Betting on language

Kevin Scannell
February 15, 2017

"... the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." -- Noam Chomsky

Language models

- Given a sequence of words w_1 w_2 .... w_n
- Estimate probability of any continuation w_{n+1}
- Write this as conditional prob P(w|w_1 ... w_n)
- It’s a discrete prob distribution over all words
- (I didn’t say “words of English”, that’s baked in)
- DEMO: test your intuition, place your bets!
- Sounds easy right?

AI

- It’s not.
- If you could do this as well as humans do, you could pass the Turing test, (a|the) def. of AI!
- Use preceding conversation as w_1...w_n
- Generate only high probability responses
- => hard for all the reasons language/AI is hard
- Examples later...

Applications

- Predictive text
- Search engines (model per document)
- Dialogue systems / spamming (DEMO)
- Spelling/grammar checking
- Text normalization (e.g. modernization)
- OCR correction
- Handwriting recognition
- Speech recognition
- Machine translation

Noisy channels

- Think about spellchecking the following way:
- Correctly-spelled text is in your brain, but is distorted somehow in the process of typing it
- Think of this as a “noisy channel”
- Spellchecking is then a kind of “decoding”
- Maximize P(correct|observed), or via Bayes law, P(observed|correct)*P(correct)
- Channel model: P(yore|your), P(tour|your), ...

Machine translation

- Almost all of the other applications fit this mold, e.g. MT
- If you want to translate from English to Irish, take the following insane perspective on English...
- It’s really just Irish that’s been “encoded” in a bizarre way by transmitting thru a noisy channel
- Now, let’s proceed to decode it.
- Same mathematical setup; want to maximize P(source|target)*P(target)
- In a nutshell why translating *into* English is easier

(Almost) SotA

- Estimate P(road|why did the chicken cross the) as P(road|cross the)
- Estimate P(road|cross the) by collecting huge amounts of text and counting!
- We lose tons of important context
- Grammar becomes hard:
  P(is |guy with the glasses) vs.
  P(are |guy with the glasses)
- Google 5-gram dataset from *trillion* word corpus!

Difficulties

- Data sparsity, “curse of dimensionality”
- Moving target; languages change
- Domain adaptation
- Real-world knowledge
- Ambiguities
- Syntactic structure
- Scaling up to many languages (Crúbadán: 2214)
- Morphologically complex languages

Evaluations

- Let’s say you have a language model implemented as a program to compute P(w|C)
- How can you tell how good your model is?
- End-to-end eval of applications! Or...
- Collect a bunch of text not used for training
- Compute -log_2 P(w|w1...) for each w, average.
- This is a “cross-entropy”, model vs. test set
- Best (neural) approaches < 5.4 bits/word