Introducing the Welsh Summarisation Dataset and Baseline Systems

Ignatius Ezeani[1], Mahmoud El-Haj[1], Jonathan Morris[2], Dawn Knight[2]
{i.ezeani, m.el-haj}@lancaster.ac.uk, {knightd5, morrisj17}@cardiff.ac.uk
[1]UCREL, School of Computing and Communications, Lancaster University, [2]Cardiff University

Abstract

Welsh is spoken by about 884,300 people (29.2% of Welsh population) and is estimated increase its speakers (Census, 2011). However, Welsh remains a minority language being revitalised by the Welsh Government and stakeholders. As part of this effort, this paper introduces the Welsh summarisation dataset for research in advancing the Welsh automatic text summarisation. Welsh speakers were asked to manually summarise Welsh Wikipedia articles and the benchmark summarisation systems were implemented, evaluated and discussed in this paper.

Keywords: Welsh, summarisation, extractive, corpus, word-embeddings.

Methodology

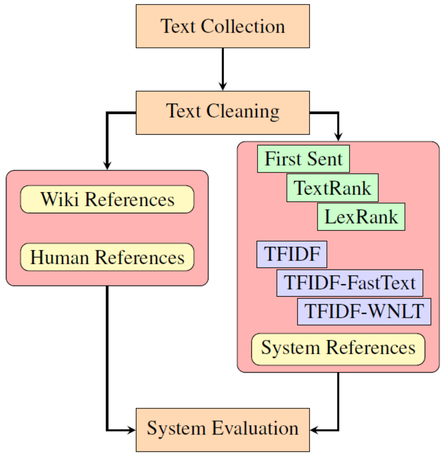

Figure 1: Project Plan

Key steps include:

- Wikipedia text collection + cleaning
- Reference summary creation
- System summary creation
- Evaluation and deployment

Results

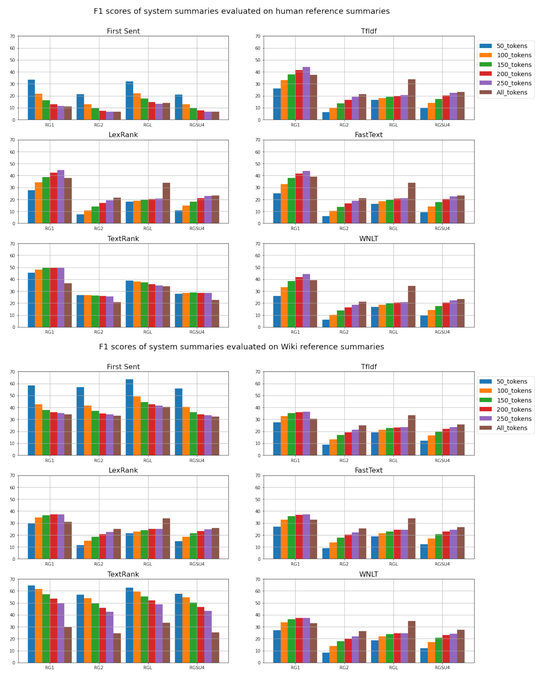

Figure 2: F1: System vs Reference Summaries

Experiments and some key findings:

- Bars show a different maximum length setting - 50 to 250 and None
- `TextRank` achieved best overall score for controlled token length evals.
- All summarizers beat the bottom line model, `First Sent`
- Wiki summaries created with automatic methods gave high precision for `First Sent`
- In general, small reference size leads to low Rouge recall scores
- Contrary to other systems, `TextRank`'s scores dropped as summary size increased.
- Wiki vs human summaries show performance drop confirming variation in inherent qualities - coherence, consistency, fluency and relevance.

Future Work

State-of-the-art abstractive summarization systems use variants of generative models based on popular architectures such as *Bidirectional Encoder Representations from Transformers* - **BERT**, *Bidirectional and Auto-Regressive Transformers* - **BART**, and *Text-to-Text Transfer Transformer* - **T5**

Our current effort focuses on building and evaluating abstractive summarizers based on pretrained mT5 model - multilingual 'Text-to-Text Transfer Transformer' models fine-tuned with Welsh texts.

References

[1] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Experimenting with automatic text summarisation for arabic. In *Language and Technology Conference*, pages 490–499. Springer, 2009.

[2] Ignatius Ezeani, Scott SL Piao, Steven Neale, Paul Rayson, and Dawn Knight. Leveraging pre-trained embeddings for welsh taggers. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 270–280, 2019.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Key Contribution

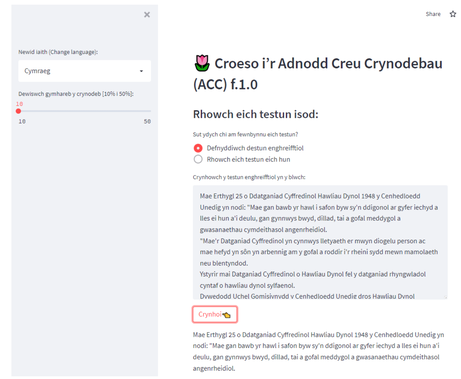

Figure 3: Demo of the Welsh Summarizer tool

Our contributions include:

- Demo of Welsh Extractive summarizer Tool

Figure 4: Try the Demo...

- Dataset of Welsh Wikipedia articles and reference summaries
- Python code and 'how-to-use' instructions

Acknowledgement

We are grateful to the **Government of Wales** for funding this project

Code+data: `https://github.com/UCREL/welsh-summarization-dataset`

Web: `https://corcencc.org/acc/`

Corresponding author: `i.ezeani@lancaster.ac.uk`