# A semi-automatic structure learning method for language modeling

Vitor Pera

September 11, 2019

Faculdade de Engenharia da Universidade do Porto (FEUP)

## Outline

# Linguistic Classes Prediction Model (LCPM)

- Multiclass-dependent Ngram ($M > N > 1$)

$$
\begin{aligned}
P(\omega_t|\omega_{1:t-1}) &= \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{1:t-1}) P(c_t|\omega_{1:t-1}) \\
&\approx \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1})
\end{aligned}
$$

# Linguistic Classes Prediction Model (LCPM)

- Multiclass-dependent Ngram ($M > N > 1$)

$$P(\omega_t|\omega_{1:t-1}) = \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{1:t-1}) P(c_t|\omega_{1:t-1})$$
$$\approx \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1})$$

- LCPM (FLM formalism)

$$P(c_t|c_{t-M+1:t-1}) \xrightarrow{c \leftrightarrow f^{1:K}} P(f_t^{1:K}|f_{t-M+1:t-1}^{1:K})$$

# Linguistic Classes Prediction Model (LCPM)

- Multiclass-dependent Ngram ($M > N > 1$)

$$P(\omega_t|\omega_{1:t-1}) = \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{1:t-1}) P(c_t|\omega_{1:t-1})$$
$$\approx \sum_{c_t \in C(\omega_t)} P(\omega_t|c_t, \omega_{t-N+1:t-1}) P(c_t|c_{t-M+1:t-1})$$

- LCPM (FLM formalism)

$$P(c_t|c_{t-M+1:t-1}) \xrightarrow{c \leftrightarrow f^{1:K}} P(f_t^{1:K}|f_{t-M+1:t-1}^{1:K})$$

- LCPM structure learning (Goal)
  - accurate and simple
  - two steps method

# LCPM's Structure Learning Method - Step 1: Intro

- Given
  - The need for a LCPM to compute $P(f_t^{1:K} | f_{t-M+1:t-1}^{1:K})$ (factors not known, yet)
  - Common knowledge on Linguistics
  - Full knowledge of the specific language interface

# LCPM's Structure Learning Method - Step 1: Intro

- Given
  - The need for a LCPM to compute $P(f_t^{1:K} | f_{t-M+1:t-1}^{1:K})$ (factors not known, yet)
  - Common knowledge on Linguistics
  - Full knowledge of the specific language interface
- Solve (non-automatically)
  - Which linguistic features use?
  - Which linguistic features exhibit some special statistical independence property?

# LCPM's Structure Learning Method - Step 1: Procedure

1. Choose the linguistic features ($\rightarrow f^{1:K}$)
   - Informative to model $P(\omega_t | f_t^{1:K}, \omega_{t-N+1:t-1})$
   - Adequate to data resources (annotation and robustness)

# LCPM's Structure Learning Method - Step 1: Procedure

**1.** Choose the linguistic features ($\rightarrow f^{1:K}$)

- Informative to model $P(\omega_t | f_t^{1:K}, \omega_{t-N+1:t-1})$
- Adequate to data resources (annotation and robustness)

**2.** Make the (credible) assumption:

$f_t^n$ is statistically independent of any other factors, given its own history, iff $1 \leq n \leq J$

(accordingly, split $f^{1:K} \rightarrow f^{1:J} +\!\!+ f^{J+1:K}, \quad 1 \leq J < K$)

# LCPM's Structure Learning Method - Step 1: Procedure

1. Choose the linguistic features ($\rightarrow f^{1:K}$)
   - Informative to model $P(\omega_t | f_t^{1:K}, \omega_{t-N+1:t-1})$
   - Adequate to data resources (annotation and robustness)

2. Make the (credible) assumption:

   $f_t^n$ is statistically independent of any other factors, given its own history, iff $1 \leq n \leq J$

   (accordingly, split $f^{1:K} \rightarrow f^{1:J} +\!\!+ f^{J+1:K}, \quad 1 \leq J < K$)

LCPM factorization

$$\left[\prod_{i=1}^{J} P(f_t^i | f_{t-M+1:t-1}^i)\right] \underbrace{P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K})}_{\text{Step 2}}$$

## LCPM's Structure Learning Method - Step 1: Example

Given some application and a corpus annotated by multiple tags

1. Admit the following tags are judged as the most appropriate:
   - *Part-of-speech* (POS)
   - Semantic tag (ST)
   - Gender inflection (GI)

## LCPM's Structure Learning Method - Step 1: Example

Given some application and a corpus annotated by multiple tags

1. Admit the following tags are judged as the most appropriate:
   - *Part-of-speech* (POS)
   - Semantic tag (ST)
   - Gender inflection (GI)
2. Assuming that from these three LFs only ST can be predicted based uniquely on its own history:
   - ST $\rightarrow f^1$
   - (POS,GI) $\rightarrow f^{2:3}$

## LCPM's Structure Learning Method - Step 1: Example

Given some application and a corpus annotated by multiple tags

1. Admit the following tags are judged as the most appropriate:
   - *Part-of-speech* (POS)
   - Semantic tag (ST)
   - Gender inflection (GI)
2. Assuming that from these three LFs only ST can be predicted based uniquely on its own history:
   - ST $\rightarrow f^1$
   - (POS,GI) $\rightarrow f^{2:3}$

Results the LCPM approximation:

$$P(f_t^{1:3}|f_{t-M+1:t-1}^{1:3}) \approx P(f_t^1|f_{t-M+1:t-1}^1)P(f_t^{2:3}|f_t^1, f_{t-M+1:t-1}^{1:3})$$

# LCPM's Structure Learning Method - Step 2: Intro

- Goal is to learn the structure of statistical model to compute $P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$, more precisely ...

# LCPM's Structure Learning Method - Step 2: Intro

- Goal is to learn the structure of statistical model to compute $P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$, more precisely ...

- Determine automatically $Z \subset f_{t-M+1:t-1}^{1:K}$ such that
  - $|Z|$ is fixed and $|Z| << |f_{t-M+1:t-1}^{1:K}|$ (robustness constraint)
  - and $P(f_t^{J+1:K} | f_t^{1:J}, Z)$ approximates the original conditional probabilities according to Information Theory based criteria

## LCPM's Structure Learning Method - Step 2: Intro

- Goal is to learn the structure of statistical model to compute $P(f_t^{J+1:K} | f_t^{1:J}, f_{t-M+1:t-1}^{1:K})$, more precisely ...
- Determine automatically $Z \subset f_{t-M+1:t-1}^{1:K}$ such that
  - $|Z|$ is fixed and $|Z| << |f_{t-M+1:t-1}^{1:K}|$ (robustness constraint)
  - and $P(f_t^{J+1:K} | f_t^{1:J}, Z)$ approximates the original conditional probabilities according to Information Theory based criteria

Notation simplification (hereafter):

$$X = f_t^{1:J}; \quad Y = f_t^{J+1:K}; \quad Z \subset W = f_{t-M+1:t-1}^{1:K}; \rightarrow P(Y|X, Z)$$

## LCPM's SL Method - Step 2: Rules to determine $Z$

- Information Theory measures
  - Conditional entropy, $H(Y|X)$
  - Conditional mutual information (CMI), $I(Y;Z|X)$
  - Cross-context conditional mutual information (CCCMI), $I_{X_l}(Y;Z|X_m)$

## LCPM's SL Method - Step 2: Rules to determine $Z$

- Information Theory measures
  - Conditional entropy, $H(Y|X)$
  - Conditional mutual information (CMI), $I(Y;Z|X)$
  - Cross-context conditional mutual information (CCCMI), $I_{X_l}(Y;Z|X_m)$
- Possible/experimented rules ($\rightarrow P(Y|X,Z)$ w/ $Z \subset W$)
  - To discard $Z^*$

    If $I(Y;Z^*|X) < \eta\, H(Y|X)$ then $Z^*$ is non-relevant

## LCPM's SL Method - Step 2: Rules to determine $Z$

- Information Theory measures
  - Conditional entropy, $H(Y|X)$
  - Conditional mutual information (CMI), $I(Y;Z|X)$
  - Cross-context conditional mutual information (CCCMI), $I_{X_l}(Y;Z|X_m)$
- Possible/experimented rules ($\rightarrow P(Y|X,Z)$ w/ $Z \subset W$)
  - To discard $Z^*$

    If $I(Y;Z^*|X) < \eta\, H(Y|X)$ then $Z^*$ is non-relevant
  - To determine $Z^*$

$$Z^* = \underset{\substack{Z \subset W \\ |Z| = \zeta}}{\operatorname{argmax}} \{I(Y;Z|X)\}$$

# LCPM's SL Method - Step 2: Rules to determine $Z$ (cont.)

- Rule to determine $Z^*$ using the *"Utility"* measure $N_\lambda$

$$Z^* = \underset{\substack{Z \subset W \\ |Z| = \zeta}}{\operatorname{argmax}} \{N_\lambda(Y; Z|X)\}, \quad 0 < \lambda \leq 1$$

# LCPM's SL Method - Step 2: Rules to determine $Z$ (cont.)

- Rule to determine $Z^*$ using the *"Utility"* measure $N_\lambda$

$$Z^* = \underset{\substack{Z \subset W \\ |Z| = \zeta}}{\operatorname{argmax}} \{N_\lambda(Y; Z|X)\}, \quad 0 < \lambda \leq 1$$

where $N_\lambda(Y; Z|X)$ represents

$$\sum_{X_m} P(X_m) \Big[ I(Y; Z|X_m) - \lambda \sum_{X_l \neq X_m} P(X_l) I_{X_l}(Y; Z|X_m) \Big]$$

## LCPM's SL Method - Step 2: Rules to determine $Z$ (cont.)

- Rule to determine $Z^*$ using the *"Utility"* measure $N_\lambda$

$$Z^* = \underset{\substack{Z \subset W \\ |Z| = \zeta}}{\operatorname{argmax}} \{N_\lambda(Y; Z|X)\}, \quad 0 < \lambda \leq 1$$

where $N_\lambda(Y; Z|X)$ represents

$$\sum_{X_m} P(X_m) \Big[ I(Y; Z|X_m) - \lambda \sum_{X_l \neq X_m} P(X_l) I_{X_l}(Y; Z|X_m) \Big]$$

and $I_{X_l}(Y; Z|X_m)$ represents

$$\sum_Y \sum_Z P(Y, Z|X_l) \log \frac{P(Y, Z|X_m)}{P(Y|X_m) P(Z|X_m)}$$

# LCPM's SL Method - Step 2: Example

Problem: Choose $Z^1$ or $Z^2$ to model $P(Y|X, Z)$;

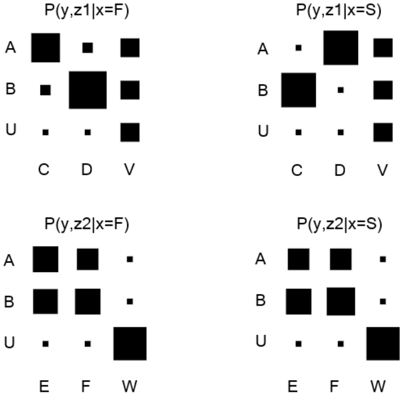
$X \in \{F, S\}$, $Y \in \{A, B, U\}$, $Z^1 \in \{C, D, V\}$, $Z^2 \in \{E, F, W\}$

## LCPM's SL Method - Step 2: Example

Problem: Choose $Z^1$ or $Z^2$ to model $P(Y|X,Z)$;

$$X \in \{F,S\},\ \ Y \in \{A,B,U\},\ \ Z^1 \in \{C,D,V\},\ \ Z^2 \in \{E,F,W\}$$

Data: $P(X=F) = P(X=S)$

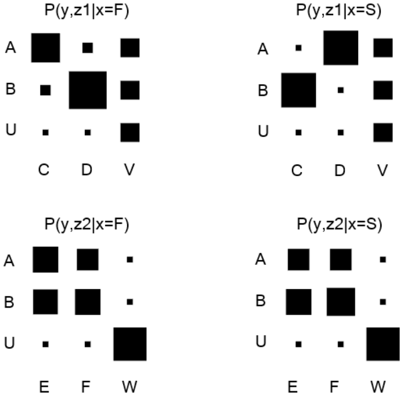


# LCPM's SL Method - Step 2: Example

Problem: Choose $Z^1$ or $Z^2$ to model $P(Y|X,Z)$;

$$X \in \{F,S\},\ \ Y \in \{A,B,U\},\ \ Z^1 \in \{C,D,V\},\ \ Z^2 \in \{E,F,W\}$$

Data: $P(X=F) = P(X=S)$




*"Utility"* & Solutions:

$$N_0(Y;Z^1|X) < N_0(Y;Z^2|X)$$

(near equality)

$$\therefore \lambda = 0 \Rightarrow \text{choose } Z^2$$

$$N_1(Y;Z^1|X) > N_1(Y;Z^2|X)$$

$$\therefore \lambda = 1 \Rightarrow \text{choose } Z^1$$

# LCPM's SL Method - Step 2: ALgorithm to define $Z$

**Input:** $f^{1:K}_{t-M+1:t}$, $J$, $K$, $M$, $\zeta$, $\lambda$, $\gamma$, $\eta$, $Data$
**Output:** Set of factors: $Z$
**for** *each* $z \in f^{1:K}_{t-M+1:t-1}$ **do** `// factors relevance`
  **if** $I(f^{J+1:K}_t; z|f^{1:J}_t) < \gamma H(f^{J+1:K}_t|f^{1:J}_t)$ **then**
    Remove $z$ from $f^{1:K}_{t-M+1:t-1}$
  **end**
**end**
Sort $f^{1:K}_{t-M+1:t-1}$ by descending order of $N_{(\lambda)}(f^{J+1:K}_t; z|f^{1:J}_t)$
$Z \leftarrow \emptyset$
**repeat** `// factors redundancy`
  $z \leftarrow$ next non-processed element in $f^{1:K}_{t-M+1:t-1}$
  **if** $I(f^{J+1:K}_t; z|f^{1:J}_t) > \eta I(z; r|f^{1:J}_t), \forall r \in Z$ **then**
    Add $z$ to $Z$
  **end**
**until** $|Z| = \zeta$ *or all elements of* $f^{1:K}_{t-M+1:t-1}$ *are processed*
Output $Z$

## Preliminary Results

- Text corpus (vocab-size$\approx 200K$) which annotations include:
  m - Part-of-speech (#13: ADJ, ADV, ...)
  g - Gender inflection (#3: M, F, N)
  n - Number inflection (#3: S, P, U)

  Select $Z \subset W = \{n_t, m_{t-1}, g_{t-1}, n_{t-1}, m_{t-2}, g_{t-2}, n_{t-2}, ...\}$
  maximizing the *Utility*, $N_\lambda(g_t; Z|m_t) \quad (\rightarrow P(g_t|m_t, Z))$

- Text corpus (vocab-size$\approx 200K$) which annotations include:
  m - Part-of-speech (#13: ADJ, ADV, ...)
  g - Gender inflection (#3: M, F, N)
  n - Number inflection (#3: S, P, U)

  Select $Z \subset W = \{n_t, m_{t-1}, g_{t-1}, n_{t-1}, m_{t-2}, g_{t-2}, n_{t-2}, ...\}$
  maximizing the *Utility*, $N_\lambda(g_t; Z|m_t) \quad (\rightarrow P(g_t|m_t, Z))$

- Results

| Cases | $\lambda$ | $Z$ sorted by decreasing $N_\lambda$ |
|---|---|---|
| $g \neq$ N and $n \neq$ U | 0 | $\{g_{t-1}, g_{t-2}, m_{t-1}, \ldots\}$ |
| | 1 | $\{g_{t-1}, m_{t-1}, g_{t-2}, \ldots\}$ |
| Whole data | 0 | $\{n_t, g_{t-1}, m_{t-1}, \ldots\}$ |
| | 1 | $\{g_{t-1}, n_{t-2}, g_{t-2}, \ldots\}$ |

## Conclusions

- Method for learning LCPM structure

## Conclusions

- Method for learning LCPM structure
- Guidelines:
  Seek accurate and simple structure (FLM approach: keep just the relevant and non-redundant factors and dependencies)

# Conclusions

- Method for learning LCPM structure
- Guidelines:
  Seek accurate and simple structure (FLM approach: keep just the relevant and non-redundant factors and dependencies)
- Process:
  Step 1 - manually set initial structure (Linguistic knowledge)
  Step 2 - automatically "prune" structure (data-driven algorithm based on Information Theory concepts)

## Conclusions

- Method for learning LCPM structure
- Guidelines:
  Seek accurate and simple structure (FLM approach: keep just the relevant and non-redundant factors and dependencies)
- Process:
  Step 1 - manually set initial structure (Linguistic knowledge)
  Step 2 - automatically "prune" structure (data-driven algorithm based on Information Theory concepts)
- Preliminary results seem promising; larger experiments are needed to get conclusive results

# References


1. J. Bilmes, “Natural Statistical Models for Automatic Speech Recognition”, PhD Thesis, 1999, Berkley, Cal, Intl. Computer Science Institute.
2. K. Kirchhoff, J. Bilmes, and K. Duh, “Factored Language Model Tutorial”, Tech. Report, 2008, Dept. Electrical Engineering, Univ. of Washington.
3. Helmut Schmid, “Improvements in Part-of-Speech Tagging with an Application to German”, Proc. ACL SIGDAT-Workshop, 1995. Dublin, Ireland.
4. D. Santos, and P. Rocha, “Evaluating CETEMPúblico, a free resource for Portuguese”, Proc. 39th Annual Meeting of the Association for Computational Linguistics, 2001, Stroudsburg, PA, USA.