Initial Plan:

Welsh Natural Language Knowledge-Base

Elliot Howells
Student ID: 1110117

Supervisor: Professor Alun Preece

Module Code: CM3203
Module Title: One Semester Individual Project
Credits: 40

28th January 2016

Project Description

There is an ever-increasing, interdisciplinary, focus on human-computer collaboration (HCC) and the ways in which users can best communicate with machines. Historically, there have been two main approaches to HCC; giving computers human-like abilities, usually focusing on language, or exploiting a machine's unique abilities to complement humans. Research suggests that the most effective way to encourage human-computer collaboration is through a unified approach, incorporating insights from both methods [1].

This presents a challenge; humans prefer natural language and images, which are difficult for machines to process and lead to ambiguity and miscommunication [2]. The use of a controlled natural language (CNL) provides a compromise. CNL is a subset of a natural language using restricted vocabulary, designed to be processed easily by machines, while also being human-readable and writable [2]. It is possible for humans to communicate in CNL, providing they have received training. However, unrestricted natural language is more preferable and the advised approach is to give human users the choice of using unrestricted natural language or CNL. Many CNLs have been defined, including a form of controlled English known as ITA Controlled English (CE), developed by IBM [3].

ITA CE is a controlled language defined by the International Technology Alliance, and is less strict in terms of precision than some other controlled languages. It favours machine-processability as the form of English is low-complexity and is unambiguous to a machine [4].

Enabling human-machine conversations with CE as a form of restricted English is Controlled English Node (CENode) [5]; a lightweight natural language knowledge-based system implemented using JavaScript. It is designed to run effectively in a wide variety of contexts, from mobile devices to servers and other ubiquitous computing devices.

CENode-based applications allow users to input queries and pieces of information in a conversational protocol, referred to as 'ask and tell'. The applications can also 'ask and tell' the user things. For example, CENode-based applications include Moira (Mobile Information Reporting Agent) for obtaining real-time reports. This has particular benefits for emergency services' professionals on patrol, for example. Another example is SHERLOCK (Simple Human Experiment Regarding Locally Observed Collective Knowledge) for crowdsourcing knowledge bases using this 'ask and tell' method of input.

These applications have clear benefits, allowing users to interact with machines in a language that is easy to understand for both human users and the device. However, they are currently restricted to English-language input, given that the knowledge-base is written in Controlled English, and no other language. This project aims to create an alternative version of this language, based on Welsh, and to adapt CENode to process "Controlled Welsh".

It is the hope that by developing a version of "Controlled Welsh", lessons will be learnt that can be applied to the development of CNLs in other languages that are

used by minority populations. This will allow such technologies to be used in areas that speak languages other than English, extending their uses and benefits to new communities. The project will also research the existing uses of natural language knowledge-based systems and consider how lessons can be learnt to identify applications for “Controlled Welsh”.

In addition to the research and development of a technical system, this project will consider the current existence of, and demand for, Welsh language technology and how this existence and demand informs Welsh speakers’ use of such technologies. As part of this, the project will look at how existing, and potentially future, Welsh language policy may influence the development of Welsh language technologies, and systems such as those based on natural language knowledge-bases.

The project aims to develop at least one application of “Controlled Welsh” and provide an opportunity to demonstrate and evaluate this application. Based on the evaluation and lessons learnt during the demonstration of this application, the project will consider what a fully bilingual natural language knowledge-base may look like and the potential uses for such systems in Welsh and beyond.

The ultimate goal of this project is to develop an understanding of how to design a non-English version of Controlled English and in turn, how this could be done for a language used by a minority population, such as Welsh.

Project Aims and Objectives

The following aims and objectives describe what this project sets out to specifically achieve. They are included in no order of priority, with varying weights of workload, but will be used to evaluate the progress and success of the project over its period.

- An understanding of existing natural language knowledge-bases.
    - Conduct secondary research into the natural language knowledge-bases that have been developed in languages similar to Welsh, in their linguistic and usage characteristics, and consider their applications.

- An understanding of the technical needs of Welsh speakers.
    - Conduct research into what influences a Welsh speaker's decision to use a piece of technology, reviewing existing research and surveying Welsh speakers determining if bilingualism is a major factor in this use.

- An understanding how current and future Welsh language policy may influence the development of technologies.
    - Conduct secondary research in to Welsh language policies that may have an impact on technologies such as "Controlled Welsh".

- A translation of SHERLOCK into Welsh to demo with Welsh speaking users.
    - Identify strings that need translation from English into Welsh.
    - Translate strings and amend the code to test translated application.
    - **Deliverable 1: APP**

- A demonstration of the translated application.
    - Demonstrate translated SHERLOCK to a group of selected individuals.
    - Gather feedback from users on the application.
    - Monitor interaction with the application and evaluate.
    - Make changes based on lessons learnt from the demonstration.
    - Demonstrate this translated and evaluated application, or another, to a non-selected group of individuals at a larger event for evaluation.
    - **Deliverable 2: DEMO**

- An exploration of what a fully-bilingual SHERLOCK would look like.
    - Using lessons learnt from the previously demonstrated application, explore how SHERLOCK could be fully translated into a bilingual Welsh and English application. **Deliverable 3: APP**
    - Consider how Welsh language policies may be introduced for the use of such a system.
    - Establish whether use of such applications could nudge behaviour change between in the use of Welsh vs English.

- A detailed report of the findings of the project.
    - Produce a report covering the overall project background, approach to delivering the goals and objectives including details of findings.
    - **Deliverable 4: REPORT**

Ethical Considerations

Having reviewed Cardiff University School of Computer Science & Informatics' ethical guidelines [6], it is my understanding that the planned work does not require review by the School's Ethics Committee. No personal or personally-identifiable information will be collected from users of the prototype application(s) and I will ensure any users of the system(s) developed are aware of the project and its aims and objectives.

Should this change, the appropriate approval and documentation will be acquired.

Work Plan

When developing this work plan, I have ensured adequate time is given to the research, development and implementation of each task and where appropriate, allocated time to evaluate. I have also ensured that consideration is given to other commitments such as coursework deadlines, examinations and periods of holiday. The final report will be worked on throughout the project period and dates for submission have been included below.

Note: The following periods are labelled as weeks from the beginning of the project, and not University teaching weeks.

Pre-Week 1
03/11/15 – 24/01/16

- Background reading around the project's aims and objectives has been done.
- Meetings have taken place with Dr Jeremy Evas (Cardiff University School of Welsh) and Gareth Morlais (Welsh Language Technology, Welsh Government).
- Supervisor meetings have begun, focusing on the scope of the project and potential directions in which the project could take.

Weeks 1 – 2
25/01/16 – 07/02/16

- Write and submit initial plan. Deadline: 31/01/16.
- Research natural language knowledge-bases and their applications.
- Conduct research into use of Welsh language technologies.
- Identify strings in SHERLOCK that need translation from English into Welsh.

Weeks 3 – 4
08/02/16 – 21/02/16

- Continue research into use of Welsh language technologies.
- Translate SHERLOCK strings in preparation for implementation.
- Amend application code.
- Test application in preparation for demonstration.
- **Deliverable 1: APP**

Weeks 5 – 6
22/02/16 – 06/03/16

- Demo translated SHERLOCK application to selected individuals. Date: TBA.
- Monitor interactions with the application.
- Evaluate success of the demonstration.
- **Deliverable 2: DEMO**

Weeks 7 – 8
07/03/16 – 20/03/16

- Write lessons-learnt report in preparation for future developments.
- Explore how SHERLOCK could be fully translated into Welsh and used bilingually, nudging use between English and Welsh.
- Consider how Welsh language policies may be introduced for the system.

Weeks 9 – 10
21/03/16 – 03/04/16

- This period will be primarily spent writing up in preparation for the final report.
- Note: I am out of the country during these weeks and while work will continue remotely, progress will be significantly slower during these periods.

Weeks 11 – 12
04/04/16 – 17/04/16

- Potentially demonstrate an amended version of the application, or a second application, at a larger event to non-selected individuals.
- Monitor interaction with the application for further evaluation and establish whether the ultimate aim of the project has been achieved.
- **Deliverable 3: APP**

Weeks 13 – 14
18/04/16 – 01/05/16

- Produce draft version of final report.
- Make amends to draft version.

Week 15 Onwards
02/05/16 – 10/06/16

- Finalise and submit final report. Deadline: 06/05/16.
- Prepare for Viva. Date: TBA
- **Deliverable 4: REPORT**

Supervisor Meetings

I will have regular meetings with my project supervisor ensuring work is on track and to give me the opportunity to test any developed applications. In between formal meetings, Slack [7] will be used for continuous communication to ensure updates are communicated efficiently.

Provisional dates for these meetings are set below:

- Friday, 15th January 2016
- Tuesday, 2nd February 2016
- Tuesday, 16th February 2016
- Tuesday, 1st March 2016
- Tuesday, 15th March 2016
- Tuesday, 12th April 2016
- Tuesday, 26th April 2016
- Tuesday, 10th May 2016

Gantt Chart

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	>15
W/C	25/01/16	01/02/16	08/02/16	15/02/16	22/02/16	29/02/16	07/03/16	14/03/16	21/03/16	28/03/16	04/04/16	11/04/16	18/04/16	25/04/16	02/05/16
Task															
Initial Plan															
Natural Language Knowledge Base Research															
Welsh Technology Research															
SHERLOCK Translation															
SHERLOCK Testing															
Welsh SHERLOCK Demonstration															
Demonstration Evaluation															
Bilingual SHERLOCK Research															
Welsh Policies Research															
Bilingual SHERLOCK Development															
Bilingual SHERLOCK Evaluation															
Report Writing															
Viva															

References

[1] L. Terveen. Overview of human-computer collaboration. Knowledge-Based Systems, 67–69, 1995.

[2] A. Preece, C. Gwilliams, C. Parizas, D. Pizzocaro, J. Z. Bakdash, D. Braines. Conversational Sensing, 5-7, 2014. [Online]. Available: http://arxiv.org/pdf/1406.1907.pdf. [Accessed: 26th Jan 2016].

[3] IBM, IBM Controlled Natural Language Processing Environment. [Online]. Available: https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=558d55b6-78b6-43e6-9c14-0792481e4532. [Accessed: 26th Jan 2016].

[4] T. Kuhn, A Survey and Classification of Controlled Natural Languages. [Online]. Available: http://www.aclweb.org/anthology/J14-1005. [Accessed: 27th Jan 2016].

[5] Cardiff University, IBM UK, CENode.js. [Online]. Available: http://cenode.io. [Accessed: 26th Jan 2016].

[6] I. Spasic, Cardiff University. Research Ethics. [Online]. Available: http://users.cs.cf.ac.uk/I.Spasic/ethics/. [Accessed: 26th Jan 2016].

[7] Slack. [Online]. Available: https://slack.com. [Accessed: 26th Jan 2016].