Welsh Automatic Text Summarisation

Jonathan Morris[1], Ignatius Ezeani[2], Ianto Gruffydd[1], Katharine Young[1,3], Lynne Davies[1], Mahmoud El-Haj[2], Dawn Knight[3]

[1]School of Welsh, Cardiff University, [2]School of Computing and Communications, Lancaster University, [3]School of English, Communication and Philosophy, Cardiff University

morrisj17@cardiff.ac.uk i.ezeani@lancaster.ac.uk gruffyddiw@cardiff.ac.uk youngks2@cardiff.ac.uk DaviesJL17@cardiff.ac.uk m.el-haj@lancaster.ac.uk knightd5@cardiff.ac.uk

Abstract

Text summarisation is a digital approach to summarising 'key' information contained within texts, and the creation of shortened versions of texts based on this content. Text summarisation function is to provide succinct and coherent summaries to users, something that is often time-consuming and difficult to conduct manually. This is useful in the modern digital world where the creation and sharing of text is ever-increasing, as it enables users to navigate, and make sense of, the dearth of digital information that is available, with ease. This paper reports on work on a project which aims to develop an online Automatic Text Summarisation tool for the Welsh language, ACC (Adnodd Creu Crynodebau). This paper contextualises the need for this text summarisation tool, underlines how a dataset for training and testing the methods was created, and outlines plans for the development of the summariser.

Keywords: text summarisation, Welsh language, dataset extraction, creation and evaluation

1. Context

Work on automatic text summarisation has a long history in Natural Language Processing (NLP). This work originally focused only on English, as a global lingua franca, but is now used in a range of other language contexts, including French, Spanish, Hindi, Arabic, amongst others. The 'MultiLing' project and associated conference series, for example, are a noteworthy champion of developing text summarisation in a range of the world's 7000+ different languages. The website, http://multiling.iit.demokritos.gr provides an open repository for summarisation tasks test/training data, model summaries, amongst others.

Missing from current summarisation resources are tools that effectively work with the Welsh language. The development of ACC contributes to work on text summarisation in minority languages and contributes to the technological resources available to Welsh speakers.

2. Welsh Language Online

There exists a relatively low use of Welsh language websites and e-services, despite the fact that numerous surveys suggest that Welsh speakers would like more opportunities to use the language, and that there has been an expansive history of civil disobedience in order to gain language rights in the Welsh language context (Evas & Cunliffe 2016: 64).

One reason for the relatively low take-up of Welsh-language options on websites is the assumption that the language will be too complicated (Evas & Cunliffe 2016: 83).Concerns around the complexity of public-facing Welsh language services and documents are not new. A series of guidelines on creating easy-to-read documents in Welsh are outlined in *Cymraeg Clir* (Williams 1999). Williams (1999: Preface) notes that the need for simplified versions of Welsh is arguably greater than for English considering (1) many Welsh public-facing documents are translated from English, (2) the standard varieties of Welsh are further removed from local dialects compared to English, and (3) newly-translated technical terms are more likely to be familiar to the reader. The principles outlined in *Cymraeg Clir* therefore include the use of shorter sentences, everyday words rather than specialised terminology, and a neutral (rather than formal) register (Williams 1999: 46).

ACC will provide the means for summarising and simplifying digital language sources which will help to address the fears of Welsh speakers that language online is too complicated.

ACC will also contribute to the digital infrastructure of the Welsh language. The most recent Welsh Government strategy for the revitalisation of Welsh has infrastructure (and particularly digital infrastructure) as a main theme (along with increasing the number of speakers and increasing language use, Welsh Government 2017: 5). The aim is to 'ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts' (Welsh Government 2017: 71). Given the introduction of Welsh Language Standards (see Carlin and Mac Giolla Chríost 2016) and a concerted effort to both invest in Welsh language technologies and improve the way in which language choice is presented to the public, the development of ACC will complement the suite of Welsh language technologies (e.g. Canolfan Bedwyr 2021) for both content creators and Welsh readers.

It is also envisaged that ACC will contribute to Welsh-medium education by allowing educators to create summaries for use in the classroom as pedagogical tools. Summaries will also be of use to Welsh learners who will be able to focus on understanding the key information within a text.

3. Data Extraction

The first stage of the development process is to develop a small corpus (dataset) of target language data that will subsequently be summarised and evaluated by human annotators and used to develop and train the automated summarisation models (i.e. acting as a 'gold-standard' dataset).

Wikipedia[1] was selected as the primary source of data for creating the Welsh language dataset for ACC. This was owing to the fact that an extensive number of Welsh language texts exist on this website (over 133,000 articles), all of which are available under GNU Free Documentation license. To ensure that pages that contained a sufficient quantity of text were extracted for use, a minimum threshold of 500 tokens per article and a target of at least 500 articles was established at the outset. A selection of 800 most accessed Wikipedia pages in Welsh were initially extracted for use. An additional 100 Wikipedia pages were included from the WiciAddysg project organised by the National Library of Wales and Menter Iaith Môn[2]. However, it was observed that more than 50% of the articles from this original list of Wikipedia pages did not meet the minimum-token threshold of 500. To mitigate this, a list of 20 Welsh keywords was used to generate an additional 100 Wikipedia pages per keyword (which was provided by the first author and contained words synonymous with the Welsh language, Welsh history and geography). This was added to the list of 100 most-edited Welsh Wikipedia pages and pages from the WiciAddysg project.

The data extraction applied a simple iterative process and implemented a Python script based on the WikipediaAPI[3] that takes a wikipedia page; extracts key contents (article text, summary, category) and checks whether the article text contains a minimum number of tokens. At the end of this process, the dataset was created from a total of 513 Wikipedia pages that met the set criteria. The extracted dataset contains a file for each Wikipedia page with the following structure and tags:

<title>Article Title on Wikipedia</title>
<text> Article Text </text>
<category>Article Categories </category>.

These files are available in both plain text and html file formats.

4. Dataset Creation

A total of 19 undergraduate and postgraduate students from Cardiff University were recruited to create, summarise and evaluate the dataset. Of these students, 13 were undertaking an undergraduate or postgraduate degree in Welsh which involved previous training on creating summaries from complex texts. The remaining six students were undergraduate students on other degree programmes in the Humanities and Social Sciences at Cardiff University and had completed their compulsory education at Welsh-medium or bilingual schools.

Students were asked to complete a questionnaire prior to starting work which elicited biographical information. A total of 17 students had acquired Welsh in the home. One student acquired the language via Welsh-medium immersion education and one student had learned the language as an adult. The majority of students came from south-west Wales (*n*=11). This region included the counties of Carmarthenshire, Ceredigion, Neath Port Talbot, and Swansea. A further five students came from north-west Wales which comprised the counties of Anglesey and Gwynedd. One student came from south-east Wales (Cardiff), one from mid Wales (Powys), and one from north-east Wales (Conwy).

A broad distinction can be made between northern and southern Welsh. The two varieties (within which further dialectal differences exist) exhibit some differences at all levels of language structure although all varieties are mutually intelligible. Students were asked four questions which elicited information on the lexical, grammatical, and phonological variants they would ordinarily use. The results largely corresponded to geographical area: 11 students used southern forms and seven students used northern forms (including the student from mid Wales). One student, from Cardiff, used a mixture of both northern and southern forms.

Students were given oral and written instructions on how to complete the task. Specifically, they were told that the aim of the task was to produce a simple summary for each of the Wikipedia articles (allocated to them) which contained the most important information. They were also asked to conform to the following principles:

- The length of each summary should be 230 - 250 words.
- The summary should be written in the author's own words and not be extracted (copy-pasted) from the Wikipedia article.
- No information which is not included in the article should be included in the summary.
- Any reference to a living person in the article should be anonymised in the summary (to conform to the ethical requirements of each partner institution).
- All summaries should be proofread and checked using spell checker software prior (Cysill[4]) to submission.

Further instruction was given on the register to be used in the creation of summaries. Students were asked to broadly conform to the principles of *Cymraeg Clir* (Williams 1999) and, in particular, avoid less common short forms of verbs and the passive mode, and use simple vocabulary where possible instead of specialised terms.

Each student completed between 60 - 100 summaries between July and October 2021. The median amount of time spent on each summary was 30 minutes. The complete dataset comprises 1,461 summaries with the remaining 39 summaries not being completed due to one student prematurely dropping out of the project and some instances of unsuitable articles (e.g. lists of bullet points).

[1] Welsh Wikipedia: https://cy.wikipedia.org/wiki/Hafan (Wicipedia)
[2]WikiAddysg: https://cy.wikipedia.org/wiki/Categori:Prosiect_WiciAddysg
[3] Wikipedia API: https://pypi.org/project/wikipedia/
[4] Cysill: https://www.cysgliad.com/cy/cysill/

Three of the postgraduate students recruited were also asked to evaluate the summaries by giving a score between one and five. Table 1 shows the marking criteria.

Score	*Criteria*
5	- Very clear expression and very readable style. - Very few language errors. - Relevant knowledge and a good understanding of the article; without significant gaps.
4	- Clear expression and legible style. - Small number of language errors. - Relevant knowledge and a good understanding of the article, with some gaps
3	- Generally clear expression, and legible style. - Number of language errors. - The knowledge and understanding of the article is sufficient, although there are several omissions and several errors.
2	- Expression is generally clear but sometimes unclear. - Significant number of language errors. - The knowledge and understanding of the article is sufficient for an elementary summary, but there are a number of omissions and errors.
1	- Expression is often difficult to understand. Defective style. - Persistently serious language errors. - The information is inadequate for summary purposes. Obvious deficiencies in understanding the article.

Table 1: Criteria for the marking of summaries

Both the mean and median scores for the summaries were 4. Evaluators were instructed to fix common language errors (such as mutation errors and spelling mistakes) but not to correct syntax.

5. Summarisation Tool Description

The second phase of this summarisation project is to use the corpus dataset to inform the iterative development and evaluation of digital summarisation tools. The main approaches to text summarisation include extraction-based summarisation and abstraction-based summarisation. The former extracts specific words/phrases from the text in the creation of the summary, while the latter works to provide paraphrased summaries (i.e. not directly extracted) from the source text. The successful extraction/abstraction of content, when using summarisation tools/approaches, depends on the accuracy of automatic algorithms (which require training using hand-coded gold-standard datasets).

As an under-resourced language with limited literature on Welsh summarisation, applying summarisation techniques from the literature helps in having initial results that can be used to benchmark the performance of other summarisers on the Welsh language. In this project we are to develop a combination of extractive and abstractive single-document summarisation methods. The process will start by implementing and evaluating basic baseline systems that are frequently used in the literature as bench-lines. These will be followed by more complex state of the art summarisation models as well as hybrid systems as toplines

5.1 Baselines

The sections below provide an overview of the summarisation systems that this project will be focusing on currently as well as within the life of the project.

5.1.1 First Sentence Summariser

Rather than using a document's title or keywords, some summarisers tend to use the first sentence of an article to identify the topic to be summarised. The justification behind selecting the first sentence as being representative of the relevant topic is based on the belief that in many cases, especially in news articles or articles found on Wikipedia, the first sentence tends to contain key information about the content of the entire article (Radev et al., 2004; Fattah and Ren, 2008; Yeh et al.,2008).

5.1.2 TextRank

This summarisation technique was introduced by Rada Mihalcea and Paul Tarau (2004). This was the first graph-based automated text summarisation algorithm that is based on the simple application of the PageRank algorithm. PageRank is used by Google Search to rank web pages in their search engine results (Brin and Page, 1998). TextRank utilises this feature to identify the most important sentences in an article.

5.1.3 LexRank

Similar to TextRank, LexRank uses a graph-based algorithm for automated text summarisation (Erkan and Radev, 2004). The technique is based on the fact that a cluster of documents can be viewed as a network of sentences that are related to each other. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. Like TextRank, LexRank too uses the PageRank algorithm for extracting top keywords. The key difference between the two baselines is the weighting function used for assigning weights to the edges of the graph. While TextRank simply assumes all weights to be unit weights and computes ranks like a typical PageRank execution, LexRank uses degrees of similarity between words and phrases and computes the centrality of the sentences to assign weights (Erkan and Radev, 2004).

5.2 Toplines

As the project progresses, we will develop more complex summarisers and evaluate their performance by comparing the summarisation results of the three baselines mentioned above. The purpose of the topline summarisers is to prove that using language related technology to

summarise Welsh documents will improve the results of those produced by the baseline summarisers.

5.2.1 TF.IDF Welsh Summariser

A summariser using Text Frequency Inverse Document Frequency (TF.IDF) works on findings words that have the highest ratio of those words frequency in the to be summarised document in comparison to their occurrence in the full set of documents to be summarised (Salton and McGill, 1986). TF.IDF is a simple numerical statistic which reflect the importance of a word to a document in a text collection or corpus and is usually used as a weighing factor in information retrieval, thus using it to find important sentences in extractive summarisation (Hajime and Manabu, 2000; Wolf et al., 2004).

The summariser will work on finding key and important words in the documents to be summarised in an attempt to produce relevant summaries. Using TF.IDF in the Welsh language is not new. Arthur et al. (2019), used a social network that they built using Twitter's geo-locations to identify contiguous geographical regions and identify patterns of communication within and between them. Similarly, we will use TF.IDF to identify important sentences based on patterns detected between the summarised document and the summaries corpus.

5.2.2 TF.IDF Welsh Summariser with Welsh Word Embeddings

In order to improve the similarity measure between sentences, we use pre-trained word embedding features combined with the previously mentioned TF.IDF features. For that we use the FastText Welsh pre-trained word vectors (Joulin et al., 2016). FastText is an extension of the word2vec (Mikolov et at, 2013) model where instead of learning vectors for words directly, FastText represents each word as an n-gram of characters, which helps in capturing the meaning for shorter words and allow the embeddings to understand suffixes and prefixes. Ezeani et al. (2019) leveraged existing language models such as Welsh FastText for multi-task classification of Welsh part of speech and semantic tagging. We will repeat the experiment but this time using Welsh word embeddings created by Corcoran et al. (2021) where they used word2vec and FastText, to automatically learn Welsh word embeddings taking into account syntactic and morphological idiosyncrasies of this language. We will build upon those two previous efforts and harness the language models towards enriching the performance of the TF.IDF summariser in 5.2.1.

5.3 State of the Art Welsh Summarisers

The final stage of the project is to use state of the art summarisation technologies to summarise Welsh documents. This will include building Extractive and Abstractive summarisers using deep neural network machine learning techniques or what is known as Deep Learning. The summarisation state of the art literature shows a great shift towards using deep learning to create extractive and abstractive supervised and unsupervised summarisers using deep learning models such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) and many others (Song et al., 2019; Zmandar et al., 2021a; Zmandar et al., 2021b; Magdum et al., 2021). In this project we will combine the use of the aforementioned Welsh word embeddings to try and improve the results and create Welsh summarisation systems that are on par with other English and European state of the art summarisers.

6. Evaluation

The gold-standard summaries created by the human summarisers as described in Section 4 will be used to automatically evaluate any system summaries generated by the models developed in Section 5. The system summaries will be evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics (Lin, 2004). ROUGE is a set of metrics used for evaluating automatic summarisation in natural language processing. The metrics compare an automatically produced summary against gold-standard summaries.

As an additional level of evaluation, a sample of system summaries generated by our best performing summarisation model (see Section 5.2) will be manually evaluated by native Welsh speakers in order to measure the quality of those summaries. The final stages of this project will include the development of a freely-available user-friendly web-based user interface that can be used by users from all age groups. The system will allow users to define the level of compression (e.g. a summary of no more than 200 words). The summariser will also be available as open-source Python packages to allow developers to work on enhancing the summarisers in the future.

7. Conclusion

The released version of ACC will contribute to the automated tools available in the Welsh language and facilitate the work of those involved in document preparation, proof-reading, and (in certain circumstances) translation. The tool will also allow professionals to quickly summarise long documents for efficient presentation. For instance, the tool will allow educators to adapt long documents for use in the classroom. It is also envisaged that the tool will benefit the wider public, who may prefer to read a summary of complex information presented on the internet or who may have difficulties reading translated versions of information on websites. To keep up to date with developments on this tool, please visit the main project website at: https://corcencc.org/resources/#ACC

8. Acknowledgements

This research was funded by the Welsh Government, under the Grant "Welsh Automatic Text Summarisation". We are grateful to Jason Evans, National Wikimedian at the National Library of Wales, for this initial advice.

9. References

Arthur, R. and Williams, H.T. (2019). The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales. *PloS one*, 14 (4):e0214466.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107-17.

Canolfan Bedwyr. (2021). *Cysgliad: Help i ysgrifennu yn Gymraeg*. Online: https://www.cysgliad.com/cy/

Carlin, P. and Chríost, D.M.G. (2016). A standard for language? Policy, territory, and constitutionality in a devolving Wales. In Durham, M. & Morris, J. (eds), *Sociolinguistics in Wales*. London: Palgrave Macmillan, pp. 93-119.

Citizens Advice Bureau. (2015). English by default - Understanding the use of non-use of Welsh language services. Online: https://www.citizensadvice.org.uk/Global/Migrated_Documents/corporate/english-by-default--march-2015.pdf

Corcoran, P., Palmer, G., Arman, L., Knight, D. and Spasić, I. (2021). Creating Welsh language word embeddings. *Applied Sciences,* 11(15): 6896.

Cunliffe, D., Morris, D. and Prys, C. (2013). Young bilinguals' language behaviour in social networking sites: The use of Welsh on Facebook. *Journal of Computer-Mediated Communication*, *18*(3), pp.339-361.

Erkan, G. and Radev, D.R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22: 457–479.

Evas, J. and Cunliffe, D. (2016). Behavioural Economics and Minority Language e-Services—The Case of Welsh. In Durham, M. and Morris, J. (eds), *Sociolinguistics in Wales*. London: Palgrave Macmillan, pp. 61-91.

Ezeani, I., Piao, S.S., Neale, S., Rayson, P. and Knight, D. (2019). Leveraging pre-trained embeddings for Welsh taggers. In Proceedings of the *4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 270-280, Florence.

Fattah, M. and Ren, F. (2008). Automatic Text Summarization. In *Proceedings of World Academy of Science*, 27, World Academy of Science, pp. 192–195.

Government Digital Service and Wales Office. (2015). *The Welsh experience on Gov.uk - a qualitative research study. Online: https://userresearch.blog.gov.uk/2015/09/15/the-welsh-experience-on-gov-uk-a-qualitative-research-study/*

Hajime, M. and Manabu, O. (2000). A Comparison of Summarization Methods based on Taskbased Evaluation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T. (2016). *Fasttext. zip: Compressing text classification models*. arXiv preprint arXiv:1612.03651.

Lin, C-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.

Magdum, P.G. and Rathi, S. (2021). A Survey on Deep Learning-Based Automatic Text Summarization Models. *Advances in Artificial Intelligence and Data Engineering*, pp. 377-392. Springer, Singapore.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Radev, D., Jing, H., Sty, M. and Tam, D. (2004). Centroid-based Summarization of Multiple Documents. *Information Processing and Management*, 40: 919–938.

Salton G. and McGill M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Song, S., Huang, H. and Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications,* 78(1): 857-75.

Welsh Government. (2017). *Cymraeg 2050 - A million Welsh speakers. Cardiff: Welsh Government. Online*: https://gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf

Williams, C. (1999). *Cymraeg Clir: Canllawiau Iaith*. Bangor: Gwynedd Council, Welsh Language Board and Canolfan Bedwyr.

Wolf, C., Alpert, S., Vergo, J., Kozakov, L. and Doganata, Y. (2004). Summarizing Technical Support Documents for Search: Expert and User Studies. *IBM Systems Journal,* 43(3): 564–586.

Yeh, J., Ke, H. and Yang, W. (2008). iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Systems with Applications*, 35(3):1451–1462.

Zmandar, N., El-Haj, M., Rayson, P., Litvak, M., Giannakopoulos, G. and Pittaras N. (2021b). The Financial Narrative Summarisation Shared Task FNS 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pp. 120-125, Lancaster University.

Zmandar, N., Singh, A., El-Haj, M. and Rayson, P. (2021a). Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pp. 99-105, Lancaster University.