gaHealth: An English–Irish Bilingual Corpus of Health Data

Séamus Lankford [1], Haithem Afli [1], Órla Ní Loinsigh [2], Andy Way [2]

[1] Department of Computer Science, Munster Technological University, Cork, Ireland.
[2] Adapt Centre, Dublin City University, Dublin, Ireland
{seamus.lankford, haithem.afli}@mtu.ie
{andy.way, orla.niloinsigh}@adaptcentre.ie

Abstract

Machine Translation is a mature technology for many high-resource language pairs. However in the context of low-resource languages, there is a paucity of parallel data datasets available for developing translation models. Furthermore, the development of datasets for low-resource languages often focuses on simply creating the largest possible dataset for generic translation. The benefits and development of smaller in-domain datasets can easily be overlooked. To assess the merits of using in-domain data, a dataset for the specific domain of health was developed for the low-resource English to Irish language pair. Our study outlines the process used in developing the corpus and empirically demonstrates the benefits of using an in-domain dataset for the health domain. In the context of translating health-related data, models developed using the *gaHealth* corpus demonstrated a maximum BLEU score improvement of 22.2 points (40%) when compared with top performing models from the LoResMT2021 Shared Task. Furthermore, we define linguistic guidelines for developing *gaHealth*, the first bilingual corpus of health data for the Irish language, which we hope will be of use to other creators of low-resource data sets. *gaHealth* is now freely available online and is ready to be explored for further research.

Keywords: Health data, parallel corpus, machine translation, Irish

1. Introduction

Improvements in performance in natural language processing (NLP) tasks are typically to be seen when deep learning models are used. However, deep learning requires large amounts of data for model training. Consequently, the availability of large amounts of textual data has become fundamental to the success of NLP applications, such as language modelling (Buck et al., 2014) and Neural Machine Translation (NMT) (Bahdanau et al., 2014; Sennrich et al., 2016).

A popular method of developing such corpora for Machine Translation (MT) tasks is to crawl and parse bilingual web pages to general parallel corpora. However, given the nature of low-resource languages, there are often insufficient web sites available in both the languages of study. Accordingly, a lack of web content typically hinders the development of NLP applications for low-resource languages.

The motivation underpinning our present work comes from the challenges we faced in developing high-performing MT models in low-resource settings (Afli et al., 2017). In this work, we developed the first bilingual corpus of health data for the English-Irish pair. A procedure was created to extract, clean and select appropriate sentences to build a bilingual corpus. In addition, we built a high-performing MT model for translating in-domain health data.

2. Related work

2.1. Transformer

Transformer (Vaswani et al., 2017) is an architecture for transforming an input sequence into an output sequence via an Encoder and Decoder without relying on Recurrent Neural Networks. Transformer models use attention to focus on previously generated tokens. This approach allows models to develop a long memory which is particularly useful in the domain of language translation.

2.2. Transformer Hyperparameter Optimization

Hyperparameter optimization of Transformer models in translating the low-resource English-Irish language pair has been evaluated in previous studies (Lankford et al., 2021b). Carefully selecting the appropriate subword model has been shown to be an important driver of translation performance. A Transformer architecture, using a 16k BPE SentencePiece subword model, demonstrated optimal performance.

2.3. Neural Machine Translation

Using large bilingual corpora, NMT approaches require the training of neural networks to learn a statistical model for machine translation. The technique has demonstrated state-of-the-art translation performance on many benchmarks. However, one of the key factors in enabling the development of high performing NMT models is the availability of large amounts of parallel data (Koehn and Knowles, 2017; Sennrich and Zhang, 2019).

3. Proposed Approach

3.1. Sources for *gaHealth* Development

To build a bilingual corpus of health data, we selected multiple sources of professionally translated documents from within the Irish government, all of which are publicly available. In particular, the bilingual strategy statements and annual reports of the Irish Department of Health since 2010 were chosen.

Furthermore, a dataset of Covid-related data, developed for a previous study (Lankford et al., 2021a), was incorporated into a larger health dataset. Given the pace at which Covid-19 data was being published, translated Irish website content often lagged the English-language counterpart. Web-

Type	Sentence
EN-1	The Programme for Government makes strong commitments to strengthen community-based care, including primary care and social care, and sees this as fundamental to advancing Sláintecare reforms.
GA-1	Tá gealltanais láidre sa Chlár Rialtais maidir le cúram pobalbhunaithe a neartú, cúram príomhúil agus cúram sóisialta san áireamh, agus breathnaítear air sin mar chuid bhunri-achtanach d'athchóirithe Sláintecare a chur ar aghaidh.
EN-2	The Hepatitis C Compensation Tribunal (Amendment) Act 2006 established a statutory scheme to address insurance difficulties experienced by persons infected with Hepatitis C and HIV through the administration within the State of blood and blood products.
GA-2	Faoin Acht um Binse Cúitimh i ndáil le Heipitíteas C (Leasú) 2006, bunaíodh scéim reachtúil chun dul i ngleic le deacrachtaí árachais a bhí ag daoine a bhí ionfhabhtaithe le Heipitíteas C agus VEID trí fhuil agus táirgí fola a tugadh dóibh laistigh den Stát.
EN-3	To provide virological evidence on the presence and extent of undetected community transmission of covid-19 and monitor positivity rates among individuals presenting ili or acute respiratory tract infections to primary care.
GA-3	Fianaise víreolaíoch a sholáthar maidir le láithreacht agus méid an tarchuir pobail anaithnid de covid-19 agus monatóireacht a dhéanamh ar rátaí dearfacha i measc daoine aonair a bhfuil ionfhabhtuithe ili nó conaire riospráide géarmhíochaine orthu chuig cúram príomhúil.

Table 1: Extracts from the *gaHealth* corpus are illustrated in this table. The sentences, EN-1 / GA-1, are drawn from strategy statements, EN-2 / GA-2 are taken from annual reports whereas EN-3 / GA-3 are from Covid sources.

Documents	Source	Lines
Strategy Statement 2020	HSE	3k
Strategy Statement 2017	HSE	2.5k
Strategy Statement 2015	HSE	3k
Annual Report 2020	HSE	2k
Annual Report 2019	HSE	2k
Annual Report 2017	HSE	2k
Website (Covid)	Citizen's Advice	4k
Publications (Covid)	HSE	4k

Table 2: Sources used in corpus development

site snapshots taken by the WayBack Machine [1] (Arora et al., 2016) proved particularly useful in creating good parallel data from unaligned parallel websites. Extracts from the corpus are illustrated in Table 1.

This amalgamated corpus, *gaHealth*, consists of 16,201 lines of parallel text files. The combined English and Irish vocabulary size is 19,269 unique words. The constituent elements of the dataset, prior to applying the toolchain, are outlined in Table 2.

3.2. Toolchain used for *gaHealth* Development

The HSE PDF and Word documents were pre-processed using a toolchain currently under development as part of the Irish Language Resource Infrastructure project (ILRI), funded by the Department of the Gaeltacht. This toolchain has been written to accept primarily data that originates in public administration organizations, i.e. relatively formal text for which the translation quality is assumed to be high, the structure/formatting to be reasonably consistent, and the potential for noise to be low.

The source material consisted of a combination of twelve input files: six in English (all PDF) and six in Irish (five PDF, one Word); all PDFs had a text layer. They ranged from relatively short (30-40 A4 pages) to more substantial (ca. 200 pages). PDFs in particular can be problematic for creating high-quality corpora for a variety of reasons. In other words, while the quality of the input content in this case can be said to be high, the quality of the input medium is low.

The process used for developing the corpus is illustrated in Figure 1. The toolchain consists of a set of components run in sequence over a set of input documents, in order to convert it from raw content to a sentence-aligned corpus. Several of the components listed below have different implementations depending on the source type and intended output.

Text extractors These are wrappers around external components that extract text based on input type; for the *gaHealth* file types, wrappers were written for LibreOffice and pdftotext.

Unicode normalizer This is used to achieve Unicode equivalence, and optionally to substitute certain (e.g. corrupted) characters.

Language detector In this toolchain, the language detector is a wrapper around langdetect, itself a port from language-detection by Nakatani Shuyo (Nakatani, 2010). The wrapper was written to allow this to be run conveniently either on a string or on an entire file.

Sentence splitters These are custom-written components to reconstruct sentence boundary information. For editable file types like plain text and Word, this process is relatively straightforward. However, PDFs present particular challenges in this regard. Along with ordering issues, the absence of sentence boundary information is one of the biggest reasons why it is so difficult to construct a high-quality corpus from PDFs. A custom sentence splitter was used to determine sentence boundaries from text extracted from PDFs specifically, primarily using capitalization and language-specific lists of abbreviations to determine where sentences should be broken.
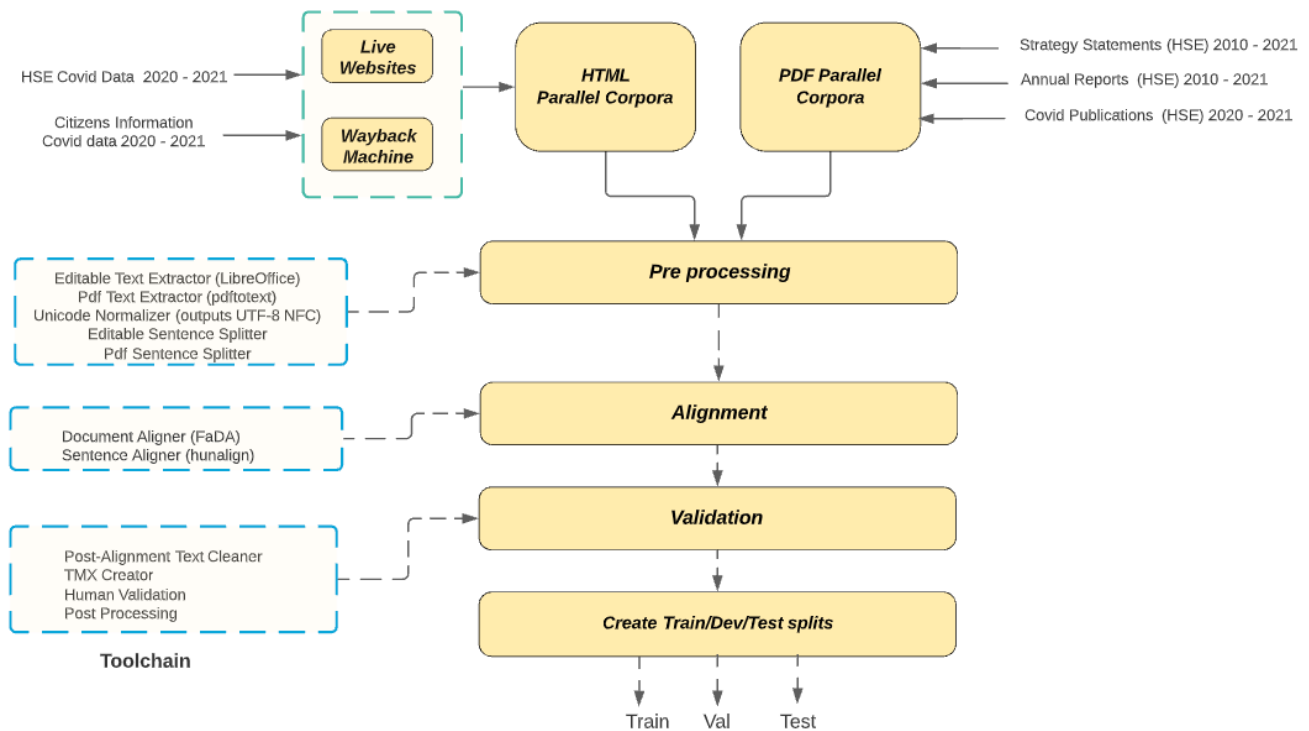
---

[1] https://archive.org/web/

HSE Covid Data 2020 - 2021 →

Citizens Information Covid data 2020 - 2021 →

Live Websites

Wayback Machine

HTML Parallel Corpora

PDF Parallel Corpora

← Strategy Statements (HSE) 2010 - 2021
← Annual Reports (HSE) 2010 - 2021
← Covid Publications (HSE) 2020 - 2021

Editable Text Extractor (LibreOffice)
Pdf Text Extractor (pdftotext)
Unicode Normalizer (outputs UTF-8 NFC)
Editable Sentence Splitter
Pdf Sentence Splitter

Pre processing

Document Aligner (FaDA)
Sentence Aligner (hunalign)

Alignment

Post-Alignment Text Cleaner
TMX Creator
Human Validation
Post Processing

Validation

Toolchain

Create Train/Dev/Test splits

Train   Val   Test

Figure 1: Corpus development process. In developing the corpus, the key steps of data collection, pre-processing, alignment and validation were followed. The role of the toolchain at various stages is highlighted.

Document aligner This aligns sets of files whose languages have been identified. A wrapper was written around an external component called FaDA (Lohar et al., 2016) to adapt it to the toolchain. As FaDA always names an alignment for each input file, sometimes even mapping two different files to the same one, it was necessary to put some selection logic in here, as well as a mechanism for determining when there is no appropriate mapping. Constraints may be put on the relative size of the files to accept an alignment, and the process may be re-run multiple times, with previously rejected files being run again.

Sentence aligner This aligns pairs of files at the sentence level. This was a wrapper around the external component hunalign (Varga et al., 2005). No special parameters were used, as default settings produced results of high quality.

Text cleaners These remove sentence pairs that are believed to be incorrect alignments, such as empty segments and those with obviously mismatched content.

3.3. Guidelines

With the above considerations in mind, the following set of rules was decided upon when processing the *gaHealth* dataset. Many of these could be specified as parameters to the toolchain, while others were hard-coded into the system.

1. Unicode standard: normalize all characters to Unicode UTF-8 NFC. Remove any byte order marks.

2. Whitespacing and capitalization: merge sequences of whitespace characters into a single space. Do not perform tokenization or truecasing.

3. File language detection: scan the first 50 lines, and then every 100th line.

4. Document alignment: assume that specific patterns like a line beginning with a single letter in parentheses or a number followed by a full stop indicate a sentence break from the previous line. Ensure each document is 0.75-1.33 times the size of the document it is being aligned with. Run for a maximum of three iterations.

5. Sentence alignment: allow one-to-many alignments.

6. Cleaning: remove any pairs where source or target:

   - is empty
   - contains no non-alphabetical characters
   - is of an incorrect language. This will remove most untranslated segments. The language is only to be detected for segments that have at least 40 characters

3.4. Transformer architecture

All EN-GA and GA-EN models, trained with the *gaHealth* corpus, were developed using a Transformer architecture. A reduction in the number of hidden layer neurons and increasing dropout significantly improved performance. Furthermore, within the context of low-resource English to Irish translation, using a 16k BPE submodel resulted in the highest performing models. Optimal hyperparameters were selected in line with these findings from our previous and are outlined in Table 3.

Hyperparameter	Values
Learning rate	0.1, 0.01, 0.001, **2**
Batch size	1024, **2048**, 4096, 8192
Attention heads	**2**, 4, **8**
Number of layers	5, **6**
Feed-forward dimension	**2048**
Embedding dimension	128, **256**, 512
Label smoothing	**0.1**, 0.3
Dropout	0.1, **0.3**
Attention dropout	**0.1**
Average Decay	0, **0.0001**

Table 3: Hyperparameter optimization for Transformer models. Optimal parameters are highlighted in bold (Lankford et al., 2021b).

4. Empirical Evaluation

In addition to developing the *gaHealth* corpus, the effectiveness of the dataset was evaluated by training models for English-Irish and Irish-English translation in the Health domain. All experiments involved concatenating source and target corpora to create a shared vocabulary and a shared SentencePiece (Kudo and Richardson, 2018) subword model. The impact of using separate source and target subword models was not explored.

To benchmark the performance of our models, the EN-GA and GA-EN test datasets from the LoResMT2021 Shared Task (Ojha et al., 2021) were used. These test datasets enabled the evaluation of the *gaHealth* corpus, and associated models, since this shared task focused on an application of the health domain i.e. the translation of Covid-related data. Furthermore, using a shared task test dataset enables the comparison of gaHealth models' performance with models entered by other teams.

The results from the IIITT (Puranik et al., 2021) and UCF (Chen and Fazio, 2021) teams are included in Tables **??** and 7 so the performance of the *gaHealth* models can be easily compared with the findings of LoResMT2021. IIITT fine-tuned an Opus MT model from Helsinki NLP on the training dataset. UCF (Chen and Fazio, 2021) used transfer learning, unigram and subword segmentation methods for English–Irish and Irish–English translation.

4.1. Infrastructure

Rapid prototype development was enabled through a Google Colab Pro subscription using NVIDIA Tesla P100 PCIe 16 GB graphic cards and up to 27GB of memory when available (Bisong, 2019). Our MT models were trained using the Pytorch implementation of OpenNMT 2.0, an open-source toolkit for NMT (Klein et al., 2017).

4.2. Metrics

Automated metrics were used to determine the translation quality. All models were trained and evaluated using the BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and ChrF (Popović, 2015) evaluation metrics. Case-insensitive BLEU scores, at the corpus level, are reported. Model training was stopped after 40k training steps or once

an early stopping criteria of no improvement in validation accuracy for four consecutive iterations was recorded.

4.3. Results: Automatic Evaluation

The hyperparameters used for developing the models are outlined in Table 3. The details of the train, validation and test sets used by our NMT models are outlined in Table 4 and Table 5. In all cases, 502 lines were used from the LoResMT2021 validation dataset whereas the test dataset used 502 lines for EN-GA translation and 250 lines for GA-EN translation. Both were independent health-specific Covid test sets which were provided by LoResMT2021. There was one exception, due to a data overlap between test and training data, a reduced test set was used when testing the *gaHealth* en2ga* system.

Team	System	Train	Dev	Test
adapt	covid_extended	13k	502	500
adapt	combined_domains	65k	502	500
IIITT	en2ga-b	8k	502	500
UCF	en2ga-a	8k	502	500
gaHealth	en2ga	24k	502	500
gaHealth	en2ga*	24k	502	338

Table 4: EN-GA Train, Dev and Test dataset distributions. The baseline *gaHealth* system was augmented with an 8k Covid dataset provided by LoResMT2021. A smaller test set was used when evaluating *gaHealth* en2ga* due to an overlap with the training data. An alternative approach of removing the overlap from the *gaHealth* corpus, prior to training, was also carried out to produce the *gaHealth* en2ga system.

Team	System	Train	Dev	Test
IIITT	ga2en-b	8k	502	250
UCF	ga2en-b	8k	502	250
gaHealth	ga2en	24k	502	250

Table 5: GA-EN Train, Dev and Test dataset distributions. The baseline *gaHealth* system was augmented with an 8k Covid dataset provided by LoResMT2021. All overlaps were removed from the *gaHealth* corpus prior to training the *gaHealth* ga2en model.
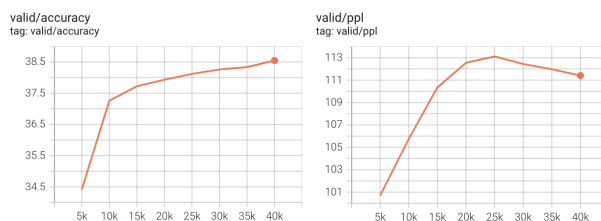

Figure 2: *gaHealth* en2ga* system: training *EN-GA* model with combined 16k gaHealth corpus and 8k LoResMT2021 covid corpus achieving a max validation accuracy of 38.5% and perplexity of 111 after 40k steps. BLEU score: **37.6**.

Experimental results achieved using a Transformer architecture, are summarized in Table **??** and Table 7. In the
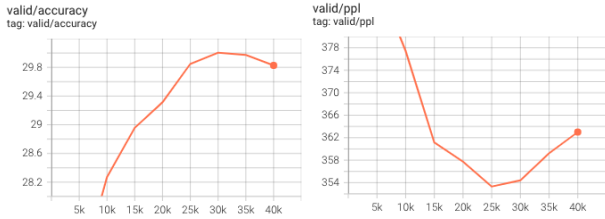
Figure 3: adapt covid_extended system: training *EN-GA* model with 8k LoResMT2021 covid corpus achieving a max validation accuracy of 30.0% and perplexity of 354 after 30k steps. BLEU score: **36.0**.
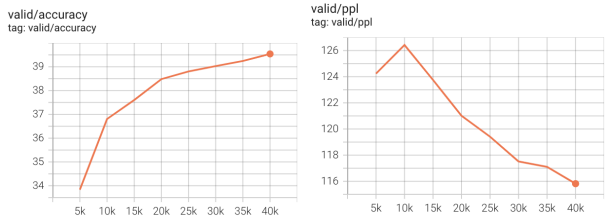

Figure 4: *gaHealth* ga2en system: training *GA-EN* model with combined 16k gaHealth corpus and 8k LoResMT2021 covid corpus achieving a max validation accuracy of 39.5% and perplexity of 116 after 40k steps. BLEU score: **57.6**.

LoResMT2021 Shared Task, the highest-performing EN-GA system was submitted by the ADAPT team. The system uses an extended Covid dataset (13k, which is a combination of the MT summit Covid baseline and a custom DCU Covid dataset. This Transformer model, with 2 heads, performs well across all key translation metrics (BLEU: 36.0, TER: 0.531 and ChrF3: 0.6).

Team	System	BLEU ↑	TER ↓	ChrF3 ↑
UCF	en2ga-b	13.5	0.756	0.37
IIITT	en2ga-b	25.8	0.629	0.53
adapt	combined	32.8	0.590	0.57
gaHealth	en2ga	33.3	0.604	0.56
adapt	covid_extended	36.0	0.531	0.60
gaHealth	en2ga*	**37.6**	0.577	0.57

Table 6: EN-GA *gaHealth* system compared with LoResMT 2021 EN-GA systems.

Validation accuracy, and model perplexity, in developing the *gaHealth* models are illustrated in Figure 2 and Figure 4 whereas Figure 3 illustrates model training on just the covid_extended dataset. Rapid convergence was observed while training the *gaHealth* models such that little accuracy improvement occurs after 30k steps. Only marginal gains were achieved after this point and it actually declined in the case of the system trained using the covid_extended dataset. Perplexity (PPL) shows how many different, equally probable words can be produced during translation. As a metric for translation performance, it is important to keep low scores so the number of alternative translations is reduced. Of the models developed by the ADAPT team, the worst-performing model uses a larger 65k dataset. This is not surprising given the dataset is from a generic domain of which only 20% is health related. The performance of this higher-resourced 65k line model lags the augmented *gaHealth* model which was developed using just 24k lines.

Team	System	BLEU ↑	TER ↓	ChrF3 ↑
UCF	ga2en-b	21.3	0.711	0.45
IIITT	ga2en-b	34.6	0.586	0.61
gaHealth	ga2en	**57.6**	0.385	0.71

Table 7: GA-EN *gaHealth* system compared with LoResMT 2021 GA-EN systems.

For translation in the GA-EN direction, the best performing model for the LoResMT2021 Shared Task was developed by IIITT with a BLEU of 34.6, a TER of 0.586 and ChrF3: 0.6. This effectively serves as the baseline by which our GA-EN model, developed using the *gaHealth* corpus, can be benchmarked. The performance of the *gaHealth* model offers an improvement across all metrics with a BLEU score of 57.6, a TER of 0.385 and a CHrF3 result of 0.71. In particular, the 40% improvement in BLEU score is very significant.

5. Discussion

Although the main objective of this work is to develop the first bilingual corpus of English–Irish data, we conduct initial experiments on the effectiveness of such datasets in training MT models. We have used our *gaHealth* dataset to train an MT model on test data from the LoResMT2021 Shared task to evaluate how the system performs translating health data in both the EN-GA and GA-EN direction. Our systems, developed using the *gaHealth* corpus achieved significantly higher scores.

6. Conclusion and Future Work

The main contribution of this work is to present an ongoing translation project that aims at building the first ever parallel corpus of health data for the Irish language – *gaHealth* – by fully utilising freely available parallel documents.
Due to the issues encountered during conversion of PDF documents, we developed guidelines in order to aid in the conversion process. In addition to developing the *gaHealth* corpus, we trained and evaluated translation models for in-domain health data.
In our experiments, the models achieved a BLEU score of 37.6 (Table **??**) for translating EN-GA test data and 57.6 (Table 7) for translating in the GA-EN direction, which is encouraging performance given this is the beginning of our work on *gaHealth*. There is no such corpus available according to the best of our knowledge, so *gaHealth* will become a useful resource in the NLP community, especially for those working with the Irish language domain.
For future work, we intend to extend the corpus, as more Irish language documents become available. Upon extension we will refine our models. One important aspect which needs further investigation is to understand why the EN-GA model (*gaHealth* en2ga), tested with the full test set, performed worse than the model (*gaHealth* en2ga*) which was tested with the reduced test set. A deep linguistic investigation involving a sentence level BLEU analysis will be conducted as part of a future study.

In addition, we aim to build in-domain datasets for other key domains such as Education and Finance. We will also apply deep learning techniques to further refine our in-domain models. We have released the *gaHealth* corpus online[2] to facilitate further research on this data set.

7. Acknowledgements

8. Bibliographical References

Afli, H., Maguire, S., and Way, A. (2017). Sentiment Translation for Low-resourced Languages: Experiments on Irish General Election Tweets.

Arora, S. K., Li, Y., Youtie, J., and Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: a methodological resource. *Journal of the Association for Information Science and Technology*, 67(8):1904–1915.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bisong, E. (2019). Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Springer.

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram Counts and Language Models from the Common Crawl. In *LREC*, volume 2, page 4. Citeseer.

Chen, W. and Fazio, B. (2021). The UCF Systems for the LoResMT 2021 Machine Translation Shared Task. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 129–133.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Lankford, S., Afli, H., and Way, A. (2021a). Machine Translation in the Covid domain: an English-Irish case study for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 144–150.

Lankford, S., Afli, H., and Way, A. (2021b). Transformers for low-resource languages: Is féidir linn! In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 48–60.

Lohar, P., Ganguly, D., Afli, H., Way, A., and Jones, G. J. (2016). FaDA: fast document aligner using word embedding. *Prague Bulletin of Mathematical Linguistics*, 106:169–179.

Nakatani, S. (2010). Language detection library for java. https://github.com/shuyo/language-detection.

Ojha, A. K., Liu, C.-H., Kann, K., Ortega, J., Shatam, S., and Fransen, T. (2021). Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-resource Languages. *arXiv preprint arXiv:2108.06598*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Puranik, K., Hande, A., Priyadharshini, R., Durairaj, T., Sampath, A., Thamburaj, K. P., and Chakravarthi, B. R. (2021). Attentive fine-tuning of transformers for translation of low-resourced languages@ loresmt 2021. *arXiv preprint arXiv:2108.08556*.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, Borovets, Bulgaria.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

---

[2]https://github.com/seamusl/gaHealth