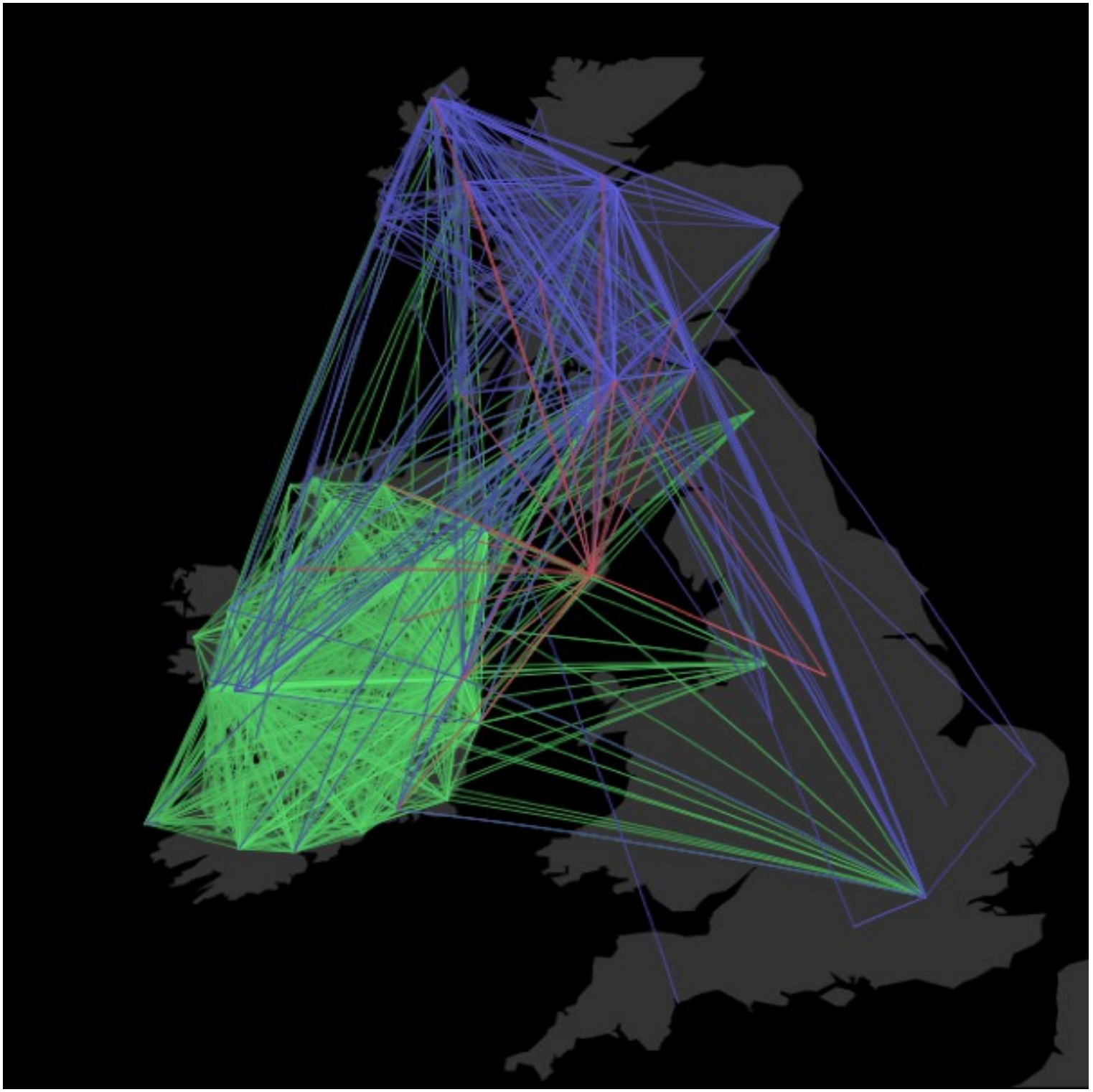
Scottish Gaelic language technologies based on web corpora

Kevin Scannell
Saint Louis University
4 June 2015



Why Machine Translation?

- One big community from 2 (or 3) small ones
- Break down communication barriers
- Open new markets for Gaelic writers
- Tools for learners from another Gaelic lang.
- “Transfer” Irish language tech to Gaelic
- Enables comparative linguistic study
- Rehabilitate a much-maligned technology!

Example 1

- gd: “Bha e fhéin 'na sheasamh a-measg a' bhuntàta an uair a chunnaic e iad le 'n gunnachan...”

Example 1

- gd: “Bha e fhéin 'na sheasamh a-measg a' bhuntàta an uair a chunnaic e iad le 'n gunnachan...”
- ga: “Bhí sé féin ina sheasamh i measc na bprátaí nuair a chonaic sé iad lena gcuid gunnaí...”

Example 2

- gd: “Chunnacas fo sgàil craobh na dòrainn
  a' coiseachd sràidean Pharais gu lòghmhor na
  seann siùrsaichean beaga breòite a chunnaic
  Baudelaire 'na ònrachd.”

Example 2

- gd: “Chunnacas fo sgàil craobh na dòrainn a' coiseachd sràidean Pharais gu lòghmhor na seann siùrsaichean beaga breòite a chunnaic Baudelaire 'na ònrachd.”
- ga: “Chonacthas faoi scáth chrann an doilíosa ag siúl sráideanna Pháras go soilseach na striapaigh aosta bheaga bhuailte a chonaic Baudelaire ina uaigneas”
- Somhairle MacGill-Eain, aistr. Paddy Bushe

Parallel corpus

- Aligned Irish-Scottish Gaelic texts
- A little bit of everything!
- Software translations, Bible texts, tweets
- Wikipedia articles, poems, prayers, ...
- 130k segments, ~1M words on each side
- Unusual in that very little is direct translation

Statistical Machine Translation

- We extract translation pairs from parallel corpus
- Even easier when pairs are often cognates
- Model incorporates rule-based spelling changes
- sg- → sc, -chd- → -cht-, etc.
- Each SG word then maps to (usu. many) IG words

How to translate?

- Work through the source sentence left-to-right
- Incorporate limited reordering via a “phrase table”
- e.g. “mun cuairt oirnn” → “inár dtimpeall”
- Each word/phrase has multiple possible translations
- Hence each sentence has *many* possible translations
- 20 word sentence, ~2 translations/word => ~1000000!
- Goal is to find the *most probable* translation
- Prune the possibilities via a “Markov model”

Guessing Game, I

- Probability of word depends only on prev two
- “the two ___”
- P(men|the two) = 0.0413
- P(of|the two) = 0.0338
- P(countries|the two) = 0.0298
- P(sides|the two) = 0.0204
- P(groups|the two) = 0.0164

Guessing Game, II

- “the fact _____”
- P(that|the fact) = 0.8698
- P(is|the fact) = 0.0312
- P(of|the fact) = 0.0241
- P(remains|the fact) = 0.0092
- P(was|the fact) = 0.0050
- P(they|the fact) = 0.0043

Guessing Game, III

- “the united ___”
- P(states|the united) = 0.5240
- P(kingdom|the united) = 0.3129
- P(nations|the united) = 0.0859
- P(arab|the united) = 0.0075
- P(front|the united) = 0.0061
- P(democratic|the united) = 0.0024

Guessing Game, IV

- “button fell ____”
- Doesn't appear at all in a 100M word corpus
- “Backoff smoothing”
- Estimate P(w|button fell) using P(w|fell)
- Or get a bigger corpus!

Web Corpora

- No linguistic analysis of source lang needed
- No statistics for source language either!
- Entirely driven by statistics of target (Irish)
- Of which there is A LOT online
- About 150 million words of Irish in total
- Some care needed, but basically more=better
- Crúbadán project supported by NSF grant 1159174

Disambiguation I: Lexical Gaps

- “ach coiseachd an iar tron Mhunadh Gheal”
- “am biodh a' ghaoth an iar leotha...”
- “air a' chosta an iar...”
- Give “siar”, “aniar”, “thiar” respectively

Disambiguation II: Function Words

- gd: Bhitheamaid gun sgillinn ruadh nan dèanamaid sin
- “nan” -> na, ina, or dá
- ga: Bheimis gan pingin rua dá ndéanfaimis sin

Disambiguation III: Initial mutations

- gd #1: “Chroch sibh an radan”
- ga: “{Chroch,gCroch} {sibh,tú} {a,an,in} {francach,fhrancach,bhfrancach}”
- gd #2: “Beir air an radan”
- ga: “{Beir,mBeir} {air,ar} {a,an,in} {francach,fhrancach,bhfrancach}”
- 2×2×3×3=36 possible translations in each case
- Best for #1: “Chroch sibh an francach” (or “tú”?)
- Best for #2: “Beir ar an bhfrancach”

Deliverables

- Gàidhlig-Gaeilge dictionary: 16107 headwords
- Gàidhlig twitter stream for Irish speakers
- InterGaelic.com, with Michal Boleslav Měchura
- Integrated into Clilstore and Multidict
- Open source code and data for translator
- Translations via a web service
- Coming soon... Manx → Irish

Ar Ghuaillí na bhFathach

- Michael Bauer
- Caoimhín Ó Donnaíle
- Donncha King
- Ciarán Ó Duibhín
- Ciarán Dunbar
- Phil Kelly
- Michal Boleslav Měchura