20 Years a’ Growing: Past, Present, and Future of Irish NLP

Kevin Scannell
Saint Louis University
The Irish language

- First official language in the Republic of Ireland
- One of 24 official languages of the European Union
- About 75,000 daily speakers (outside of the education system)
- Only about 20,000 of these in the Gaeltacht areas
- I also work on the other Gaelic (Q-Celtic) languages:
  - Scottish Gaelic (58,000 total speakers)
  - Manx Gaelic (less than 2000 total speakers)
- Goal: support speakers in language revitalization efforts through technology
<table>
<thead>
<tr>
<th>Irish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiche bliain ag fás.</td>
<td>Twenty years growing.</td>
</tr>
<tr>
<td>Fiche bliain faoi bhláth.</td>
<td>Twenty years in bloom.</td>
</tr>
<tr>
<td>Fiche bliain ag cromadh.</td>
<td>Twenty years declining.</td>
</tr>
<tr>
<td>Fiche bliain gur cuma ann nó as.</td>
<td>Twenty years when it doesn’t matter whether you’re there or not.</td>
</tr>
</tbody>
</table>
Decline of Ireland’s native Irish speakers (1800-2000)

1800
1850 (Great Hunger)
1900
2000
Quick survey

- Spelling and grammar checkers
- Machine translation engines
- Dictionaries and thesauri
- Web-crawled corpora for 2500 languages
- Indigenous Tweets project
- Diacritic restoration
- Dependency parsing (Manx, Irish)
- Software localizations
- https://cadhan.com/
Outline of talk

- Case study of one particular problem I’ve worked on for 20+ years
- My solutions have evolved and improved with advances in the field
- Good illustration of what’s achievable and what’s not; and where “AI” helps
- I’ll conclude with some lessons I’ve learned through experience
Case Study: Grammatical error correction

- I’ll focus on a small subset of Irish grammar: correcting “initial mutations”

<table>
<thead>
<tr>
<th>Téacs le seiceáil:</th>
<th>Teanga an chomhéadain:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tá an bean sin anseo arís</td>
<td>Afracáinis (af) Mongóilis (mn)</td>
</tr>
<tr>
<td></td>
<td>Béarla (en_US) Ollainnis (nl)</td>
</tr>
<tr>
<td></td>
<td>Breathnais (cy) Rómánis (ro)</td>
</tr>
<tr>
<td></td>
<td>Dannhairgis (da) Sínis (zh_CN)</td>
</tr>
<tr>
<td></td>
<td>Esperanto (eo) Slóvaicis (sk)</td>
</tr>
<tr>
<td></td>
<td>Fionlainnis (fi) Spáinnis (es)</td>
</tr>
<tr>
<td></td>
<td>Francis (fr) Sualainnis (sv)</td>
</tr>
<tr>
<td></td>
<td>Gaeilge (ga) Ungáiris (hu)</td>
</tr>
<tr>
<td></td>
<td>Gearmainnis (de) Vítneamais (vi)</td>
</tr>
<tr>
<td></td>
<td>Indínéisis (id)</td>
</tr>
</tbody>
</table>

1: Tá an bean sin anseo arís
Séimhiú ar iarraidh
Celtic initial mutations

- Celtic languages have initial mutations usually triggered by context
- Today will focus on Irish, but approach works for Scottish Gaelic too
- \( \text{bád seoil} \) “sailboat”, \( \text{mo bhád seoil} \) “my sailboat”, \( \text{ár mbád seoil} \) “our sailboat”
- Gender: \( \text{fear} \) “man”, \( \text{an fear bocht} \) “the poor man”, but:
  - \( \text{bean} \) “woman”, \( \text{an bhéan bhocht} \) “the poor woman”
- Dative case: \( \text{ar an mbád seoil} \) “on the sailboat” (or, \( \text{ar an bhád seoil} \))
- Genitive plural: \( \text{leithreas na bhfear} \) toilet DET.GEN.PL men.GEN.PL “the men’s toilet”
A 20 year obsession

• Produce an algorithm that accurately predicts initial mutations in context
• Original motivation: context-sensitive spell checking and grammar checking
• More recently, this has proved useful in statistical language modeling
• Why is this hard?
  ○ In many cases, like the examples on the previous slide, it’s not!
  ○ But mutations do sometimes carry important information, so are hard or impossible to predict
  ○ Semantic: “Thug Kim a b(h?)eannacht do…”
  ○ Syntactic: “Bhí an bhean g(h?)nóthach sa bhaile”
  ○ There are “rules” as part of the written standard in Irish, but no one follows them completely
  ○ Some natural variation across dialects
  ○ Want to handle “real world” texts robustly; code-switching, pre- or non-standard spellings, etc.
A 20 year obsession

- Produce an algorithm that accurately predicts initial mutations in context
- Original motivation: context-sensitive spell checking and grammar checking
- More recently, this has proved useful in statistical language modeling
- Why is this hard?
  - In many cases, like the examples on the previous slide, it’s not!
  - But mutations do sometimes carry important information, so are hard or impossible to predict
  - Semantic: “Thug Kim a bheannacht do dhí-armáil núicléach iomlán leithinis na Cóiré”
  - Syntactic: “Bhí an bhean gnóthach sa bhaile”
  - There are “rules” as part of the written standard in Irish, but no one follows them completely
  - Some natural variation across dialects
  - Want to handle “real world” texts robustly; code-switching, pre- or non-standard spellings, etc.
Rule-based system

- This initial attempt was based on explicit rules
- Perform part-of-speech tagging, and then pattern-matching rules
- Exceptions, and exceptions to the exceptions, etc. (2814 rules in all)
- *Bhí Ó Baoill cúpla samhradh ag iascaireacht ar an bád
- Rules detect the error here, but just suggest some mutation
- *Chaith an sagart tamall ar an Mór-Roinn ina saighdiúir
- Error here is essentially impossible to encode this way

- “Resource-light”: gather statistics from untagged corpus to make predictions
- Need to hand-craft features to allow the model to make useful generalizations
- *snideog mór → snideog mhór*
- Target word and n-1 previous words
- Suffix of previous word
- First one or two letters of target word
- Suffix of target word
- Quickly run into issues of data sparsity
Parallel backoff

- If we’ve never seen a context before, *backoff* becomes critical
- Basically, simplify the context until it’s one you have seen before
- bhreathnaíonn an bean > ______ an bean
- But what about: ar an *crannstruchtúr*
- Two problems! Rare word, so simple n-gram backoff doesn’t help
- But even if we had seen “an crannstruchtúr”, that would give the wrong answer
- What we really want is to backoff to: ... ar an c________
- Similarly: *bhí sí ina uachtarán*
- “Generalized parallel backoff” (Bilmes and Kirchhoff, 2003)
Neural model: eliminates the hard parts of the statistical approach
LSTM layer(s), BiLSTM at character level
No need to hand-select features; no complicated backoff schemes
Achieves much higher accuracy than previous approaches
Character-based component learns gender other relevant features ("snideog")
Word-based component learns sometimes subtle contextual clues ("Ó Baoill")
Towards a Version 4

- Version 3 is trained only on plain text
- This leads to two major problems:
  - Complicated cases where some syntactic analysis seems to be required
  - Lack of **explainability** — critical for using these tools in educational contexts
- Mutations are often triggered by head/dependent relationships
- We use treebanks and parsers in the *Universal Dependencies* (UD) framework
Universal Dependencies


Be-PAST the day very-hot and be-PAST every person parched with the thirst
‘The day was very hot and everyone was parched with the thirst’
Celtic UD treebanks

- Scottish Gaelic: [Colin Batchelor](#) (2019)
- Manx Gaelic: [Scannell](#) (2020)
- Welsh: [Heinecke and Tyers](#) (2019)
- Breton: [Tyers and Ravishankar](#) (2018)
- Cornish: ????

Considerable effort has gone into harmonizing annotation schemes

Allows cross-linguistic comparison and transfer learning
Bhí sí naoi mbliana agus leathchéad.
Strategy for Version 4

- Use dependency parser to recognize contexts where mutations should occur
- Generate synthetic training examples by varying mutations in context
- Nuair nach mbíonn an bean ildánach seo...
- For version 3, the label on this example would indicate just the mutation (“S”)
- Explainability: “enrich” this tag set to include reference to a standard grammar
- Example above would be tagged “S/10.2.1.a”

10.2.1 InDaidh an Ailt

Cuírtear séimhiú ar an ainmfhocal i ndiaidh an ailt (mura d, t nó s an túschonsan)--
(a) san ainmneach uatha baininscneach, e.g., an chathair, an ghloine, an fhuascailt.

Tá an bhean ag canadh.
Ar dhún sé an fhuinneog?
Minority languages: Collecting “everything”

- The Crúbadán project (crubadan.org), c. 2000 - present
- Indigenous Tweets (indigenoustweets.com), 2011 - present
- RSS feeds
- Public Facebook posts
- Feedback loops + crawling
- Around 250 million words of Irish online, before cleaning
- Adding around 1 million words per month; will reach 1 billion circa...
- ...2083
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

- 250 million words of Irish might sound like a lot!
- Recent models for English have been trained on as many as 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
- May be impossible to achieve “human parity” results, ever
- 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 take tremendous care in selecting the training text for my models
- But most commercial systems are built with random text from the web
- For Irish, as much as 10% of the text in standard datasets is machine translated!
- Another 5-10% written by learners without a strong command of the language
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
- I suspect no one at the big tech companies is worrying about this
Long-term impact

● What impact will your digital work have in 50 or 100 years?
● The hard truth: no one will use any of your code, algorithms, or architectures
● Your data might survive and be useful in 100 years
  ○ Put it in the public domain or under a permissive license like CC-BY
  ○ Put many copies online, in standardized plain text format
  ○ Include your data in a “software pool” (Streiter et al 2007)
  ○ Incorporate into linked open data efforts, e.g. https://lod-cloud.net/
  ○ Document your data thoroughly, e.g. through a “data statement” (Bender and Friedman, 2018)
Thank you! / Go raibh maith agaibh!

- https://cs.slu.edu/~scannell/
- https://cadhan.com/
- https://github.com/kscanne/