A roundtable discussion to promote a strategic vision for Celtic Language Technologies

Report compiled by Delyth Prys and Indeg Williams, Bangor University, August 2019

Discussion led by Delyth Prys, Caoimhín O Donnaile, Kevin Scannell, Colin Batchelor, Meghan Dowling.

Notes taken by Indeg Williams.

Contents

I Sharing Information

1. Background

A roundtable discussion was held by the Celtic Language Technologies Group as part of the Celtic Congress held at Bangor University on Friday 26 of July, 2019. The session was well-attended, with about 20 participants present. Members of the CLT group who were unable to attend were encouraged to send in relevant questions and comments ahead of time, and these were gathered under relevant thematic headings for discussion. This document attempts to summarise the discussions and present them for further development of a common strategic vision and action by the Celtic Language Technologies community.

2. The Celtic Language Technologies Group

The CLTG is an informal gathering of people interested in Language Technologies for all Celtic languages. It exists as an informal discussion list open to all. To join go to https://groups.google.com/forum/#!forum/celtic-language-technology.

The CLTG encourages academic research in LT for the Celtic languages and has so far organised three workshops allied to high profile international conferences to further this aim. Proceedings for the first two can be found on-line and the third will be available soon.

- Proceedings of the first CLTW Workshop 2014 https://aclweb.org/anthology/W14-4600
- Proceedings second CLTW Workshop 2016 https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V06-CLTW.pdf
- 3rd Workshop on Celtic Language Technologies – forthcoming, Dublin 19th August 2019 https://www.mtsummit2019.com/workshops

Together, our six Celtic languages can make more of an impact than they can on their own. There are many other benefits in working together, both in terms of mutual help and support and shared research on closely related languages.

Participants were encouraged to join the group as a low cost way of contributing together to the development of CLT.

3. Recent Relevant Strategic Documents

Attention was drawn to three important strategic documents published in the latter half of 2018 that are relevant to CLT research. We were encouraged to become familiar with their content, and to refer to them in any forthcoming grant applications to help make the case for funding for our research. Participants were asked if they knew of any other relevant documents to add to the list, but none were added. The three documents are detailed below.

a. The Digital Language Survival Kit

http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf

The Digital Language Survival Kit was one of the outputs of the Digital Language Diversity Project, headed by Claudia Soria of Pizza University and funded by the EU under its Erasmus+ Programme. Its aim is to provide recommendations to improve digital vitality for regional and minority languages. It provides an instrument for communities to self-assess the vitality of their language and to engage in concrete actions and initiatives to improve this level of vitality.

b. European Parliament Resolution on Language Equality in the Digital Age

http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html?redirect

The resolution on Language Equality in the Digital Age was proposed to the European Parliament by Jill Evans, MEP for Wales, with the support of the Greens/European Free Alliance Group, and adopted by the Eurpean Parliament at Strasbourg on the 11th of September 2018 with 592 votes for, 45 against, and 44 abstentions. It details current obstacles to achieving language equality in the digital age in Europe; ways to improve the institutional framework for LT policies at EU level; recommendations for EU research policies; education policies to improve the future of LTs in Europe; and list benefits for both private companies and public bodies in the use of LTs.

c. Welsh Language Technology Action Plan

https://gov.wales/welsh-language-technology-and-digital-media-action-plan

This action plan was published in October 2018 to help in realizing the Welsh government's ambition to double the number of Welsh speakers to one million by 2050 (see https://gweddill.gov.wales/docs/dcells/publications/170711-welsh-language-strategy-eng.pdf). Digital technology and linguistic infrastructure form an important part of this strategy, and the Action Plan details the necessary steps needed as including specifically Machine Translation, Speech Technology and Conversational AI. Although it is aimed primarily at the Welsh language there are many useful pointers in the document for the other Celtic languages.

4. Information on New and Current Projects

Kevin Scannell was congratulated on winning a Fulbright Scholarship where he will spend 6 months in Ireland developing computer resources for the Irish language.

Delyth Prys reported that the Language Technologies Unit at Bangor University had just won a Welsh Government SMART partnership with a local translation company to develop neural net Machine Translation for the Welsh/English translation pair.

Meghan Dowling noted both her research with Irish MT and her voluntary work as editor of the Irish language Wikipedia.

Caomhín Ó Donnaíle reported on a number of Scottish Gaelic project he was engaged with, mainly lexicographical ones, and his hosting of several websites at Sabhal Mòr Ostaig.

Johannes Heinecke referred to his research on a syntax-treebank for Welsh, on which he had given a paper at the Celtic Congress.

Indeg Williams explained her PhD research with Welsh speech technology and experiments to compare results using various different speech recognition platforms.

Delyth Prys mentioned the new Cornish<>English on-line dictionary which was based on the Maes T platform for developing Welsh terminology, on which she also had given a paper at the Celtic Conference.

Theodorus Fransen reported on his forthcoming move to Galway University and his research on computational approaches to historical Irish. He also presented a paper on this at the Celtic Congress.

II Questions and discussion points

5. Needs and Priorities for Celtic Language Technologies

Amongst the most urgent problems the following were noted:

- The difficulties in finding skilled people. Getting the necessary combination of Irish+linguistic+technical knowledge in one person is difficult. The same is true of other Celtic languages also.
- Gathering sufficient data for LT projects, e.g. for MT. It is necessary to educate people on what data is, on its value, and how it can benefit them to share translation data, for example in order to develop new translation tools for their use.
- Getting speakers to use what's already available. For example, Scottish Gaelic is fairly well off in terms of localized software and dictionaries thanks to one engineer in Glasgow, but these resources are not used as much as they could be. Getting Gaelic

schools to use the available tools and resources is a challenge because the necessary Gaelic characters aren't even present on the computer keyboard.

- The need for semantically annotated data, and big corpora for training purposes. Big cross-lingual projects tend not to include Celtic languages and this needs to be rectified.
- Copyright issues. Scarcity of data is compounded by lack of appropriately licensed tools and resources and ignorance on licensing issues.

6. Upskilling people in terms of technology in the context of Digital Humanities

a. Courses

There seemed to be a general lack of suitable courses in language technologies, especially cross-over ones to teach computing skills to linguists and linguistic skills to computer scientists. Courses in Computational Linguistics, Speech and/or Translation Technology or Information Sciences are available internationally but with little reference to the Celtic and other morphologically rich languages.

Delyth Prys reported that a new Masters in Language Technologies was under development at Bangor University but had not been approved yet. Meghan Dowling reported that her undergraduate course was relevant but that it had very low numbers. LT projects in Ireland were instead looking for computer science graduates who speak Irish to fill in the gap. In Wales it seemed easier get linguists but harder to find computer scientists interested in language.

Gruffudd Prys reported on how he had been training interns 'on the job' and breaking down NLP tasks to create simpler tasks for them. This method seemed to be successful in generating interest in the field and hopefully producing students who would like to undertake courses in LT in future.

b. Workshops

The forthcoming Celtic Language Technologies Workshop in Dublin was again referred to but other than that few formal workshops were planned for the near future. If we were thinking of holding small training sessions for our own local needs, maybe we could extend those to include participants working with other Celtic languages and circulate information on them through the CLT group. For example, Welsh workshops were currently only advertised in Wales, but with a little effort could be opened to other Celtic participants.

Johannes Heinke noted the need for training for our annotators as there was currently a shortage of them. Others agreed that workshops was one way of accomplishing this. We should look therefore at specific topics that needed to be addressed and find ways of working together to provide training.

7. Sustainability of resources

Some participants had been to the session on the eDIL project (electronic Dictionary of the Irish Language) at the Celtic Congress. We had heard that it's coming to an end at the end of this summer, with no long term sustainability plan yet in place. If this is true of eDIL, it must be true of many of our resources. This was confirmed by other participants from Wales and Scotland. Both the LTU at Bangor University and Sabhal Mòr Ostaig on Skye had experience of keeping looking after old repositories and data sets after project end so that they would not cease to be available.

Caoimhín Ó Donnaíle noted that this was a problem also with archiving recordings and other types of data. This could be improved with better funding but often programmers have moved on, and there is a continuing problem for maintenance of old websites, software updates and now also security problems which cannot be ignored.

Colin Batchelor noted that focusing on safeguarding data rather than code (as the data is less brittle) would help.

The use of international repositories such as github was recommended, as they are suited for long term publication and unlikely to be withdrawn soon. Use of Docker for containerized resources, and Metashare, a European project, suitable for data storage, were also recommended. It was noted that the Welsh National Technologies Portal (techiaith.cymru) was really a brochure website with the data and code actually kept in github, Docker and Metashare. A data management plan should be part of any grant application, and is a requirement in many science and software grant schemes such as those by the European Commission. Kevin Scannell noted that a data management/archiving plan is a strict requirement in the US for research projects.

It was noted that a workshop on how to create data management and archiving plans for the digital humanities/academics could be a useful contribution by members of this group. Some academics from linguistic and humanities background still need to learn what is a repository, what is a data set etc. even though these issues are well understood by people from a more scientific background.

8. New and improved methodologies

We discussed how to streamline methods and frameworks and how to use technology existing for modern Celtic varieties, and make them compatible; how to establish a healthy balance between manual annotation on the one hand and automatic morphological/ syntactic parsing technologies on the other; and the potential benefit of language-independent machine learning.

Theodorus Fransen shared his perspective from working with Old Irish where he was using mainly manual annotation at the moment. Whilst working, he had to decide if he should create a rule for specific linguistic phenomena or continue manually, with this being a constant challenge/problem. What was true of working with Old Irish is also true of the wider context. Older versions of our Celtic languages were problematic in that their orthography was not standardized. On the other hand they had more lexical and grammatical features in common.

Semi-automated solutions for annotation were discussed. David Chan suggested that there was a difference between contexts where features need to work perfectly and those where we can accept partial success with some examples of failure. Prioritizing important elements could also be important. Gruffudd Prys noted that in his research he found that automatic methods were not good enough and that he had to spend too much time fixing things himself for it to be worth it. Evaluation of time saved though automated processes might help.

The newer neural approaches were discussed, as well as hybrid methods, especially given the lack of sufficient ‘big data’ in some areas. There was a feeling that they might be helpful in some cases if not in all. There may be a payoff between adequacy and fluency, some situations favour one over the other, and both are useful in different contexts. Kevin Scannell reminded us that the big companies are publishing the ‘best’ results, but these are English language developments. The research is optimized for English, and we were right to be skeptical for the use of some of the newer methods with Celtic languages/other languages with complicated morphology. There may be a danger that we’re left behind with the fast pace of new developments. What may be state of the art for Celtic may not be the same worldwide, other larger languages may have moved on. We might be better off if we come up with our own models instead of adapting English models. We needed PhD programmes in language technology, e.g. neural net machine translation. Meghan Dowling noted that the automatic metrics are all geared towards English. They are trained on English so the program is surprised to see something different, such as different forms etc. Since the Celtic languages share certain features, learning from other Celtic languages might be useful. This was echoed by Camhín Ó Donnaíle who observed that it is often possible to tell if a machine has translated the work in Scots Gaelic, not because of errors, but because of because of features like word order. A sentence may often be ‘correct’ but not what a native speaker would produce.

Delyth Prys mentioned current efforts to form a research proposal ‘Big Data for Small Languages’ between Wales and Ireland but looking at transfer learning between all the Celtic languages. Maybe the P-Celtic and Q-Celtic sub-groups could usefully be treated together for some tasks, and the issue of time periods (looking at earlier forms of our languages when they were closer together) could also be explored. Any interested participants were invited to continue the discussion after the session.

9. Crowdsourcing

We discussed volunteering and crowdsourcing in order to obtain speech and text data from our language communities. This is now a popular method of working even for major languages, due to the cost of paying informants for the very large datasets needed for many LT tasks. Although our language communities are small, volunteers are often happy to help due to their desire to see our languages survive and prosper in a digital environment.

Some international platforms such as Wikipedia and Mozilla’s Common Voice project facilitate crowdsourcing and sharing of resources and make it easier for small language communities to engage with volunteers.

Rhoslyn Prys outlined his experience of getting other volunteers to contribute recordings of their voices to Mozilla’s Common Voice project. Originally launched for English only about 2 years ago, it added other languages, including Welsh, a little over a year ago. Anyone can ask for other languages to be included, but that language community has to supply its own sentences to serve as recording prompts, and run its own publicity campaigns to get its members to contribute. Contributors read aloud predetermined sentences which are recorded and stored on Mozilla servers. Mozilla process the data and make the data files available for download. Although their initial use is for their own DeepSpeech project, the data is released on a CC-0 license and may be used by other users for any other purpose. In Wales many organizations have contributed to the project, e.g. the Mentrau Iaith (local language ventures), National Library of Wales, Welsh Government, Gwynedd Council. Welsh has 56 hours of speech recordings after one year, with Mozilla aiming for tens of thousands of hours. This is challenging for all small languages and we are working with Mozilla to introduce gamification methods to help keep contributors interested and increase participation.

Wikipedia is another international platform which provides valuable datasets for many small languages who can’t obtain data any other way. The CC-BY-SA license is restrictive for the development of commercial products, but Wikipedia is still a useful source of data for many situations.

Meghan Dowling talked about her experience as a Wikipedia for Irish. Language quality is often an issue, but other editors can help in correcting articles where the many linguistic errors. There was a lively debate on the usefulness or otherwise of Wikipedia articles that contain grammatically incorrect sentences (a problem for all Celtic languages, not only Irish). In some instances it is better to have something than nothing at all. David Chan pointed out that this is a particular problem that some minority languages have, due to a large number of learners contributing, and is not true of all lesser resourced languages. Looking at Chinese/Indian minorities, people contributing to Wikipedia have full command of their languages and errors made by less fluent speakers is not a problem. It is a challenging issue to solve it at a Wikipedia level as it only affects a small group of languages.

For some uses such as deep learning, quantity is better than quality, so variable quality is not such an issue. In terms of using it as training data, it is possible to be more lenient. It may also be possible to use software that assigns a quality level to filter out lower quality linguistic data.

III Conclusions

In conclusion, participants felt that this had been a very useful discussion. Key points included:

- The sharing of information across researchers working on individual Celtic languages
- Working together to improve training and providing courses in Language Technologies
- Developing transfer learning methodologies and common language models for our language group.
- The urgent need for sustainability and long-term solutions to maintain our resources so that they continue to be available and accessible to us after projects end.

We agreed that we enjoy working together across the Celtic languages. Individually, our language communities are very small, and joining together helps us achieve a critical mass for developing research projects and providing mutual support. We look forward to increased collaboration between members of our group.