Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language

Dewi Bryn Jones, Sarah Cooper
Bangor University
{d.b.jones, s.cooper}@bangor.ac.uk

Commercial Digital Intelligent Assistants

- Using digital assistants via intelligent speech interfaces increasingly popular
- Used for command and control as well as receiving answers to questions voiced in natural language
- Commercial products: Apple Siri, Google Now, Microsoft Cortana and Amazon Alexa
- Developers and third-parties are encouraged to extend their capabilities via APIs (Application Programming Interfaces) and SDKs (Software Development Kits)
- Mainly provided in English, and to a lesser extent some other major languages
- Unlikely to extend to support smaller languages such as Welsh, in the near future
- No means for external developers to adapt systems for new languages
- Undermines digital linguistic diversity in lesser resourced language communities

‘Seilwaith Cyfathrebu Cymraeg’ Project

- ‘Welsh Language Communications Infrastructure’ Project
- 8 month project funded by the Welsh Government and S4C (Welsh language public service television channel)
- Aim is to include Welsh speakers in the evolution of human computer interaction by:
  - Developing a prototype intelligent digital assistant for Welsh speakers
  - Improving language technologies, especially speech recognition for Welsh
  - Applying these technologies in the prototype
  - Make all resources openly available, with no restrictions, in order to encourage further developments in intelligent digital assistants for Welsh and other lesser resourced langauges

Towards supporting Welsh

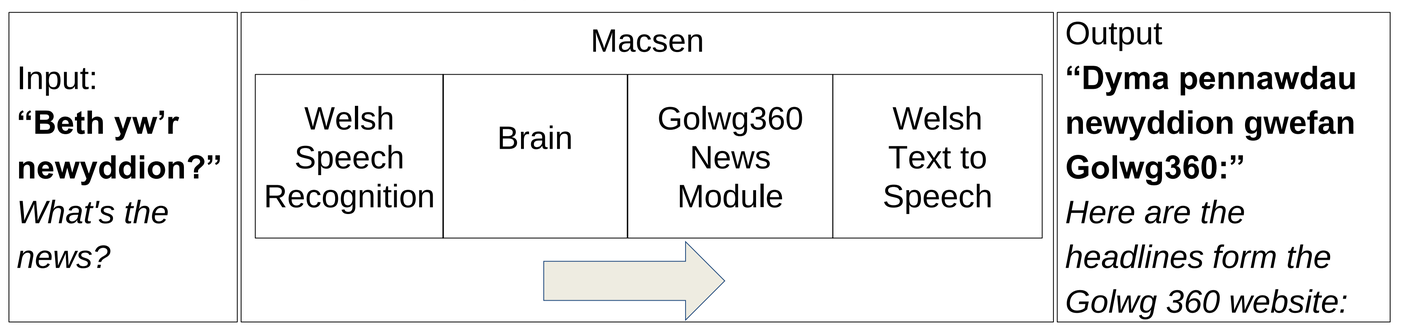
- Commercial architectures have speech recognition and natural processing in one super-component
- Building a Welsh language digital assistant is feasible only with more granular and open architectures:
  - A Welsh language speech recognition engine can be integrated
  - The NLP for understanding requests can be adapted or replaced
  - Welsh responses can be provided via text-to-speech
  - Capabilities rooted in English can still be used

Improving Speech Recognition

- Previous work on Welsh language speech recognition provided:
  - Welsh letter to sound rules
  - a crowd sourced speech corpus (via the iOS/Android based app Paldaruo)
  - a basic robotic arm command and control demo built with HTK and Julius
- To further this we:
  - Trained acoustic models for all Welsh language phones from the entire Paldaruo speech corpus
  - Encapsulated our HTK based training environment in Docker
    - training scripts cater for the fact the the Paldaruo speech corpus has not been humanly evaluated and edited
    - Paldaruo does not provide a test corpus: training data used for evaluation
      - Beneficial in automatically identifying and filtering out low quality or erroneous speech recordings
      - Used word loop grammar as basis for testing
  - Used our HTK Docker environment to improve acoustic models word accuracies: from 20% (from all 410 speakers’ contributions) to 92% (from 88 speakers’ contributions)
  - Developed a comprehensive language-specific HTK decision tree clustering script file (tree.hed) which improved word and sentence accuracies by 1%

‘Macsen’ Prototype Welsh-language Intelligent Digital Assistant

- Simple task grammar and vocabulary files for Julius developed
- Based on the ‘Jasper’ open source project for building your own digital assistant
- Capabilities can be extended with modules: Modules indicate if they can service the request
- Uses natural sounding text to speech voices by Ivona
- Runs on Raspberry Pi
- Supports answering questions and fulfilling tasks in the domains of news, weather, time, proverbs and jokes



Further work

- Improve speech recognition for Welsh with more recent tools such as Kaldi
- Create language models from corpora: 30 million word Cysill Ar-lein Corpus
- Investigate application of machine translation for consuming English services
- All models, scripts, code and data available via the Welsh National Language Technologies Portal and GitHub: http://techiaith.cymru/macsen

This project was supported financially by the Welsh Government, through its Technology and Digital Media in the Welsh Language fund, and S4C, the Welsh language television channel.
We would also like to thank all hackers and communities of users that assisted us on the project.

S4C

Noddir gan **Lywodraeth Cymru**
Sponsored by **Welsh Government**

PRIFYSGOL BANGOR UNIVERSITY