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/](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
Experiments: lemmatization, tagging, and parsing

<table>
<thead>
<tr>
<th>Model</th>
<th>— Standard —</th>
<th>— Older —</th>
<th>— Oldest —</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lem</td>
<td>POS</td>
<td>Feat</td>
</tr>
<tr>
<td>UD</td>
<td>95.8</td>
<td>94.4</td>
<td>82.1</td>
</tr>
<tr>
<td>Projecting</td>
<td>95.0</td>
<td>94.3</td>
<td>81.3</td>
</tr>
<tr>
<td>Silver</td>
<td>90.8</td>
<td>91.0</td>
<td>76.0</td>
</tr>
<tr>
<td>UD+100%</td>
<td>94.6</td>
<td>94.8</td>
<td><strong>83.9</strong></td>
</tr>
<tr>
<td>”+MUSE</td>
<td>94.6</td>
<td>94.8</td>
<td><strong>83.9</strong></td>
</tr>
<tr>
<td>UD+25%</td>
<td>95.3</td>
<td>94.7</td>
<td>83.4</td>
</tr>
<tr>
<td>UD+Lex</td>
<td><strong>95.9</strong></td>
<td><strong>94.9</strong></td>
<td>83.6</td>
</tr>
</tbody>
</table>
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](https://paperswithcode.com/area/natural-language-processing)
- [http://nlpprogress.com/](http://nlpprogress.com/)
- [https://huggingface.co/datasets](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:
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!