Relating Conditional Entropy to Mutual Intelligibility

Matthew Smith

Methodology and Statistics for Linguistic Research, Rijksuniversiteit Groningen

May 15th, 2012

Outline

- Mutual Intelligibility
- Entropy and Conditional Entropy
- My Project and Data
- Results
- Final Remarks

Mutual Intelligibility

- the ability of speakers of different languages to understand one another
- not necessarily symmetric: speakers of language A may understand language B better than speakers of language B understand language A
  - level of exposure and attitude towards the other language explain part of the asymmetry
  - but other factors may contribute, such as the difficulty involved in mapping phonemes from one language to the other

What is Entropy?

- concept from information theory
- Entropy is a measure of the uncertainty associated with a random variable
- $H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$ where n is the number of possible outcomes
- Note that the formula considers both the number of outcomes and the probability of each outcome
- The higher the uncertainty, the higher the entropy
- The lowest possible value for entropy is 0 (no uncertainty). There is no upper bound.

Entropy: Examples

- Example 1: outcome of flipping a coin
- $H(X) = -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i)$, where x_1 = heads and x_2 = tails
-
-
-

Entropy: Examples

- Example 1: outcome of flipping a coin
- $H(X) = -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i)$, where x_1 = heads and x_2 = tails
- $H(X) = -(p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2))$
-
-

Entropy: Examples

- Example 1: outcome of flipping a coin
- $H(X) = -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i)$, where $x_1 =$ heads and $x_2 =$ tails
- $H(X) = -(p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2))$
- $H(X) = -(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2})$
-

Entropy: Examples

- Example 1: outcome of flipping a coin
- $H(X) = -\sum_{i=1}^{2} p(x_i) \log_2 p(x_i)$, where $x_1 =$ heads and $x_2 =$ tails
- $H(X) = -(p(x_1) \log_2 p(x_1) + p(x_2) \log_2 p(x_2))$
- $H(X) = -(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2})$
- $H(X) = -(\frac{1}{2}(-1) + \frac{1}{2}(-1)) = 1$

- What if we replace the coin with a six-sided die?

Entropy: Examples

- Example 2: outcome of rolling a six-sided die
- $H(X) = -\sum_1^6 \frac{1}{6} \log_2 \frac{1}{6} = 2.58$
- What if the die is loaded so that half the time it lands on 6?
- What about the entropy of two coin tosses? n coin tosses?

What is Conditional Entropy?

- measure of the uncertainty associated with a random variable Y when the value of random variable X is known
- It can be applied to the mutual intelligibility of words if X is the set of phonemes from language 1 and Y is the set of phonemes from language 2
- $H(Y|X) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log_2 p(y_j|x_i)$
- $p(x_i, y_j)$ is joint probability: the probability that the outcome will be this pair
- $p(y_j|x_i)$ is conditional probability: the probability that the outcome will contain y_j as the second component of the pair if it is known that x_i is the first

Joint Probability and Conditional Probability

- What is the probability of rolling a 1 with a first die and 3 with a second?
-
-
-
-

Joint Probability and Conditional Probability

- What is the probability of rolling a 1 with a first die and 3 with a second?
- There are 36 possible outcomes, so $p(1,3) = \frac{1}{36}$
-
-
-

Joint Probability and Conditional Probability

- What is the probability of rolling a 1 with a first die and 3 with a second?
- There are 36 possible outcomes, so $p(1,3) = \frac{1}{36}$
- What is the probability of rolling a 6 with a second die if you rolled 1 with the first?
-

Joint Probability and Conditional Probability

- What is the probability of rolling a 2 with a first die and 3 with a second?
- There are 36 possible outcomes, so $p(2,3) = \frac{1}{36}$
- What is the probability of rolling a 4 with a second die if you rolled 2 with the first?
- There are 6 possible outcomes for the second die, so $p(4|2) = \frac{1}{6}$

More Conditional Entropy

- Conditional probability and conditional entropy only become interesting when the variables are dependent on one another.
- This is not the case for dice rolls.
- But it is the case for the relationship between phonemes occurring in cognate words in two different languages that are closely related to one another.
- An *f* in German is much more likely to appear in a word if its Dutch cognate contains an *f* in the same spot than it would be in a word whose Dutch cognate you know nothing about.

My Project

- Goal of project: calculate the conditional entropy from Dutch to German and from German to Dutch on the phoneme level using cognate words as the data
  - hypothesis: conditional entropy from German to Dutch should be lower than from Dutch to German
  - motivation: mutual intelligibility is not symmetric
  - A similar project has already been carried out for continental Scandinavian languages with that suggested a correlation between mutual intelligibility and conditional entropy, but more work needs to be done.
- the scope of my project is limited to the intelligibility of individual spoken words

Dataset

- 768 pairs of frequent cognates with 3576 sounds
- Non-cognates were not used since they cannot be understood by non-speakers of that language.
- The data I started with consisted of phonetic transcriptions of each of the cognates, but I needed a mapping of German sounds to Dutch sounds for each word.
- I did this myself manually

Examples of Mapping

- Example 1 boek/buch:

b	u	k
b	u	x

- Example 2 straat/strasse:

s	t	r	a	t	Null
s	t	r	a	s	ə

- Almost all cases were as straightforward as these two

Calculate Conditional Entropy

- $H(Y|X) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log_2 p(y_j|x_i)$
- here we will do the part of the calculation that involves pairs containing the Dutch sound *z*. There are three such pairs. We need to know how many times they occurred as well as two other pieces of information:
  - (z,ʃ) occurs 6 times
  - (z,s) occurs 3 times
  - (z,z) occurs 35 times
  - the total number of pairs containing Dutch *z* is $6 + 3 + 35 = 44$
  - the total number of pairs is 3576

Calculate Conditional Entropy

- (z,ʃ) occurs 6 times
- (z,s) occurs 3 times
- (z,z) occurs 35 times
- The total number of pairs containing Dutch *z* is $6 + 3 + 35 = 44$
- The total number of pairs is 3576
- What is p(z,ʃ)? p(ʃ|z)? p(z,s)? p(s|z)? p(z,z)? p(z|z)?

Calculate Conditional Entropy

- (z,ʃ) occurs 6 times
- (z,s) occurs 3 times
- (z,z) occurs 35 times
- The total number of pairs containing Dutch *z* is 6 + 3 + 35 = 44
- The total number of pairs is 3576
- What is p(z,ʃ)? p(ʃ|z)? p(z,s)? p(s|z)? p(z,z)? p(z|z)?
- $\frac{6}{3576}, \frac{6}{44}, \frac{3}{3576}, \frac{3}{44}, \frac{35}{3576}, \frac{35}{44}$

Calculate Conditional Entropy

- $-\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\log_2 p(y_j|x_i)$
- $-(p(z,ʃ)\log_2 p(ʃ|z) + p(z,s)\log_2 p(s|z) + p(z,z)\log_2 p(z|z))$
- $-(\frac{6}{3576}\log_2\frac{6}{44} + \frac{3}{3576}\log_2\frac{3}{44} + \frac{35}{3576}\log_2\frac{35}{44}) = .0113$
-

Calculate Conditional Entropy

- $H(Y|X) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\log_2 p(y_j|x_i)$
- $H(Y|X) = -(p(z,\int)\log_2 p(\int|z) + p(z,s)\log_2 p(s|z) + p(z,z)\log_2 p(z|z))$
- $H(Y|X) = -(\frac{6}{3576}\log_2\frac{6}{44} + \frac{3}{3576}\log_2\frac{3}{44} + \frac{35}{3576}\log_2\frac{35}{44}) = .0113$
- We can't infer much from this partial result. The final CE will be much higher than this. But by looking at the distribution of German phonemes given Dutch z, what can we say?

Results

- German to Dutch conditional entropy: .98
- Dutch to German conditional entropy: 1.02
- very close to one another

Next Step: N-grams?

- instead of looking at correspondences between individual sounds, we can look at sequences of more than one sound
- bigrams are sequences of two sounds
- trigrams are sequences of three sounds
- n-grams are often used in linguistics
- good way of taking part of the context into account
- you need much more data because each bigram will usually appear much less often than each individual sound

Final Remarks

- Still need to interpret results further
- Also need to determine whether sample size is sufficient (Cronbach's alpha). May need to add more data
- Conditional entropy between cognates is just one measure related to mutual intelligibility of spoken words. Another factor is which percentage of words are cognates between the two languages, because this is also asymmetrical
- Conditional entropy might be better applied to cases where speakers are exposed to the other language but cannot speak it, for example Scandinavian languages and Western Slavic languages