Session 5: Digital Resources for the Medieval Gaelic World

Kevin Scannell
Saint Louis University

Overview

- I develop software that supports speakers of under-resourced languages
- Primarily Irish, Scottish and Manx Gaelic, but many others also
- End goal is to strengthen these languages in the computing domain
  - Software localizations
  - Proofing tools
  - Dictionaries and thesauri
  - Machine translation
  - See https://cadhan.com/
- There is a research component to this, but any linguistic or sociolinguistic insights are secondary to the engineering and resource-building

Natural Language Processing

- For 25+ years, the field has been dominated by machine learning approaches
- Through the 1990's and early 2000's, statistical/probabilistic techniques
- Around 2011-2012, neural networks/deep learning began to take over
- e.g. Google Translate switched from statistical to neural MT in 2016
- Like many other fields, we're caught up in "hype" around AI/Deep Learning

Neural Machine Translation Enabling Human Parity Innovations In the Cloud

Posted on ***June 17, 2019*** *by* ***Microsoft Translator***

In March 2018 we **announced** (**Hassan et al. 2018**) a breakthrough result where we showed for the first time a Machine Translation system that could perform as well as human translators (in a specific scenario – Chinese-English news translation). This was an exciting breakthrough in Machine Translation research, but the system we built for this project was a complex, heavyweight research system, incorporating multiple cutting-edge techniques. While we released the output of this system on several test sets, the system itself was not suitable for deployment in a real-time machine translation cloud API.

Today we are excited to announce the availability in production of our latest generation of neural Machine Translation models. These models incorporate most of the goodness of our research system and are now available by default when you use the Microsoft Translator API. These new models are available today in Chinese, German, French, Hindi, Italian, Spanish, Japanese, Korean, and Russian, from and to English. More languages are coming soon.

Human parity?

- One language pair
- Two *extremely* well-resourced languages
- Translation in one direction
- In a single domain
- Too big to deploy “live”
- Evaluated at sentence level

Téacs Doiciméid

BÉARLA - AIMSITHE GO HUATHOIBRÍOCH ÚCRÁ BÉARLA RÚISIS GAEILGE

Keep dogs on leads

18/5000

Coinnigh madraí ar luaidhe

Seol aiseolas

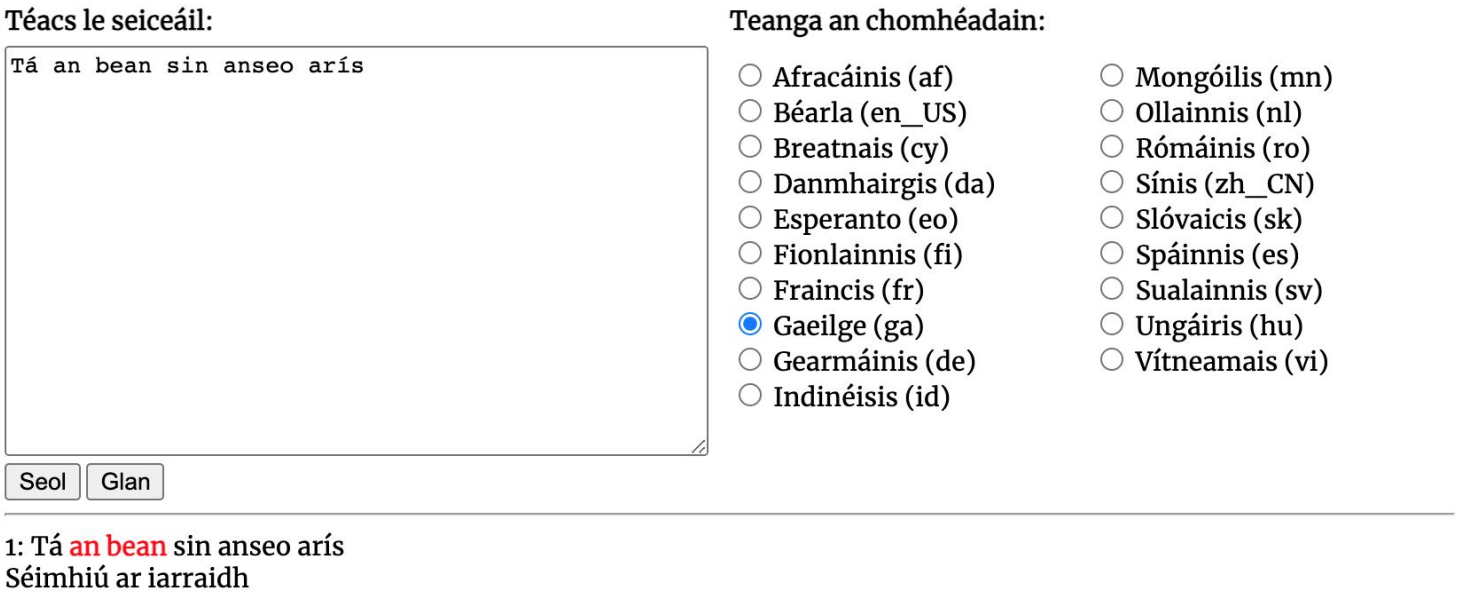
Image source: https://twitter.com/MiseAine/status/1158048916081905665

Three short case studies

- Two problems I've been working on for 15-20 years; one "new" project
- 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 takeaway lessons from these projects

Case Study #1: Grammatical error correction

- I'll focus on a small subset of Irish grammar: correcting initial mutations



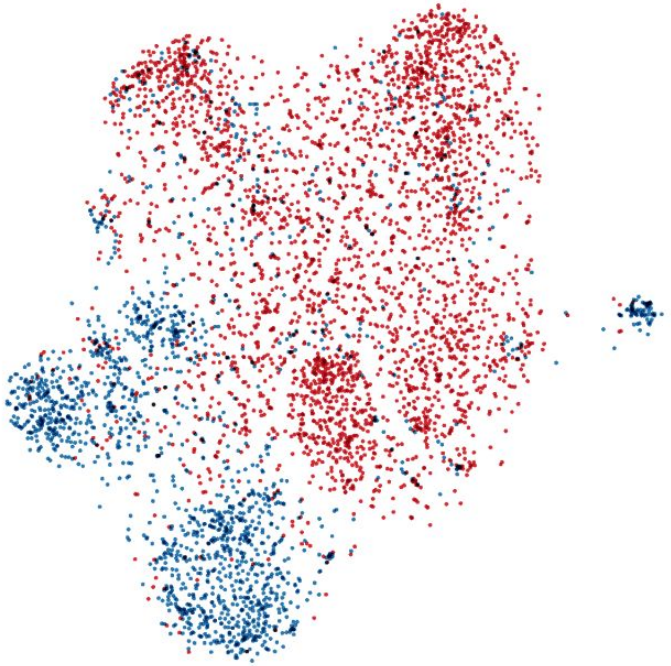
Case Study #1: Grammatical error correction

- My initial attempt (2002-2003) 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

Statistical approach

- “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`
- Promising! But still tricky to get right
- If we’ve never seen a context before, we use a trick known as “backoff”
- Basically, shorten the context until it’s one you have seen before
- But what about: `*bhí sé ar an crannstruchtúr`
- Likely to have seen a context like: `... ar an c_____`
- Similarly: `*bhí sí ina uachtarán`
- “Generalized parallel backoff” (Bilmes and Kirchhoff, 2003)

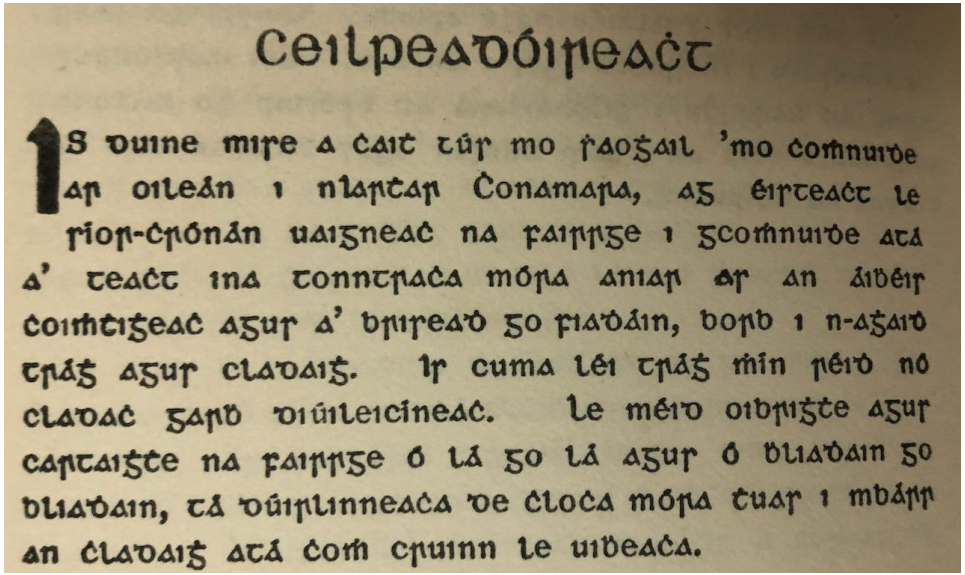
Neural network approach



- Eliminates the hard parts of the statistical approach
- 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”)

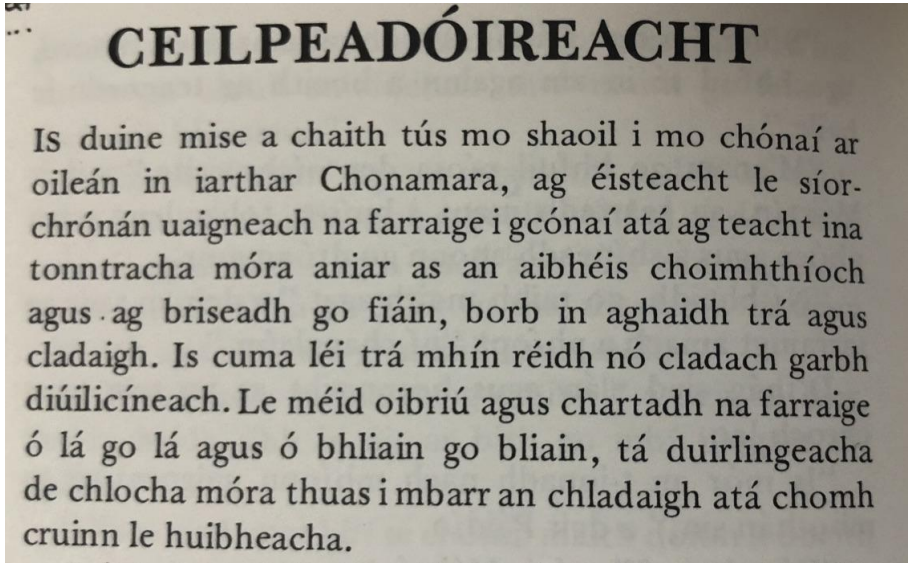
Case Study #2: Irish standardization

- Simple idea: a program the “translates” from pre-standard to standard Irish



Ceilpeadóireacht

Is duine mise a chaith tús mo shaoil i mo chónaí ar oileán in iarthar Chonamara, ag éisteacht le síor-chrónán uaigneach na farraige i gcónaí atá ag teacht ina tonntracha móra aniar as an aibhéis choimhthíoch agus ag briseadh go fiáin, borb in aghaidh trá agus cladaigh. Is cuma léi trá mhín réidh nó cladach garbh diúilicíneach. Le méid oibriú agus chartadh na farraige ó lá go lá agus ó bhliain go bliain, tá duirlingeacha de chlocha móra thuas i mbarr an chladaigh atá chomh cruinn le huibheacha.



CEILPEADÓIREACHT

Is duine mise a chaith tús mo shaoil i mo chónaí ar oileán in iarthar Chonamara, ag éisteacht le síor-chrónán uaigneach na farraige i gcónaí atá ag teacht ina tonntracha móra aniar as an aibhéis choimhthíoch agus ag briseadh go fiáin, borb in aghaidh trá agus cladaigh. Is cuma léi trá mhín réidh nó cladach garbh diúilicíneach. Le méid oibriú agus chartadh na farraige ó lá go lá agus ó bhliain go bliain, tá duirlingeacha de chlocha móra thuas i mbarr an chladaigh atá chomh cruinn le huibheacha.

Applications

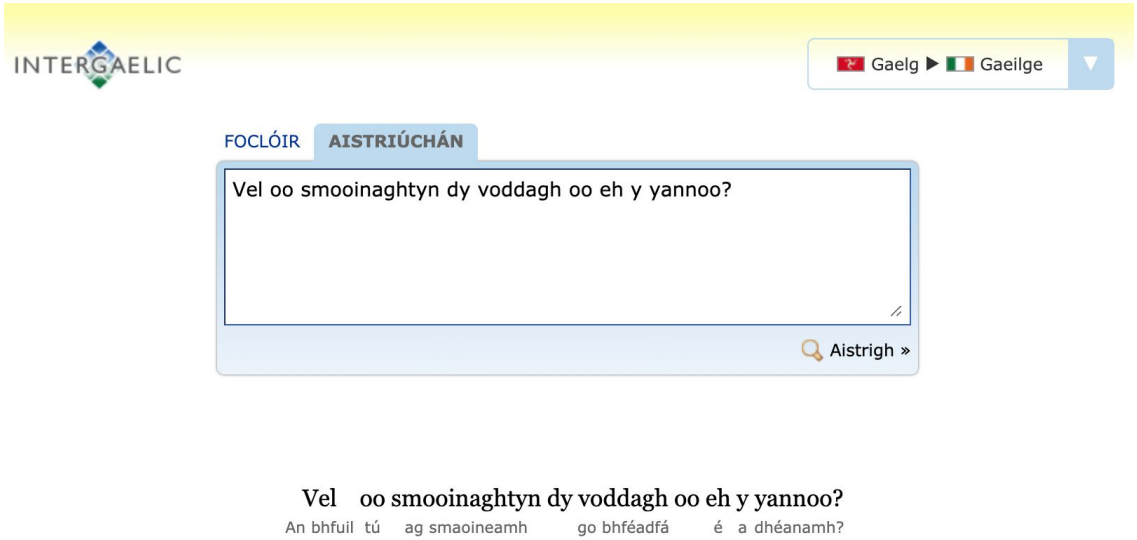
- `Ní rabh  sé de ghnáthas ag Buck na páipéir nuaidheachta a léigheamh.`
  `Ní raibh sé de ghnás    ag Buck na páipéir nuachta      a léamh.`
- Saves huge amount of manual work to edit older texts for modern readers
- More importantly, provides an easy way to “bridge” NLP tools for modern Irish to older forms of the language: POS tagging, dependency parsing
- Effective searching of pre-standard corpora by linguists and lexicographers

History

- Similar trajectory going back about 15 years
- First version was rule-based; database of pre-standard/standard word pairs
- 900+ spelling rules: `mhth→f, mhuint→úint, eóchamaoid→eoimid,` etc.
- Version 2.0 closer to traditional statistical MT, but spelling rules preserved
- Currently experimenting with neural network architectures, but very difficult to achieve results comparable to version 2.0!
    - Almost no “parallel text” between pre-standard and *strictly* standard Irish
    - Difficult to incorporate the resources we do have: great dictionaries, data on spelling changes

Case Study #3: Manx NLP

- A few years ago I developed machine translation engines and bilingual glossaries for the three Gaelic languages: http://www.intergaelic.com/



NLP from scratch

- A major obstacle in the case of Manx was the lack of any annotated corpora or NLP tools (part-of-speech tagger, parser, etc.)
- Fortunately the Manx language community has been energetic in digitizing and publishing texts online, and participating in social media
- Crawled a corpus of about 8 million words of Manx from the web
- Manually annotated a subset in Universal Dependencies format
- Just 6k words is enough for high quality POS tagging, “good” syntactic parsing
- Lays the foundation for future work on Manx corpus linguistics, lexicography

NLP in under-resourced settings

- The field is badly overfit to one Germanic language with almost no morphology
- Incorporate the resources you do have, even if new approaches required
- “Big data, small linguistics; Small data, big linguistics” — Fran Tyers
- May be impossible to achieve “human parity” results, ever
- Push performance as best we can, but then use tools with limitations in mind

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)