*Proceedings of the CLTW 4 @ LREC2022* , pages 14–21
Marseille, 20-25 June 2022


# Creation of an Evaluation Corpus and Baseline Evaluation Scores for Welsh Text Summarisation

**Mahmoud El-Haj[1], Ignatius Ezeani[1], Jonathan Morris[2] and Dawn Knight[3]**
[1]UCREL NLP Group, Lancaster University,
[2]School of Welsh, [3]School of English, Communication and Philosophy, Cardiff University
{m.el-haj, i.ezeani}@lancaster.ac.uk, {knightd5, morrisj17}@cardiff.ac.uk

**Abstract**

As part of the effort to increase the availability of Welsh digital technology, this paper introduces the first human vs metrics Welsh summarisation evaluation results and dataset, which we provide freely for research purposes to help advance the work on Welsh summarisation. The system summaries were created using an extractive graph-based Welsh summariser. The system summaries were evaluated by both human and a range of ROUGE metric variants (e.g. ROUGE 1, 2, L and SU4). The summaries and evaluation results will serve as benchmarks for the development of summarisers and evaluation metrics in other minority language contexts.

**Keywords:** summarisation, Welsh, evaluation, corpus, annotators

## 1. Introduction

Work on automatic text summarisation has a long history in Natural Language Processing (NLP). The majority of research on text summarisation was originally focused only on English, as a global lingua franca (Goldstein et al., 2000; Svore et al., 2007; Svore et al., 2007; Litvak and Last, 2008; El-Haj et al., 2011; El-Haj and Rayson, 2013). Recently this started to change with researchers shifting their focus towards a range of other language contexts, including French, Spanish, Hindi, Arabic, amongst others. Research community efforts such as the 'MultiLing' (Giannakopoulos et al., 2011) project and its associated workshop series, for example, are a noteworthy champion of developing text summarisation in a range of the world's 7000+ different languages. The MultiLing website[1] provides an open repository for summarisation tasks test/training data, model summaries, amongst others.

The development of the Adnodd Creu Crynodebau (ACC) project[2] contributes to both the development of summarisation tools in minority languages more generally and to the digital infrastructure of Welsh. Improving digital infrastructure for the Welsh language is a cornerstone of current Welsh Government policy designed to safeguard and promote the language[3]. Specifically, the Welsh Government's 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).

The development of an automatic summarisation tool contributes to this aim insofar as it will facilitate the preparation of summaries among professional content creators which can be made available online. From the user's perspective, ACC gives the reader agency to create easy-to-read summaries of long texts which enables the use of Welsh on the internet.
Table 1 shows a sample of a text in Welsh and a system summary that was generated using the Welsh Text Summary Creator (ACC) v.1.0[4] (Ezeani et al., 2022). The article in Table 1 can be found on Wikipedia both in Welsh[5] and English[6].

In this paper, we focus on the evaluation process of summaries created by ACC. Specifically, we compare the results of human evaluation with those produced using the ROUGE summarisation metric. Evaluating the output of summarisation tools using metrics such as ROUGE is a common practice in the field, but using this metric relies on comparison data. As ACC is the first summariser for Welsh, comparison data were not available and therefore human evaluation was needed. The evaluation metrics used were ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-SU4. In addition, we provide results for human evaluation for summaries generated by our best performing summariser.

The remainder of the paper presents more context on the Welsh language and the development of the tool, before we turn to the methodology used to compare the human and ROUGE metrics and the results. The dataset and the code we used to create the summarisers are available on the Welsh Summarisation Project

[1] http://multiling.iit.demokritos.gr

[2] English translation from Welsh: "Welsh Summary Creator": http://wp.lancs.ac.uk/acc/

[3] Welsh Government: Cymraeg 2050 - A million Welsh speakers: https://gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf

[4] https://share.streamlit.io/ucrel/welsh_summarizer/main/app/app.py

[5] https://cy.wikipedia.org/wiki/Yr_hawl_i_iechyd

[6] https://en.wikipedia.org/wiki/Right_to_health

GitHub page[7].

| |
|---|
| **Welsh text:** *Mae Erthygl 25 o Ddatganiad Cyffredinol Hawliau Dynol 1948 y Cenhedloedd Unedig yn nodi: Mae gan bawb yr hawl i safon byw sy'n ddigonol ar gyfer iechyd a lles ei hun a'i deulu, gan gynnwys bwyd, dillad, tai a gofal meddygol a gwasanaethau cymdeithasol angenrheidiol. Mae'r Datganiad Cyffredinol yn cynnwys lletyaeth er mwyn diogelu person ac mae hefyd yn sôn yn arbennig am y gofal a roddir i'r rheini sydd mewn mamolaeth neu blentyndod. Ystyrir mai Datganiad Cyffredinol o Hawliau Dynol fel y datganiad rhyngwladol cyntaf o hawliau dynol sylfaenol. Dywedodd Uchel Gomisiynydd y Cenhedloedd Unedig dros Hawliau Dynol Navanethem Pillay fod y Datganiad Cyffredinol o Hawliau Dynol yn ymgorffori gweledigaeth sy'n gofyn am gymryd yr holl hawliau dynol - sifil, gwleidyddol, economaidd, cymdeithasol neu ddiwylliannol - fel cyfanwaith anwahanadwy ac organig, anwahanadwy a rhyngddibynnol.* |
| **System Summary:** *Mae Erthygl 25 o Ddatganiad Cyffredinol Hawliau Dynol 1948 y Cenhedloedd Unedig yn nodi: Mae gan bawb yr hawl i safon byw sy'n ddigonol ar gyfer iechyd a lles ei hun a'i deulu, gan gynnwys bwyd, dillad, tai a gofal meddygol a gwasanaethau cymdeithasol angenrheidiol. Dywedodd Uchel Gomisiynydd y Cenhedloedd Unedig dros Hawliau Dynol Navanethem Pillay fod y Datganiad Cyffredinol o Hawliau Dynol.* |

Table 1: Example text with an automatically generated summary.

## 2. The Welsh Language

It is estimated that over a quarter (29.2%) of the population in Wales aged over 3 consider themselves to be Welsh speakers[8]. This estimate represents an increase in the proportion of the population who reported speaking Welsh at the (2011) census[9] and can be attributed, at least in part, to the ongoing attempts by Welsh Government and its stakeholders to safeguard the language and promote its use among the population (Carlin and Chríost, 2016).

Despite the promotion of Welsh in various domains, the use of Welsh language websites and e-services remains relatively low, despite the fact that numerous surveys suggest that Welsh speakers would like more opportunities to use the language, and that there has been extensive campaigning in order to gain language rights in the Welsh language context (Cunliffe et al., 2013). One reason for the relatively low take-up of Welsh-language options on websites is the assumption that the language used in such resources will be too complicated (Cunliffe et al., 2013).

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 (Arthur and Williams, 2019). Williams (1999) notes that the need for simplified versions of Welsh is arguably greater than for English in Wales 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).

Whilst the Welsh language is not necessarily more structurally complex than other languages for which automatic summarisation tools have been developed, there are sociolinguistic considerations which do need to be considered. In addition to the various dialects, there are differences in register between formal and informal varieties of Welsh, with informal registers formally found mainly in spoken Welsh now increasingly appearing also in written text. This has led to increased morphosyntactical and lexical differences between written varieties. As is shown below, this was considered when formulating guidance for those involved with the preparation of the human gold-standard summaries but does not necessarily mean that variation is not present in the dataset.

Our work will contribute to the digital infrastructure of the Welsh language. Given the introduction of Welsh Language Standards (Carlin and Chríost, 2016), which places requirements on public institutions to provide fully bilingual web content, 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 and evaluation of ACC will complement the suite of Welsh language technologies (e.g. Canolfan Bedwyr 2021[10]) 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

[7] https://github.com/Welsh-Summarization-Project

[8] https://gov.wales/welsh-language-data-annual-population-survey-july-2020-june-2021

[9] https://statswales.gov.wales/Catalogue/Welsh-Language/Census-Welsh-Language. The results of the 2021 Census are not yet released.

[10] Cysgliad: Help i ysgrifennu yn Gymraeg. Online: https://www.cysgliad.com/cy/

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. Methods

Figure 1 shows the four key processes involved in the creation and evaluation of the Welsh summarisation dataset i.e. **a.** collection of the text data; **b.** creation of the reference (human) summaries; **c.** building summarisers and generating system summaries and **d.** evaluating the performance of the summarisation systems outputs on the reference summaries both using automatic metrics and human effort.

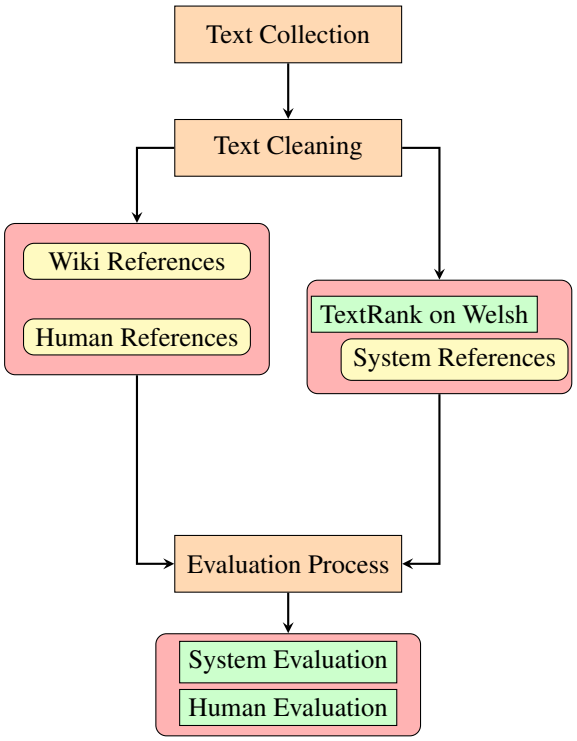


Figure 1: An overview of the process diagram

### 3.1. Text Collection

In order to be able to automatically evaluate the generated system summaries, we needed to first create reference human summaries (gold–standards). To do so we started by collecting 513 Wikipedia articles from the Welsh Wikipedia[11]. We then pre-processed the articles in order to extract the textual content. The data extraction applied a simple iterative process and implemented a Python script based on the WikipediaAPI[12] 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. Figure 2 shows token counts of the 513 Wikipedia articles used for training of system summarisers as well as the average counts of the articles and the summaries. The majority of the articles (about 80%) contain between 500 and 2000 tokens. A total of 28 articles contain more than 5000 tokens. The extracted dataset contains a file for each Wikipedia page with the following structure and tags[13]:

```
<title>Article Title</title>
<text>Article Text</text>
<category>Article Categories</category>
```

The data files are also available in `plain text`, `.html`, `.csv` and `.json` file formats.

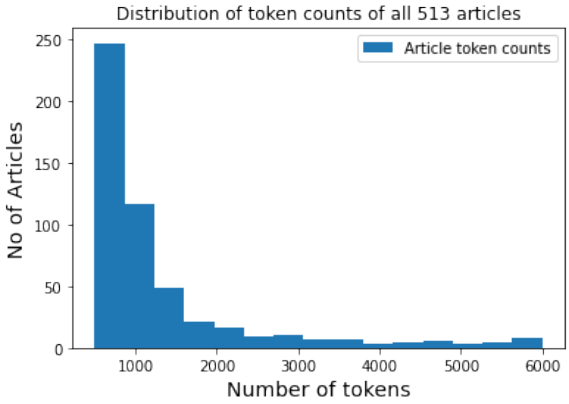


Figure 2: Distribution of tokens count

### 3.2. Reference Summaries Creation

In this work, two sources were used: a) the Wikipedia summaries extracted using the Wikipedia API[14] during the text collection stage and b) the summaries created by the human participants. A total of 19 undergraduate and postgraduate students from Cardiff University were recruited to create, summarise and evaluate the generated summaries, 13 of them 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 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. 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:

[11]Welsh Wikipedia: `https://cy.wikipedia.org/wiki/Hafan` (Wicipedia)

[12]https://pypi.org/project/Wikipedia-API/

[13]The tags are there to help users find and extract part of the data they are interested in.

[14]Class WikipediaPage has property summary, which returns a description of a Wikipedia page `https://pypi.org/project/Wikipedia-API/`

- 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.
- The summary should not include any information that is not contained in the article
- 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 (Cysill) prior to submission[15].

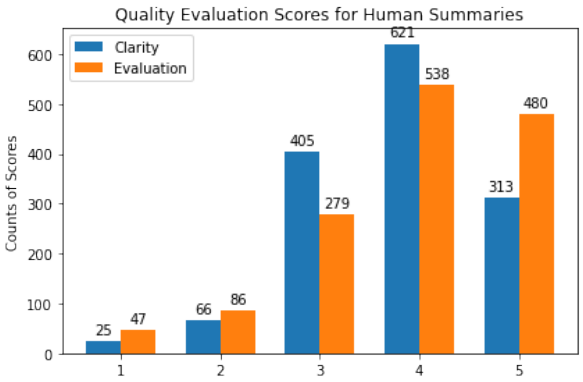


Figure 3: Distribution of the readability (clarity) and overall quality evaluation scores

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. In total the participants generated a number of 1,430 human summaries with an average of 3 summaries per article. In addition, three of the postgraduate students recruited were asked to evaluate the human summaries by giving a score between one and five.

Each summary was evaluated only once (by 1 participant) as the process here was to double check the summaries are according to the given instructions. Figure 3 shows the distribution of the readability (clarity) and overall quality evaluation scores for all the 1,430 currently available in the Welsh Summarisation Dataset. The mean and median scores for the human summaries were 4. The evaluators were instructed to fix common language errors (such as mutation errors and spelling mistakes) but not to correct syntax. All the participants were duly paid an approved legal wage for their work. Table 2 shows the marking criteria. The same criteria were later used when evaluating the system summaries.

[15]Cysill: `www.cysgliad.com/cy/cysill`

| Score | Criteria |
|---|---|
| 5 | • Very clear expression and very readable style.<br>• Very few language errors.<br>• Relevant knowledge and a good understanding of the article; without significant gaps. |
| 4 | • Clear expression and legible style.<br>• Small number of language errors.<br>• Relevant knowledge and a good understanding of the article, with some gaps. |
| 3 | • Generally clear expression, and legible style.<br>• Number of language errors.<br>• 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.<br>• Significant number of language errors.<br>• 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.<br>• Persistently serious language errors.<br>• The information is inadequate for summary purposes. Obvious deficiencies in understanding the article. |

Table 2: Criteria for the marking of summaries

### 3.3. Building Summariser Systems

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 approaches used in this work is extraction-based summarisation. The successful extraction 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 implemented and evaluated basic single-document extractive summarisation systems. That included the use of first-sentence-summary and a simple TF.IDF approach, but when evaluating the summaries using ROUGE we found that TextRank consistently outperformed the others systems when generating summaries of no longer than 250 words. In this paper we only focus on summaries generated using TextRank. The evaluation process took into consideration the human reference summaries as well as the Wikipedia summary (see Section 3.2). The summaries and their ROUGE evaluation results are explained in details in (Ezeani et al., 2022).

TextRank technique was introduced by Radev et al. (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.

## 4. Evaluation Methodology

The performance evaluation of the system summarisers was carried out using variants of the ROUGE[16] metrics as well as human evaluators by scoring summaries generated by the best performing summariser (TextRank in our case (Erkan and Radev, 2004)). ROUGE measures the quality of the system generated summaries as compared with the reference summaries created or validated by humans (see Section 3.2). The current work uses the ROUGE variants that are commonly applied in literature: *ROUGE-N* (where N= 1 or 2) which considers N-gram text units i.e. unigrams and bigrams; *ROUGE-L* which measures the longest common sub-sequence in both system and reference summaries while maintaining the order of words; and *ROUGE-SU4* is an extended version of *ROUGE-S*[17] that includes unigrams. In this work we focus on ROUGE-1 as it was found to correlate particularly well with human judgement(Lin and Hovy, 2003).

Common implementations of ROUGE (Ganesan, 2018) typically produce three key metric scores precision, recall and F1-score as described below.

$$precision = \frac{count(overlapping\ units)}{count(system\ summary\ units)}$$

$$recall = \frac{count(overlapping\ units)}{count(reference\ summary\ units)}$$

$$f1 = (1 + \beta^2) * \frac{recall * precision}{(\beta^2 * precision) + recall}$$

where the value of $\beta$ is used to control the relative importance of *precision* and *recall*. Larger $\beta$ values give more weight to *recall* while $\beta$ values less than 1 give preference to *precision*. In the current work, $\beta$ is set to 1 making it equivalent to the harmonic mean between *precision* and *recall*. The term '*units*' as used in the equation refers to either words or n-grams.

It is possible to achieve very high recall or precision scores if the system generates a lot more or fewer words than in the reference summary respectively. While we can mitigate that with F1 score to achieve a more reliable measure, we designed our evaluation scheme to investigate the effect of the summary sizes on the performance of the systems. We achieved this by varying the lengths of the system-reference summary pairs[18] during evaluation with `tokens` = [`50`, `100`, `150`, `200`, `250` and `None`] where `tokens` indicates the maximum tokens included in the summary and None signifies using the summary as it is. More details on the All reported scores are averages of the individual document scores over all the 513 Wikipedia documents used in the experiment.

In addition, we hired three undergraduate students at Cardiff University to perform the human evaluation of some of the summaries generated by TextRank[19]. In total 80[20] system summaries were evaluated with each summary being scored by each of the evaluators. The participants are two females and one male all aged 20 from Ceredigion, Denbighshire, and Gwynedd in Wales. All are native Welsh speakers. The evaluators followed the same scoring criteria shown in Table 2. In order to avoid bias, they were not told whether those are human or system summaries.

## 5. Results and Discussion

To measure the degree of agreement among the raters we asked the three annotators to blind score the 80 summaries generated by TextRank, all the summarised documents are articles collected from the Welsh Wikipedia as explained earlier (see Section 3.1). Each summary was scored by each of the annotators. To calculate inter-rater agreement we used Pearson Correlation Coefficient and Spearman's Rank Coefficient results, both coefficients were used in previous research to investigate the correlation between ROUGE metrics and hu-

[16]Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004)

[17]Default *ROUGE-S* uses skip-gram co-occurrence which considers any pair of words in a sentence allowing for arbitrary gaps while maintaining the order.

[18]Note that the reference summaries have a length between 230 and 250 words as explained in Section 3.2. Therefore, studying a varying number of smaller lengths helps us in understanding the effect of summary size on the evaluation process.

[19]TextRank generated Welsh summaries of no longer than 250 words each.

[20]With only three evaluators, we were only able to manually evaluate 15% of the generated summaries. The summaries were chosen randomly.

man evaluations (Liu and Liu, 2008; Murray et al., 2005).

The correlation results in Table 3 show low agreements between the human evaluators[21], which is expected given that there is no ideal summary, especially that each evaluator would have personal perspectives and preferences on what to consider key information despite following the same guidelines (El-Haj et al., 2010; El-Haj et al., 2009). The table shows consistent correlations between Pearson and Spearman's, which shows that the evaluators did not agree most of the time, having said that the results are not suggesting zero relationship between the scores given by the human evaluators. Although this might sound negative in a way, we still believe the results are important to shed light on the complexity of the automatic summarisation task in general and in particular (e.g. Welsh text summarisation).

| Evaluators | Pearson | Spearman's |
|---|---|---|
| **E1 vs E2** | 0.170 | 0.161 |
| **E1 vs E3** | 0.325 | 0.355 |
| **E2 vs E3** | 0.327 | 0.233 |
| **R1 vs E1** | 0.154 | 0.168 |
| **R1 vs E2** | 0.007 | 0.117 |
| **R1 vs E3** | 0.014 | 0.201 |

Table 3: Inter-rater agreement scores (Pearson Correlation Coefficient and Spearman's Rank Coefficient). E: Evaluator; R: ROUGE-1.

In addition, we calculate the correlation between the human scores and ROUGE metrics, taking as a use case the results of ROUGE-1. As reported by (Lin and Hovy, 2003), ROUGE-1 was found to correlate particularly well with human judgement. The results in Table 3 show less correlation between ROUGE-1 (R1) and each of the human evaluators, especially when it comes to Pearson's linear relationship correlation, which seems to contradict to the findings reported by Lin and Hovy (2003). This disagreement could be due to the fact that the human evaluations originally run by the Document Understanding Conference (DUC)[22], was performed on news corpora and those are known to be shorter and less informative than Wikipedia articles. The correlation scores could also suggest that ROUGE may be less suited for summaries written in Welsh or languages other than English.

Table 4 shows the distribution of scores in terms of agreement/disagreement. This is shown between the human evaluators themselves as well as between them and ROUGE-1 scores. The results show low agreement between the given scores, again confirming with the correlation results from Table 3.

[21]Note that due to the notion of Pearson and Spearman's formulas, we observe scores > 0.0 despite the lack of agreement between Evaluator 1 and Evaluator 3.

[22]`https://duc.nist.gov/`

| Evaluators | Agree | Disagree | % |
|---|---|---|---|
| **E1 vs E2** | 4 | 76 | 5% |
| **E1 vs E3** | 0 | 80 | 0% |
| **E2 vs E3** | 34 | 46 | 43% |
| **R1 vs E1** | 31 | 49 | 39% |
| **R1 vs E2** | 7 | 73 | 9% |
| **R1 vs E3** | 2 | 78 | 3% |

Table 4: Scores agreement between the raters and ROUGE. E: Evaluator; R: ROUGE-1.

Table 5, shows the breakdown of the Likert Scale scores given by the human evaluators. In addition, we show the ROUGE-1 scores transformed into the same 1-5 Likert Scale for comparison purposes. As shown in the table, ROUGE-1 scores seem to alternate between a scale of 2 and 3, which is expected given the notion of ROUGE's similarity measure, which uses n-grams overlap. This would suggest that it will be difficult for a summary to have a score of zero and again, and given the lack of idealism in summarisation, would also mean that a score of 5 (total overlap) is near impossible since the human (reference/gold-standard) summaries were created using abstractive human summarisation method as explained in Section 3.2. It is also worth noting that the length of the generated summaries is no longer than 250 words but also not less than 10% of the original document, this is to avoid bias towards shorter summaries.

The results show that the human evaluators were more keen to give scores that are either 1 or > 3, which seems to be difficult to achieve using ROUGE. Figure 4 plots that distribution showing a somehow similar pattern between the second (E2) and third (E3) evaluators. On the other hand and given that the first evaluator (E1) scores are confined between 1 and 3, we can examine a pattern between those scores and the ones given by ROUGE-1 (R1).

| Evaluators | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| **E1** | 38 | 29 | 13 | 0 | 0 | 80 |
| **E2** | 1 | 2 | 17 | 34 | 26 | 80 |
| **E3** | 0 | 2 | 5 | 33 | 40 | 80 |
| **R1** | 0 | 50 | 30 | 0 | 0 | 80 |

Table 5: Evaluation scores given by each of the raters and ROUGE. E: Evaluator; R1: ROUGE-1.

## 6. Conclusion and future work

This work shows the creation and evaluation of the first publicly available and freely accessible high-quality Welsh text summarisation dataset. Given that Welsh is considered low-resourced with regards to NLP, this dataset will enable further research works in Welsh automatic text summarisation systems as well as Welsh language technology in general. Overall, the development of the automated tools for Welsh

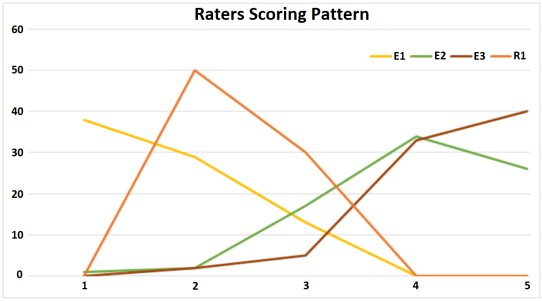


Figure 4: Evaluation (Likert) scores pattern for each of the human raters and ROUGE-1.

language and facilitate the work of those involved in document preparation, proof-reading, and (in certain circumstances) translation. In addition, providing a comparison between human and automatic evaluation results for Welsh summaries should help researchers in developing evaluation metrics that work for complex languages, where there is a less chance of overlapping n-grams between system and human summaries. The correlation results we got are consistent with correlation results in previous research applied on summaries written in English (Liu and Liu, 2008; Murray et al., 2005), which may suggest that the lack of correlation between ROUGE and human evaluations is consistent across different languages. Of course more research is required to fulfil this claim.

We are currently focusing on leveraging the existing state-of-the-art transformer based models for building and deploying Welsh text summariser model. 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 and Rathi, 2021).

In our future work we will examine the correlation between a larger set of system summaries generated using more complex and state-of-the-art summarisation methods as explained earlier and work on recruiting a large group of evaluators to try and match the previous effort by DUC conference.

## 7. 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.

## 8. Bibliographical 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):107–117.

Carlin, P. and Chríost, D. M. G. (2016). A standard for language? policy, territory, and constitutionality in a devolving wales. In *Sociolinguistics in Wales*, pages 93–119. Springer.

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):339–361.

El-Haj, M. and Rayson, P. (2013). Using a keyness metric for single and multi document summarisation. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 64–71.

El-Haj, M., Kruschwitz, U., and Fox, C. (2009). Experimenting with automatic text summarisation for arabic. In *Language and Technology Conference*, pages 490–499. Springer.

El-Haj, M., Kruschwitz, U., and Fox, C. (2010). Using mechanical turk to create a corpus of arabic summaries.

El-Haj, M., Kruschwitz, U., and Fox, C. (2011). Multi-document arabic text summarisation. In *2011 3rd Computer Science and Electronic Engineering Conference (CEEC)*, pages 40–44. IEEE.

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.

Ignatius Ezeani, et al., editors. (2022). *Introducing the Welsh Text Summarisation Dataset and Baseline Systems*, Marseille, France, 20-25 June. The 13th Language Resources and Evaluation Conference, LREC 2022.

Ganesan, K. (2018). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., and Varma, V. (2011). Tac 2011 multiling pilot overview.

Goldstein, J., Mittal, V. O., Carbonell, J. G., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of*

*the association for computational linguistics*, pages 150–157.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Litvak, M. and Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, pages 17–24.

Liu, F. and Liu, Y. (2008). Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, short papers*, pages 201–204.

Magdum, P. and Rathi, S. (2021). A survey on deep learning-based automatic text summarization models. In *Advances in Artificial Intelligence and Data Engineering*, pages 377–392. Springer.

Murray, G., Renals, S., Carletta, J., and Moore, J. (2005). Evaluating automatic summaries of meeting recordings.

Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Song, S., Huang, H., and Ruan, T. (2019). Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications*, 78(1):857–875.

Svore, K., Vanderwende, L., and Burges, C. (2007). Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 448–457.

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

Zmandar, N., El-Haj, M., Rayson, P., Litvak, M., Giannakopoulos, G., Pittaras, N., et al. (2021a). The financial narrative summarisation shared task fns 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 120–125.

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