Using context and phonetic features in models of etymological sound change

H. Wettig, K. Reshetnikov and R. Yangarber (2012)

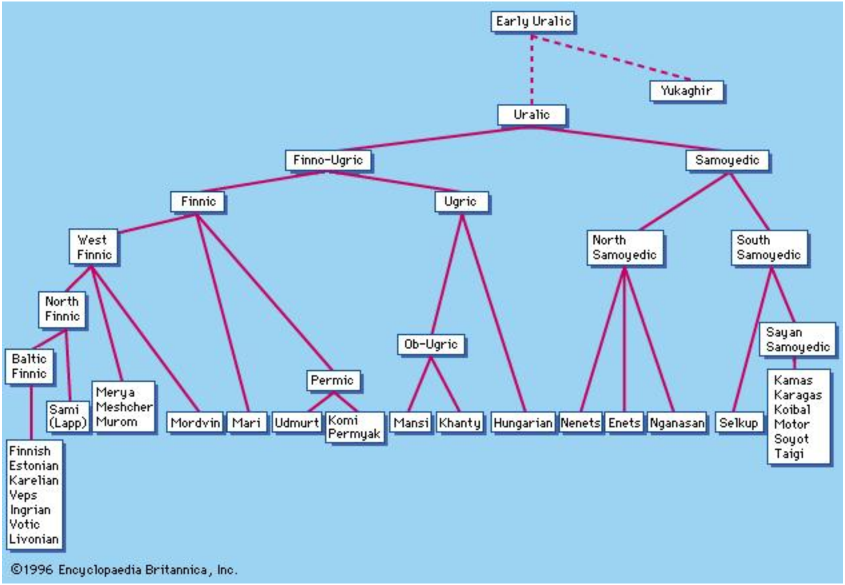
Verena Blaschke

Unsupervised Learning in Computational Linguistics
WS 16/17

December 06, 2017

Outline

Uralic languages



Source: [Encyclopaedia Britannica]

Symbol-level word alignment

source alphabet Σ incl. the empty string: $\Sigma_{\cdot}$
target alphabet T incl. the empty string: $T_{\cdot}$
empty string: $\cdot$

symbol pair $(\sigma : \tau)$ s.t. $\sigma \in \Sigma_{\cdot}$ and $\tau \in T_{\cdot}$ but not $\sigma = \cdot \wedge \tau = \cdot$

v	u	o	s	i		v	u	o	s	i	etc.
a	l	·	·	·		·	a	·	l	·	

Decision trees

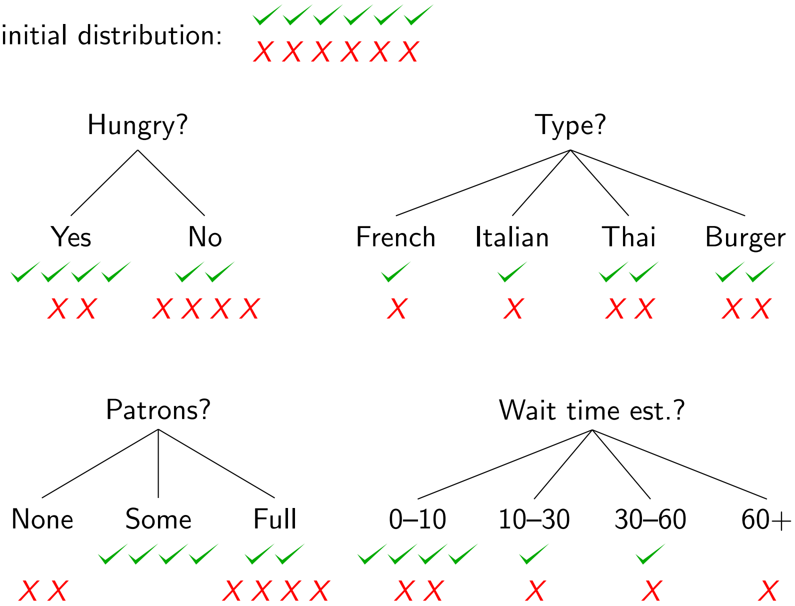
observation	attributes				outcome
	hungry	**patrons**	**type**	**wait est**	
x_1	Yes	Some	French	0–10	✓
x_2	Yes	Full	Thai	30–60	✗
x_3	No	Some	Burger	0–10	✓
x_4	Yes	Full	Thai	10–30	✓
x_5	No	Full	French	60+	✗
x_6	Yes	Some	Italian	0–10	✓
x_7	No	None	Burger	0–10	✗
x_8	Yes	Some	Thai	0–10	✓
x_9	No	Full	Burger	60+	✗
x_{10}	Yes	Full	Italian	10–30	✗
x_{11}	No	None	Thai	0–10	✗
x_{12}	Yes	Full	Burger	30–60	✓

Adapted from [Russell and Norvig].

Decision trees

initial distribution: ✓✓✓✓✓✓ ✗✗✗✗✗✗

Decision trees



Decision trees

entropy $H(X) = -\sum_x P(x) log P(x)$

assuming logarithm base 2 and $0 * log(0) = 0$

✓✓✓✓✓✓ $\quad H(X) = -0.5 * log(0.5) - 0.5 * log(0.5)$

$X\ X\ X\ X\ X\ X \quad = -0.5 * (-1) - 0.5 * (-1) = 1$

✓✓✓✓✓✓ $\quad H(X) = -1 * log(1) - 0 * log(0)$

$= -0 - 0 = 0$

Decision trees

entropy $H(X) = -\sum_x P(x) log P(x)$

assuming logarithm base 2 and $0 * log(0) = 0$

✓✓✓✓✓✓ $H(X) = -0.5 * log(0.5) - 0.5 * log(0.5)$
$X\ X\ X\ X\ X\ X$ $= -0.5 * (-1) - 0.5 * (-1) = 1$

✓✓✓✓✓✓ $H(X) = -1 * log(1) - 0 * log(0)$
$= -0 - 0 = 0$

weighted entropy of m child nodes $c_1 ... c_m$ of parent node n after splitting by attribute a:

$$H(X|a) = \sum_{i=1}^{m} \frac{|c_i|}{|n|} H(X_{c_i})$$

Decision trees

entropy $H(X) = -\sum_x P(x) log P(x)$

assuming logarithm base 2 and $0 * log(0) = 0$

✓✓✓✓✓✓ $H(X) = -0.5 * log(0.5) - 0.5 * log(0.5)$

X X X X X X $= -0.5 * (-1) - 0.5 * (-1) = 1$

✓✓✓✓✓✓ $H(X) = -1 * log(1) - 0 * log(0)$

$= -0 - 0 = 0$

weighted entropy of m child nodes $c_1 \ldots c_m$ of parent node n after splitting by attribute a:

$$H(X|a) = \sum_{i=1}^{m} \frac{|c_i|}{|n|} H(X_{c_i})$$

information gain $=$ reduction in entropy caused by the split:

$$IG(X;a) = H(X) - H(X|a)$$

Decision trees

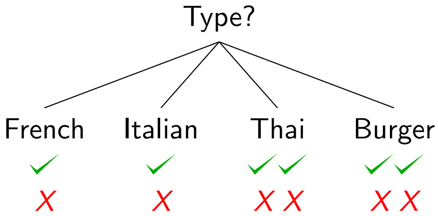
$\checkmark\ \checkmark\ \checkmark\ \checkmark\ \checkmark\ \checkmark$

$X\ X\ X\ X\ X\ X \qquad H(X) = 1$

Decision trees

✓ ✓ ✓ ✓ ✓ ✓

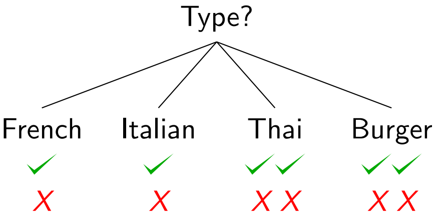
X X X X X X $\quad H(X) = 1$



Decision trees

✓✓✓✓✓✓

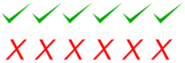
X X X X X X $H(X) = 1$



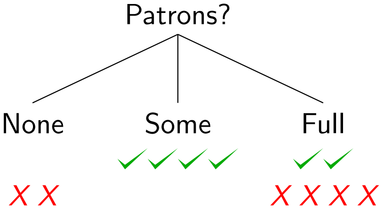
$$H(X|\,Type) = \frac{2}{12} * H(0.5, 0.5) + \frac{2}{12} * H(0.5, 0.5) + \frac{4}{12} * H(0.5, 0.5) + \frac{4}{12} * H(0.5, 0.5) = 1$$

$$IG(X; Type) = 1 - 1 = 0$$

Decision trees



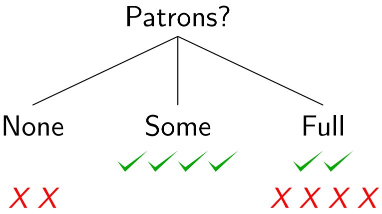
$H(X) = 1$



Decision trees

✓✓✓✓✓✓

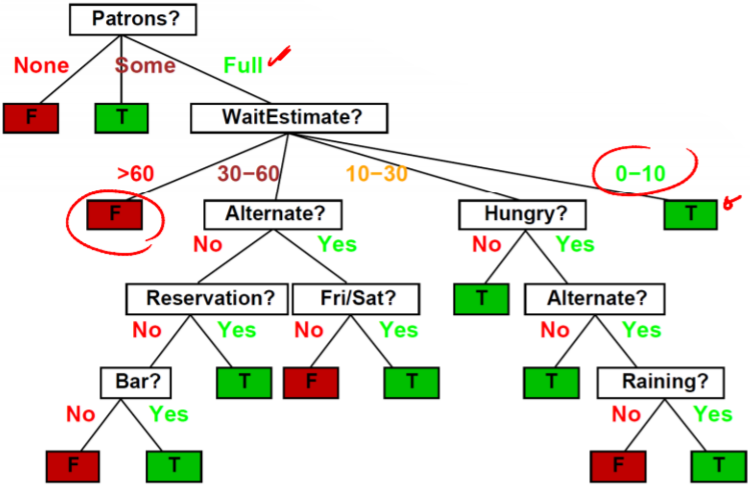
X X X X X X $\quad H(X) = 1$



$$H(X|Patrons) = \frac{2}{12} * H(0,1) + \frac{4}{12} * H(1,0) + \frac{6}{12} * H(\frac{1}{3}, \frac{2}{3})$$
$$\approx \frac{2}{12} * 0 + \frac{4}{12} * 0 + \frac{6}{12} * 0.9183 \approx 0.4591$$

$$IG(X; Patrons) \approx 1 - 0.4591 \approx 0.5408$$

Decision trees



Source: [Freitas]

Features

	Type	consonant K, vowel V, empty string ·, word boundary #

Consonant articulation

M	**Manner**	plosive, fricative, glide, …
P	**Place**	labial, dental, …, velar
X	**Voiced**	–, +
S	**Secondary**	–, affricate, aspirate, …

Vowel articulation

V	**Vertical**	high–low
H	**Horizontal**	front–back
R	**Rounding**	–, +
L	**Length**	1–5

Contexts

Level: source, target

Position:

I	itself
–P	previous position
–S	previous non-dot symbol
–K	previous consonant
–V	previous vowel
+S	previous or self non-dot symbol
+K	previous or self consonant
+V	previous or self vowel

candidate context (*Level*, *Position*, *Feature*)

Decision forest

18 decision trees: one for each feature and level

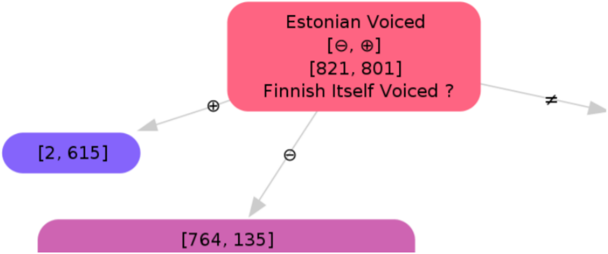
example: target-**X**
What is the value of the current sound in the target language for the feature "voice"?

split by contexts

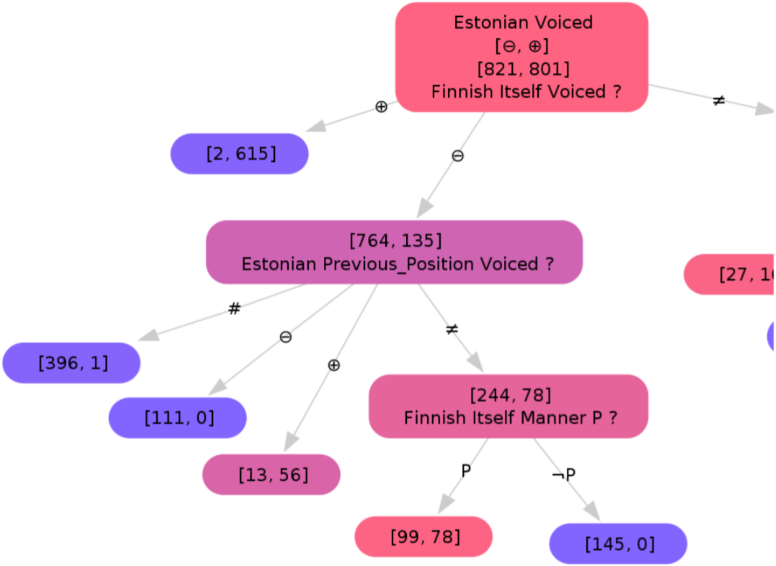
features in addition to type, consonant-related features, vowel-related features:

$\neq$	not applicable
$\#$	not applicable (word boundary)

Target-X tree



Target-X tree



Objective function

cost $\widehat{=}$ entropy in leaf nodes

Normalized Maximum Likelihood code-length
leaf node N that contains n instances
tree has λ levels and describes feature F
F has k values that are distributed s.t. n_i instances have value $i \in \{1, ..., k\}$

$$L_{NML}(\lambda; F; N) = -logP_{NML}(\lambda; F; N) = -log\frac{\prod_{i=1}^{k}(\frac{n_i}{n})^{n_i}}{C(n, k)}$$

$$C(n, k) = \sum_{n'_1+\ldots+n'_k=n} \prod_{i=1}^{k} (\frac{n'_i}{n})^{n'_i}$$

Symbol-level word alignment

	$-$	τ_1	$\dots$	τ_j	$\dots$	τ_m
$-$	0					
σ_1						
$\dots$						
σ_i						
$\dots$						
σ_m						$\star$

source word
$\vec{\sigma} = [\sigma_1 ... \sigma_n] \in \Sigma^*$
target word
$\vec{\tau} = [\tau_1 ... \tau_m] \in T^*$
matrix V

$$V(i,j) = min \begin{cases} V(i,j-1) & +L(\cdot : \tau_j) \\ V(i-1,j) & +L(\sigma_i : \cdot) \\ V(i-1,j-1) & +L(\sigma_i : \tau_j) \end{cases}$$

L $=$ change in code length that would be caused by adding this instance to the corresponding leaf nodes

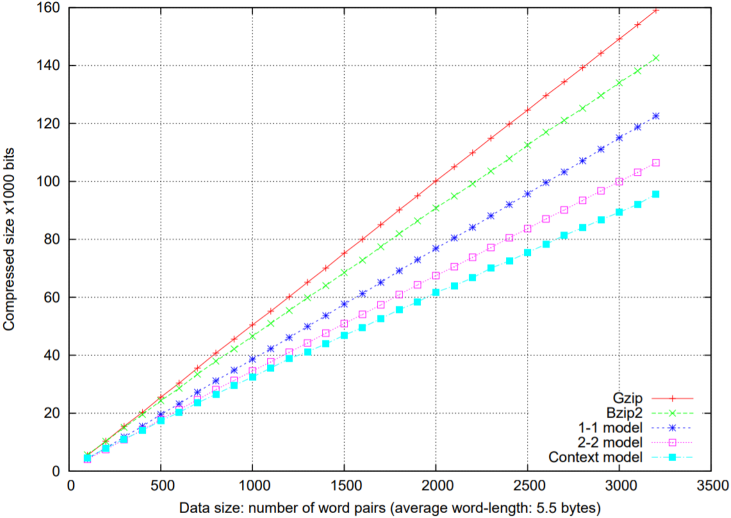
Putting the parts together

1. randomly align each pair of words
2. (re-)build all decision trees for this alignment
3. re-align all word pairs

   repeat steps 2 and 3 until convergence

Evaluation and results

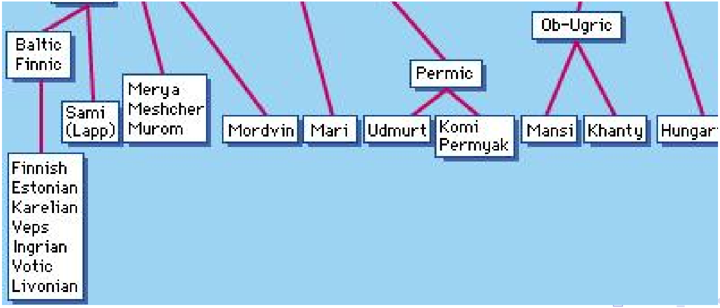
1. gold-standard alignments
2. rules of correspondence
3. compression



Evaluation and results

4 imputation (summed edit distances between imputed and actual L_2 words, normalized by size of true L_2 data)

	fin	*khn*	*kom*	*man*	*mar*	*mrd*	*saa*	*udm*	*ugr*
est	**0.26**	0.66	0.64	0.65	0.61	0.57	0.57	0.62	0.62
fin		0.63	0.64	0.65	0.59	0.56	0.50	0.62	0.63
khn			0.65	**0.58**	0.69	0.64	0.67	0.66	0.66
kom				0.63	0.68	0.66	0.70	**0.39**	0.66
man					0.68	0.65	0.72	0.62	0.62
mar						0.65	0.69	0.65	0.66
mrd							0.58	0.66	0.63
saa								0.67	0.70
udm									0.65



Questions

word alignment

1. I'm curious about the realignment as described in 4.4 and 4.5. It's randomized, so it's not *guaranteed* to converge (even though of course in reality it can be expected to), but more importantly is it the most efficient way to initialize the values? They say at the beginning of section 4 that they've shown it to be an effective method, but it's not really clear to me why that would be so.

   –Peter

2. Will the maximum change, if the random values for initialization change?

   –Le Duyen

Questions

features

1. What do you think about possible pros and cons of using sound features comparing to just using sound symbols? And do you think that using sound feautures can help solving many problems? For example, when I had the presentation of the paper by Rama et al. (2017) we saw that some models had problems dealing with Chinese and other languages that use tones. In the same time it looks the sound features can be one of the solutions of this problem.

–Maxim

Questions

evaluation: compression

1. I find the [compression] approach quite appealing, trying to see if the own method finds more regularities and can therefore generate a smaller output. But did I get it right that the use of decision trees would also guarantee a 100% successful decompression?

–Andi

Questions

evaluation: imputation

2 [about imputation] They compare the Levenshtein distances normalized by the true L2 data. I find their argumentation in principle convincing, but they don't give any real numbers. So if model B has a smaller NED than model A I accept that B is probably better than A, but how good is it actually? how far off are the Levenshtein distances in general? Or do we only need the general information?

–Andi

3 I like the way they evaluated their model, but is Normalized Edit Distance really "the ultimate test of the model's quality" as they say? It seems to me like you lose a lot of information with it. For instance, you wouldn't know by just looking at this one number if a few outliers (say with very high edit distances) are skewing the result.

–Becca

Questions

evaluation: gold standard

4 [The] context models also discover rules of sound changes: [T]o which extend does this happen compared to an already known set of such rules (lesser rules, exactly the same number, or maybe even more rules which have not been noticed so far) or does this depend on the data?

–Samantha

5 They said that it was extremely difficult to obtain a gold standard for Uralic. Wouldn't it be possible to use the already known sound change rules between languages as gold standard and evaluate the model by comparing in which degree it was able to find these expected rules?

–Samantha

Questions

discussion

1. I can't help wondering if [the paper's] applications aren't a bit too limited. If I understood it correctly, the model can evaluate existing etymological data but it can't really do anything without already selected cognate pairs? So I guess my question is, can it be used to make progress in identifying yet unknown relationships between languages?

   –Luana

2. If we would (-throw all the advantages of this objective/unbiased model overboard and-) add further linguistic assumptions, would it perform even better?

   –Le Duyen

Sources & Resources

H. Wettig, K. Reshetnikov and R. Yangarber
Using context and phonetic features in models of etymological sound change.
Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCL (pp. 108-116). Association for Computational Linguistics. 2012.

S. Russel and P. Norvig
Artificial Intelligence: A Modern Approach.
3rd ed., Prentice-Hall, 2010.

Nandos de Freitas
Decision Trees. CPSC540 Machine Learning. University of British Columbia.
[slides][video]

Encyclopaedia Britannica
Baltic-Finnic languages: family tree. [Illustration]
www.britannica.com/topic/Finnish-language?oasmId=2100