Saving Languages with Statistics

Kevin Scannell
Saint Louis University
January 14, 2011

Endangered Languages

- About 7000 living languages (ethnologue.com)
- UNESCO Atlas of the World's Languages in Danger
- At least 43% are endangered; not spoken by children or used only in restricted domains
- Longer term: Between 50-90% of the 7000 languages spoken today will be gone by 2100
- Loss of indigenous knowledge and world views; loss to linguistic science

Language Revitalization

- Reasons for language shift are complex and varied; globalization, imposed “national” languages
- One near-universal is the perception that local languages are “irrelevant” in the modern world
- Sadly, not far from true in the computing domain
- Speakers of only about 40 of the 7000 languages can use a computer in their native language
- Firefox 3.6 in about 70 languages
- Spellcheckers for about 120 languages

Taking Revitalization Online

- Most endangered languages have small population bases, often geographically scattered
- Online communities, blogs, social networks allow small language groups to communicate and be creative in their native language
- But: still only looking at maybe 1000 languages: more than half (80%?) of the 7000 languages in the world have no written tradition
- For many others, limited literacy, or no electricity let alone internet connectivity

Basic Computing Resources

- Localized software
- Keyboard input methods, especially for phones
- Spelling and grammar checkers
- Online dictionaries and thesauri
- Translation software; especially from global to local language
- Plus things I don't work on, esp. speech recognition

Project Scope

- All work done in collaboration with native speakers
- Focus on resources with an immediate impact
- All software we release is free and open source
- Open source is critical for languages with limited resources: reusable components, no reinventing the wheel, and no need to rely on for-profit companies: spirit of "community ownership"
- N.B. Not all groups want their language online, or for language materials available to outsiders, or even for the language to be written

Global Reach

- Assamese (India)
- Azerbaijani (Azerbaijan)
- Chichewa (Malawi)
- Frisian (Friesland)
- Haitian Creole (Haiti)
- Hawaiian (Hawaii)
- Hiligaynon (Philippines)
- Irish (Ireland)
- Kashubian (Poland)
- Kinyarwanda (Rwanda)
- Kirghiz (Kirghistan)
- Kurdish (Kurdistan)
- Lingala (D.R.C.)
- Malagasy (Madagascar)
- Manx Gaelic (Isle of Man)
- Mongolian (Mongolia)
- Oromo (Ethiopia)
- Samoan (Samoa)
- Scottish Gaelic (Scotland)
- Setswana (Botswana)
- Somali (Somalia)
- Songhay (Mali)
- Tagalog (Philippines)
- Tetum (East Timor)
- Turkmen (Turkmenistan)
- Welsh (Wales)
- Many more in progress...

Statistical Language Processing

- Modern approaches to machine translation, speech recognition, etc. rely on statistical machine learning
- Learn from “corpora”, monolingual or bilingual
- Most problems can be cast in standard ways:
- Classification problems: part-of-speech tagging, word sense disambiguation, diacritic restoration (e.g. kookan→kò ò kan), spam filtering
- Search problems: MT, speech recognition, OCR, parsing; “noisy channel model”

Let's Learn Irish

- Q: What can we learn from (just) a bilingual corpus?

Bhris sé clocha	He broke rocks
D'ith sé clocha	He ate rocks
Bhris sí clocha	She broke rocks
Bhris sé a lámh	He broke his hand
Bhris sí a lámh	She broke her hand
D'ith sé a arán	He ate his bread
D'ith sí a harán	She ate her bread

Translation Models

- A: We can learn lexical translation probabilities and also “word alignment” probabilities
- $t(g|e)$ = probability that English word “e” translates to Irish word “g”
- t(arán|bread) ≈ 0.763, t(harán|bread) ≈ 0.032, t(n-arán|bread) ≈ 0.051, t(aráin|bread) ≈ 0.123, ...
- Easy for humans on small scale; how does the computer do it on a grand scale?

Expectation Maximization Algorithm

- Chicken and egg problem:
- If you knew probabilities of different word alignments, computing the translation probabilities would be trivial (just a weighted count)
- If you knew the translation probabilities, you could compute the probability of any alignment
- Start with uniform probabilities and iterate!
- This is a standard setup in machine learning; it's fair to say that the EM algorithm drives the whole field of statistical MT

Example Corpus: Initial State

	ate	bread	broke	hand	he	her	his	rocks	she
a	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
arán	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
bhris	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
clocha	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
d'ith	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
harán	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
lámh	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
sé	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
sí	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111

Example Corpus: Iteration 1

	ate	bread	broke	hand	he	her	his	rocks	she
a	0.167	0.250	0.125	0.250	0.125	0.250	0.250	0.000	0.167
arán	0.083	0.125	0.000	0.000	0.062	0.000	0.125	0.000	0.000
bhris	0.000	0.000	0.292	0.250	0.146	0.125	0.125	0.222	0.194
clocha	0.111	0.000	0.167	0.000	0.167	0.000	0.000	0.333	0.111
d'ith	0.278	0.250	0.000	0.000	0.146	0.125	0.125	0.111	0.083
harán	0.083	0.125	0.000	0.000	0.000	0.125	0.000	0.000	0.083
lámh	0.000	0.000	0.125	0.250	0.062	0.125	0.125	0.000	0.083
sé	0.194	0.125	0.146	0.125	0.292	0.000	0.250	0.222	0.000
sí	0.083	0.125	0.146	0.111	0.000	0.250	0.000	0.111	0.278

Example Corpus: Iteration 2

	ate	bread	broke	hand	he	her	his	rocks	she
a	0.143	0.280	0.088	0.267	0.092	0.294	0.310	0.000	0.147
arán	0.074	0.144	0.000	0.000	0.045	0.000	0.151	0.000	0.000
bhris	0.000	0.000	0.420	0.246	0.114	0.069	0.073	0.197	0.179
clocha	0.064	0.000	0.142	0.000	0.148	0.000	0.000	0.481	0.065
d'ith	0.435	0.297	0.000	0.000	0.129	0.081	0.075	0.063	0.041
harán	0.070	0.136	0.000	0.000	0.000	0.143	0.000	0.000	0.072
lámh	0.000	0.000	0.118	0.359	0.032	0.102	0.106	0.000	0.051
sé	0.175	0.066	0.109	0.063	0.440	0.000	0.285	0.197	0.000
sí	0.040	0.077	0.123	0.064	0.000	0.311	0.000	0.063	0.446

Example Corpus: Iteration 10

	ate	bread	broke	hand	he	her	his	rocks	she
a	0.010	0.352	0.002	0.182	0.001	0.645	0.678	0.000	0.020
arán	0.007	0.259	0.000	0.000	0.045	0.000	0.224	0.000	0.000
bhris	0.000	0.000	0.989	0.029	0.001	0.000	0.000	0.007	0.009
clocha	0.000	0.000	0.001	0.000	0.002	0.000	0.000	0.986	0.000
d'ith	0.970	0.142	0.000	0.000	0.001	0.000	0.000	0.000	0.000
harán	0.007	0.246	0.000	0.000	0.000	0.227	0.000	0.000	0.007
lámh	0.000	0.000	0.007	0.789	0.000	0.003	0.003	0.000	0.000
sé	0.006	0.000	0.000	0.000	0.994	0.000	0.094	0.007	0.000
sí	0.000	0.000	0.001	0.000	0.000	0.126	0.000	0.000	0.964

Example Corpus: Iteration 1000

	ate	bread	broke	hand	he	her	his	rocks	she
a	0.000	0.007	0.000	0.003	0.000	0.994	0.994	0.000	0.000
arán	0.000	0.495	0.000	0.000	0.000	0.000	0.005	0.000	0.000
bhris	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
clocha	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
d'ith	1.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
harán	0.000	0.495	0.000	0.000	0.000	0.005	0.000	0.000	0.000
lámh	0.000	0.000	0.000	0.997	0.000	0.000	0.000	0.000	0.000
sé	0.000	0.000	0.000	0.000	1.000	0.000	0.001	0.000	0.000
sí	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.964

Translating New Sentences

- Given an unseen English sentence ***e***, we need to choose the Irish sentence ***g***, maximizing ***P(g|e)***
- Bayes' Law: ***P(g|e) = P(e|g)P(g)/P(e)***
- ***P(e)*** is constant for all candidate translations; ignore
- ***P(e|g)*** measures "fidelity", ***P(g)*** measures "fluency"
- (Very) naïve estimate of ***P(e|g)*** using t(e|g) probabilities summed over all possible alignments; "bag of words"
- Translation amounts to a search in the space of all possible sentences; standard pruning techniques

Language Modeling

- “The notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term” -Chomsky
- $P(w_1 ... w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)...P(w_n|w_1 ... w_{n-1})$
  $\approx P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)...P(w_n|w_{n-2} w_{n-1})$
- “n-gram” model, often n=3, but 4,5,... if you are Google (recently released massive 5-gram data set for English)
- Easily trainable using big monolingual corpora
- Can also be thought of as a linguistically-naive generative model; indeed you've likely gotten spam generated this way
- Upshot: lots of text needed for training...

An Crúbadán

- Web crawler that seeks out texts written in endangered languages, runs 24/7
- Started in 2003 for the Celtic languages
- Project has now grown to 487 languages
- Language of newly-found text is determined using a statistical classifier based on character sequences
- New language models are bootstrapped from a small amount of training text
- Models are refined (dialects, variant orthographies) with the help of an army of volunteers

Statistics and Endangered Languages

- Endangered languages have been left out because they lack the necessary training data
- Traditional alternative is a "rule-based" approach; can be labor-intensive; requires trained linguists and rich lexicographical resources; resulting systems tend to be less robust
- Given a large enough literate speaker base, we can "crowd-source" creation of bilingual corpora (and resulting data can be of independent usefulness, e.g. by translating Wikipedia articles)

Future Prospects

- How many languages are on the web? 1000?
- By 2015: open source spell checkers and morphological analyzers for 200 languages
- Incorporating more linguistic structure into statistical MT, especially syntax which is critical for English/Irish (VSO)
- Handling complex morphology on the target side in statistical MT (English/Bantu)