cause of death categories used for the MDS data.