Gender Assignment of French Nouns: A Maxent & SVM Model

Yiran Jiang (yj2462)

Background and Literature Review

- Cognitive science & Machine Learning:
  Modeling the learning process of languages.

- Linguistics point of view:
  Second language acquisition theory.

- Computer scientists point of view:
  Models: Connectionist model(Neural Network), Maximum Entropy Model. Support Vector Machine.

- Gender assignment:
  Reveal the cognition of inherent structure of the language.

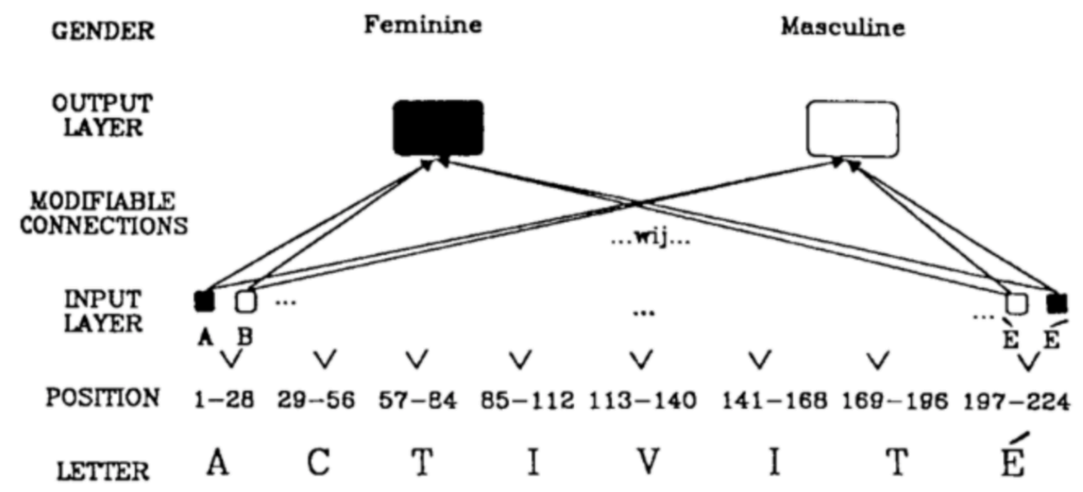


Fig. 1. Visualization of the connectionist model. Reprinted from *Sokolik*(1992)

Problem Formulation

- What does this paper do?
- A model for the cognitive process of gender assignment.

- How does this paper do?
- Features of language as input of models.

- What features?
- Orthography, Morphology, Syntax …

- How to extract features?
- Morphologic knowledge.
  Word vector training.

- How to extract orthography features?
- Last 1~2 letters of words.

- How to extract "context" features?
- Training word2vec on **<u>Stemmed</u>** Corpus.

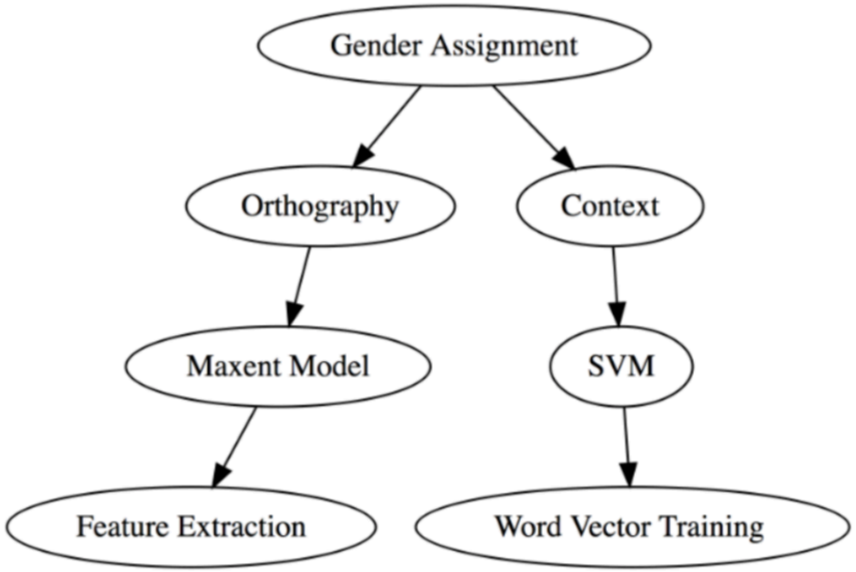


Fig. 2. The main structure and process of this paper

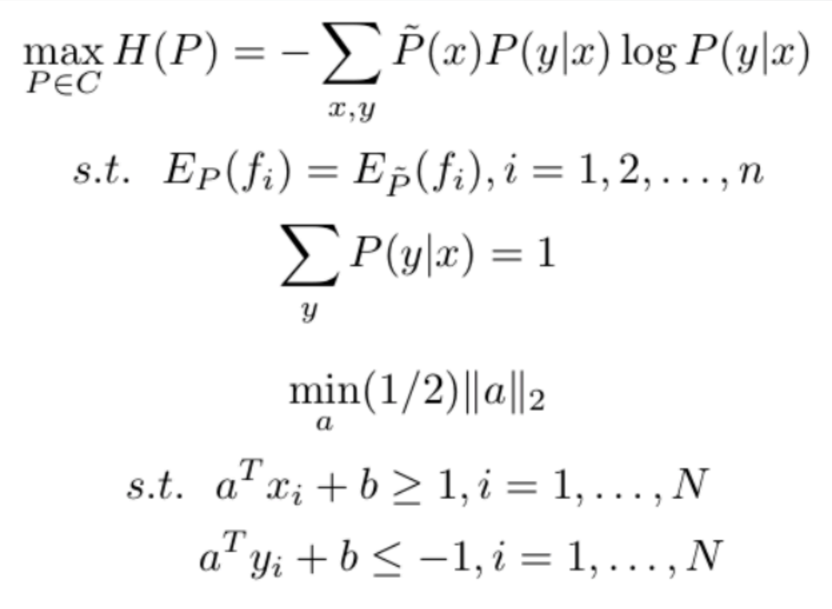
$$f(x, y) = \begin{cases} 1 & \text{if } y = en \text{ and } April \text{ follows } in \\ 0 & \text{otherwise} \end{cases}$$

Solutions

- What are the labels?
- Binary labels: feminine or masculine.

- How to solve the Model?
- **<u>Convex Optimization.</u>**

- How to solve the Convex Optimization problem?
- The Maxent: IIS(Improved Iterative Scaling)
- The SVM: Stochastic gradient descent on its dual.



$$\max_{P \in C} H(P) = -\sum_{x,y} \tilde{P}(x)P(y|x)\log P(y|x)$$

$$s.t. \quad E_P(f_i) = E_{\tilde{P}}(f_i), i = 1, 2, \ldots, n$$

$$\sum_y P(y|x) = 1$$

$$\min_a (1/2)\|a\|_2$$

$$s.t. \quad a^T x_i + b \geq 1, i = 1, \ldots, N$$

$$a^T y_i + b \leq -1, i = 1, \ldots, N$$

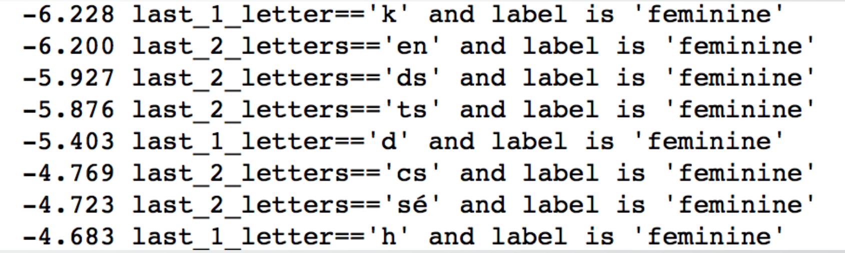
Experiment Results

- Package source:
  Wordvec2, NLTK toolkit

- Data source:
- http://www.lexique.org (Université Savoie Mont Blanc)

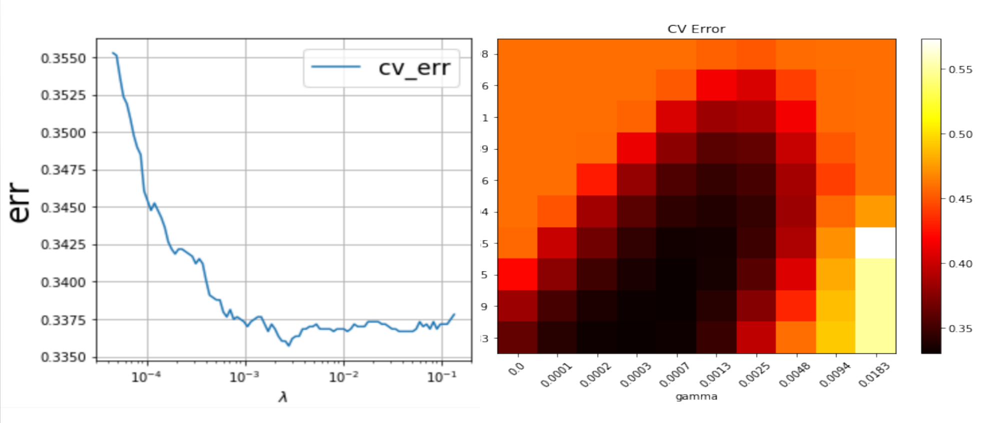
- Methodology:
- 10% test data, 5-fold CV for both model.

UNIVERSITÉ
SAVOIE
MONT BLANC



```
-6.228 last_1_letter=='k' and label is 'feminine'
-6.200 last_2_letters=='en' and label is 'feminine'
-5.927 last_2_letters=='ds' and label is 'feminine'
-5.876 last_2_letters=='ts' and label is 'feminine'
-5.403 last_1_letter=='d' and label is 'feminine'
-4.769 last_2_letters=='cs' and label is 'feminine'
-4.723 last_2_letters=='sé' and label is 'feminine'
-4.683 last_1_letter=='h' and label is 'feminine'
```

Model	Parameters	CV Error	Test Error
Maxent	None	0.171	0.178
SVM(Linear)	C = 0.0028	0.336	0.322
SVM(RBF Kernel)	C = 2.7183, gamma = 0.0003	0.350	0.434



Q&A

Thank you!

https://github.com/YiranJiang/Course-Project-for-EEOR-E4650
(Currently Empty)