Learning OT constraint rankings using a maximum entropy model

Goldwater, S. & Johnson, M. (2003)

Pilar Oplustil Gallegos

Topics in Natural Language Processing

University of Edinburgh

Contents

Introduction

- Context of the paper:

  Statistical models + phonology = computational phonology.

  Learning/acquisition of phonology.

- Goal of the paper:

  Apply a MaxEnt model to learn different types of phonological grammars.

  Compare it to Boersma’s probabilistic OT with Gradual Learning Algorithm (GLA) (Boersma, 1997).

Background: OT phonology

- Standard OT model (Prince & Smolensky, 2003):

Underlying form (phonemes): /deɪta/

$\downarrow$ **?**

Surface form (allophones): [deɪrə]

Background: OT phonology

- Standard OT model:

Lexicon → /deɪta/

↓ ?

[deɪrə]

Background: OT phonology

- Standard OT model:

Lexicon → /deɪta/

↓

Gen

… [deɪrə] [deɪta] [deɪtə] …

Background: OT phonology

- Standard OT model:

Lexicon → /deɪta/

↓

Gen

... [deɪrə] [deɪta] [deɪtə] ...

Con → H - Eval

C1 = no [t] between vowels, if the first one is stressed

C2 = keep the surface similar to the underlying form

Background: OT phonology

- Standard OT model:

Lexicon → /deɪta/

↓

Gen

↙ ↓ ↘

$\underbrace{\text{… [deɪrə] [deɪta] [deɪtə] …}}$

Con → H – Eval → [deɪrə] Optimal candidate

C1 = no [t] between vowels, if the first one is stressed

C2 = keep the surface similar to the underlying form

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right)$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y, x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y, x))$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right)$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right)$$

Background: MaxEnt model

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

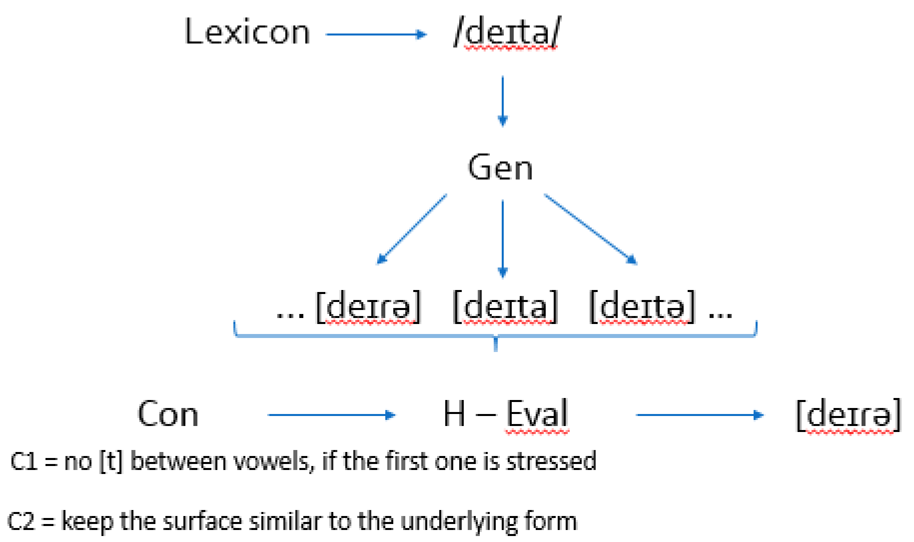
MaxEnt + OT

- How do we map the OT model with the MaxEnt model?

$$\Pr(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x)), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp(\sum_{i=1}^{m} w_i f_i(y,x))$$

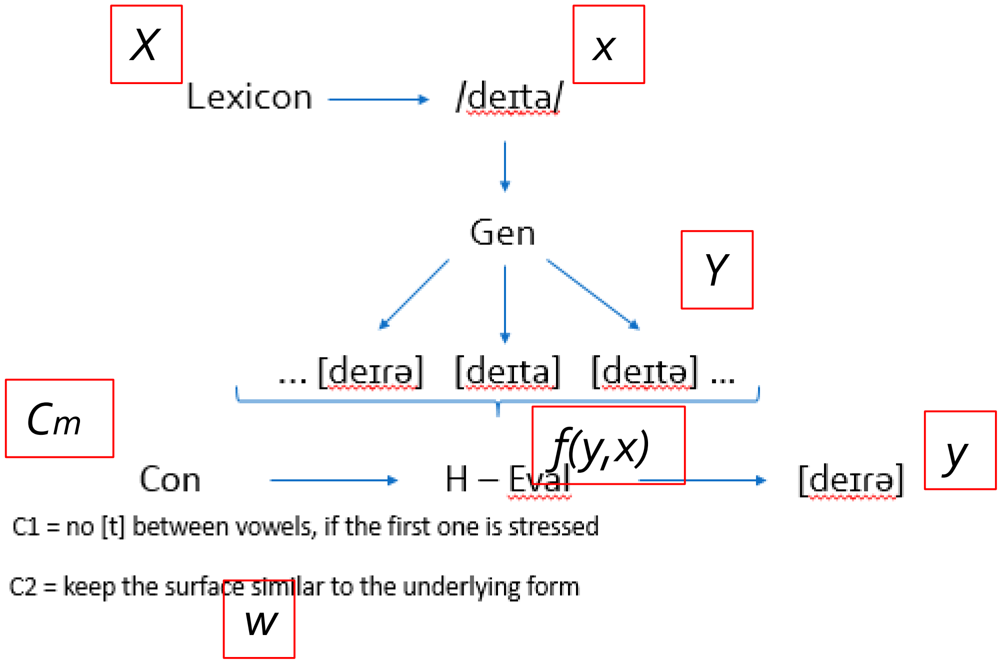
?



MaxEnt + OT

$$\Pr(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right), \text{ where}$$

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} \exp\left(\sum_{i=1}^{m} w_i f_i(y,x)\right)$$



Methodology

- Objective:

  Find the constraint ranking = parameters of the model (*w*)

  Model can learn the phonological grammar of the language.

1) Task 1: Categorical grammar
2) Task 2: Stochastic grammar

Compare the results to Boersma's model GLA

Supervised training

$$\mathrm{PL}_{\bar{w}}(\bar{y}|\bar{x}) = \prod_{j=1}^{n} \mathrm{Pr}_{\bar{w}}(Y = y_j | x(Y) = x_j)$$

Supervised training

- Optimise the parameters: Conjugate Gradient algorithm (Johnson, et al. 1999)

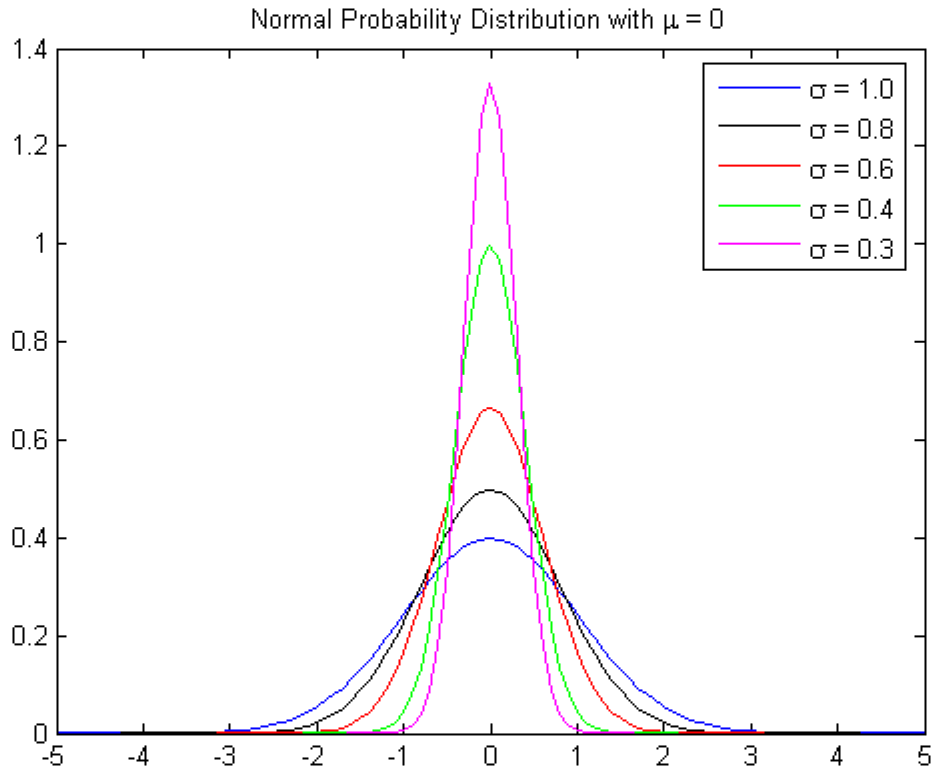
- Prevent overfitting: Gaussian distribution

- Final objective function:

$$\log \mathrm{PL}_{\bar{w}}(\bar{y}|\bar{x}) - \sum_{i=1}^{m} \frac{(w_i - \overset{= \text{zero}}{\mu_i})^2}{2\underset{= \sigma}{\sigma_i^2}}$$

Supervised training

- Setting σ



https://explorable.com/images/normal-probability-distribution.png

Task 1: learn a categorical grammar

Data: Wolof tongue-root grammar

- Set of constraints

*RTRHI: High vowels must not have a retracted tongue root (rtr).
*ATRLO: Low vowels must not have an advanced tongue root (atr).
PARSE[RTR]: If an input segment is [rtr], it must be realized as [rtr] in the output.
PARSE[ATR]: If an input segment is [atr], it must be realized as [atr] in the output.
GESTURE[CONTOUR]: Do not change from [rtr] to [atr], or vice versa, within a word.

- Set of 36 underlying forms
- Set of 10.000 surface forms

Task 1: results

- MaxEnt:

Constraint	Weight
*RtrHi	33.89
Parse[Rtr]	17.00
Gesture[Contour]	10.00
Parse[Atr]	3.53
*AtrLo	0.41

- Average error over input forms:

GLA: 0.2%

MaxEnt: 0.19%

0% (increase σ)

Task 2: learning a stochastic grammar

Data: Finnish genitive plurals

- Set of constraints:

C_1 (STRESS-TO-WEIGHT): Stressed syllables must be heavy.
C_2 (WEIGHT-TO-STRESS): Heavy syllables must bear stress.
C_3, C_4, C_5 (*Í, *Ó, *Á): No stressed syllables with underlying high/mid/low vowels.
C_6, C_7, C_8 (*Ĭ, *Ŏ, *Ă): No unstressed syllables with underlying high/mid/low vowels.
C_9 (*H.H): No consecutive heavy syllables.
C_{10} (*L.L): No consecutive light syllables.
C_{11} (*LAPSE): No consecutive unstressed syllables.

- 5698 tokens divided in 8 classes with different patterns of constraining candidates

Task 2: results

Class	Tokens	% Majority	GLA	MaxEnt
1	1097	100	99.5	99.6
2	1000	100	100.0	100.0
3	923	100	100.0	100.0
4	873	70.7	69.5	69.4
5	821	98.4	100	99.8
6	457	99.6	99.4	98.0
7	436	82.1	81.6	80.5
8	91	50.5	58.0	55.3

Task 2: results

Class	Tokens	% Majority	GLA	MaxEnt
1	1097	100	99.5	99.6
2	1000	100	100.0	100.0
3	923	100	100.0	100.0
4	873	70.7	69.5	69.4
5	821	98.4	100	99.8
6	457	99.6	99.4	98.0
7	436	82.1	81.6	80.5
8	91	50.5	58.0	55.3

Task 2: results

Class	Tokens	% Majority	GLA	MaxEnt
1	1097	100	99.5	99.6
2	1000	100	100.0	100.0
3	923	100	100.0	100.0
4	873	70.7	69.5	69.4
5	821	98.4	100	99.8
6	457	99.6	99.4	98.0
7	436	82.1	81.6	80.5
8	91	50.5	58.0	55.3

Discussion

1) Critics to GLA:

- They don't have a clear objective function to maximise.
- They apply an arbitrary learning scheme.
- They have two parameters to tune.
- Ad hoc model.

Discussion

2) MaxEnt advantages:

- General and mathematically well-motivated model.
- Initial State: interpret prior as initial state of acquisition.
- Can apply any algorithm to it, not just Conjugate Gradient.

Discussion

3) Generalization:

- The typical scheme training 90/10 testing, can’t be used here.
- The model is based on classes, not words.
- The constraints are already given (Hayes & Wilson, 2008).

References

- Goldwater, S. & Johnson, M. (2003) 'Learning OT constraint rankings using a maximum entropy model'. *In Proceedings of the Workshop on Variation within Optimality Theory*. pp. 111-120.

- Boersma, P. (1997) "How we learn variation, optionality and probability", *Proceedings* 21, 43-58.
- Collins, M. (2000) *Log-Linear Models*. http://www.cs.columbia.edu/~mcollins/loglinear.pdf
- Hayes, B. & Wilson, C. (2008). "A Maximum Entropy Model of Phonotactics and Phonotactic Learning", *Linguistic Inquiry*, Volume 39, Number 3, Summer 2008, 379–440
- Johnson, M., Geman, S., Canon, S., Chi, Z., & Riezler,S. (1999). "Estimators for stochastic 'unification-based' grammars". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.
- Prince, A. & Smolesky, P. (2003) "Optimality Theory: constraint interaction in Generative Grammar" in *Optimality theory in Phonology*, ed. Mc. Carthy, J. Blackwell: Oxford.

Thank you!

Questions?