Feature-based Generalization

Jeffrey Heinz (heinz@udel.edu) and Cesar Koirala (koirala@udel.edu)

1. Contributions

1. Question the extent of feature-based generalization in Hayes and Wilson (2008)
2. Present a model which generalizes on the basis of phonological features. This model:
   (a) Is provably correct, and provably efficiently estimable
   (b) Integrates into Strictly Local (n-gram) or Strictly Piecewise models
   (c) Assumes statistical independence of individual features
   (d) Captures intuition that sounds with like features have like distributions

2. Expositional Feature Chart

We demonstrate with this feature system, but nothing hinges on it. This proposal accommodates privative and multi-valued feature systems.

	F	G
a	+	-
b	+	+
c	-	+

3. Expressivity of maxent models

Theorem 1. Every maxent model with featural constraints which describes a distribution is describable by one with segmental constraints.

Proof sketch.

Grammar 1		Grammar 2	
constraint	weight	constraint	weight
*[+F][+G]	w_1	*ab	w_1
*[-G][-F]	w_2	*ac	$w_1 + w_2$
		*bb	w_1
		*bc	w_1

For each constraint C with weight w (e.g. *[+X] or *[+X][+Y]), add w to the weight of all segmental sequences violating C, (adding more segmental constraints with weight w if needed). This procedure ensures maxent grammars G_1 and G_2 assign the same maxent scores to all words.

4. Features in Hayes & Wilson (2008)

The table shows the correlation (Spearman's r) between Hayes & Wilson's maxent grammars obtained with their learner on CMU English onsets and Schole's (1966) experimental results. Are their results due to features or use of complement classes?

Hayes and Wilson maxent models	r
features & complement classes	0.95
no features, complement classes	**0.94**
no features & no complement classes	0.88

5. Feature-based Strictly 2-Local (Bigram) Probability Distributions

Let $w = a_1 a_2 \dots a_n$ and let $\mathbb{F}$ be a feature system. Then

$$P(w) \stackrel{\text{def}}{=} P(a_1 \mid \#) \times P(a_2 \mid a_1) \times \dots \times P(a_n \mid a_{n-1}) \times P(\# \mid a_n) \quad (1)$$

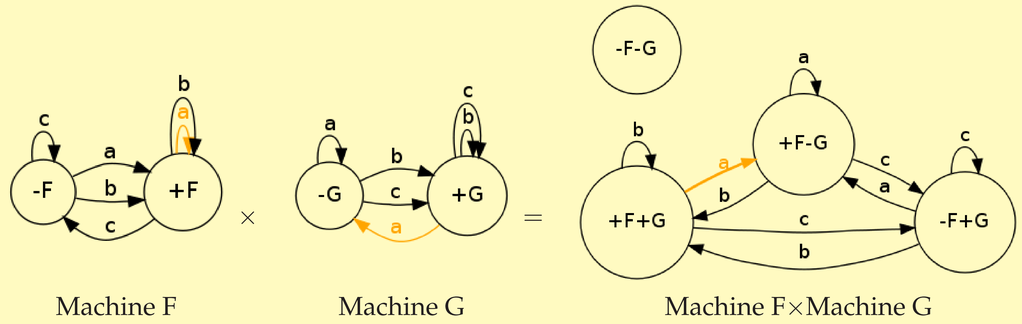
$$P(a \mid b) \stackrel{\text{def}}{=} P(a \mid \mathbb{F}_b) = \frac{\Pi_{f \in \mathbb{F}_b} P(a \mid f)}{\sum_{a' \in \Sigma \cup \{\#\}} \Pi_{f \in \mathbb{F}_b} P(a' \mid f)} \quad (2)$$

Theorem 2. Equations 1 and 2 define a well-formed probability distribution over Σ^*.

Corollary 1. There are $|\Sigma| \times |\mathbb{F}|$ parameters of the distribution. They are, for all $a \in \Sigma$ and $f \in \mathbb{F}$, $P(\text{a} \mid \text{f})$.

Corollary 2. These parameters can be estimated by finding the Maximum Likelihood Estimate using standard techniques for probabilistic finite-state machines (de la Higuera, in press).

Proof sketch of Theorem 2.



Machine F Machine G Machine F×Machine G

$$P(a \mid b) = P(a \mid [+F, +G]) = \frac{P(a \mid [+F]) \times P(a \mid [+G])}{\sum_{x \in \{a,b,c,\#\}} P(x \mid [+F]) \times P(x \mid [+G])}$$

For all $x \in \{a, b, c\}$, $P(\text{x} \mid \text{[+F]})$ and $P(\text{x} \mid \text{[+G]})$ are parameters of the model. Parameters are estimated by parsing the data sample with Machines F and G (and not their product), counting the transitions traversed, and then normalizing each state.

References and Acknowledgments

[1] Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:9–41.
[2] Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.
[3] de la Higuera, Colin. in press. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press.

This work was supported by a University of Delaware Research Fund grant for 2008-2009 and a General University Research grant from the University of Delaware for summer 2009. We thank James Rogers and Robert Wilder for valuable discussion.

6. Feature-Based Bigrams

Corpus = { aaab, caca , acab , cbb}

Segment-based generalization

P(x|y)		x			
		a	b	c	#
y	a	0.29	0.29	0.29	0.14
	b	**0.**	0.25	0.	0.75
	c	0.75	0.25	0.	0.
	#	0.5	**0.**	0.5	0.

Feature-based generalization

P(x|y)		x			
		a	b	c	#
y	a	0.22	0.33	0.22	0.22
	b	**0.25**	0.25	0.	0.5
	c	0.81	0.18	0.	0.
	#	0.33	**0.33**	0.33	0.

7. Word Initial Velar Nasals

Since nasals like [m,n] and velars like [k,g] begin words, the model infers [ŋ] ought to as well.

x	$P(x \mid \#)$
ŋ	0.0005
n	0.001
m	0.0014
k	0.0694
g	0.0291

(Features from Hayes and Wilson (2008) and the training data is theirs from CMU Dictionary.)

$$\text{Expected}(X) = P(X) \times N$$
$$\text{Expected}(\#\text{ŋ}) = 0.0005 \times 31,641 = 15.8$$
$$\text{Observed}(\#\text{ŋ}) = 0.$$

This is instructive!

> [There are] …two stages of evaluation: a preliminary initial assessment of probability of segment combinations and subsequent grammatical evaluation… Albright (2009)

We expect comparing expected values given by the feature-based distributions to observed values provides a platform for the inference of constraints.