Conditional entropy as a measure of intelligibility

Jens Moberg

Department of Alfa-Informatica
Rijksuniversiteit Groningen

7th April 2006

Outline

- The languages in concern are the three mainland Scandinavian languages: Swedish, Danish, Norwegian
- Can be considered dialects which are separate languages only for political and historical reasons
- Enables semi-communication: Inter-Scandinavian communication in each speaker's own language
- Tendency that mutual intelligibility has deteriorated
- How can you measure this deterioration, i.e. the linguistic distance between the languages?

- The languages in concern are the three mainland Scandinavian languages: Swedish, Danish, Norwegian
- Can be considered dialects which are separate languages only for political and historical reasons
- Enables semi-communication: Inter-Scandinavian communication in each speaker's own language
- Tendency that mutual intelligibility has deteriorated
- How can you measure this deterioration, i.e. the linguistic distance between the languages?

- The languages in concern are the three mainland Scandinavian languages: Swedish, Danish, Norwegian
- Can be considered dialects which are separate languages only for political and historical reasons
- Enables semi-communication: Inter-Scandinavian communication in each speaker's own language
- Tendency that mutual intelligibility has deteriorated
- How can you measure this deterioration, i.e. the linguistic distance between the languages?

- The languages in concern are the three mainland Scandinavian languages: Swedish, Danish, Norwegian
- Can be considered dialects which are separate languages only for political and historical reasons
- Enables semi-communication: Inter-Scandinavian communication in each speaker's own language
- Tendency that mutual intelligibility has deteriorated
- How can you measure this deterioration, i.e. the linguistic distance between the languages?

- The languages in concern are the three mainland Scandinavian languages: Swedish, Danish, Norwegian
- Can be considered dialects which are separate languages only for political and historical reasons
- Enables semi-communication: Inter-Scandinavian communication in each speaker's own language
- Tendency that mutual intelligibility has deteriorated
- How can you measure this deterioration, i.e. the linguistic distance between the languages?

Outline

- Measure phonetic distances among Scandinavian languages
- Discover regularities in phonetic correspondences
- Knowledge of correspondences enables word prediction
- How can you measure the difficulty of these predictions?
- By using conditional entropy (CE)!

- Measure phonetic distances among Scandinavian languages
- Discover regularities in phonetic correspondences
- Knowledge of correspondences enables word prediction
- How can you measure the difficulty of these predictions?
- By using conditional entropy (CE)!

- Measure phonetic distances among Scandinavian languages
- Discover regularities in phonetic correspondences
- Knowledge of correspondences enables word prediction
- How can you measure the difficulty of these predictions?
- By using conditional entropy (CE)!

- Measure phonetic distances among Scandinavian languages
- Discover regularities in phonetic correspondences
- Knowledge of correspondences enables word prediction
- How can you measure the difficulty of these predictions?
- By using conditional entropy (CE)!

- Measure phonetic distances among Scandinavian languages
- Discover regularities in phonetic correspondences
- Knowledge of correspondences enables word prediction
- How can you measure the difficulty of these predictions?
- By using conditional entropy (CE)!

Outline

- Given a corpus of cognate word pairs..
- ..probabilities of sound correspondences can be measured
- The minimum number of bits needed to encode one language based on the other (the conditional entropy) can then be calculated
- Conditional entropy also captures assymetry, so you can model the idea that Danes understand Swedish better than Swedes understand Danish

- Given a corpus of cognate word pairs..
- ..probabilities of sound correspondences can be measured
- The minimum number of bits needed to encode one language based on the other (the conditional entropy) can then be calculated
- Conditional entropy also captures assymetry, so you can model the idea that Danes understand Swedish better than Swedes understand Danish

- Given a corpus of cognate word pairs..
- ..probabilities of sound correspondences can be measured
- The minimum number of bits needed to encode one language based on the other (the conditional entropy) can then be calculated
- Conditional entropy also captures assymetry, so you can model the idea that Danes understand Swedish better than Swedes understand Danish

- Given a corpus of cognate word pairs..
- ..probabilities of sound correspondences can be measured
- The minimum number of bits needed to encode one language based on the other (the conditional entropy) can then be calculated
- Conditional entropy also captures assymetry, so you can model the idea that Danes understand Swedish better than Swedes understand Danish

Outline

Statistical entropy, overview

- Entropy measures the amount of information in a random variable..
- ..i.e. the degree of freedom in a given situation
- Great freedom in the choice of unit (phoneme,letter etc) means few limitations and high entropy.

Statistical entropy, overview

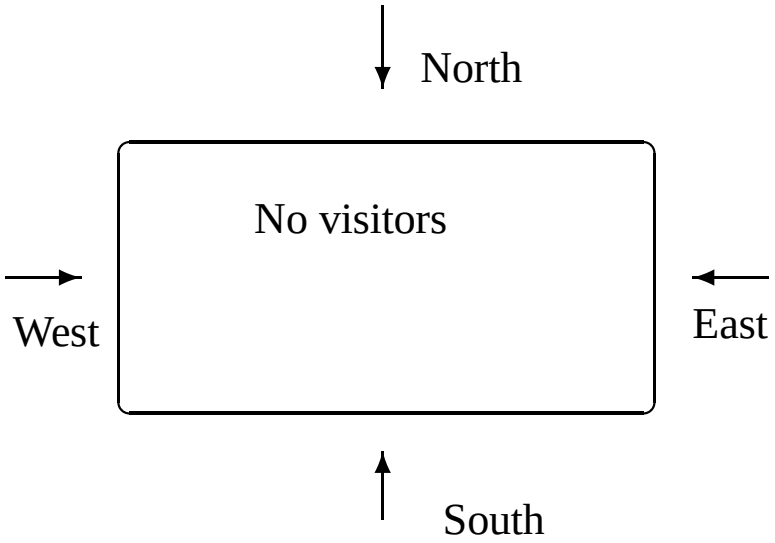
- Entropy measures the amount of information in a random variable..
- ..i.e. the degree of freedom in a given situation
- Great freedom in the choice of unit (phoneme,letter etc) means few limitations and high entropy.

Statistical entropy, overview

- Entropy measures the amount of information in a random variable..
- ..i.e. the degree of freedom in a given situation
- Great freedom in the choice of unit (phoneme,letter etc) means few limitations and high entropy.

The lookout example

If a visitor is approaching, the lookout needs to report the direction the visitor is coming from, i.e. one of five messages:



These options can be thought of as bits: 000, 001, 010, 011 and 100. All codes have three bits.

Example: code length

Assuming that every message is equally likely, the binary log gives us the code length:

$$\text{code length} = \lceil \log_2 |M| \rceil, \text{ where } M \text{ are the messages}$$

What happens when you add frequency?

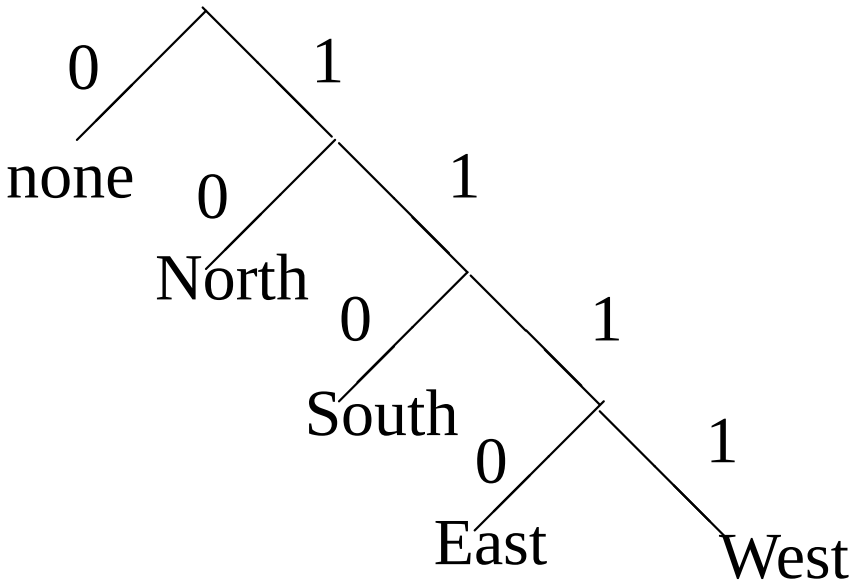
message	rel. freq.
no visitor	99%
North	0.5%
South	0.25%
East, West	0.125%

Example: expected Code Length

We now calculate the code length:

message	code length	rel. freq.	expected bit length
no visitor	1	0.99	0.99
North	2	0.005	0.01
South	3	0.0025	0.0075
East	4	0.00125	0.005
West	4	0.00125	0.005
Total			1.0175

Example: a code tree



message	code
no visitor	0
North	10
South	110
East	1110
West	1111

Entropy

The optimal code cannot be compressed further than the **entropy** (informational uncertainty) of the dataset:

$$H(S) = -\sum_{i \in S} p_i \log_2 p_i$$

message	p_i	$-\log p_i$	$p_i \log p_i$
no visitor	0.99	0.004	0.0044
North	0.005	2.3	0.0115
South	0.0025	2.6	0.0065
East	0.00125	2.9	0.0036
West	0.00125	2.9	0.0036
Total			0.021

- Think of entropy as the ”20 questions” game! You need to ask 0.021 yes/no questions on average to identify the message

Entropy reduction

- By adding knowledge to the system, one reduces the uncertainty. The information gain can be quantified by comparing the total entropies of the original system and the final system.
- Suppose visitors never come on Mondays. Then adding information about the day of the week will reduce the entropy:

Day	P	Entropy
Mondays	0.143	0
Other	0.857	0.021
Total		0.018

Outline

- The joint entropy of a pair of random variables is the amount of info needed on average to specify both their values:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) log p(x,y)$$

Outline

Overview and explanation on conditional entropy

- CE is always calculated in relation to other information
- CE relies on conditional probabilities
- CE of Y given X is the joint entropy of X and Y minus the entropy of X:

$$H(Y \mid X) = H(X, Y) - H(X)$$

- As opposed to joint entropy, CE is not symmetrical:

$$H(Y \mid X) \neq H(X \mid Y)$$

Overview and explanation on conditional entropy

- CE is always calculated in relation to other information
- CE relies on conditional probabilities
- CE of Y given X is the joint entropy of X and Y minus the entropy of X:

$$H(Y \mid X) = H(X, Y) - H(X)$$

- As opposed to joint entropy, CE is not symmetrical:

$$H(Y \mid X) \neq H(X \mid Y)$$

Overview and explanation on conditional entropy

- CE is always calculated in relation to other information
- CE relies on conditional probabilities
- CE of Y given X is the joint entropy of X and Y minus the entropy of X:

$$H(Y \mid X) = H(X, Y) - H(X)$$

- As opposed to joint entropy, CE is not symmetrical:

$$H(Y \mid X) \neq H(X \mid Y)$$

Overview and explanation on conditional entropy

- CE is always calculated in relation to other information
- CE relies on conditional probabilities
- CE of Y given X is the joint entropy of X and Y minus the entropy of X:

$$H(Y \mid X) = H(X, Y) - H(X)$$

- As opposed to joint entropy, CE is not symmetrical:

$$H(Y \mid X) \neq H(X \mid Y)$$

Measuring CE

- $H(X \mid Y)$ is the uncertainty in X given knowledge of Y.
- CE measures how much entropy a random variable Y has remaining if the value of a second random variable X is known
- This means that in a linguistic context, CE can be used to measure the difficulty of predicting a unit in the source language given a corresponding unit in the related language

Measuring CE

- $H(X \mid Y)$ is the uncertainty in X given knowledge of Y.
- CE measures how much entropy a random variable Y has remaining if the value of a second random variable X is known
- This means that in a linguistic context, CE can be used to measure the difficulty of predicting a unit in the source language given a corresponding unit in the related language

Measuring CE

- $H(X \mid Y)$ is the uncertainty in X given knowledge of Y.
- CE measures how much entropy a random variable Y has remaining if the value of a second random variable X is known
- This means that in a linguistic context, CE can be used to measure the difficulty of predicting a unit in the source language given a corresponding unit in the related language

Example: Danish realizations of a Swedish phoneme

Table: Conditional probabilities for Danish phonemes given Swedish /a/

Danish →	ə	a	ɒ	Others
Swedish ↓				
a	0.45	0.14	0.10	0.31
o				
u				
..etc				

Using probabilities to calculate CE

- Entropy H $(P(D \mid a))$

$$H = -\sum_{d \in D, a} p(d, a) log_2(d \mid a)$$

$$H = -(0.45 * log_2 0.45) + (0.14 * log_2 0.14) + (0.10 * log_2 0.10) + (0.31 * log_2 0.31)$$

- H(D|a) = 1.775 bits of information
- If this is done for all phonemes, about 30 in Swedish, you can predict where the biggest intelligibility problems are, i.e. where errors are most likely to be made

Using probabilities to calculate CE

- Entropy H $(P(D \mid a))$

$$H = -\sum_{d \in D, a} p(d, a) log_2(d \mid a)$$

$$H = -(0.45 * log_2 0.45) + (0.14 * log_2 0.14) + (0.10 * log_2 0.10) + (0.31 * log_2 0.31)$$

- H(D|a) = 1.775 bits of information
- If this is done for all phonemes, about 30 in Swedish, you can predict where the biggest intelligibility problems are, i.e. where errors are most likely to be made

Using probabilities to calculate CE

- Entropy H $(P(D \mid a))$

$$H = -\sum_{d \in D, a} p(d, a) log_2(d \mid a)$$

$$H = -(0.45 * log_2 0.45) + (0.14 * log_2 0.14) + (0.10 * log_2 0.10) + (0.31 * log_2 0.31)$$

- H(D|a) = 1.775 bits of information
- If this is done for all phonemes, about 30 in Swedish, you can predict where the biggest intelligibility problems are, i.e. where errors are most likely to be made

Outline

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

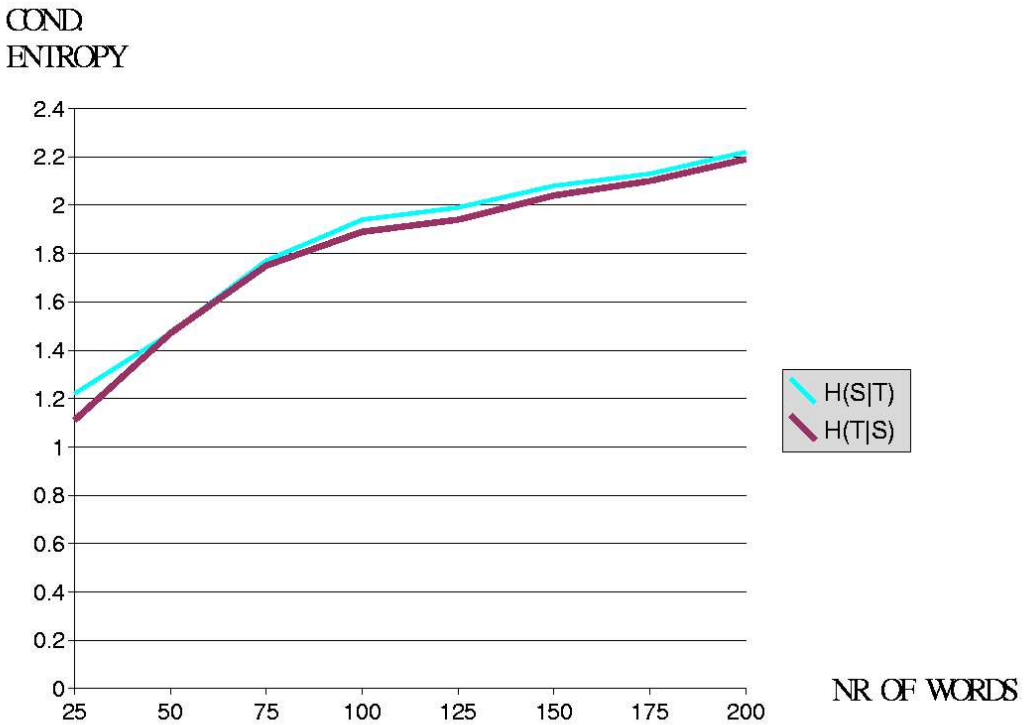
- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

- The sample corpus: 206 Danish-Swedish word pairs, some non-cognates
- Due to insertions and deletions, the word length is sometimes different. I use a filler symbol to account for this.
- S means source(Swedish) and T target (Danish)
- H(S|T) = 2.29
- H(T|S) = 2.22
- With filler-symbol:
- H(S|T) = 2.25
- H(T|S) = 2.23

Outline

Uniform distribution of CE



- For 25 words: H(S|T) = 1.22, H(T|S) = 1.11
- 200 words: H(S|T)= 2.22, H(T|S)= 2.19

What is then a suffi cient data size?

- Hard to tell..
- But every phoneme should occur at least 10 times(roughly)
- By then, the probabilities should have stabilized and be reliable measures for the languages as a whole

What is then a suffi cient data size?

- Hard to tell..
- But every phoneme should occur at least 10 times(roughly)
- By then, the probabilities should have stabilized and be reliable measures for the languages as a whole

What is then a suffi cient data size?

- Hard to tell..
- But every phoneme should occur at least 10 times(roughly)
- By then, the probabilities should have stabilized and be reliable measures for the languages as a whole

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Future work

- Complete the corpus: a core corpus with 3 parts: function words, formal content words and informal content words
- Add information about date of introduction into the language
- For the entropy experiments: extend the corpus by adding data from a Swedish speech corpus that is phonetically transcribed
- Look at the correspondence between bigrams
- Analyze the results: how does the style of language affect entropy?
- Look at date of introduction: how does this affect the difference between the languages in terms of entropy?

Acknowledgements

- Charlotte Gooskens, John Nerbonne, Jorg Tiedemann, Leonoor van der Beek, Thomas Zastrow