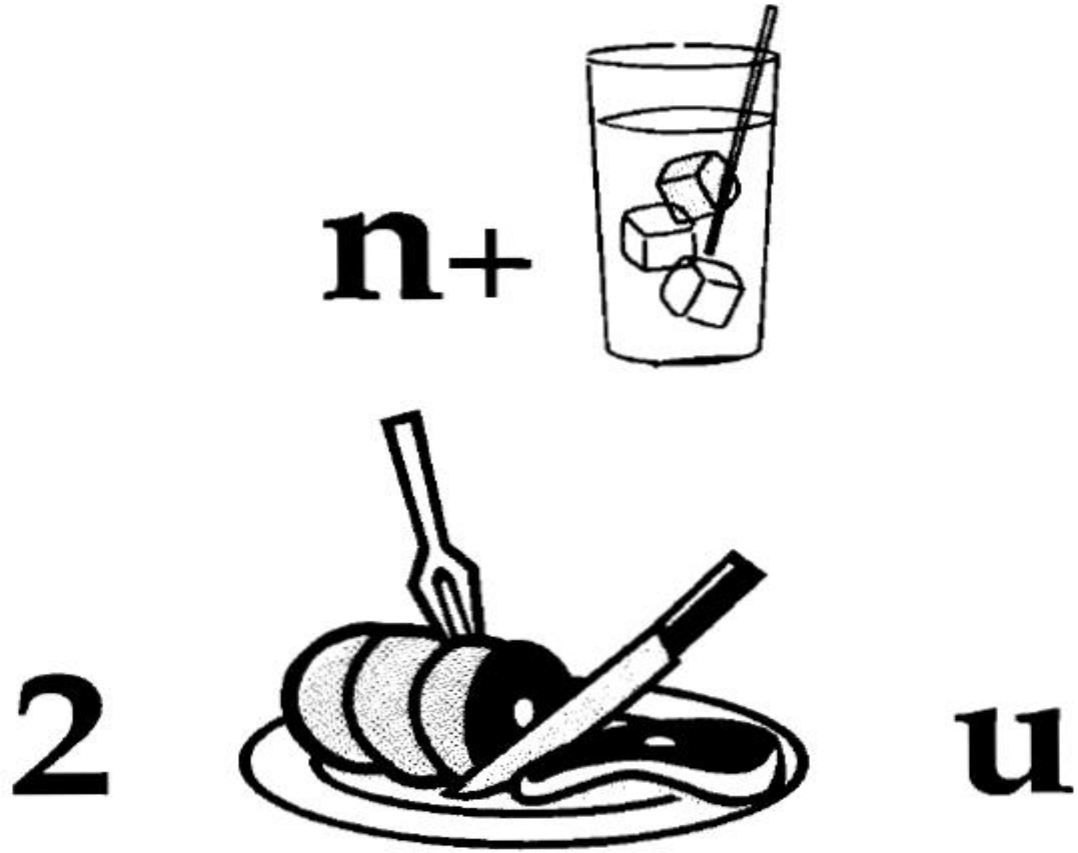
Constrained Conditional Models

Mausam

Slides by Ming-Wei Chang, Nick Rizzolo, Dan Roth, Dan Jurafsky

Nice to Meet You



ILP & Constraints Conditional Models (CCMs)

- Making global decisions in which several local interdependent decisions play a role.
- Informally:
  - Everything that has to do with constraints (and learning models)
- Forr
  -
  - s:
  - ons

Issues to attend to:

- **While we formulate the problem as an ILP problem, Inference can be done multiple ways**
  - Search; sampling; dynamic programming; SAT; ILP
- **The focus is on joint global inference**
- **Learning may or may not be joint.**
  - Decomposing models is often beneficial

- CCMs make predictions in the presence of /guided by constraints

Constraints Driven Learning and Decision Making

- Why Constraints?
  - **The Goal: Building a good NLP systems easily**
  - **We have prior knowledge at our hand**
    - How can we use it?
    - We suggest that knowledge can often be injected directly
      - **Can use it to guide learning**
      - **Can use it to improve decision making**
      - **Can use it to simplify the models we need to learn**

- How useful are constraints?
  - **Useful for supervised learning**
  - **Useful for semi-supervised & other label-lean learning paradigms**
  - **Sometimes more efficient than labeling data directly**

Motivation: IE via Hidden Markov Models

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis. DIKU , University of Copenhagen, May 1994 .

Prediction result of a trained HMM

[AUTHOR]	Lars Ole Andersen . Program analysis and
[TITLE]	specialization for the
[EDITOR]	C
[BOOKTITLE]	Programming language
[TECH-REPORT]	. PhD thesis .
[INSTITUTION]	DIKU , University of Copenhagen , May
[DATE]	1994 .

Unsatisfactory results !

Strategies for Improving the Results

- (Pure) Machine Learning Approaches
  - Higher Order HMM/CRF?
  - Increasing the window size?
  - Adding a lot of new features
    - Requires a lot of labeled examples
  - What if we only have a few labeled examples?

Increasing the model complexity

Can we keep the learned model simple and still make expressive decisions?

- Any other options?
  - Humans can immediately detect bad outputs
  - The output does not make sense

Information extraction without Prior Knowledge

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis. DIKU , University of Copenhagen, May 1994 .

Prediction result of a trained HMM

[AUTHOR]	Lars Ole Andersen . Program analysis and
[TITLE]	specialization for the
[EDITOR]	C
[BOOKTITLE]	Programming language
[TECH-REPORT]	. PhD thesis .
[INSTITUTION]	DIKU , University of Copenhagen , May
[DATE]	1994 .

Violates lots of natural constraints!

Examples of Constraints

- Each field must be a consecutive list of words and can appear at most once in a citation.
- State transitions must occur on punctuation marks.
- The citation can only start with *AUTHOR* or *EDITOR*.
- The words *pp., pages* correspond to *PAGE.*
- Four digits starting with 20xx and 19xx are *DATE*.
- Quotations can appear only in *TITLE*
- …….

Easy to express pieces of "knowledge"

Non Propositional; May use Quantifiers

Information Extraction with Constraints

- Adding constraints, we get correct results!
  - Without changing the model

- [AUTHOR] Lars Ole Andersen .
  [TITLE] Program analysis and specialization for the C Programming language .
  [TECH-REPORT] PhD thesis .
  [INSTITUTION] DIKU , University of Copenhagen ,
  [DATE] May, 1994 .

Constrained Conditional Models Allow:

- **Learning a simple model**
- **Make decisions with a more complex model**
- **Accomplished by directly incorporating constraints to bias/re-ranks decisions made by the simpler model**

Constrained Conditional Models (aka ILP Inference)

Penalty for violating the constraint.

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)})$$

(Soft) constraints component

Weight Vector for “local” models

Features, classifiers; log-linear models (HMM, CRF) or a combination

How far y is from a “legal” assignment

CCMs can be viewed as a general interface to easily combine domain knowledge with data driven statistical models

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an exact solution.

Search techniques are also possible

How to train?

Training is learning the objective Function.

How to exploit the structure to minimize supervision?

Features Versus Constraints

- ϕ_i : X $\times$ Y $\rightarrow$ R; C_i : X $\times$ Y $\rightarrow$ {0,1}; d: X $\times$ Y $\rightarrow$ R;
  - In principle, constraints and features can encode the same propeties
  - In practice, they are very different

- Features
  - Local , short distance properties – to allow tractable inference
  - Propositional (grounded):
  - E.g. True if: "the" followed by a Noun occurs in the sentence"
- Constraints
  - Global properties
  - Quantified, first order logic expressions
  - E.g.True if: "all y_is in the sequence y are assigned different values."

Indeed, used differently

Encoding Prior Knowledge

- Consider encoding the knowledge that:
  - Entities of type A and B cannot occur simultaneously in a sentence
- The "Feature" Way
  - Results in higher order HMM, CRF
  - May require designing a model tailored to knowledge/constraints
  - Large number of new features: might require more labeled data
  - Wastes parameters to learn indirectly knowledge we have.

Need more training data

- The Constraints Way
  - Keeps the model simple; add expressive constraints directly
  - A small set of constraints
  - Allows for decision time incorporation of constraints

A form of supervision

CCMs are Optimization Problems

- We pose inference as an optimization problem
  - Integer Linear Programming (ILP)
- Advantages:
  - *Keep model small; easy to learn*
  - *Still allowing expressive, long-range constraints*
  - Mathematical optimization is well studied
  - Exact solution to the inference problem is possible
  - Powerful off-the-shelf solvers exist
- Disadvantage:
  - The inference problem could be NP-hard

CCM Example

- Many works in NLP make use of constrained conditional models, implicitly or explicitly.
- Next we describe one example in detail.

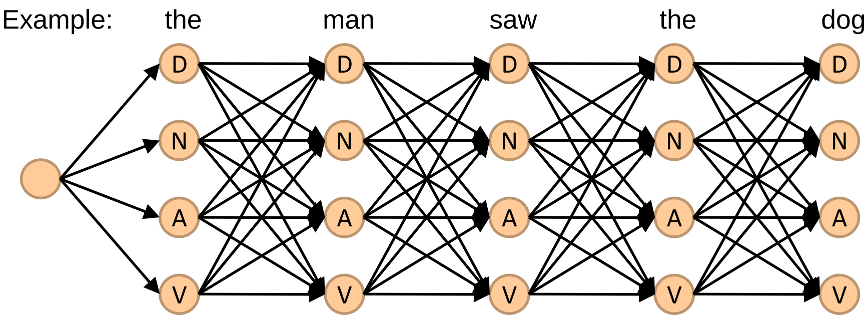
- Sequence Tagging
  - Adding long range constraints to a simple model

Example 1: Sequence Tagging

HMM / CRF:

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} P(y_0)P(x_0|y_0)\prod_{i=1}^{n-1} P(y_i|y_{i-1})P(x_i|y_i)$$

Example:



As an ILP:

L Inference Variables

$$\text{maximize} \quad \sum_{y\in\mathcal{Y}} \lambda_{0,y} 1_{\{y_0=y\}} + \sum_{i=1}^{n-1}\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} \lambda_{i,y,y'} 1_{\{y_i=y \,\wedge\, y_{i-1}=y'\}}$$

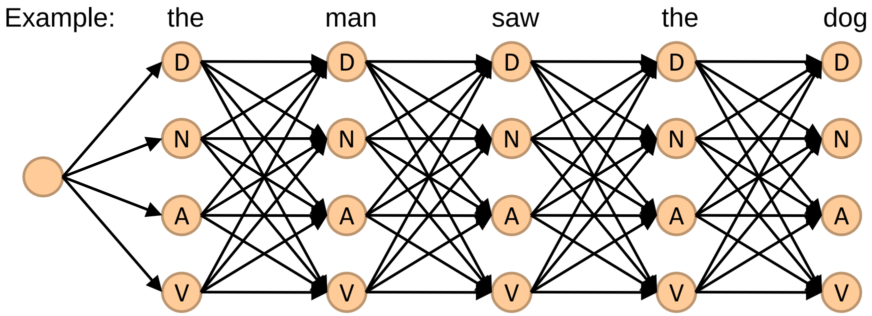
subject to

$$\lambda_{0,y} = \log(P(y)) + \log(P(x_0|y))$$
$$\lambda_{i,y,y'} = \log(P(y|y')) + \log(P(x_i|y))$$

Example 1: Sequence Tagging

HMM / CRF:

$$\mathbf{y}^* = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\, P(y_0)P(x_0|y_0)\prod_{i=1}^{n-1} P(y_i|y_{i-1})P(x_i|y_i)$$



As an ILP:

$$\text{maximize} \quad \sum_{y\in\mathcal{Y}} \lambda_{0,y} 1_{\{y_0=y\}} + \sum_{i=1}^{n-1}\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} \lambda_{i,y,y'} 1_{\{y_i=y \,\wedge\, y_{i-1}=y'\}}$$

$$\lambda_{0,y} = \log(P(y)) + \log(P(x_0|y))$$

$$\lambda_{i,y,y'} = \log(P(y|y')) + \log(P(x_i|y))$$

$$\text{subject to} \quad \sum_{y\in\mathcal{Y}} 1_{\{y_0=y\}} = 1$$

Discrete predictions

$$1_{\{y_0=\text{“NN”}\}} = 1$$

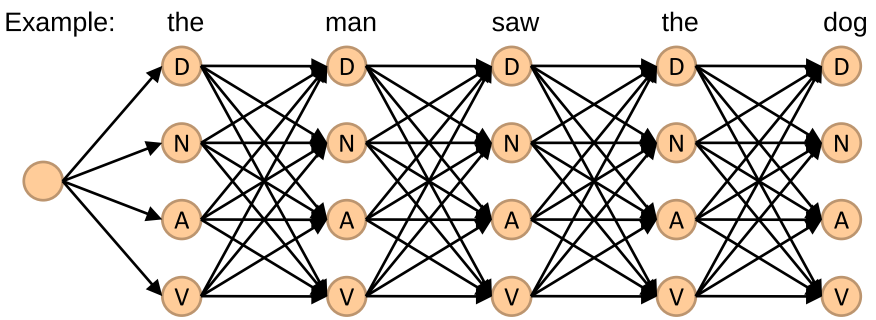
$$1_{\{y_0=\text{“VB”}\}} = 1$$

$$1_{\{y_0=\text{“JJ”}\}} = 1$$

Example 1: Sequence Tagging

HMM / CRF:

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} P(y_0)P(x_0|y_0) \prod_{i=1}^{n-1} P(y_i|y_{i-1})P(x_i|y_i)$$



As an ILP:

$$\text{maximize} \quad \sum_{y \in \mathcal{Y}} \lambda_{0,y} 1_{\{y_0 = y\}} + \sum_{i=1}^{n-1} \sum_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} \lambda_{i,y,y'} 1_{\{y_i = y \ \wedge \ y_{i-1} = y'\}}$$

$$\lambda_{0,y} = \log(P(y)) + \log(P(x_0|y))$$
$$\lambda_{i,y,y'} = \log(P(y|y')) + \log(P(x_i|y))$$

subject to

$$\sum_{y \in \mathcal{Y}} 1_{\{y_0 = y\}} = 1$$

Discrete predictions

$$\forall y, \quad 1_{\{y_0 = y\}} = \sum_{y' \in \mathcal{Y}} 1_{\{y_0 = y \ \wedge \ y_1 = y'\}}$$

$$\forall y, i > 1 \quad \sum_{y' \in \mathcal{Y}} 1_{\{y_{i-1} = y' \ \wedge \ y_i = y\}} = \sum_{y'' \in \mathcal{Y}} 1_{\{y_i = y \ \wedge \ y_{i+1} = y''\}}$$

Feature consistency

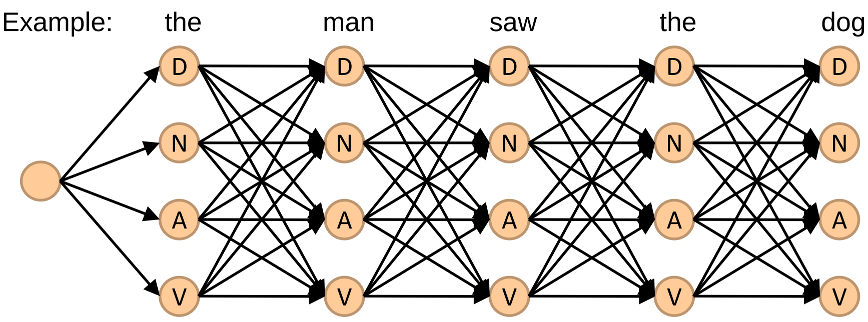
$$1_{\{y_0 = \text{“NN”}\}} = 1$$
$$1_{\{y_0 = \text{“DT”} \ \wedge \ y_1 = \text{“JJ”}\}} = 1$$

$$1_{\{y_0 = \text{“DT”} \ \wedge \ y_1 = \text{“JJ”}\}} = 1$$
$$1_{\{y_1 = \text{“NN”} \ \wedge \ y_2 = \text{“VB”}\}} = 1$$

Example 1: Sequence Tagging

HMM / CRF:

$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} P(y_0)P(x_0|y_0)\prod_{i=1}^{n-1} P(y_i|y_{i-1})P(x_i|y_i)$$



As an ILP:

$$\text{maximize} \quad \sum_{y\in\mathcal{Y}} \lambda_{0,y} 1_{\{y_0=y\}} + \sum_{i=1}^{n-1}\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} \lambda_{i,y,y'} 1_{\{y_i=y\ \wedge\ y_{i-1}=y'\}}$$

$$\lambda_{0,y} = \log(P(y)) + \log(P(x_0|y))$$
$$\lambda_{i,y,y'} = \log(P(y|y')) + \log(P(x_i|y))$$

subject to

$$\sum_{y\in\mathcal{Y}} 1_{\{y_0=y\}} = 1$$

Discrete predictions

$$\forall y, \quad 1_{\{y_0=y\}} = \sum_{y'\in\mathcal{Y}} 1_{\{y_0=y\ \wedge\ y_1=y'\}}$$

$$\forall y, i > 1 \quad \sum_{y'\in\mathcal{Y}} 1_{\{y_{i-1}=y'\ \wedge\ y_i=y\}} = \sum_{y''\in\mathcal{Y}} 1_{\{y_i=y\ \wedge\ y_{i+1}=y''\}}$$

Feature consistency

$$1_{\{y_0=\text{“V”}\}} + \sum_{i=1}^{n-1}\sum_{y\in\mathcal{Y}} 1_{\{y_{i-1}=y\ \wedge\ y_i=\text{“V”}\}} \geq 1$$

There must be a verb!

Solvers

- All applications presented so far used ILP for inference.
- People used different solvers
  - Xpress-MP
  - GLPK
  - lpsolve
  - R
  - Mosek
  - CPLEX

- Other search-based algorithms can also be used

Training Constrained Conditional Models

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y) - \sum_{i=1}^{K} \boxed{\rho_i}\, d(y, 1_{C_i(x)})$$

- Learning model
  - Independently of the constraints (L+I)
  - Jointly, in the presence of the constraints (IBT)
  - Decomposed to simpler models
- Learning constraints' penalties
  - Independently of learning the model
  - Jointly, along with learning the model
- Dealing with lack of supervision
  - Constraints Driven Semi-Supervised learning (CODL)
  - Indirect Supervision
- Learning Constrained Latent Representations

Soft Constraints

$$-\sum_{i=1}^{K} \rho_k \underline{d(y, 1_{C_i(x)})}$$

- Hard Versus Soft Constraints
  - Hard constraints: Fixed Penalty $\rho_i = \infty$
  - Soft constraints: Need to set the penalty

- Why soft constraints?
  - Constraints might be violated by gold data
  - Some constraint violations are more serious
  - An example can violate a constraint multiple times!
  - Degree of violation is only meaningful when constraints are soft!

Examples of Constraints

- Each field must be a consecutive list of words and can appear at most once in a citation.

- State transitions must occur on punctuation marks.

- The citation can only start with *AUTHOR* or *EDITOR*.

- The words *pp., pages* correspond to *PAGE.*
- Four digits starting with 20xx and 19xx are *DATE*.
- Quotations can appear only in *TITLE*
- ....

Degree of Violations

One way: Count how many times the assignment y violated the constraint

$$d(y, 1_{C(x)}) = \sum_{j=1}^{T} \phi_C(y_j)$$

$$\phi_C(y_j) = \begin{cases} 1 \text{ - if assigning } y_i \text{ to } x_i \text{ violates the constraint } C \text{ with respect to assignment } (x_1,..,x_{i-1};y_1,...,y_{i-1}) \\ 0 \text{ - otherwise} \end{cases}$$

State transition must occur on punctuations.

$\forall i, y_{i-1} \neq y_i \Rightarrow x_{i-1}$ is a punctuation

Lars	Ole	Andersen	.
AUTH	BOOK	EDITOR	EDITOR
$\Phi_c(y_1)=0$	$\Phi_c(y_2)=1$	$\Phi_c(y_3)=1$	$\Phi_c(y_4)=0$

$\sum \Phi_c(y_j) = 2$

Strategy: Independently of learning the model

- Model: (First order) Hidden Markov Model $P_\theta(x, y)$
- Constraints: long distance constraints
  - The i-th the constraint: C_i
  - The probability that the i-th constraint is violated $P(C_i = 1)$
- The learning problem
  - Given labeled data, estimate θ and $P(C_i = 1)$
  - For one labeled example,

$$\textsc{Score}(x, y) = \text{HMM Probability} \times \text{ Constraint Violation Score}$$

  - Training: Maximize the score of all labeled examples!

$$\Omega(\mathbf{x}^j, \mathbf{y}^j) = \text{HMM Probability} \times \text{Constraint Violation Score}$$

$$= P_\Theta(\mathbf{x}^j, \mathbf{y}^j) \prod_{k=1}^{m} \prod_{i=1}^{T_j} P(C_k = 1)^{c_{k,i}^j} P(C_k = 0)^{1-c_{k,i}^j},$$

where Θ are the parameters of the HMM, T_j represents the number of tokens in the sentence $\mathbf{x}^j$, $c_{k,i}^j$ is a binary variable equal to 1 if the label assignment to y_i^j violates the constraint C_k with respect to partial assignment $\mathbf{y}_{[1\ldots i-1]}^j$, and $C_k = 1$ indicates the event that the constraint C_k is violated.

$$\log \Omega(\mathbf{x}^j, \mathbf{y}^j) \equiv \hat{f}_{w,\rho}(\mathbf{x}^j, \mathbf{y}^j)$$

$$= \mathbf{w}^T \Phi(\mathbf{x}^j, \mathbf{y}^j) + \sum_{k=1}^{m} \log \frac{P(C_k = 1)}{P(C_k = 0)} \sum_{i}^{T_j} c_{k,i}^j + c$$

$$= \mathbf{w}^T \Phi(\mathbf{x}^j, \mathbf{y}^j) - \sum_{k=1}^{m} \rho_k d_{C_k}(\mathbf{x}^j, \mathbf{y}^j) + c,$$

where $\rho_k = -\log \frac{P(C_k=1)}{P(C_k=0)}$, $d_{C_k}(\mathbf{x}^j, \mathbf{y}^j) = \sum_i^{T_j} c_{k,i}^j$

Strategy: Independently of learning the model (cont.)

$$\textsc{Score}(x,y) = \text{HMM Probability} \times \text{ Constraint Violation Score}$$

- The new score function is a CCM!
  - Setting $\rho_i = -\log \frac{P(C_i = 1)}{P(C_i = 0)}$
  - New score:

$$\log \textsc{Score}(x,y) = \lambda \cdot F(x,y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)}) + c$$

- Maximize this new scoring function on labeled data
  - Learn a HMM separately
  - Estimate $P(C_i = 1)$ separately by counting how many times the constraint is violated by the training data!
- A formal justification for optimizing the model and the penalty weights separately!

Summary

- Constrained Conditional Models: Computational Framework for global inference and a vehicle for incorporating knowledge
- Direct supervision for structured NLP tasks is **expensive**
- Indirect supervision is cheap and easy to obtain
- Constrained Conditional Models combine
  - Learning conditional models with using declarative expressive constraints
  - Within a constrained optimization framework
- diverse usage CCMs have already found in NLP
  - Significant success on several NLP and IE tasks (often, with ILP)