Statistical models for text normalization and MT

Kevin Scannell
Saint Louis University
23 August 2014

Goal

- Describe two systems
- Shared statistical model and codebase
- http://github.com/kscanne/caighdean/
- Free software, GPLv3+
- Irish standardizer “An Caighdeánaitheoir”
- Scots Gaelic to Irish MT “gd2ga”

An Caighdeán Oifigiúil

- The Official Standard
- Introduced in the 1940's and 1950's
- Simplified spelling and grammar
- Widely adopted, all domains and registers

Example

- Old: “Ní rabh 'sa dearbhughadh sin acht a chuid uchtaighe, eisean, a h-Aodh féin ag teacht na h-arraicis.”
- New: Ní raibh sa dearbhú sin ach a chuid uchtaí, eisean, a hAodh féin ag teacht ina haraicis.

The Problem

- Searching the web
- Searching corpus texts for lexicography
- Language modeling
- Parallel corpora
- New literary editions for modern readership

A Solution

- Treat as an MT problem
- *Very* closely-related languages
- IBM model 1
- Some reordering via a phrase table (gd2ga mostly)
- e.g. “mun cuairt oirnn” → “inár dtimpeall”
- “Translation model” based on lexicography
- Allows rule-based spelling changes
- sg- → sc, -chd- → -cht-, etc.

Difficulty: target language model

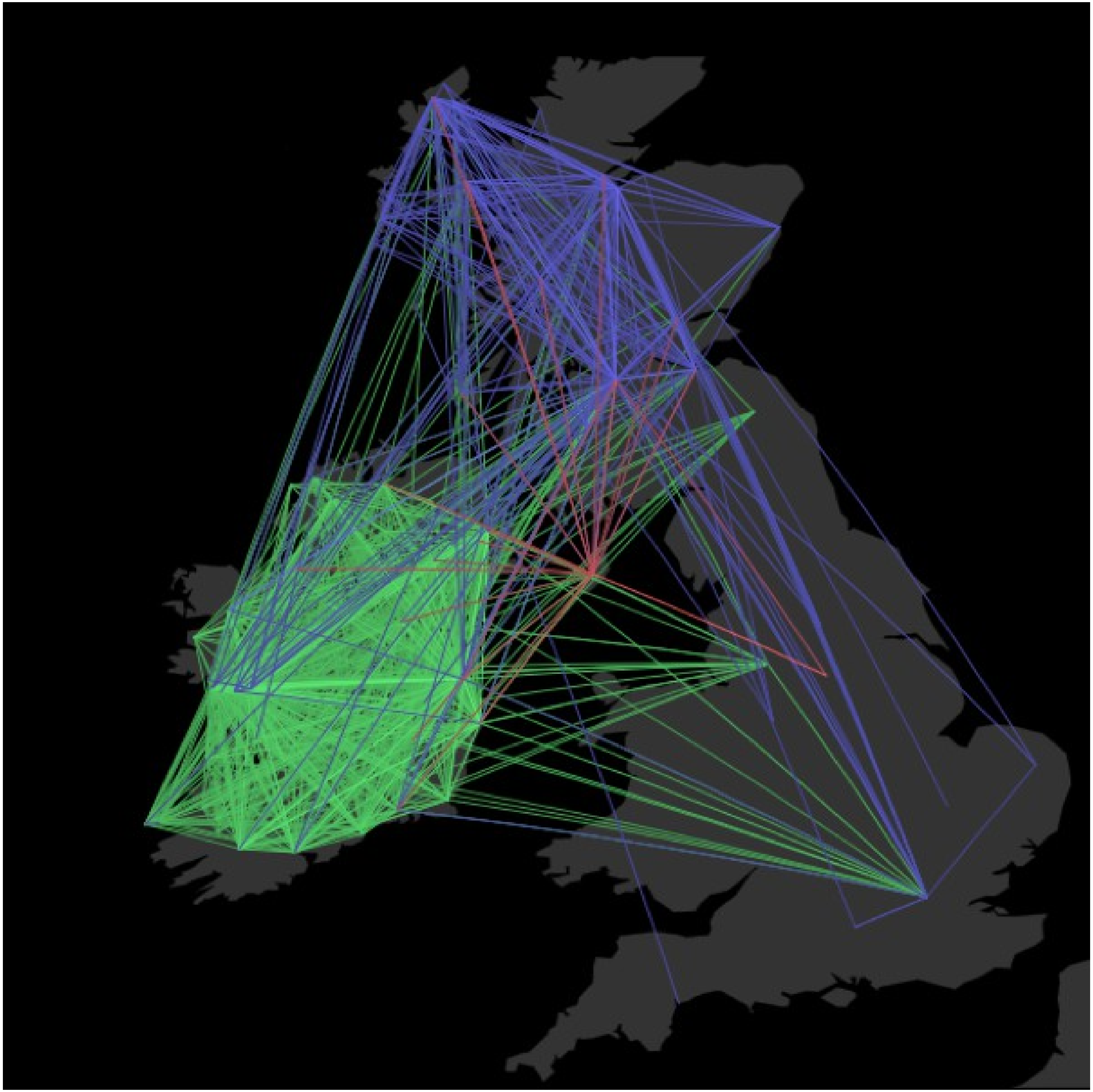
- “Cad is brí le 'caighdeánach' san aois iar-nua-aoiseach seo?”
- Target in both systems is “standard Irish”
- Built 3-gram model from 100M word corpus
- Books, newspapers, web texts
- But, almost *no* texts conform completely
- Used grammar checker to pare down to 50M
- Manual replacement of common constructions

Evaluation

- Parallel corpus: 47k segments, 2 x 800k words
- Texts from RIA and An Gúm
- WER on a 200 sentence test set: 9.86%
- Baseline WER (leave unchanged!): 27.28%

Scots Gaelic and Irish

- Shared history, literary tradition
- Pre-stnd Irish, SG orthographies similar
- Many shared grammatical constructions
- Same statistical model is effective
- Useful for xfer of Irish NLP resources to SG
- Do humans need it?



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

- Used to extract (many) translation pairs
- All pairs validated manually
- Used in evaluation too
- A little bit of everything!
- Software translation, Bible texts, tweets
- Wikipedia articles, poems, prayers, ...
- 130k segments, ~1M words on each side

Bilingual Lexicon

- 14000 headwords, manually constructed
- 96.7% coverage on running Scots Gaelic text
- Default treats source word as candidate
- Many “false friends” fall out

Statistical disambiguation

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

Evaluation

- Most parallel corpus texts translated from en
- Translated 593 sentences from SG to Irish
- WER on this eval set: 37.40%
- Baseline system: 88.09%
- Still a bit “unfair”
- Initial mutations
- “Tha mi a' tuigsinn a-nis” vs. “Tuigim anois”

Go raibh míle maith agaibh!

- Michael Bauer
- Caoimhín Ó Donnaíle
- Donncha King
- Ciarán Ó Duibhín
- Ruairí Ó hUiginn
- Máire Nic Mhaoláin
- Elaine Uí Dhonnchadha
- Brian Ó Raghallaigh