AI in minority language contexts: a new digital divide?

4 Bealtaine 2022

Languages and AI

- AI systems can perform tasks usually associated with human intelligence
- Driving cars, reading X-rays, playing chess or Go, etc.
- Language is the quintessential expression of human intelligence (Turing test)
  - Machine translation
  - Speech recognition and question answering (Siri/Alexa)
  - Text generation (news articles, also spam, student essays?, ...)
  - Proofing tools (style and grammar correction)

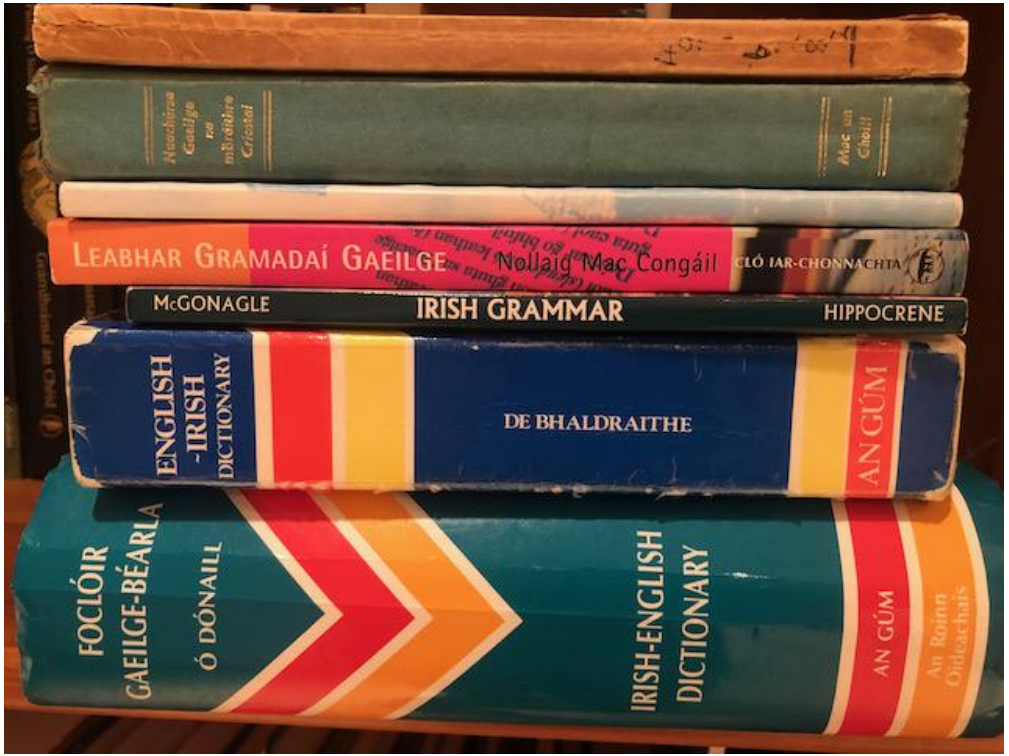
Outline

- How does AI work? A simple example
- The landscape of AI research
- The new digital divide
- Who defines the Irish language?

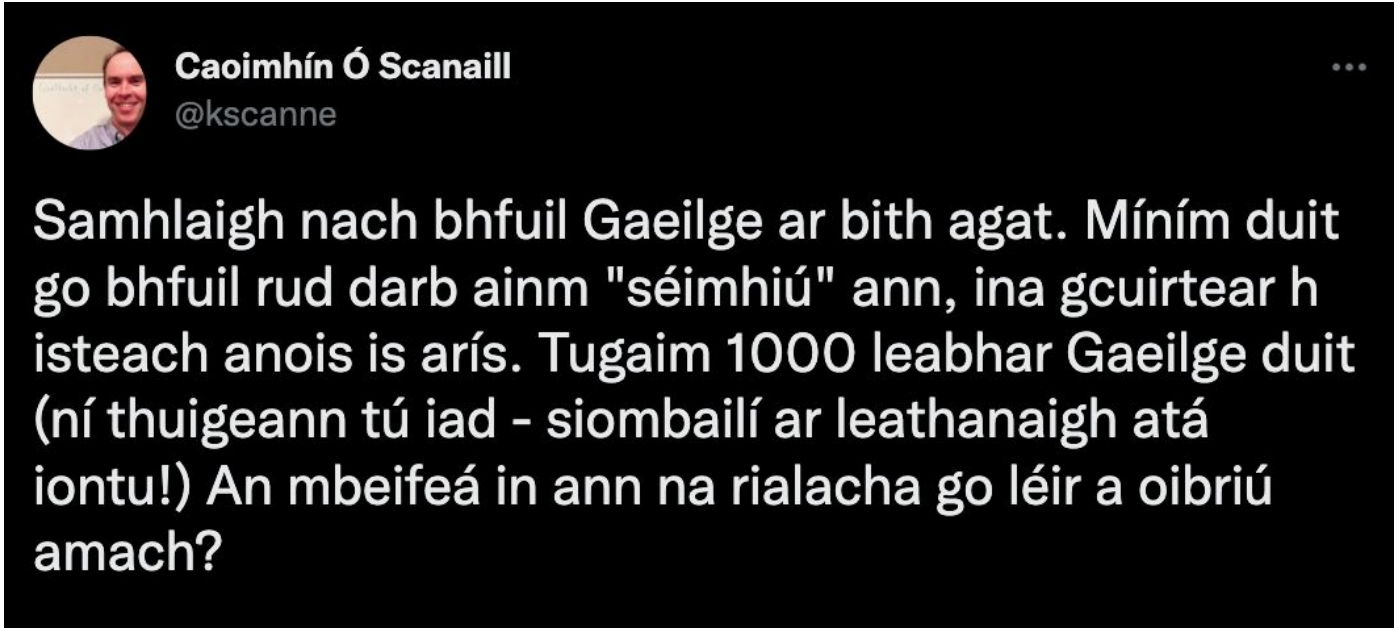
A simple example

- Irish has so-called “initial mutations”
- ***bean*** (“woman”) but ***an bhean*** (“the woman”)
- I have a computer program that will check these for you!
- This is a kind of AI; certainly a non-trivial task for humans, even fluent speakers
- In fact, I’ve written this program, from scratch, *twice* in my life

Version 1: 2000–2004



Version 2: 17–18 October 2019



How does this work?

- Makes use of a large *neural network*
- I provided about 50 million words of Irish text to the network to learn from
- Learns statistical patterns that allow it to recognize if mutations are needed
- Think about what I *didn't* provide:
  - No dictionary at all
  - No understanding of nouns, verbs, adjective, or even that such categories exist
  - No notion of masculine or feminine nouns (which play an important role in mutations)
  - Nothing about broad vs. slender consonants, dentals, etc.

Why is Version 2 better than Version 1?

- Two days vs. four years to develop
- No knowledge of Irish required (!)
- (?) Reflects only actual usage by Irish speakers
- Resulting system actually works **better**
  - Can handle new or made-up words not in Version 1's dictionary ("snideog m**h**ór")
  - Works well with older versions of Irish also
  - Most important: Version 2 often gets the right answer even in cases that require knowledge that is *too subtle to be encoded in rules*

Landscape of AI Research: Primacy of English

- Biggest advances are now driven by industry players, not by academics
- Virtually all of the research in this area is focused (implicitly!) on English
- The word “English” doesn’t even appear in many landmark papers
- Advances are sold as advances in language technologies in general

AI is driven by Big Data

- 50 million words of Irish might sound like a lot!
- Recent models for English have been trained on 270 *billion* words
- Maybe 100x more than all Irish text that's been written, printed, or typed, ever
- These approaches will *never* be accessible to minoritized languages
- This is the "new digital divide"
- Smaller language communities that can't assemble the datasets to build speech interfaces for example, will be forced to shift languages

Data Curation

- Garbage in, garbage out
- I took tremendous care in selecting the 50 million words of training text
- But most systems are built with random text crawled from the web
- As much as 10% of this text in standard datasets is Google Translated!
- Maybe another 5-10% written by learners without a strong command of Irish

What is Irish?

- Data curation raises important questions around authority and standards
- I make decisions every day to include or exclude texts from the models I build
- Implicit value judgements over what Irish is “good enough”
- Make every effort to be balanced by dialect, gender, etc.
- Still, I have huge qualms about being the arbiter of what is included/excluded
- No one at Google is worrying about this

Corporate priorities

- Google, Apple, and friends control the platforms to deliver these tools
- They decide which languages “get in” and which don’t
- No consultation with language communities on which tools they want
- Why Google Translate adding 13 new languages isn’t good news
- Potential harms that can come from these tools
- Important questions about community control over language data
- Māori are trying to save their language from Big Tech
- A new vision of artificial intelligence for the people

Prospects for 2030

- AI is a powerful tool for developing software to process human language
- Research focuses on English and approaches that don’t “scale down”
- Minoritized language communities risk being locked out of these tools
- What’s needed?
  - R&D informed by the needs of language communities
  - New approaches to AI that don’t require massive datasets to train
  - Local control and curation of high-quality datasets

Go raibh maith agaibh!