INTRODUCING THE WELSH SUMMARISATION DATASET AND BASELINE SYSTEMS

IGNATIUS EZEANI¹, MAHMOUD EL-HAJ¹, JONATHAN MORRIS², DAWN KNIGHT²

{i.ezeani, m.el-haj}@lancaster.ac.uk, {knightd5, morrisj17}@cardiff.ac.uk
¹UCREL, School of Computing and Communications, Lancaster University, ²Cardiff 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

Key steps include:
- Wikipedia text collection + cleaning
- Reference summary creation
- System summary creation
- Evaluation and deployment

RESULTS

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


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

KEY CONTRIBUTION

Our contributions include:
- Demo of Welsh Extractive summarizer Tool
- Dataset of Welsh Wikipedia articles and reference summaries
- Python code and ‘how-to-use’ instructions

Figure 1: Project Plan

Figure 2: F1: System vs Reference Summaries

Figure 3: Demo of the Welsh Summarizer tool

Figure 4: Try the Demo...

For more information, visit https://corcencc.org/acc/