Diachronic Parsing of Pre-Standard Irish

Kevin Scannell
20 June 2022

Outline

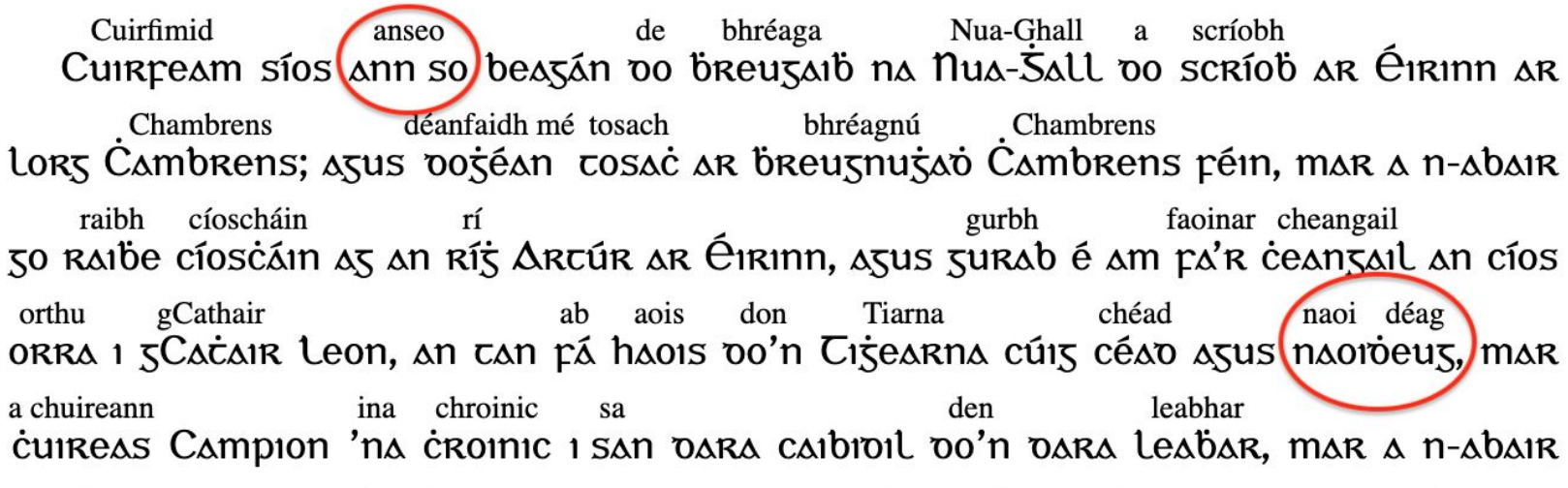
- Parsing and tagging experiments on historical Irish texts
- New repository for Irish NLP datasets and benchmarks

Standard Irish

- Official standard introduced in the 1940's and 1950's
- Significant simplifications to spelling and grammar
- NLP tools developed for modern language struggle with older texts
- Foclóir Stairiúil na Gaeilge (1600–)
- http://corpas.ria.ie/ — 3000 texts published between 1600 and 1926

Standardization

- I developed a tool for standardizing Irish texts, c. 2007 (paper @ 1st CLTW!)
- Shallow statistical MT approach; *does no annotation of pre-standard text*



Traditional processing pipeline

- Run an older text through the standardizer, outputs word-level alignments
- Tag/parse the standardized text using tools for the modern language
- “Project” the annotations back to the original text
  - One-to-many standardization (“naoidheug”): adjust tokenization of source text
  - Many-to-one standardization (“ann so”): DB of 750 most common examples + annotations
- Essentially the pipeline used for the corpas.ria.ie site
- ***But how well does this work?***

Test corpus of pre-standard Irish texts

- 150 sentences, just under 4000 tokens
- 25 sentences from three 20th c. books, one per major dialect: “Older” corpus
- 25 sentences from three very challenging texts: “Oldest” corpus
  - 1602 Irish New Testament
  - Foras Feasa ar Éirinn (1630s)
  - Cín Lae Amhlaoibh (1820s)
- Manually tagged/parsed following the Universal Dependencies guidelines
- https://github.com/UniversalDependencies/UD_Irish-Cadhan/blob/dev/ga_cadhan-ud-test.conllu

Experiments: lemmatization, tagging, and parsing

	— Standard —					— Older —					— Oldest —				
Model	Lem	POS	Feat	UAS	LAS	Lem	POS	Feat	UAS	LAS	Lem	POS	Feat	UAS	LAS
UD	95.8	94.4	82.1	81.8	74.5	80.8	85.2	74.4	77.6	67.4	63.8	72.3	56.4	61.2	46.8
Projecting	95.0	94.3	81.3	81.1	74.0	**97.9**	**96.4**	**89.8**	**84.8**	**77.3**	**89.4**	**89.7**	**77.5**	**73.0**	**63.1**
Silver	90.8	91.0	76.0	74.9	67.4	95.3	94.8	86.8	84.0	75.6	85.1	86.7	72.3	70.6	60.6
UD+100%	94.6	94.8	**83.9**	80.6	74.4	95.3	94.8	86.6	84.0	75.6	85.0	86.8	72.6	71.8	61.7
"+MUSE	94.6	94.8	**83.9**	**82.0**	**75.5**	95.3	94.8	86.6	84.4	76.4	85.0	86.8	72.6	71.8	61.4
UD+25%	95.3	94.7	83.4	81.8	75.0	92.2	93.3	84.2	81.4	72.9	80.0	83.9	68.5	70.4	58.7
UD+Lex	**95.9**	**94.9**	83.6	81.7	75.0	92.4	92.6	81.4	80.0	71.3	81.2	84.0	65.1	68.6	56.1

Observations

- Modern taggers/parsers perform poorly on older texts
- Traditional pipeline using the standardizer gives the best results
- But, promising results w/o gold training and w/o using the standardizer directly
- With a large enough training corpus, can we eliminate the standardizer?
- Then, tag/parse older texts directly, and use that to write a better standardizer!

NLP evaluation

- We all know evaluation with standard test sets is important in this field!
- https://paperswithcode.com/area/natural-language-processing
- http://nlpprogress.com/
- https://huggingface.co/datasets
- But....

“Leaderboard Culture”

- Papers publishable if and only if they achieve SOTA on some standard benchmark
- Rewards teams with bigger datasets, more GPUs, better hyperparameter searching
- If you plug your new pre-trained model into an existing algorithm, is that interesting?
- If you improve SOTA by tweaking parameters, is that interesting research?
- If you improve SOTA but your code is too slow for applications, is that a good thing?
- If you improve SOTA but pump tons of CO_2 into the atmosphere, is that a good thing?
- If you improve SOTA but your model contains harmful biases, is that a good thing?

25 years a’ growing

- I’ve been working on Irish NLP since 1997
- I am good as an open-source developer, terrible as an academic
- I have lots of code and datasets for various NLP tasks, very few papers
- Goal: jump-start research on many of these tasks by publishing datasets
- But do so in a way that transcends “leaderboard culture”
- Also, support non-researchers with simple baseline implementations
- e.g. Irish dialect identification

GBB: Giorraíonn BERT Bóthar

- New repository of benchmarks and datasets for Irish NLP
- https://github.com/kscanne/gbb/
- Brings together in one place datasets I've built over the last 20+ years
- Each task comes with one or more baseline Python implementations
- Useful as starting points for research, but also for application developers
- Eventually (work in progress) will cover the following 25 tasks:
  - Author Identification, Bilingual Lexicon Induction, Chunking, Code-switching Detection, Constituency Parsing, Conversational Agents, Dependency Parsing, Diacritic Restoration, Dialect Classification, Gender Identification, Grammar Checking, Irish-English Machine Translation, Irish-Manx Gaelic Machine Translation, Irish-Scottish Gaelic Machine Translation, Irish Standardization, Language Modeling, Lemmatization, Named Entity Recognition, Native speaker vs. Learner classification, OCR Correction, Part-of-Speech Tagging, Prediction of Initial Mutations, Question Answering, Sentiment Analysis, Topic and Genre Classification

duchas.ie
@duchas_ie

Irish Proverb

Giorraíonn beirt bóthar

(Two shorten the road, ie, a journey seems shorter when travelling with someone)

Photo: An Cheathrú Rua, Galway c.1930

Collaborative not competitive

- Move away from researchers **competing** for top spot on leaderboard
- Instead: language community **collaborating** on building the best tools possible
- This should look more like open source development, less like “research”
- New implementations (or improvements to existing) are made via pull requests
- Separate this from academic publications on the subject
- Implementation is still citable; all non-trivial contributors included
- Example: diacritic restoration

Summary

- New centralized repository for Irish NLP evaluations
- Benchmarks and datasets for 25 tasks, several released for the first time
- Strong baseline implementations, all available as open source software
- Collaborative (not competitive) leaderboards

Go raibh míle maith agaibh!