Topics Today

1. Language Modeling and Machine Learning
2. Sparse Data and Handling it: Smoothing, Backoff, Interpolation
3. Generalizing N-gram models

Generalizing N-gram models

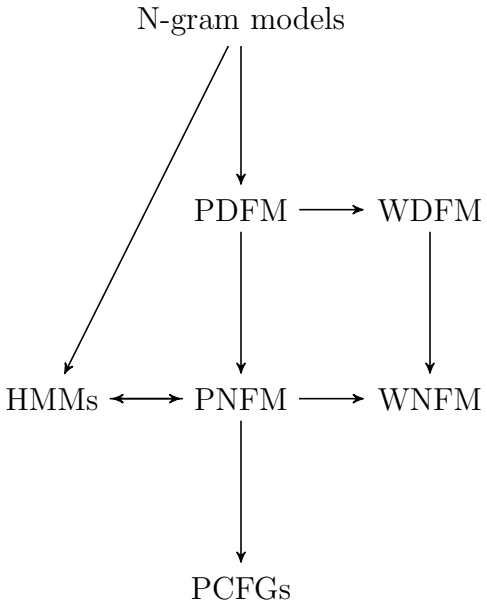

Figure 1: Language Models

Abbreviation	Model
PDFM	Probabilistic Deterministic Finite-state Machine
WDFM	Weighted Deterministic Finite-state Machine
PNFM	Probabilistic Non-deterministic Finite-state Machine
WNFM	Weighted Non-deterministic Finite-state Machine
HMM	Hidden Markov Models
PCFG	Probabilistic Context-Free Grammars

- Given a deterministic finite-state model, the maximum likelihood estimate (MLE) can be always be found.
- For PNFM, HMMs, and PCFGs, there are no guarantees but there are methods such as Expectation-Maximization which can find parameter values that work in practice.
- The probability distributions describable with HMMs are exactly the ones describable with PNFMs [Vidal et al., 2005a,b].

PDFA

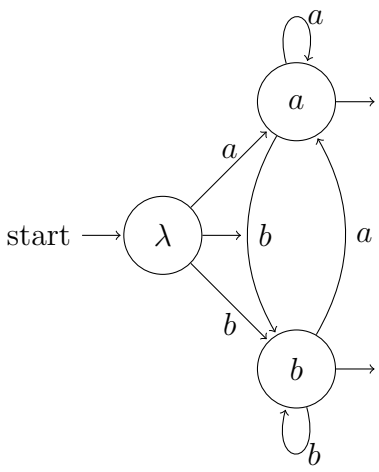

Figure 2: Example: Bigram Model

The MLE is obtained by passing the data through the deterministic finite-state machine and normalizing [Vidal et al., 2005a,b]. This is true for any deterministic finite-state model, not just ones representing n-gram. For example, suppose $D = \{ab, aabb\}$.

Parameters	counts	normalized
$\theta_{\rtimes a}$	2	1
$\theta_{\rtimes b}$	0	0
$\theta_{\rtimes \ltimes}$	0	0
θ_{aa}	1	1/3
θ_{ab}	2	2/3
$\theta_{a\ltimes}$	0	0
θ_{ba}	0	0
θ_{bb}	1	1/3
$\theta_{b\ltimes}$	2	2/3

References

Enrique Vidal, Franck Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. Probabilistic finite-state machines-part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005a. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.147.

Enrique Vidal, Frank Thollard, Colin de la Higuera, Francisco Casacuberta, and Rafael C. Carrasco. Probabilistic finite-state machines-part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1026–1039, 2005b. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.148.