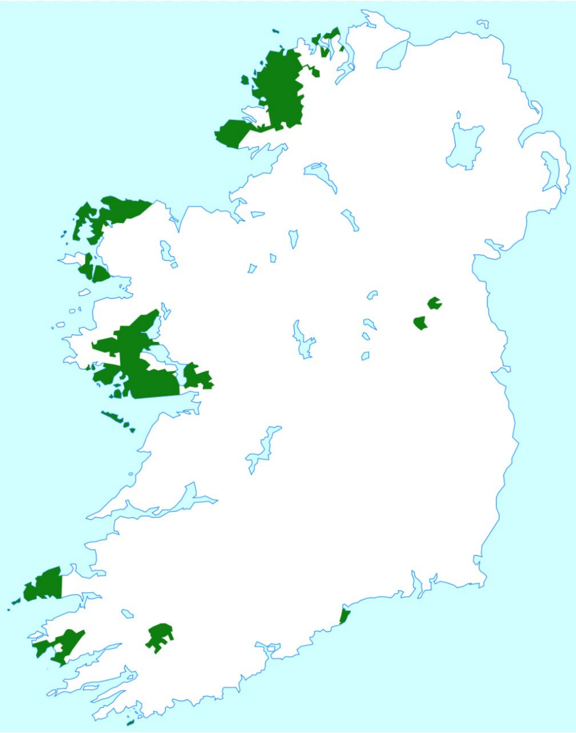
Neural language technology in an under-resourced setting

Kevin Scannell
Saint Louis University
September 26, 2019

Irish language



Irish Language Technology

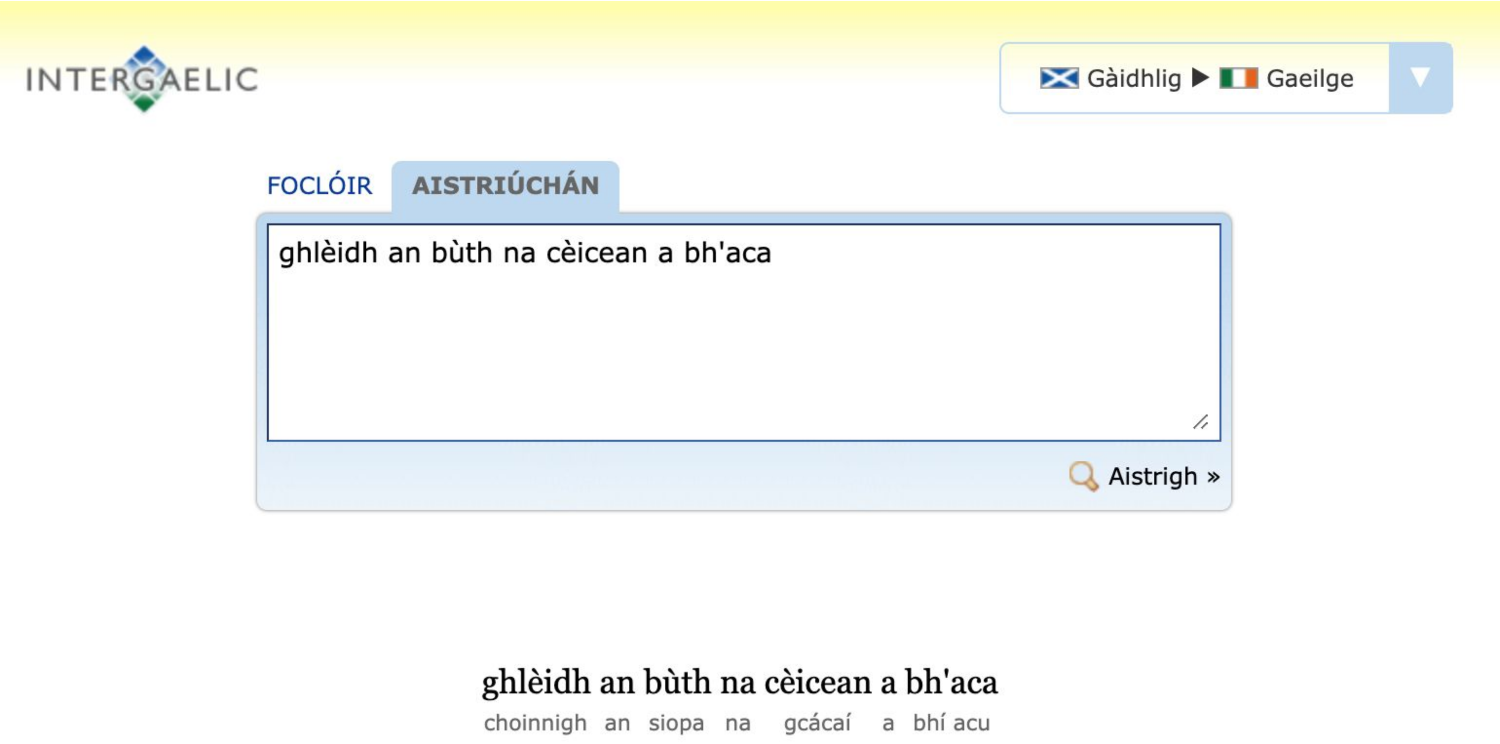
- Spelling and grammar checkers
- Dictionaries, semantic network
- Machine translation engines
- Part-of-speech taggers
- Dependency parser
- Speech synthesis
- Standardization tool
- Plain text corpora (> 200M words)
- Parallel corpora

“Praistriúchán”

New Irish portmanteau word:

“praiseach” = “a mess, a botch job”, “aistriúchán” = “translation”

intergaelic.com



Language modeling

- A language model (LM) is a probability distribution over sequences of words
- If S = “colorless green ideas…”, a language model assigns this a prob P(S):
- P(S) = P(colorless|^) P(green|colorless) P(ideas|colorless green) …
- Usually formulated and computed this way (word prob given history)
- LMs capture a lot! Pragmatics, syntax, real-world knowledge, …
- P(Friday|My party is this coming) > P(Tuesday|My party is this coming)
- P(is|The man with the glasses) > P(are|The man with the glasses)

Applications

- Almost all important language technologies use LMs at some level!
- Can be used generatively
- MT, ASR, etc. fundamentally generate text, conditioned on input
- Conversational agents (Turing test)
- Traditional probabilistic models ("noisy channel"): $P(T \mid S) = P(S \mid T)\, P(T) / P(S)$
- Strong LM alone can do question answering, summarization, … (GPT-2) !
- Better language models give better end-to-end performance, generally

Intrinsic and extrinsic evaluation

- Extrinsic: incorporate in language tech and evaluate end-to-end
- Intrinsic: what probability does the model assign to a big test corpus?
- $S = w_1 w_2 w_3 \ldots w_N$ where N is in the millions or more
- Average log-prob of over words (units are bits/word)

$$[- \sum \log_2 P(w_j \mid w_1 w_2 w_3 \ldots w_{j-1})] / N$$

- Approximation of cross-entropy between language and the model
- Best non-neural "n-gram" models for English just over 6 bits/word
- State-of-the-art neural models for English now *under* 4 bits/word

Neural language models

- A flood of recent papers on neural language modeling, big leaps forward
- Originally, feed-forward neural networks (Bengio et al, 2003)
- Various refinements + regularization of recurrent networks (LSTMs, etc.)
- Most recently the Transformer architecture (Vaswani et al, 2017)
- OpenAI's recently announced GPT-2 for English
- "...concerns about [the model] being used to generate deceptive, biased, or abusive language at scale"

Rank	Method	Test perplexity	Validation perplexity	Number of params	Extra Training Data	Paper Title	Year	Paper	Code
1	**Megatron-LM**	10.8		8300M	✓	Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism	2019		
2	**Transformer-XL + RMS dynamic eval**	16.4	15.8	257M	✗	Dynamic Evaluation of Transformer Language Models	2019		
3	**Transformer-XL + SGD dynamic eval**	17.0	16.3	257M	✗	Dynamic Evaluation of Transformer Language Models	2019		
4	**GPT-2 Full**	17.48		1542M	✓	Language Models are Unsupervised Multitask Learners	2019		
5	**Transformer-XL Large**	18.3	18.2	257M	✗	Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context	2019		
6	**Transformer + Adaptive inputs**	18.70	17.97	247M	✗	Adaptive Input Representations for Neural Language Modeling	2018		
7	**All-attention network - 36 layers**	20.6	19.7	133M	✗	Augmenting Self-attention with Persistent Memory	2019		

Research on English != Research on Language

- 8 tables tracking SOTA for language modeling; English datasets only
- Research almost 100% (and implicitly!) focused on English
- The word “English” isn’t used even once in these groundbreaking papers:
    - Google Brain’s landmark 2016 paper “Exploring the limits of language modeling”
    - Melis et al’s “On the state of the art of evaluation in neural language models” (2017)
    - Dai et al’s “Transformer-XL” paper (2019)
    - New SOTA “Megatron-LM” paper (up on arXiv Sept 17th!)
- “Bender Rule”
- SOTA neural models applied to many other languages actually perform worse

Celtic initial mutations

- Celtic languages have initial mutations usually triggered by context
- *bád seoil* “sailboat”, *mo b**h**ád seoil* “my sailboat”, *ár **m**bád seoil* “our sailboat”
- Gender: *fear* “man”, *an fear bocht* “the poor man”, but:
- *bean* “woman”, *an b**h**ean b**h**ocht* “the poor woman”
- Dative case: *ar an **m**bád seoil* “on the sailboat” (or, *ar an b**h**ád seoil*)
- Genitive plural:

leithreas	*na*	***bh**fear*
toilet	DET.GEN.PL	men.GEN.PL

  “the men’s toilet”
- Dozens, maybe hundreds of rules that no one knows or uses completely

Motivating examples

- This was (one of) Google’s mistakes in the earlier image:
  *_tríd an bóthar_ → _tríd an **m**bóthar_
  through the road
- And Intergaelic too, tricked by VSO:
  *_choinnigh an siopa na **g**cácaí a bhí acu_
  kept the shop the cakes that were at-them
  “the shop kept their cakes”
  (cf. _siopa na gcácaí_ “the shop of the cakes”, “the cake shop”)

Factored language models

- Word-based LMs don't see that *bád, bhád, mbád* are really the same word
- Since "bád" is most common, harder to predict collocations like "bhád seoil"
- Well-known issue in LMs for morphologically complex languages
- Standard solution: factored language models (Bilmes and Kirchhoff, 2003)
- View each word w_t as a bundle of features $f^1_t, \ldots, f^k_t$
- Factor P(w) as a product of feature probabilities conditioned on earlier features
- For mutations, e.g., P(*bhád* | ... *mo*) = P(*bád* | ... *mo*) P(**lenition** | ... mo bád)

Mutations as low-entropy features

- Celtic mutations carry very little information
- Usually determined by the previous two words and initial letter of target word
- Could remove them and one can almost always replace them unambiguously
- Using our language modeling framework we can assign a number to this!
- “Average number of bits per word carried by mutations” (claiming it’s small)
- Five mutations: **none, lenition, eclipsis, t-prothesis, h-prothesis**
- Build a model that predicts P(mutation | word history) as in the factored model
- Compute the $\log_2$ loss of this model on a test set

Which mutations carry information? (part one)

- 3rd person possessive “a”

a bád	*a b**h**ád*	*a **m**bád*
her boat	his boat	their boat

- Certain set phrases

Tá sé ar siúl	*Tá sé ar s**h**iúl*
“It is underway”	“He is away”

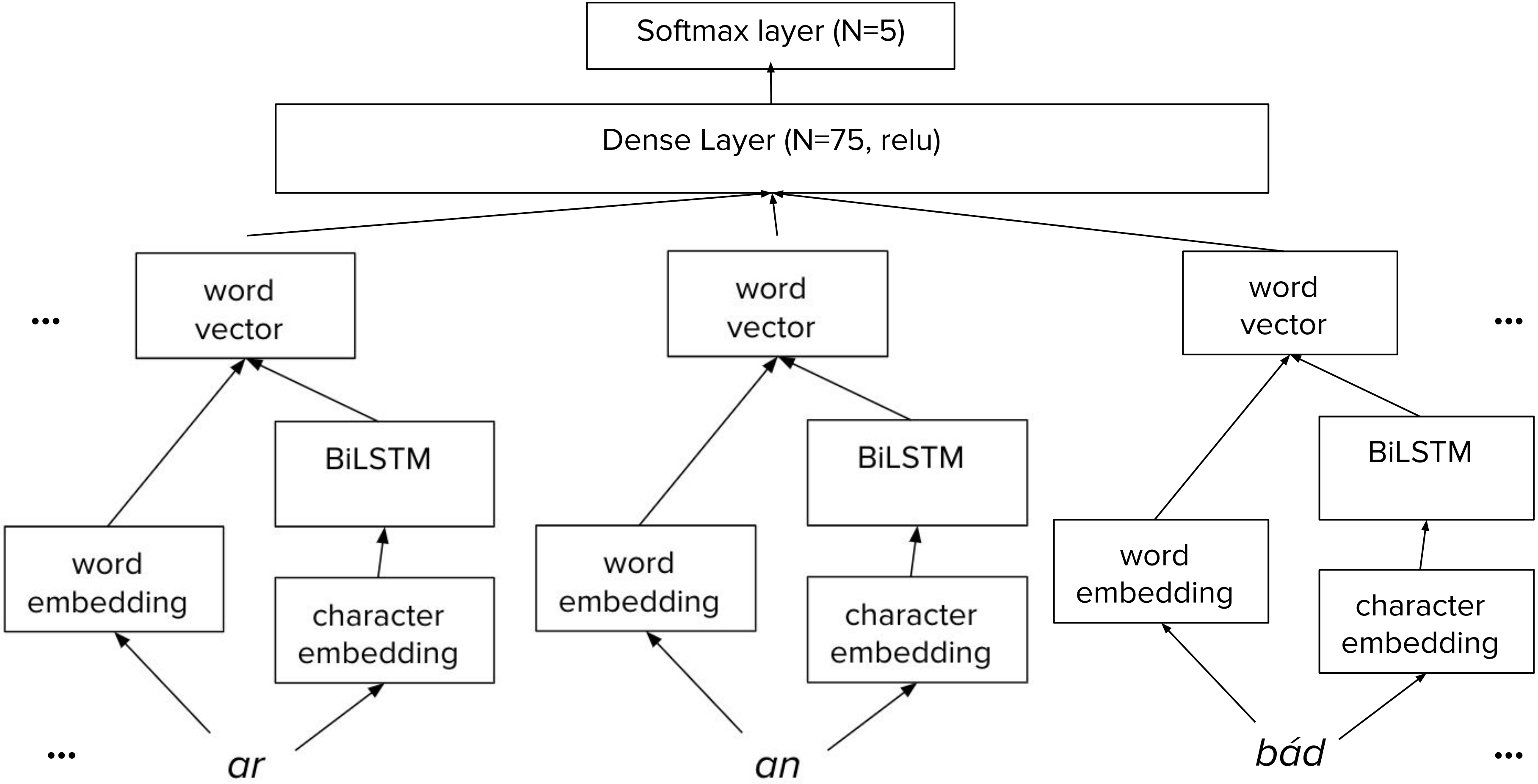
- Occasional syntactic bad luck

Tá an bhean ghnóthach ina hoifig	*Tá an bhean gnóthach ina hoifig*
“The busy woman is in her office”	“The woman is busy in her office”

- Tense: copula *gur* triggers lenition only in past tense
- Dialect: “ar an mbád” (Connacht, Munster) vs “ar an bhád” (Ulster)

Digression: orthographic transparency

- Four of the five mutations in Irish can be trivially and algorithmically removed
- h-prothesis cannot, in general: (***h****amhlaidh* vs. *hidrigin*)
- Even with a dictionary, some ambiguity: *aiste* “essay” vs. *haiste* “hatch”
- I strip all h’s and let the neural networks figure it out!
- Note this introduces issues with English, too: (h)all, (h)airline, (h)and, etc.
- Scottish Gaelic is transparent in all cases (they write h-)
- Welsh, Cornish, Breton, and Manx Gaelic are not at all transparent!



Softmax layer (N=5)
Dense Layer (N=75, relu)
...
word vector
word vector
word vector
...
BiLSTM
BiLSTM
BiLSTM
word embedding
character embedding
word embedding
character embedding
word embedding
character embedding
...
ar
an
bád
...

Results

- 2.32193 ($\log_2 5$) bits/word for random labels
- 0.78936 bits/word using label prior probabilities
- 0.50388 bits/word using unigram model (label distribution per word)
- **0.05619** bits/word: NN trained on 40M words, 40k vocabulary, 15 epochs

Applications

- Improved LM for Irish when used in a factored model on demutated words
- Hope to show end-to-end improvement on machine translation engines (WIP)
- Data-driven grammar checking which robustly handles variant spellings, etc.
- Sociolinguistics: wild divergence between official standard(s) and actual usage

Which mutations carry information? (part two)

- Data-driven answer to the question above
- Of 10000 examples I checked, correct label was assigned P<0.5 184 times
- These 184 examples contribute 72% of the total loss
- 58 are usage errors in the test file including the top 9 producing largest loss
- 37 relate to the third person possessive in one form or another
- 16 are dialect differences
- 10 were assigned low prob only because of lack of context to the right
- 8 relate to difference between past tense and imperative verbs
- 8 relate to two versions of relativizing particle “a” (one lenites, one eclipses)
- Various assorted others

Gender bias

- *tá sé/sí* *ina* *mhúinteoir/múinteoir*
  is he/she in-his/her teacher
  “he/she is a teacher”
- male bias in corpus: *cathaoirleach* (chairperson), *ceannaire* (leader), *traenálaí* (trainer), *gobharnóir* (governor), *oifigeach* (officer), *aire* (government minister)
- female bias in corpus: *déagóir* (teenager), *girseach* (girl), *cailleach* (witch), *dornálaí* (boxer), *damhsóir* (dancer), *comhstiúrthóir* (co-director)

Scaling up to 1000’s of languages

- Crúbadán project; web crawled corpora for under-resourced languages
- Now crawling 2233 languages, hundreds more queued for training
- Scaled up thanks to NSF grant 1159174; see http://crubadan.org/
- Twitter corpora for 180 languages (indigenoustweets.com), 2011-present
- RSS feeds and public Facebook groups (and hand posted links to crawler)

Thank you! / Go raibh maith agaibh!

- http://cs.slu.edu/~scannell/
- https://cadhan.com/
- http://crubadan.org/
- http://indigenoustweets.com/
- http://chuala.me/
- http://intergaelic.com/
- http://corpas.ria.ie/
- https://github.com/kscanne/