Scaling up language technology for the next 1000 languages

Kevin Scannell
Saint Louis University
October 26, 2017

Linguistic Diversity: The Numbers

- About 7100 languages spoken in the world (Ethnologue)
- Almost half are “endangered” (UNESCO)
- 2500-3000 have some online presence
- < 1000 written by the language community
- Wikipedias for about 300 languages
- Tweets in about 275 languages
- Spell checkers for about 180
- Google search interface translated into 150
- Google Translate in about 100

Language Technology

- Machine translation
- Speech recognition
- Predictive text
- Search engines
- Dialogue systems
- Spelling/grammar checking
- Text normalization (e.g. modernization)
- Optical character recognition
- Handwriting recognition

Example: Context-Sensitive Spell checking

- In your brain: B = “May I borrow your phone?”
- On your screen: S = “May I borrow yore phone?”
- Think of S as a “distortion” of B
- Job of a spell checker is to “decode” the original message B from S
- For simplicity assume we know all other words are correct
- Let S(w) be the result of substituting word w for “yore”
- Want to find the w maximizing P(S(w) | S)
- Bayes’ Law, same as maximizing P(S | S(w)) × P(S(w))
- Model first term as P(yore | w); think of w = yore, your, tore, pen, etc.
- Second term is simply the probability of the sentence: “Language Model”

Noisy Channels

- Almost all of the language technologies I mentioned can be modeled this way
- Machine translation, for example, from French to English
- Weird perspective on the French language!
- Given French sentence F, pick English sentence E maximizing $P(F \mid E) \times P(E)$
- $P(F \mid E)$ is the “translation model”; often word or phrase-based models
- Different channel (speech, OCR, etc.) => different channel model
- Second term is the same across all of these problems!
- Better language models => better language technology
- (Incidentally this is why translation *into* English is usually better…)

Language modeling

- Chomsky: “... the notion of ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.”
- Let $S = w_1 w_2 w_3 ... w_n$
- $P(S) = P(w_1|\hat{}) \; P(w_2|w_1) \; P(w_3|w_1 w_2) ... \; P(w_n|w_1 w_2 w_3 ... w_{n-1})$
- Usually formulated and computed this way (word prob. given history)
- Humans are pretty good at estimating these, at least
- 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)

Turing Test

- Language models can be used *generatively* also
- Assume you've already generated $w_1 w_2 w_3 \ldots w_{n-1}$
- Simply choose the next word from the distribution $P(w_n | w_1 w_2 w_3 \ldots w_{n-1})$
- Using full history of a conversation, generate "probable" responses
- Good enough language model => indistinguishable from a human (?)
- Language modeling == AI? (if you believe Turing's formulation)

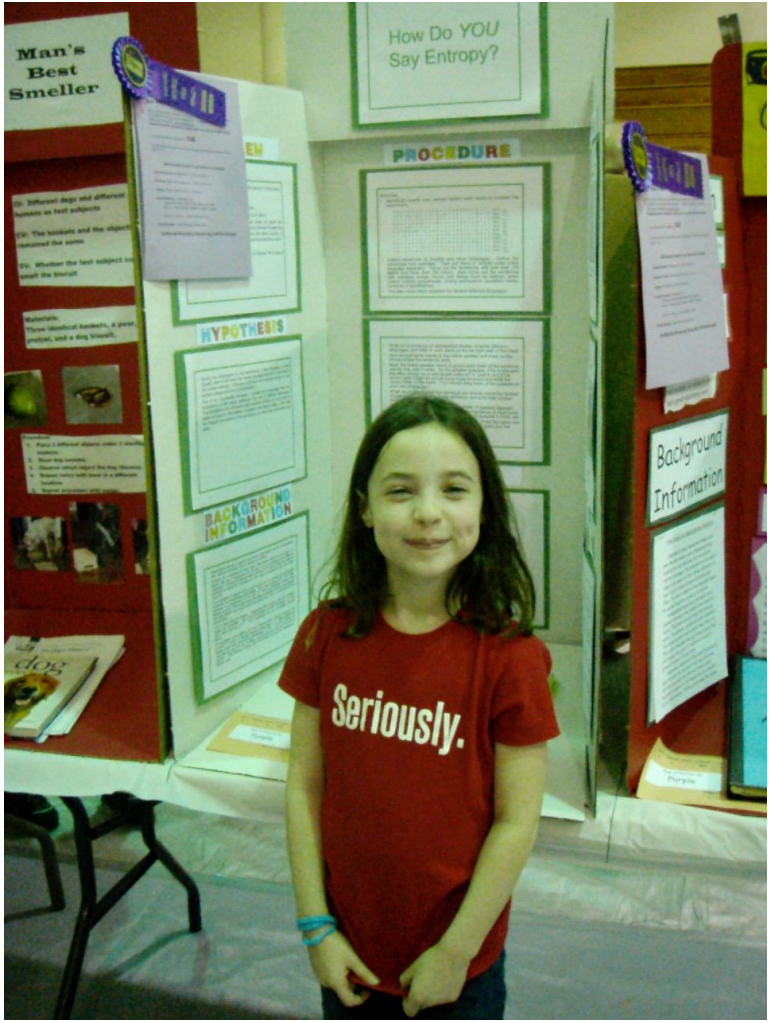
N-gram language models

- Markov assumption: $P(w_n|w_1w_2w_3...w_{n-1}) \approx P(w_n|w_{n-k+1}...w_{n-1})$
- Often k=3 (trigram modeling): need to compute $P(w_n|w_{n-2}w_{n-1})$
- Trained using large text corpora and counting trigrams
- Smooth models by backing off to lower-order n-grams
- P(Friday|this coming) >? P(Tuesday|this coming)
- P(is|the glasses) < P(are|the glasses)
- For English, trainable using Google's 5-gram dataset (*trillion* word corpus)

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 n-gram models for English just over 6 bits/word
- State-of-the-art neural models *under* 5 bits/word


Man's Best Smeller
How Do YOU Say Entropy?
PROCEDURE
HYPOTHESIS
BACKGROUND INFORMATION
Background Information
Seriously.

Issues for under-resourced languages

- Pros and cons of language independence
- Lack for large text corpora for training and evaluating models

“Praistriúchán”

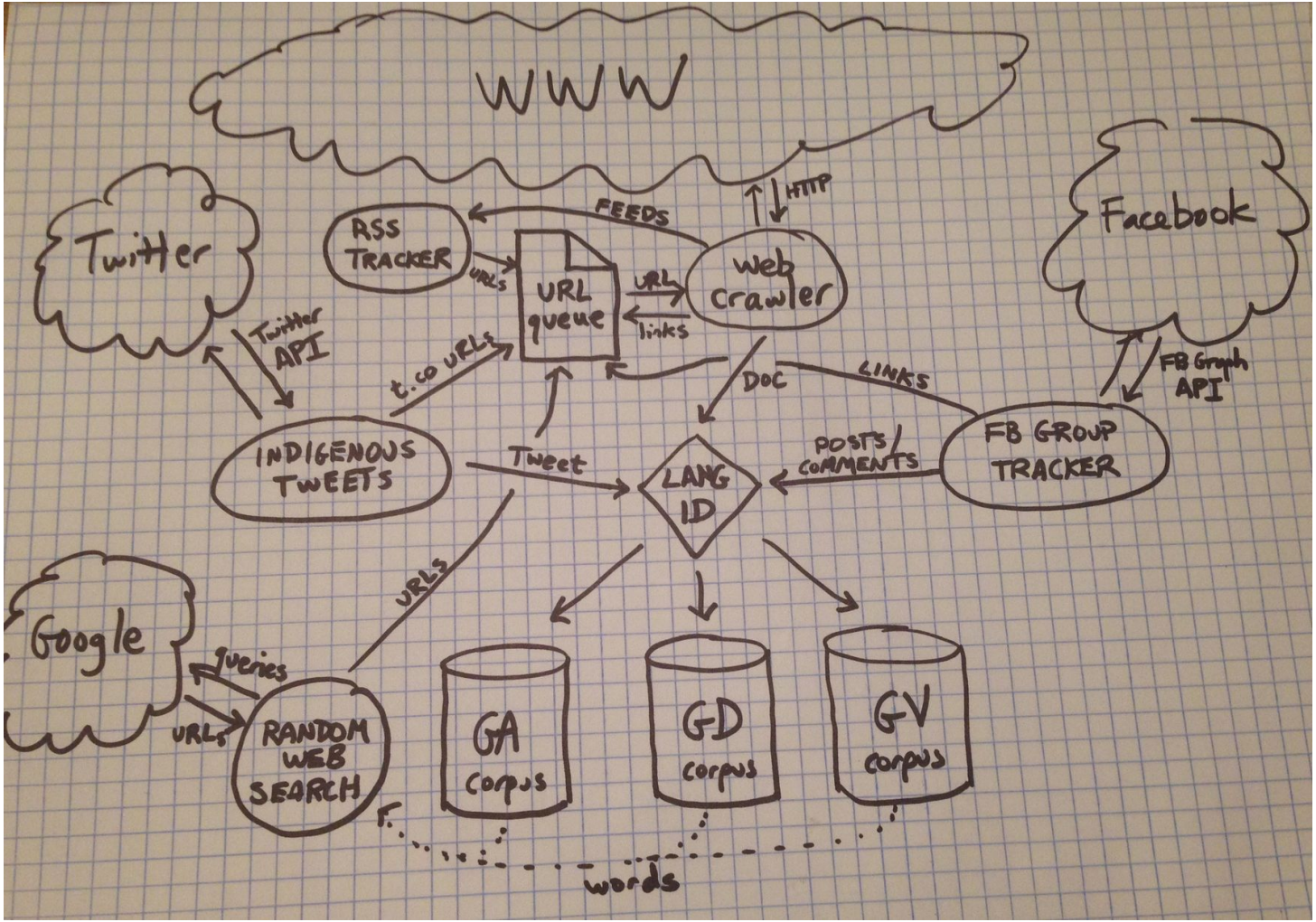
New Irish portmanteau word: “praiseach” = “a mess, a botch job”, “aistriúchán” = “translation”

Incorporating linguistic knowledge

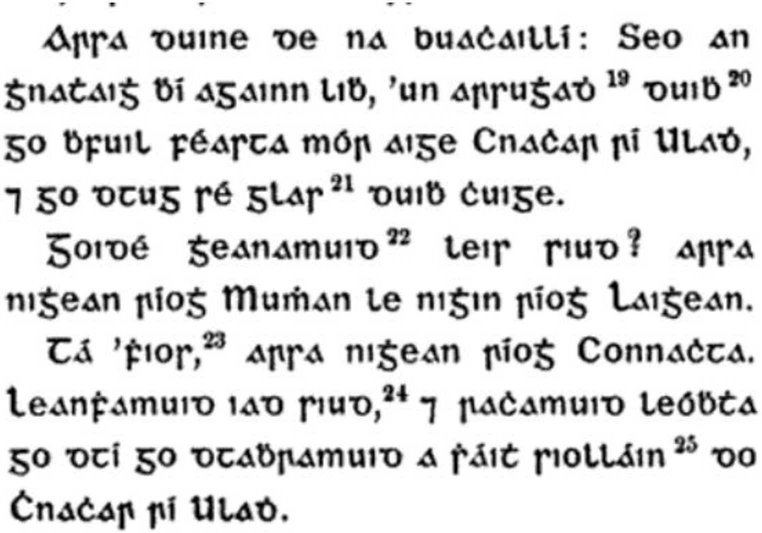
- Celtic languages have “initial mutations”
- Almost always predictable from previous two words
- *bád seoil* “sailboat”, *mo b**h**ád seoil* “my sailboat”, *ár **m**bád seoil* “our sailboat”
- N-gram models don’t “see” that these are all the same word
- If most training examples are first type, say, harder to predict collocation
- (Google even gets this wrong in previous image: *tríd an **m**bóthar*)
- Easy enough to use a *factored language model* to get better results for Irish
- Examples like this abound…

Language models via web corpora

- Crúbadán project; web crawled corpora for under-resourced languages
- Began around 2003 for the six Celtic languages
- Now crawling 2225 languages, hundreds more queued for training
- Scaled up thanks to NSF grant 1159174 (Linguistics)
- http://crubadan.org/
- For Irish, more than 100 million words (> 175 million with offline material)


WWW
Twitter
Facebook
RSS TRACKER
FEEDS
HTTP
URLs
URL queue
URL
links
Web Crawler
Twitter API
t.co URLs
DOC
LINKS
FB Graph API
INDIGENOUS TWEETS
Tweet
LANG ID
POSTS / COMMENTS
FB GROUP TRACKER
Google
URLs
queries
URLs
RANDOM WEB SEARCH
GA corpus
GD corpus
GV corpus
words

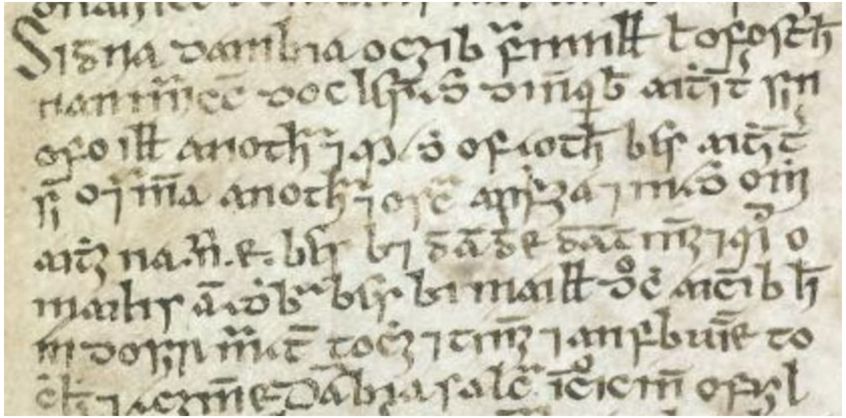
Unlocking the past: fonts and manuscripts



Arsa duine de na buaċaillí: Seo an ġnaċaiġ ḃí againn liḃ, 'un arruġaḋ[19] duiḃ[20] go ḃfuil féasta mór aige Cnaċar rí Ulaḋ, ⁊ go dtug sé glas[21] duiḃ ċuige.

Goidé geanamuid[22] leis siud? arsa nighean ríoġ Muṁan le nigin ríoġ Laighean.

Tá 'ḟios,[23] arsa nighean ríoġ Connaċta. Leanfamuid iad siud,[24] ⁊ raċamuid leóḃta go dtí go dtaḃramuid a ḟáit siolláin[25] do Ċnaċar rí Ulaḋ.



Unlocking the past: standardization

Gaeilge Gaedhilge Gaeilige Gaelige Gailge Gaoidhilge Gaoidheilge Gaelge
Gaidhlige Gaedheilge Gaoidhelge Gailege Gaielge Gaodhailge Gaedilge
Gaeidhilge Gaedhilige Gaoidilge Gaeilgele Gaedhlige Gaédhilge Gaoilge
Gaeillge Gaeilga Gaidhilge Gaelilge Gaodheilge Gaeiilge Gaedhilghe
Gadhilge Gaheilge Gaellge Gaoilaige Gaodhilge Gaedhilgé Gaeilege Gaaeilge
Gailige Gaeilgé Gaeghilge Gaedhailge Gaoidhlige Gaelgie Gaeiloge Gaeilgle
Gaeilghe Gaeilge Gaeidhlge Gaeidheilge Gaeeilge Gaoilige Gaóilge
Gaoilaga Gaoigheilge Gaoidhlge Gaoidelge Gaoideilge Gaodhéilge Gaieilge
Gaeulge Gaeuilge Gaeolge Gaeoidheilge Gaeilgi Gaeilgee Gaeílge Gaeidlge
Gaeidilge Gaeidhelge Gaehilge Gaeeilgee Gaedhlge Gaedhiilge Gaedhelga
Gaédhailge Gaedgilge Gadehilge Gaddhilge Gaoghailge Gaileige Gaidlhlige
Gaidhlge Gaeliage Gaelga Gaéilge Gaedilghe Gaedhulge Gaedhealg Gaedheilg
Gaédheilg Gaedhilg Gaedhilig Gaeilg Gaoidhealg Geadhilge Geailge

Royal Irish Academy Corpus

corpas.ria.ie - 19 million words digitized and indexed by standard forms