Towards Automating Healthcare Question Answering in a Noisy Multilingual Low-Resource Setting

Jeanne E. Daniel
Applied Mathematics
Stellenbosch University
jeanne.e.daniel@gmail.com

Willie Brink
Applied Mathematics
Stellenbosch University
wbrink@sun.ac.za

Ryan Eloff
Electrical & Electronic Engineering
Stellenbosch University
ryan.peter.eloff@gmail.com

Charles Copley
Data Science
Praekelt Foundation
charles@praekelt.org

Abstract

We discuss ongoing work into automating a multilingual digital helpdesk service available via text messaging to pregnant and breastfeeding mothers in South Africa. Our anonymized dataset consists of short informal questions, often in low-resource languages, with unreliable language labels, spelling errors and code-mixing, as well as template answers with some inconsistencies. We explore cross-lingual word embeddings, and train parametric and non-parametric models on 90K samples for answer selection from a set of 126 templates. Preliminary results indicate that LSTMs trained end-to-end perform best, with a test accuracy of 62.13% and a recall@5 of 89.56%, and demonstrate that we can accelerate response time by several orders of magnitude.

1 Introduction

MomConnect is a project led by South Africa's National Department of Health (NDOH), and is freely available at public clinics to all mothers and pregnant women who wish to sign up. The platform provides maternal support through text messaging in all 11 official languages of South Africa and has had 2.6 million registrations since 2014. MomConnect is unique in the public health sector, allowing users to pose questions to helpdesk staff who respond manually. The recent introduction of WhatsApp as a channel additional to SMS has increased the volume of questions substantially. This presents a significant challenge for the staffing complement, and median response time is currently 20 hours.

The templated-based nature of answers provides an immediate opportunity for using computational linguistics to automate the response pipeline. In a similar study, Engelhard et al. (2018) evaluated the need for and feasibility of automated message triage to improve helpdesk responsiveness to high-priority messages.

Language	Users
Afrikaans	2,578
English	51,250
Ndebele	97
N. Sotho	3,400
S. Sotho	749
Swati	287

Language	Users
Tsonga	1,361
Tswana	948
Venda	1,105
Xhosa	6,821
Zulu	17,615
Total	**86,211**

Table 1: Unique users in dataset per sign-up language.

During 2018 we worked with the NDOH to gain access to MomConnect data for research purposes. After rigorous ethical clearance and user privacy protection protocols, we obtained a static copy of about 230,000 raw textual question-answer pairs. The primary objective of this research is to investigate ways in which the burden on helpdesk staff can be reduced, which could enable MomConnect to scale towards a wider reach and a more effective service. One aspect of this is to automate the question answering process.

A language label is recorded when a user signs up at the clinic, but users are free to ask questions in any language, which poses a significant challenge to the language processing problem. However, the labels do provide a proxy of the language imbalance within the dataset, as indicated in Table 1.

The challenge of automating MomConnect gives us rare access to a fairly large dataset of closed-domain multilingual questions paired with template English answers, with many questions in low-resource languages, unreliable language labels, a prevalence of code-mixing, misspellings and use of shorthand, and some inconsistencies in

the answers. While the data cannot be made public due to its highly sensitive nature, we can report our findings and provide guidelines for future problems of a similar nature.

In this paper we describe the process of acquiring, anonymizing and filtering the dataset, deduplicating the answer set, and our first attempts towards automating the answering of questions. The work provides a unique opportunity to apply computational linguistic theories to a real-world problem for social impact.

2 Related Work

Question Answering (QA) aims to interpret natural language questions and respond appropriately with natural language answers. Current approaches achieve impressive results on factoid, list and definition questions, but struggle in real-world settings where the questions and answers are more complex (Wang and Ittycheriah, 2015). FAQ approaches aim to economically reuse previously answered questions to guide future answers (Burke et al., 1997), but a core challenge is to calculate similarity between questions with little word overlap. Proposed solutions to this challenge include machine-readable dictionaries like the semantic lexicon WordNet (Miller, 1995), manual rules or templates (Sneiders, 2002), and statistical NLP and information retrieval techniques (Berger et al., 2000). Each has its drawbacks: machine-readable dictionaries exist only for a select few languages, manual template creation is time-consuming and does not scale well, and statistical methods usually require large datasets.

Distributional semantic models learn a mapping of words in their textual form to a dense, low-dimensional vector space (Mikolov et al., 2013a,b; Pennington et al., 2014). Such methods render contextual and co-occurrence information from large open-domain text repositories. For low-resource languages there is often not sufficient data to develop useful word embedding models, and several approaches to deal with this issue have been proposed. With a collaborative filtering technique called positive-unlabeled learning, unobserved word pairs can provide valuable information, especially in low-resource settings (Jiang et al., 2018). Another approach is cross-lingual word representations, where a shared word embedding is trained from multiple languages. This approach facilitates knowledge transfer from high-resource to low-resource language models (Ruder, 2017), and can lead to more robust multilingual information retrieval (Vulić et al., 2015). Embeddings trained on multilingual data with code-mixing also seem to outperform those trained separately on monolingual data (Pratapa et al., 2018).

3 Data Acquisition and Anonymization

Data acquisition followed a rigorous ethical clearance process and, given the highly sensitive nature of the data (disclosing HIV status, for example), access was conditional on anonymization and non-distribution of the data, and granted with the sole purpose of research. A future goal of this work is to provide a process for data generated on similar platforms to be shared more widely for research purposes, without compromising any data protection rights of individuals.

Guided by the General Data Protection Regulations (GDPR), and prior to any data processing or analysis, an anonymization protocol was established to meet the "motivated intruder" test. It can be insufficient to simply remove identifiers such as name and telephone number, as identities might still be deducible from contextual information. It would be problematic if, for example, a certain clinic location and sign-up language narrowed possibilities to a handful of individuals. As such, we rank identifying information by importance to our research and algorithmically remove data in order of increasing priority, to ensure some lower bound on the size of any single distinguishable group. This approach is conceptually similar to the k-anonymization algorithm (Samarati and Sweeney, 1998; El Emam and Dankar, 2008). We also replace absolute identifiers (e.g. expected delivery date) with relative quantities (e.g. days to delivery). Age data is bucketed and district information hashed against one-time random numbers, to prevent direct identification.

In future we plan to improve the protocol with differential privacy techniques (Dwork and Roth, 2014) where applicable.

4 Answer Selection

We proceed to describe our first attempts at automating answer selection for MomConnect. We evaluate naive Bayes on bag-of-words, exact and approximate k-nearest neighbors on cross-lingual word embeddings, as well as long-short term memory networks trained end-to-end.

Question	Template Answer
What causes heartburn?	Hormone progesterone relaxes smooth muscles, thus the valve that separates the gastric acid relaxes, allowing it to go up causing heartburn. Avoid consuming too much spicy, acidic, fizzy, citrus fruits, chocolates, lots of sugar, rich or fried or fatty food. Eat small meals slowly, wear loose fitting clothes, and do not sleep immediately after eating. Drink peppermint tea and Gaviscon.
Kungani ngihlale ngigula njalo ekuseni?	Do you feel like vomiting? Morning sickness or nausea is common in the first 3 months. It should ease from 4 - 7 months. Avoid food with too much fat and spices. Eat dry bread or a dry biscuit when you wake up. If you cannot eat, drink lots of water or tea but not alcohol. Adding some ginger, mint or lemon to your tea may help to ease the nausea.
What are the signs of labour?	Signs of labour include a jelly-like discharge, your water breaking, and regular and painful labour contractions. Make sure you can get to a hospital.

Table 2: Sample question-answers pairs. Questions can be posed in any of South Africa's 11 official languages, while the template answers are currently all English.

4.1 Data Preparation

In the raw dataset of 230K question-answer pairs we discovered 42,675 unique answers, approximating a power law distribution. During initial investigations we found that many answers were near-duplicates of others, likely due to revisions and updates in the history of the manual answering process. Near-duplicates were automatically identified using a word-level Jaccard similarity index, and substituted with the more frequently occurring answers. A small sample of positive and negative word-level matches were manually verified.

This being a work-in-progress, we decided for now to focus on answers appearing at least 128 times in the dataset. This number attempts to address the need to include as many training samples per answer as possible to reduce the variance, while ensuring that under-represented languages (such as Ndebele, with only 97 registered users) are not excluded in the reduced dataset. This leaves us with 126 template answers, and account for close to 70% of all the data. Table 2 shows a few examples. The reduced set of 150K question-answer pairs were split into training, validation and test data (60:20:20).

The remaining 30% of data makes up the long tail of the frequency distribution of answers, many of which occurring only once or twice, and modelling these answers is reserved for future work.

4.2 Cross-lingual Word Embeddings

Motivated by the literature on cross-lingual word embeddings (Vulić et al., 2015; Ruder, 2017; Pratapa et al., 2018) and our data having several low-resource languages, unreliable language labels and a prevalence of code-mixing, we opt to mix all the languages together into a shared cross-lingual continuous bag-of-words embedding space (Mikolov et al., 2013a)

Characters other than Latin symbols, spaces and numerals are removed. We do not remove any stop words, in order to preserve the limited vocabulary of some of the low-resource languages, and end up with a dictionary size of 65,547. For the nearest-neighbor classifiers, the word embeddings of all the words in a question are averaged into a single vector (Wieting et al., 2015).

For a peek at what the continuous bag-of-words embedding model does, here is an example of the same word in English, Zulu and Xhosa, and their respective closest neighbors in embedding space, using cosine distance:

child: baby, bbe, babe, bby, babay

ingane: ingan, yami, ngane, umtwana

umntwana: umtwana, wam, umntana, wami

Different spellings and shorthand of the same word or concept tend to be clustered together, which is useful when working with SMS and WhatsApp messages. Note also the slight overlap in the Zulu and Xhosa examples, due to the two languages being closely related.

4.3 Classification

We perform classification on the questions to select most appropriate answers from the 126 templates. As a baseline we train a multinomial naive Bayes (MNB) classifier on bag-of-words represen-

tations of the questions, using as dictionary only the 7,000 most frequent words across the training set. We then consider k-nearest neighbor (k-NN) classification on the averaged word embeddings of questions, using uniformly weighted majority voting, and for increasing values of k. We also consider locality-sensitive hashing (LSH), an approximate nearest neighbor algorithm, which sacrifices accuracy in k-NN for efficiency. With LSH, embedding vectors are randomly hashed into short binary encodings that preserve local information, thus enabling nearest neighbor searching in sublinear time (Andoni and Indyk, 2008).

Long short-term memory (LSTM) networks have been shown to model sequential text data well (Tai et al., 2015). We train various networks end-to-end, with increasing numbers of hidden units (LSTM-k will denote a network with k hidden units). Each model takes a variable-length sequence of word IDs as input and has a softmax output layer for classification. The networks are optimized using Adam (Kingma and Ba, 2014), with a learning rate of 10^{-3}, batch size of 32, and early stopping based on validation loss. For regularization we apply 40% dropout (Srivastava et al., 2014) to the final layer of the LSTM. We experimented with using sequences of our cross-lingual word embeddings as input, but saw better performance with end-to-end training on sequences of word IDs. We also tested bidirectional LSTMs but found no improvement in performance.

5 Results

The models are evaluated by classification accuracy on the test set of 30K as yet unseen question-answer pairs. We also identify a "low-resource" part of the test set, and measure accuracy on that. Given our inclusion of the entire dictionary of words and the absence of reliable language labels, we wish to understand how the model performs on uncommon words and sentences. Thus, we rank each word in the dictionary by its frequency over the training set, as a proxy for belonging to high- or low-resource languages. Questions belonging to the test set are ranked based on how many of their words have high frequencies, and we extract the bottom 25% as our "low-resource" (LR) test set. Accuracies obtained by the various models are displayed in Table 3.

The MNB baseline performs quite well, both on the full test set and the LR test set, but possibly due to bias for the high-resource languages in its bag-of-words features. The nearest neighbor models (k-NN and k-LSH) show almost no improvement over MNB, and do worse on the LR set. The efficient LSH models perform almost the same as the NN models they approximate. The LSTM models seem to perform best. Increasing the number of hidden units improves accuracy on the full test set, but decreases accuracy on the LR set. This could again be due to slight overfitting on the high-resource languages during training.

While LSTM shows a significant improvement over the other models, it reaches an accuracy of only 62.13% on the full test set. This is understandable given the complexities of noisy data, multilinguality, and code-mixing, but succeeding only 6 times out of 10 is insufficient for a real-world implementation.

In order to gauge the feasibility of a top-5 recommender system assisting a human operator, we also measure recall@5 for the MNB baseline and LSTM models. Results are shown in Table 4. The best performance of 89.56% on the full test set and 81.23% on the LR set is encouraging, and could be considered for a real-world implementation.

Model	**Full** (%)	**LR** (%)
MNB	54.15	49.03
5-NN	53.18	42.80
25-NN	54.43	44.96
50-NN	53.71	44.27
5-LSH	51.42	42.25
25-LSH	54.33	45.18
50-LSH	52.74	44.59
LSTM-64	61.93	**56.12**
LSTM-128	61.76	55.53
LSTM-256	61.97	54.88
LSTM-512	**62.13**	54.95

Table 3: Classification accuracies (in %) of various models, on the full and low-resource (LR) test sets.

Model	**Full** (%)	**LR** (%)
MNB	82.42	77.01
LSTM-64	88.33	80.56
LSTM-128	88.73	81.15
LSTM-256	89.27	81.02
LSTM-512	**89.56**	**81.23**

Table 4: Recall@5 on the full and LR test sets.

The task of querying one of the trained models for an answer (or top five answers) to a question takes a second or two on an ordinary desktop computer. This is a significant improvement in response time over the median of 20 hours currently required by the manual answering process, and can enable MomConnect to scale.

6 Conclusion and Future Work

We described the first steps towards automating a multilingual digital helpdesk for pregnant and breastfeeding mothers in South Africa. Gaining access to data was subject to ethical clearance and data anonymization, due to the highly sensitive nature of the content and the vulnerability of individuals involved.

We considered various approaches to the answer selection problem in a noisy, multilingual, low-resource setting. LSTM networks trained end-to-end outperformed all the other models tested, achieving accuracies of about 62% and 56% on the full and low-resource test sets, respectively. The best LSTM further achieved a recall@5 of almost 90%. Such a model can serve in a semi-automated answer selection process, with a human in the loop to choose the final answer. This could significantly reduce the burden of the current staffing compliment, if approximately 70% of the queries can be dealt with in a semi-automated manner. In the case where the human does not agree with any of the suggested answers, the option can remain for the human operator to manually select the correct standardized response, as is currently done. This feedback can help improve the automated response service, and assist future research tasks.

A next step would be comprehensive error analysis for a better understanding of where the models succeed or fail in capturing semantic information, particularly for the low-resource languages. We are also working to include into our models the long tail in the distribution of template answers. We further intend to explore transfer learning techniques (Zhang et al., 2017) as well as deep architectures designed specifically for answer selection (Lai et al., 2018). There is also scope to develop language identification tools using the unreliable language labels as noisy priors. This could assist with training separate models for the low-resource languages, or provide an answer in the same language as the question.

Acknowledgments

We have gained valuable insights into the nature and processes of MomConnect, and wish to give recognition to the Praekelt Foundation and South Africa's National Department of Health for their initiative and hard work in building and maintaining this platform. We are grateful to Peter Barron for assisting in the process of acquiring the dataset, and Herman Kamper for valuable feedback. We thank the CSIR Centre for Artificial Intelligence Research for financial assistance.

References

Alexandr Andoni and Piotr Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117–122.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–199.

Robin Burke, Kristian Hammond, Vladimir Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: experiences with the FAQ FINDER system. Technical report, University of Chicago.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3):211–407.

Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

Matthew Engelhard, Charles Copley, Jacqui Watson, Yogan Pillay, Peter Barron, and Amnesty E. LeFevre. 2018. Optimising mHealth helpdesk responsiveness in South Africa: towards automated message triage. *BMJ Global Health*, 3:e000567.

Chao Jiang, Hsiang-Fu Yu, Cho-Jui Hsieh, and Kai-Wei Chang. 2018. Learning word embeddings for low-resource languages by PU learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1034.

Diederik Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *Computing Research Repository*, arXiv:1412.6980.

Tuan Manh Lai, Trung Bui, and Sheng Li. 2018. A review on deep learning techniques applied to answer selection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2132–2144.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Computing Research Repository*, arXiv:1301.3781.

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

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

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

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *Computing Research Repository*, arXiv:1706.04902.

Pierangela Samarati and Latanya Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report.

Eriks Sneiders. 2002. Automated question answering using question templates that cover the conceptual model of the database. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems*, pages 235–239.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566.

Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. *Information Processing & Management*, 51:111–147.

Zhiguo Wang and Abraham Ittycheriah. 2015. FAQ-based question answering via word alignment. *Computing Research Repository*, arXiv:1507.02628.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *Computing Research Repository*, arXiv:1511.08198.

Yuan Zhang, Regina Barzilay, and Tommi S. Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. *Computing Research Repository*, arXiv:1701.00188.