Minority Language Wikipedias in an AI-Dominated World

Kevin Scannell
Cadhan Aonair

WIKIMEDIA LANGUAGE CONFERENCE
CELTIC KNOT '24

Large language models

LLM-driven AI

- ChatGPT and friends
- Google Translate and friends
- Same underlying technology, more or less
- Driven by *text generation*, conditioned on some “prompt”

How to train your own LLM

- Gather as much text as possible!
- GPT-3: 300 *billion* tokens
- Llama 3.1: 15 *trillion* tokens (“publicly available sources”)
- 40 million GPU-hours, 11000+ tons of CO_2 emitted
- Training data usually includes text in many languages
- *Wikipedias are standardly included to train multilingual models*

Minority languages

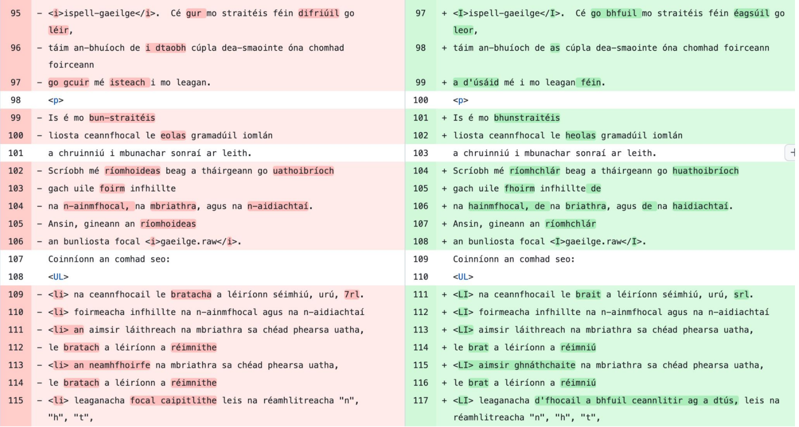
- Every word ever committed to paper or computer in Irish?
- Very likely less than 1 billion words (30000x smaller than Llama 3.1)
- The Irish data included in standard LLMs is of poor quality
    - CommonCrawl, OSCAR, etc. heavily polluted with machine translation
    - Wikipedia is wildly variable in quality (more in a moment)
- Good news: Corpus-building efforts underway in Fiontar DCU
- Highest-quality material, well-balanced, etc. — 150 million words

Role of Wikipedia

Wikipedia and ground truth

- It’s well-known that ChatGPT “makes things up” or “hallucinates”
- Again, it’s a text generation machine; no grounding in fact
- The web/social media are filling up with LLM-generated rubbish
- Greatly increases the importance of human-curated sites like WP
- Quality of WP impacts the quality of LLMs trained on it!

“Good Irish”



```
95   - <i>ispell-gaeilge</i>.  Cé gur mo straitéis féin difriúil go
       léir,
96   - táim an-bhuíoch de i dtaobh cúpla dea-smaointe óna chomhad
       foirceann
97   - go gcuir mé isteach i mo leagan.
98     <p>
99   - Is é mo bun-straitéis
100  - liosta ceannfhocal le eolas gramadúil iomlán
101    a chruinniú i mbunachar sonraí ar leith.
102  - Scríobh mé ríomhoideas beag a tháirgeann go uathoibríoch
103  - gach uile foirm infhillte
104  - na n-ainmfhocal, na mbriathra, agus na n-aidiachtaí.
105  - Ansin, gineann an ríomhoideas
106  - an bunliosta focal <i>gaeilge.raw</i>.
107    Coinníonn an comhad seo:
108    <UL>
109  - <li> na ceannfhocail le bratacha a léiríonn séimhiú, urú, 7rl.
110  - <li> foirmeacha infhillte na n-ainmfhocal agus na n-aidiachtaí
111  - <li> an aimsir láithreach na mbriathra sa chéad phearsa uatha,
112  - le bratach a léiríonn a réimnithe
113  - <li> an neamhfhoirfe na mbriathra sa chéad phearsa uatha,
114  - le bratach a léiríonn a réimnithe
115  - <li> leaganacha focal caipitlithe leis na réamhlitreacha "n",
       "h", "t",
```

```
97   + <I>ispell-gaeilge</I>.  Cé go bhfuil mo straitéis féin éagsúil go
       leor,
98   + táim an-bhuíoch de as cúpla dea-smaointe óna chomhad foirceann

99   + a d'úsáid mé i mo leagan féin.
100    <p>
101  + Is é mo bhunstraitéis
102  + liosta ceannfhocal le heolas gramadúil iomlán
103    a chruinniú i mbunachar sonraí ar leith.
104  + Scríobh mé ríomhchlár beag a tháirgeann go huathoibríoch
105  + gach uile fhoirm infhillte de
106  + na hainmfhocal, de na briathra, agus de na haidiachtaí.
107  + Ansin, gineann an ríomhchlár
108  + an bunliosta focal <I>gaeilge.raw</I>.
109    Coinníonn an comhad seo:
110    <UL>
111  + <LI> na ceannfhocail le brait a léiríonn séimhiú, urú, srl.
112  + <LI> foirmeacha infhillte na n-ainmfhocal agus na n-aidiachtaí
113  + <LI> aimsir láithreach na mbriathra sa chéad phearsa uatha,
114  + le brat a léiríonn a réimniú
115  + <LI> aimsir ghnáthchaite na mbriathra sa chéad phearsa uatha,
116  + le brat a léiríonn a réimniú
117  + <LI> leaganacha d'fhocail a bhfuil ceannlitir ag a dtús, leis na
       réamhlitreacha "n", "h", "t",
```

- Caveat!
- Training data impacts LLMs
- Garbage in, garbage out
- Big tech companies, non-Irish-speaking researchers, don’t care
- I do! *But* forces me into a role I’m not qualified for, and don’t want

WIKIMEDIA LANGUAGE CONFERENCE
CELTIC KNOT ‘24

Quality issues on Irish Wikipedia

- Dominated by stub articles; 1-2 sentences, formulaic
- Majority of articles created by two users without good Irish
- Quite a few machine-translated articles, no post-editing at all
- More worrying: impact of the content translation tool (more below)
- Upshot: poor reputation among native/fluent speakers
- Less than 1% of articles deemed acceptable for Fiontar corpus

Content translation tool

- Fast, efficient way to create new articles via translation
- Draft using Google Translate, preserves links and citations
- User can post-edit the draft translation and post to Wikipedia
- But it can even lead good speakers of Irish astray…

Machine translation shibboleths

- Many sentences awkwardly structured like English source
- Problems with phrasal verbs (“look up”, “when it comes to…”, etc.)
- Polysemous words
    - “appearances” — *cumaí*
    - “execution (criminal)” — *forghníomhú*
    - “(research) contributions” — *ranníocaíochtaí*
    - “(public) figure” — *figiúr*
    - “he refined the method” — *rinne sé scagadh ar an modh*
    - “He claimed that…” — *d’éiligh sé*
    - “bow tie” — *comhscór bogha (!)*
    - “state (of matter)” — *stát*
    - “to criticize the government” — *léirmheas a dhéanamh ar an rialtas*
    - “values and mores” — *luachanna agus tuilleadh*

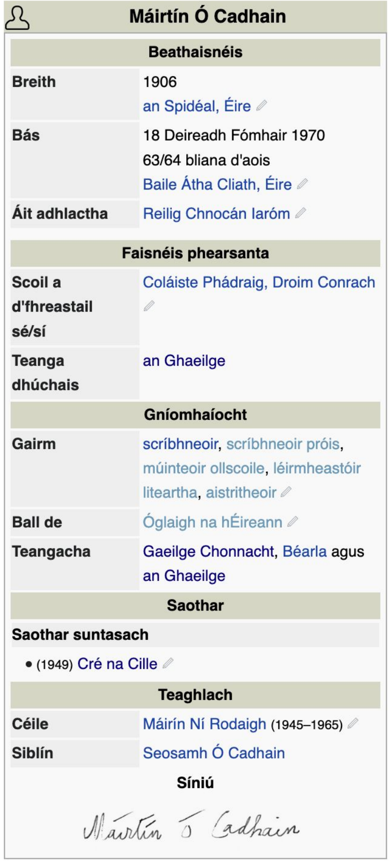
Next-generation grammar checking

- I wrote the existing Irish grammar checker in 2003 (!)
- It’s ok, but showing its age
- New approach uses LLMs and corrections mined from Wikipedia
- “Explainable” in the sense that corrections are tied to specific rules
- Not yet available publicly but have been testing on Wikipedia

Wikidata

Structured data for Wikipedia

- Wikidata contains items for all concepts corresponding to articles, and much more!
- Interlinking between items, via triples
- Q560494, P103, Q9142
- Irish Wikipedia makes heavy use of Wikidata via infoboxes as shown here



Máirtín Ó Cadhain

Beathaisnéis	
Breith	1906 an Spidéal, Éire
Bás	18 Deireadh Fómhair 1970 63/64 bliana d'aois Baile Átha Cliath, Éire
Áit adhlactha	Reilig Chnocán Iaróm
Faisnéis phearsanta	
Scoil a d'fhreastail sé/sí	Coláiste Phádraig, Droim Conrach
Teanga dhúchais	an Ghaeilge
Gníomhaíocht	
Gairm	scríbhneoir, scríbhneoir próis, múinteoir ollscoile, léirmheastóir liteartha, aistritheoir
Ball de	Óglaigh na hÉireann
Teangacha	Gaeilge Chonnacht, Béarla agus an Ghaeilge
Saothar	
Saothar suntasach	
• (1949) Cré na Cille	
Teaghlach	
Céile	Máirín Ní Rodaigh (1945–1965)
Siblín	Seosamh Ó Cadhain
Síniú	

Máirtín Ó Cadhain

WIKIMEDIA LANGUAGE CONFERENCE
CELTIC KNOT '24

LogainmBot

- Wikidata bot originally designed to link items to logainm.ie
- Similarly have added links to ainm.ie, and from species to tearma.ie
- Also adding Irish labels/descriptions to Wikidata on massive scale
- Close to 17 million additions since 2019; 4.5 million more queued
- Caution: new “multilingual” label being rolled out this month
    - Good: Q280420 is a galaxy with label “NGC 4314” in all Latin script languages
    - Bad: “Toormakeady”, “San Sebastián”, “Londonderry”, etc. and many personal names as well — potential for massive anglicization

Go raibh míle maith agaibh! (thanks)

WIKIMEDIA LANGUAGE CONFERENCE
CELTIC KNOT '24