Machine Learning in NLP

Kevin Scannell
Saint Louis University
May 15, 2013

The Noisy Channel Model

- Imagine there's a text you want to read
- Problem is, it's been distorted via a “noisy channel”
- Your job is to recover the original signal (text)
- A program which does this is called a “decoder”
- Let's look at several examples
- Source text: “He likes the color of your shirt”

Bad grammar/spelling channel

- “He like the color of yore shirt”
- A decoder is a spelling/grammar checker

Twitter channel

- He <3 the color of ur #shirt lol
- Decoder: Tweet cleaner-upper

Elizabethan Channel

- “He liketh the color of thy shirt”
- Decoder: English “modernizer”

British English Channel

- “He likes the colour of your shirt”
- Decoder: British → American translator

OCR Channel

- “Hc likes the color of vour shlrt”
- Decoder: OCR corrector

Speaker Channel

- Output is a sound wave!
- Decoder: speech recognition system

Irish (Gaelic) Channel

- “Taitníonn dath do léine leis”
- Decoder: Irish to English MT

How Computers Translate

- Given an Irish sentence ***g***, we need to choose the English sentence ***e***, maximizing ***P(e|g)***
- Bayes' Law: ***P(e|g) = P(g|e)P(e)/P(g)***
- ***P(g)*** is constant for all candidate translations; ignore
- ***P(g|e)*** measures "fidelity", ***P(e)*** measures "fluency"
- Translation amounts to two things:
- Giving reasonable estimates for ***P(g|e)*** and ***P(e)***
- Efficiently finding the best ***e*** in the space of all possible sentences ("decoding"); a search problem

Channel and Language Models

- All of the noisy channel problems are the same
- Need estimate of $\boldsymbol{P(g|e)}$: what is probability of seeing $\boldsymbol{g}$ come out of the channel if $\boldsymbol{e}$ goes in
- This is the *channel model*; different in each case
- e.g. $\boldsymbol{P(your|your)} > 0.99$; $\boldsymbol{P(yore|your)} < 0.01$, $\boldsymbol{P(tour|your)} < 0.01$, $\boldsymbol{P(crazy|your)} = 0$?
- Then you need $\boldsymbol{P(e)}$; same in each case!
- This is the *language model*; more on this later

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

- We want to learn two things: “lexical translation probabilities” and “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, ...
- Word alignments are pairing between source and target words; some more probable than others!
- Model ***P(g|e)*** as a weighted sum over all alignments

Expectation Maximization Algorithm

- This is a classic “chicken and egg” problem
- If you knew the probability of any given word alignment, computing the translation probabilities would be trivial (just a weighted count)
- If you knew the translation probabilities, you could compute the probability of any word 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

Language Modeling

- “The notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term” -Chomsky
- $P(w_1 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \ldots P(w_n|w_1 \ldots w_{n-1})$
  $\approx P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \ldots P(w_n|w_{n-2} w_{n-1})$
- “n-gram” model, often n=3, but 4,5,... if you are Google (who have released massive 5-gram data set for English)
- Easily trainable using big monolingual corpora

Guessing Game, I

- “the two ___”
- P(men|the two) = 0.0413
- P(of|the two) = 0.0338
- P(countries|the two) = 0.0298
- P(sides|the two) = 0.0204
- P(groups|the two) = 0.0164
- P(main|the two) = 0.0158
- ...

Guessing Game, II

- “the fact ____”
- $P(\text{that}|\text{the fact}) = 0.8698$
- $P(\text{is}|\text{the fact}) = 0.0312$
- $P(\text{of}|\text{the fact}) = 0.0241$
- $P(\text{remains}|\text{the fact}) = 0.0092$
- $P(\text{was}|\text{the fact}) = 0.0050$
- $P(\text{they}|\text{the fact}) = 0.0043$

Guessing Game, III

- “the united ___”
- P(states|the united) = 0.5240
- P(kingdom|the united) = 0.3129
- P(nations|the united) = 0.0859
- P(arab|the united) = 0.0075
- P(front|the united) = 0.0061
- P(democratic|the united) = 0.0024

Guessing Game, IV

- “button fell ____”
- Doesn't appear at all in the 100M word corpus
- “Backoff smoothing”
- Estimate P(w|button fell) using P(w|fell)
- Or get a bigger corpus!

Let's Generate Spam with n-grams

- You can think also of an n-gram model as a naive "generative model" of English
- What sorts of grammatical errors do we expect?
- $P(\text{has}|\text{years}) > P(\text{have}|\text{years})$!

Better Language Models?

- Linguistically speaking, n-grams are deeply flawed
- Still, they're effective in practice for languages like English
- Almost useless for morphologically complex languages
- Examples: Bantu languages, Inuktitut, Basque, Finnish, etc.
- Chichewa
- Kinyarwanda
- Syntactic language models

An Crúbadán

- Web crawler that seeks out texts written in endangered languages, runs 24/7
- Started in 2003 for the six Celtic languages
- Project has now grown to 1503 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 of the "statistical revolution" due to a lack of training data
- An alternative is a "rule-based" approach; labor-intensive; requires trained linguists and rich 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)