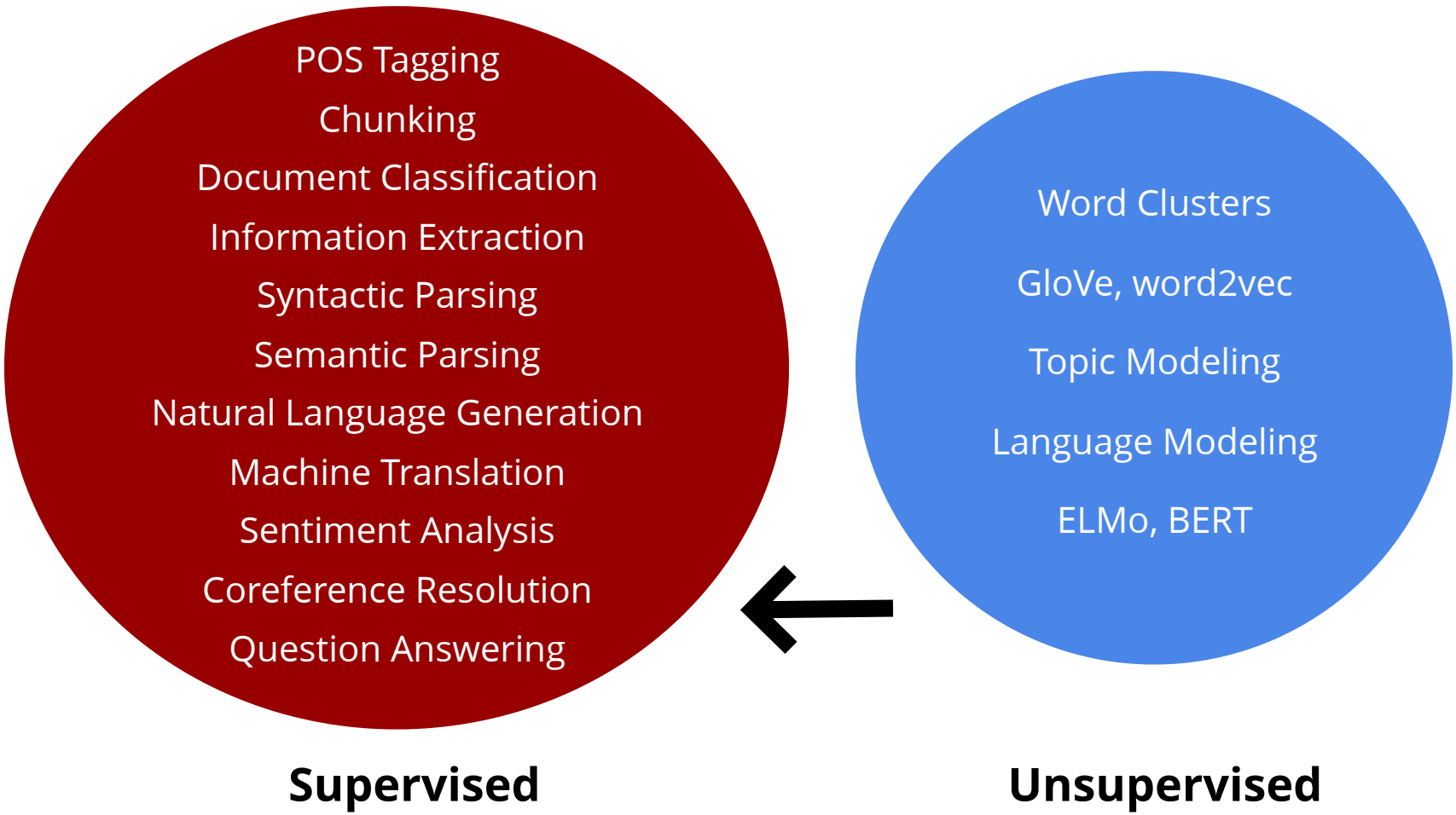
50.040
Natural Language Processing

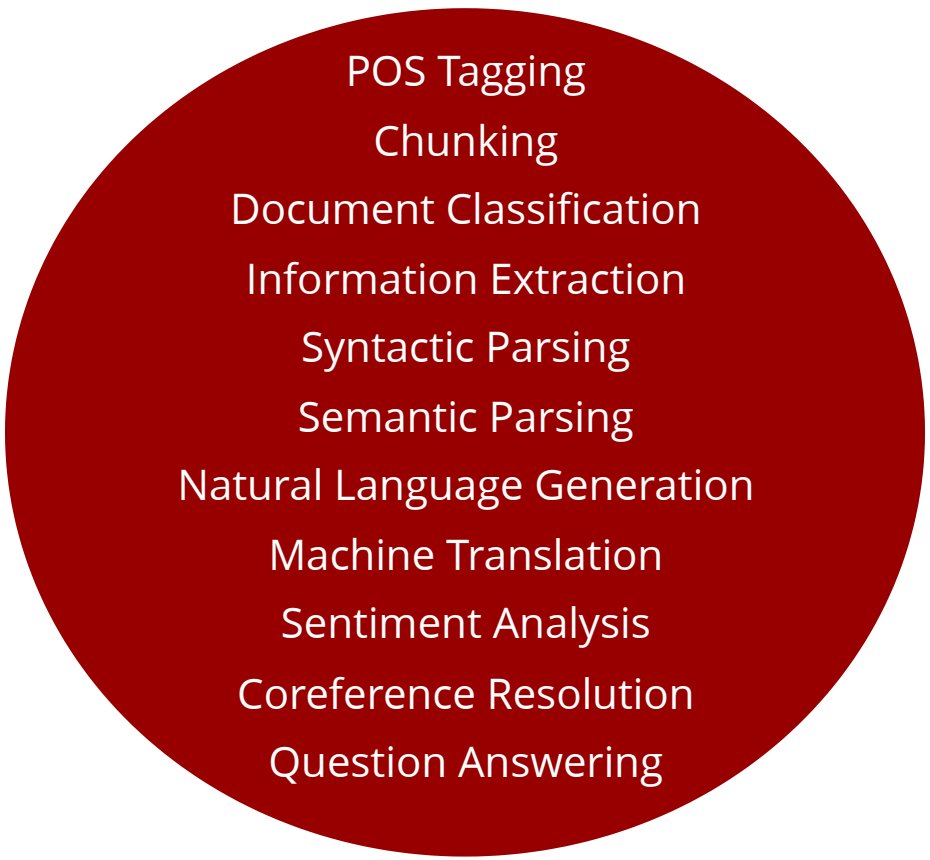
Lu, Wei

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Tasks in NLP

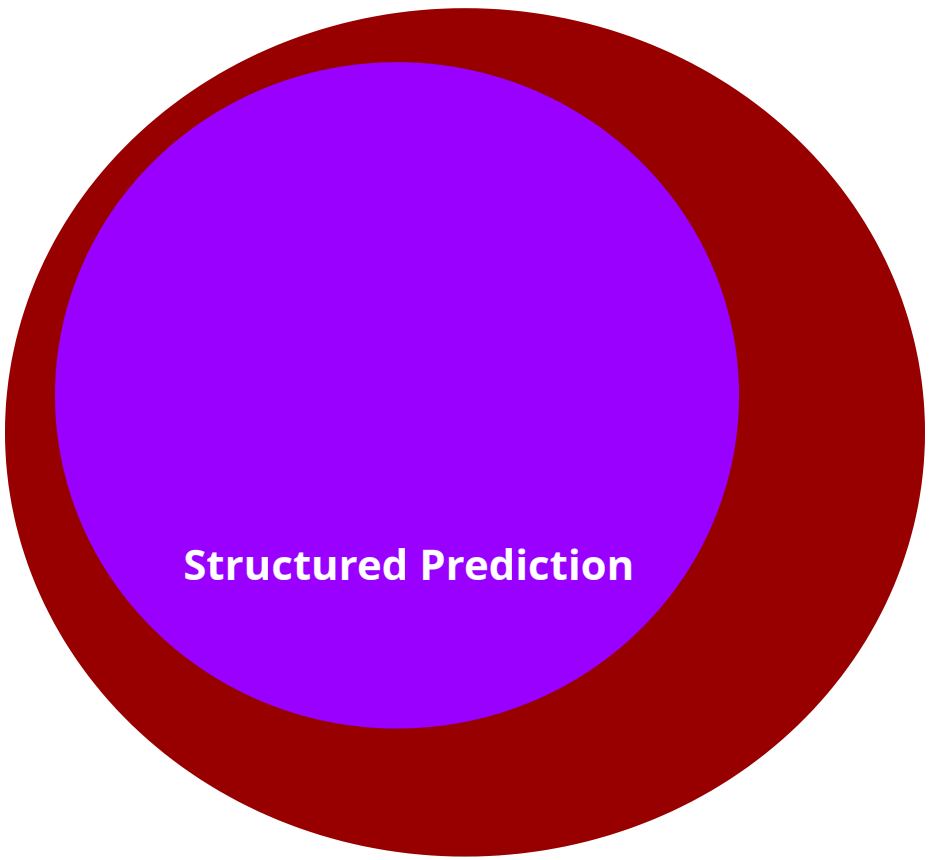

POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering
Word Clusters
GloVe, word2vec
Topic Modeling
Language Modeling
ELMo, BERT
Supervised
Unsupervised

Tasks in NLP



Supervised

Tasks in NLP



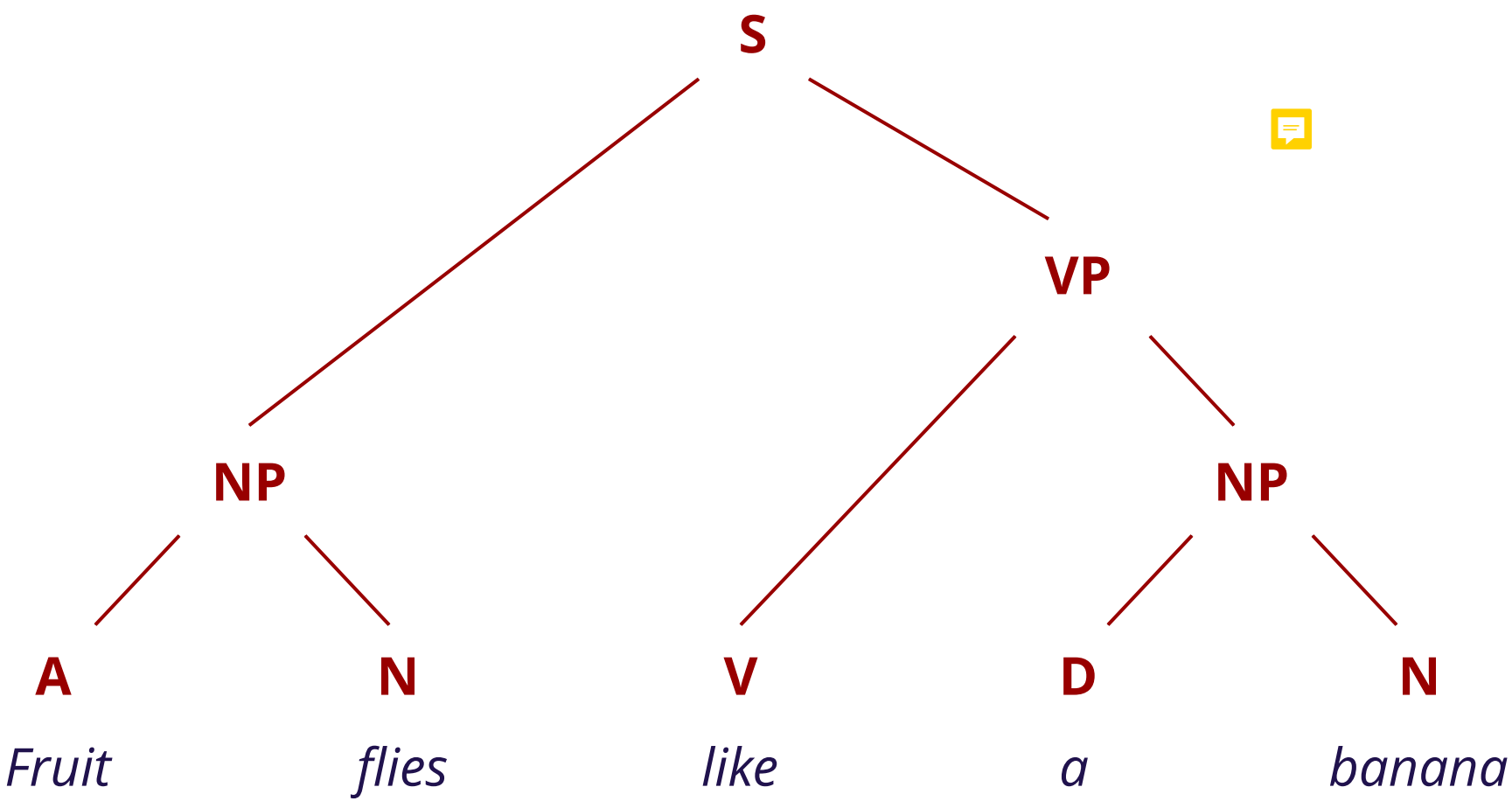
Structured Prediction

Part-of-Speech Tagging

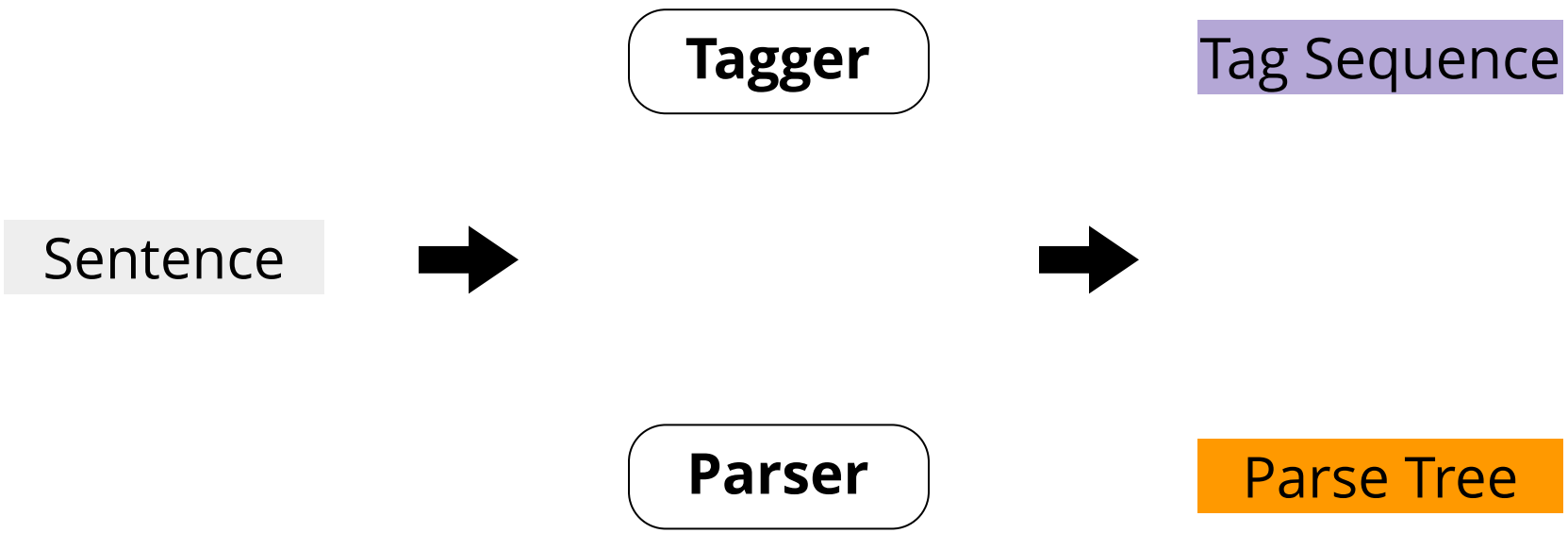
A	N	V	D	N
Fruit	*flies*	*like*	*a*	*banana*

Structured Prediction

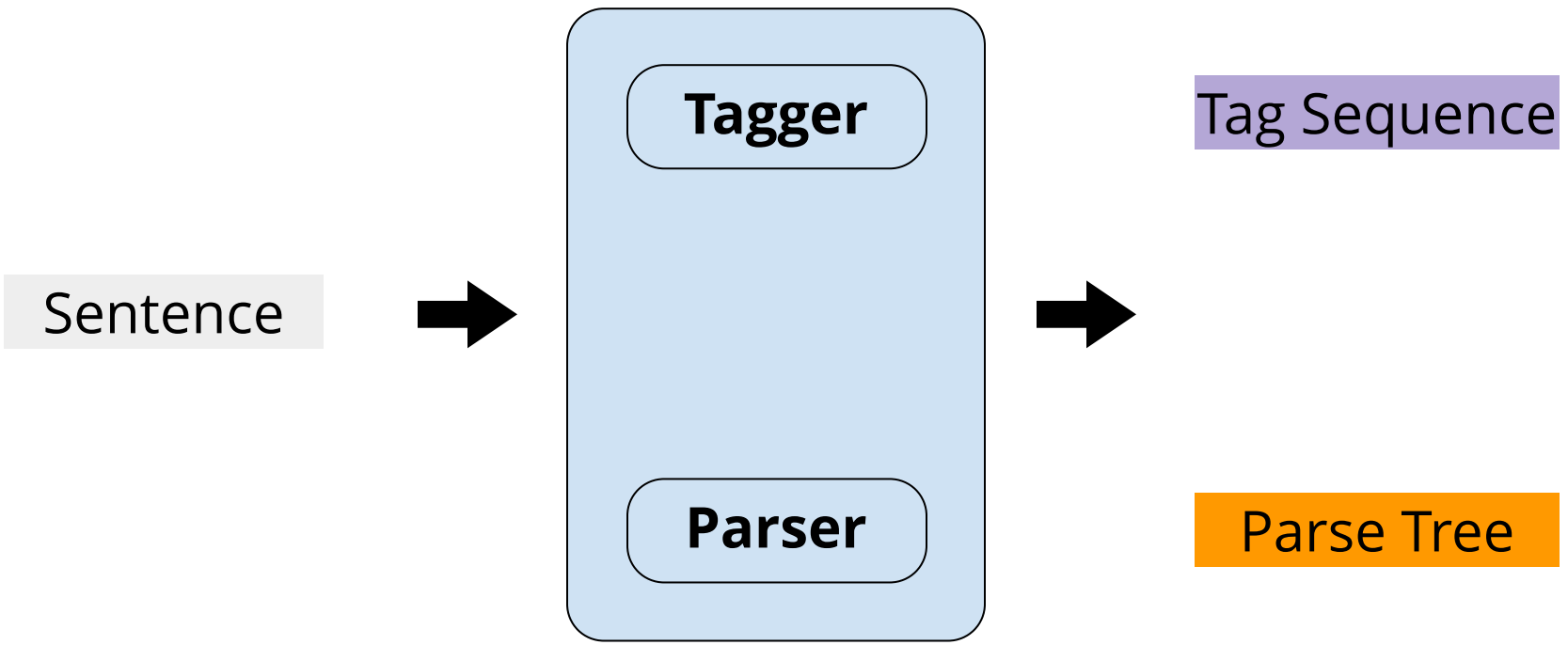
Constituency Parsing



Structured Prediction



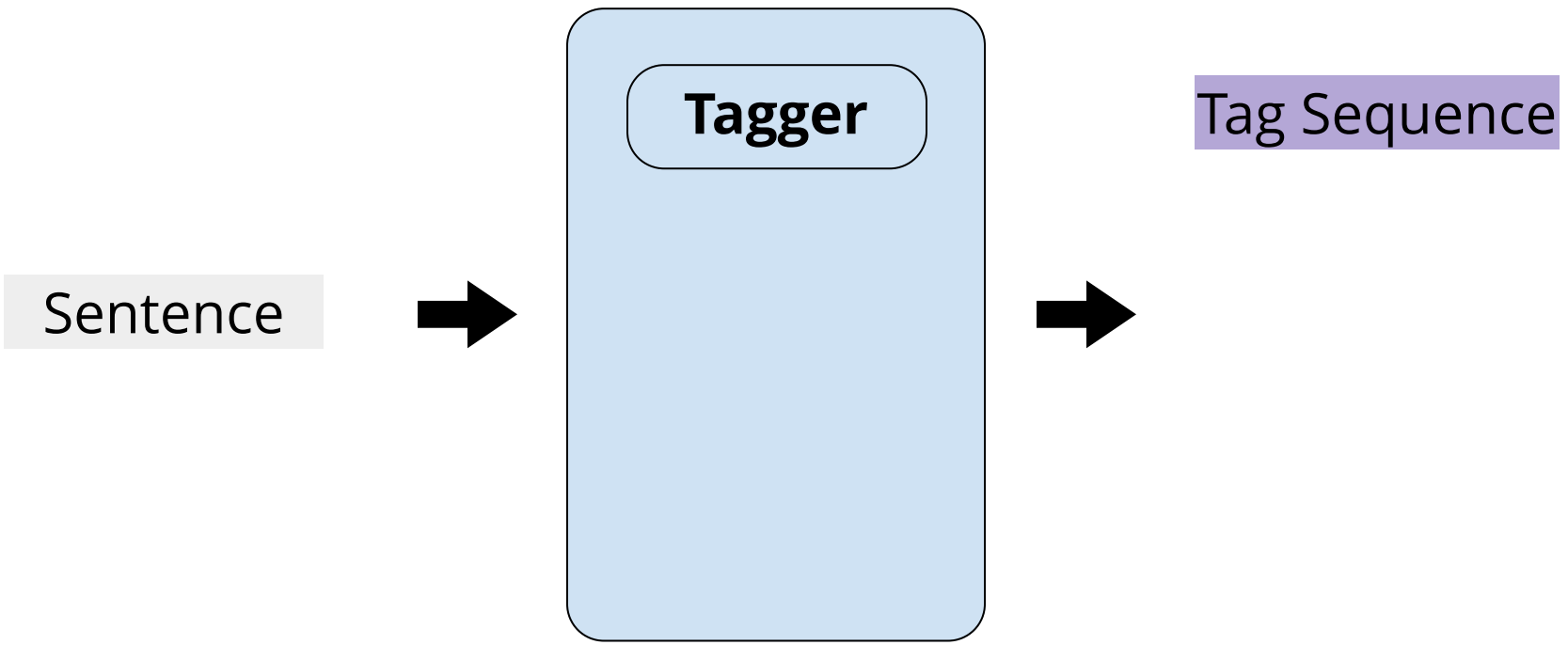
Structured Prediction



Structured Space $\mathcal{S}_i$ | Structured Prediction Model M | Structured Space $\mathcal{S}_o$

$$\mathsf{M} : \mathcal{S}_i \rightarrow \mathcal{S}_o$$

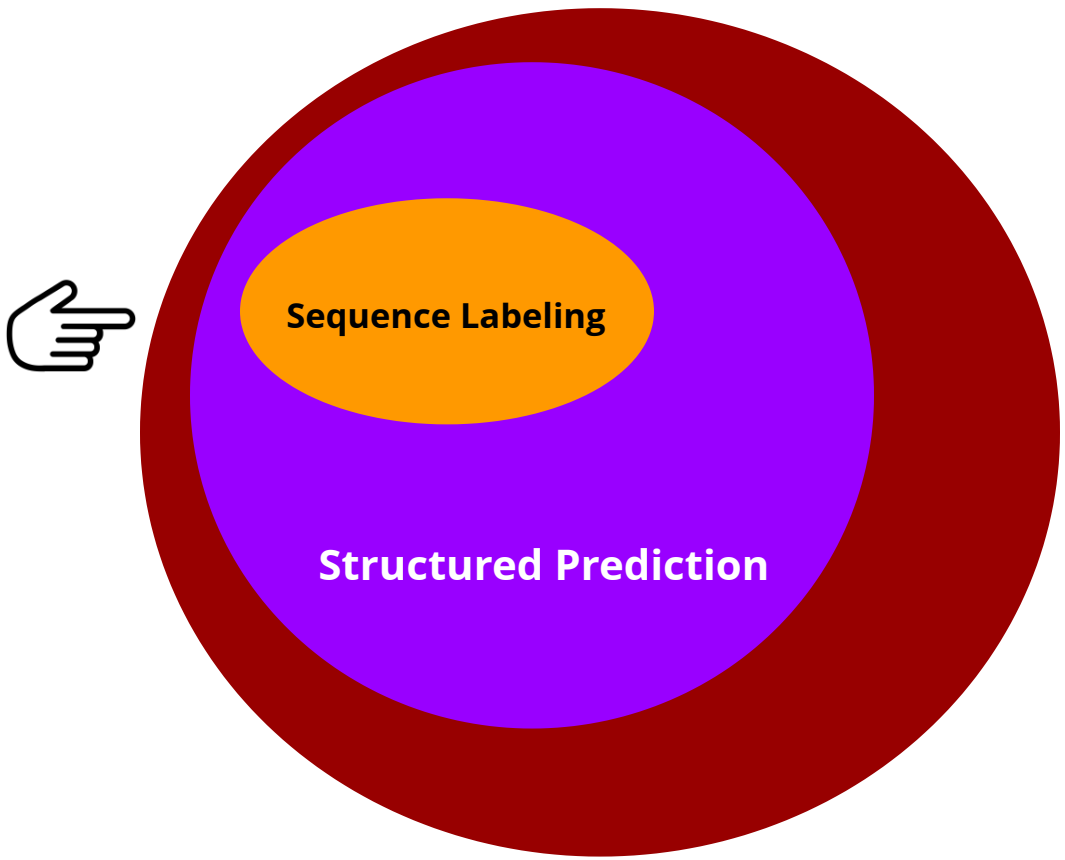
Sequence Labeling



Structured Space $\mathcal{S}_i$ | Structured Prediction Model M_ℓ | Structured Space $\mathcal{S}_o$

$$\mathsf{M}_\ell : \mathcal{S}_i \rightarrow \mathcal{S}_o$$

Tasks in NLP



Sequence Labeling

Definition

A type of machine learning task that involves assigning one discrete label to each member of a sequence of input tokens.

Sequence Labeling

Definition

A type of machine learning task that involves assigning one discrete label to each member of a sequence of input tokens.

Example Tasks

Part-of-Speech Tagging

Noun Phrase Chunking

Named Entity Recognition

Chinese Word Segmentation

Part-of-Speech Tagging

Assigning each word token a part-of-speech (POS) tag, where a POS refers to a category of words which have similar grammatical properties.

A	**N**	**V**	**D**	**N**
Fruit	*flies*	*like*	*a*	*banana*

Part-of-Speech Tagging

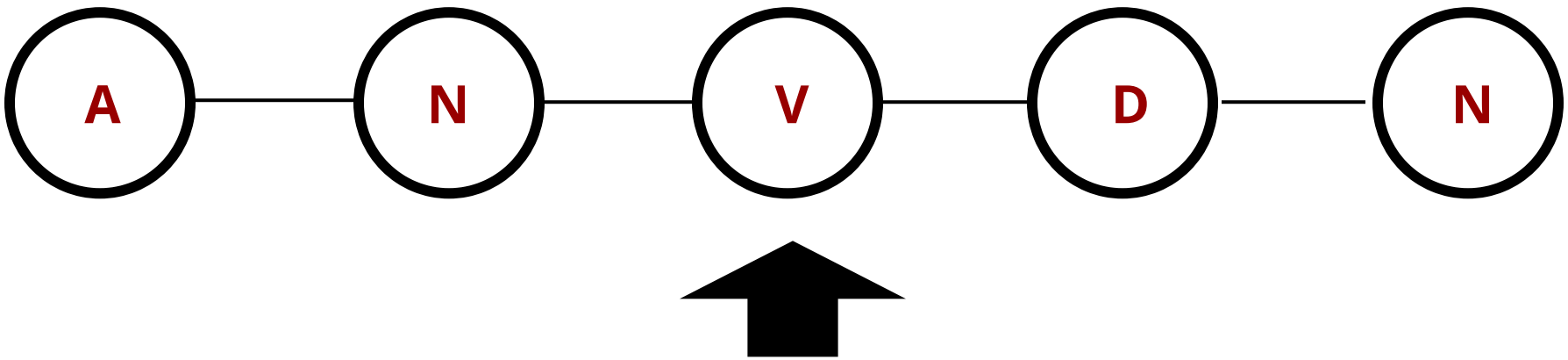
For English, there are 9 main part of speech (POS):

noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection.

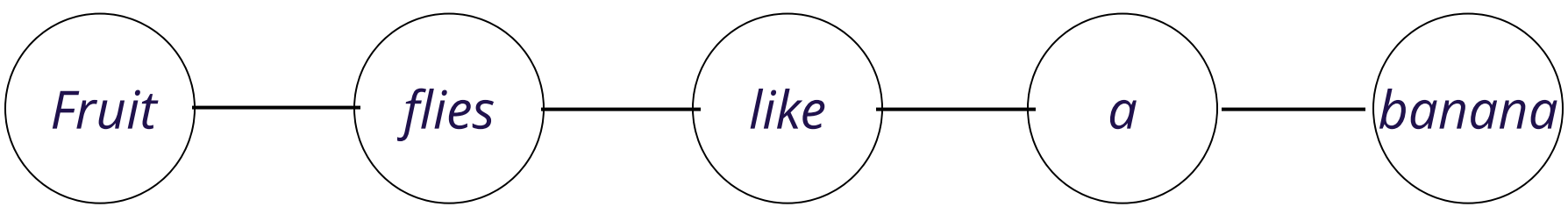
However, in practice there are far more fine-grained POS tags (50 - 150).

A	**N**	**V**	**D**	**N**
Fruit	*flies*	*like*	*a*	*banana*

Part-of-Speech Tagging



Tagger $\mathcal{M}_\ell$



Brill Tagger

(Brill, 1993)

Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging

Eric Brill*
The Johns Hopkins Univ

Recently, there has been a re
ual encoding of linguistic in
as a method of providing a
though corpus-based approa
processing, it is often the c
modelling indirectly in larg
understand and improve the
In this paper, we will descri
knowledge. This approach ha
and more direct fashion with
of this learning method appl

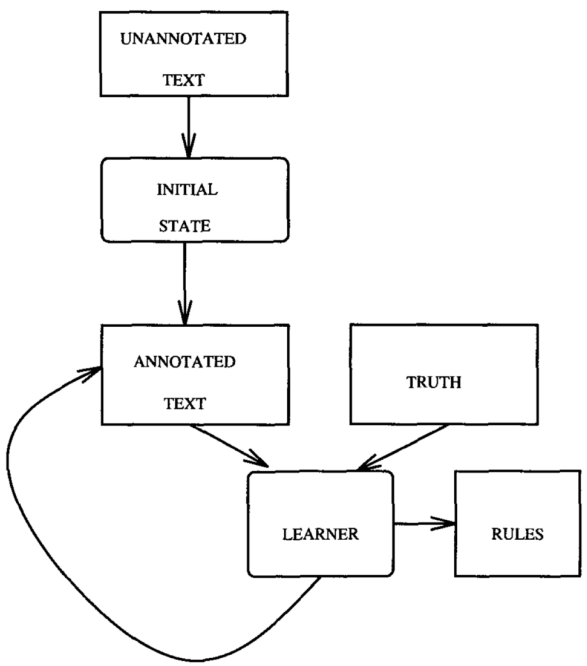


Figure 1
Transformation-Based Error-Driven Learning.

Transformation-Based Error-Driven Learning

Rules are of the form:

Change tag from *a* to *b* when *condition*

```
1. apply initial-state annotator to corpus
2. while transformations can still be found do
3.    for from_tag = tag_1 to tag_n
4.       for to_tag = tag_1 to tag_n
5.          for corpus_position = 1 to corpus_size
6.             if (correct_tag(corpus_position) == to_tag
                  && current_tag(corpus_position) == from_tag)
7.                      num_good_transformations(tag(corpus_position -1))++
8.             else if (correct_tag(corpus_position) == from_tag
                  && current_tag(corpus_position) == from_tag)
9.                      num_bad_transformations(tag(corpus_position-1 ))++
10.         find max_T (num_good_transformations(T) - num_bad_transformations(T))
11.         if this is the best-scoring rule found yet then store as best rule:
              Change tag from from_tag to to_tag if previous tag is T
12.   apply best rule to training corpus
13.   append best rule to ordered list of transformations
```

Figure 3
Pseudocode for learning transformations.

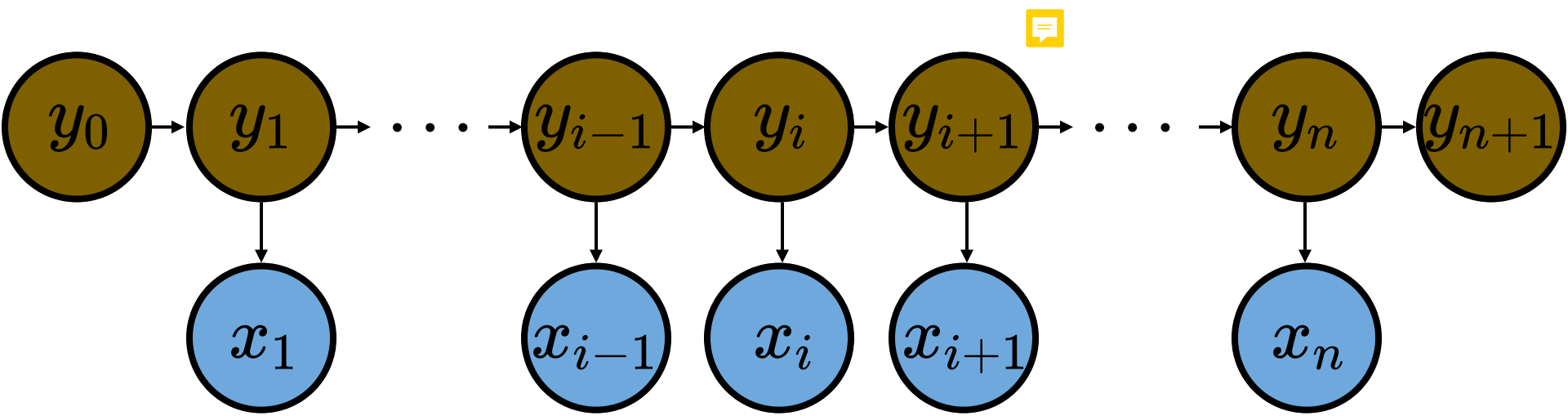
#	Change Tag From	Change Tag To	Condition
1	NN	VB	Previous tag is *TO*
2	VBP	VB	One of the previous three tags is *MD*
3	NN	VB	One of the previous two tags is *MD*
4	VB	NN	One of the previous two tags is *DT*
5	VBD	VBN	One of the previous three tags is *VBZ*
6	VBN	VBD	Previous tag is *PRP*
7	VBN	VBD	Previous tag is *NNP*
8	VBD	VBN	Previous tag is *VBD*
9	VBP	VB	Previous tag is *TO*
10	POS	VBZ	Previous tag is *PRP*
11	VB	VBP	Previous tag is *NNS*
12	VBD	VBN	One of previous three tags is *VBP*
13	IN	WDT	One of next two tags is *VB*
14	VBD	VBN	One of previous two tags is *VB*
15	VB	VBP	Previous tag is *PRP*
16	IN	WDT	Next tag is *VBZ*
17	IN	DT	Next tag is *NN*
18	JJ	NNP	Next tag is *NNP*
19	IN	WDT	Next tag is *VBD*
20	JJR	RBR	Next tag is *JJ*

Figure 4
The first 20 nonlexicalized transformations.

Hidden Markov Model

Generative Approach

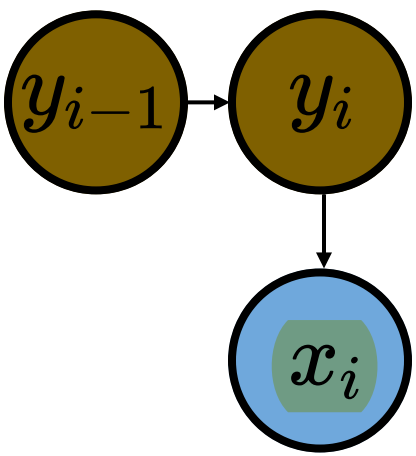
$$p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_n, y_{n+1}) = \prod_{i=1}^{n+1} p(y_i|y_{i-1}) \prod_{i=1}^{n} p(x_i|y_i)$$



Hidden Markov Model

One Limitation

$$p(y_i|y_{i-1})p(x_i|y_i)$$

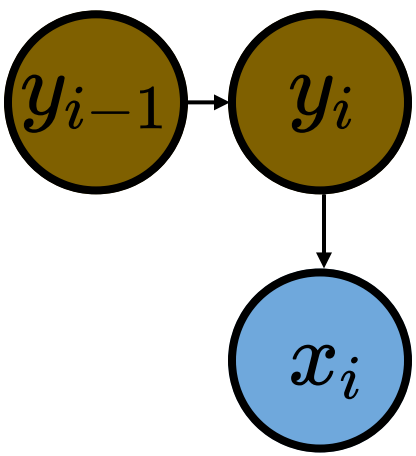


Does a word that ends with "*-ed*" or "*-ly*" likely to reveal its POS information?

Hidden Markov Model

One Limitation

$$p(y_i|y_{i-1})p(x_i|y_i)$$



How to capture the rich linguistic features within the inputs?

Conditional Model

Recall what we did in classification?

We can capture features with a discriminative approach!

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n)$$

Conditional Model

$$
\begin{aligned}
& p(y_1, \ldots, y_n \mid x_1, \ldots, x_n) \\
= \quad & \textstyle\prod_{i=1}^{n} p(y_i \mid x_1, \ldots, x_i, \ldots, x_n, y_1, \ldots, y_{i-1})
\end{aligned}
$$

Conditional Model

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n)$$

$$= \prod_{i=1}^{n} p(y_i \mid x_1, \ldots, x_i, \ldots, x_n, y_1, \ldots, y_{i-1})$$

$$= \prod_{i=1}^{n} p(y_i \mid \cancel{x_1, \ldots}, x_i, \cancel{\ldots, x_n, y_1, \ldots}, y_{i-1})$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, y_{i-1})$$

Conditional Model

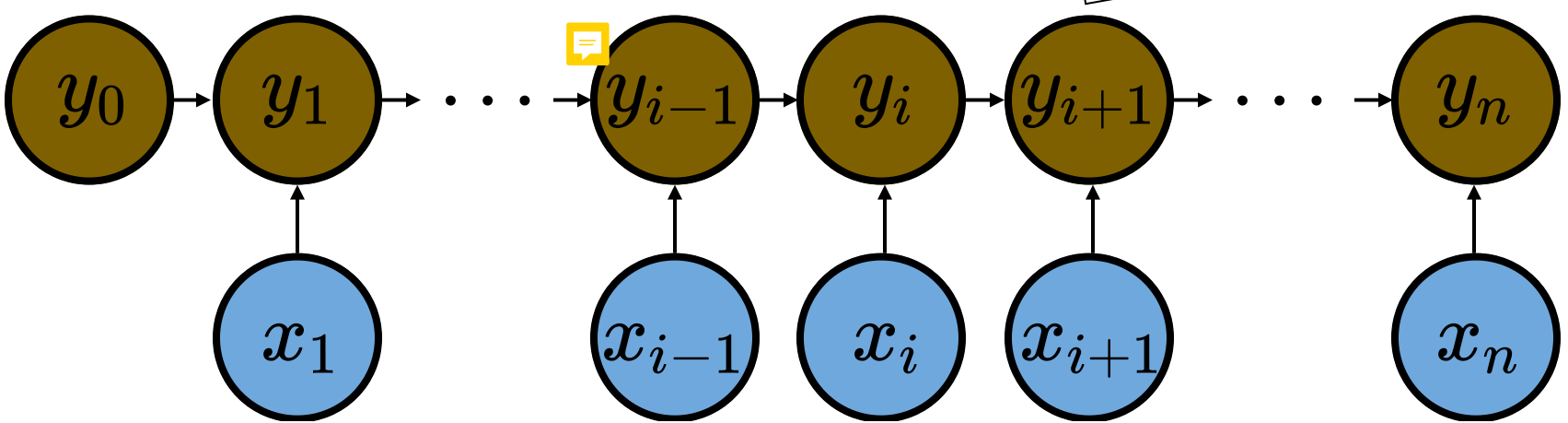
$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n)$$

$$= \prod_{i=1}^{n} p(y_i \mid x_1, \ldots, x_i, \ldots, x_n, y_1, \ldots, y_{i-1})$$

$$= \prod_{i=1}^{n} p(y_i \mid \cancel{x_1}, \ldots, x_i, \ldots, \cancel{x_n}, \cancel{y_1}, \ldots, y_{i-1})$$

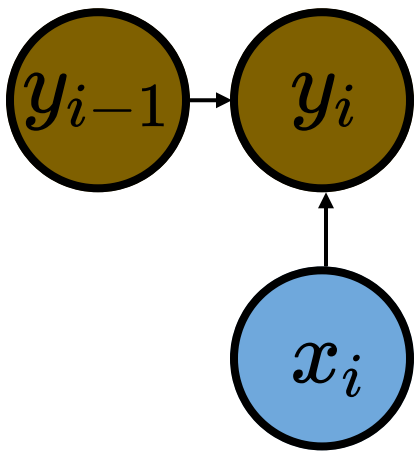
$$= \prod_{i=1}^{n} p(y_i \mid x_i, y_{i-1})$$

Markov Assumptions



Conditional Model

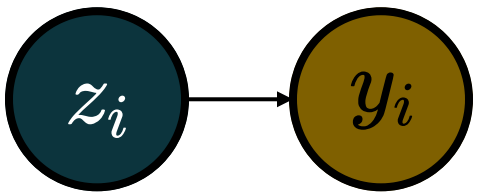
$$p(y_i \mid x_i, y_{i-1})$$



How to parameterize this conditional probability?

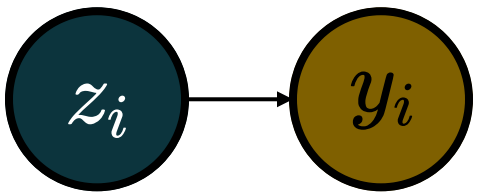
Conditional Model

$$p(y_i \mid x_i, y_{i-1}) = p(y_i \mid z_i), \text{where } z_i = \langle x_i, y_{i-1} \rangle$$



Conditional Model

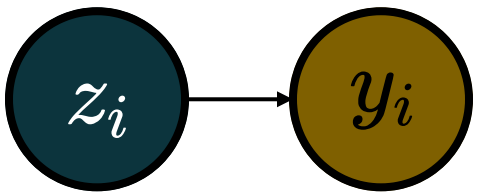
$$p(y_i \mid x_i, y_{i-1}) = p(y_i \mid z_i), \text{where } z_i = \langle x_i, y_{i-1} \rangle$$



What if the discrete variable y_i is binary?

Logistic Regression

$$p(y_i \mid x_i, y_{i-1}) = p(y_i \mid z_i), \text{ where } z_i = \langle x_i, y_{i-1} \rangle$$



weights/parameters

$$p(y_i \mid z_i) \propto \exp(-(\boldsymbol{f}(z_i) \cdot \boldsymbol{\theta} + \theta_0) y_i)$$

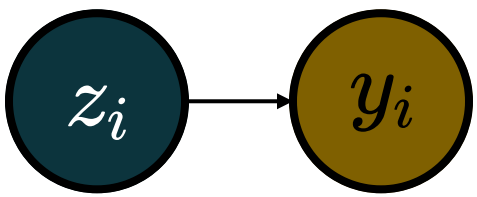
feature vector

binary variable

What if the discrete variable y_i is no longer binary?

Softmax Regression

$$p(y_i \mid x_i, y_{i-1}) = p(y_i \mid z_i), \text{ where } z_i = \langle x_i, y_{i-1} \rangle$$



weights/parameters

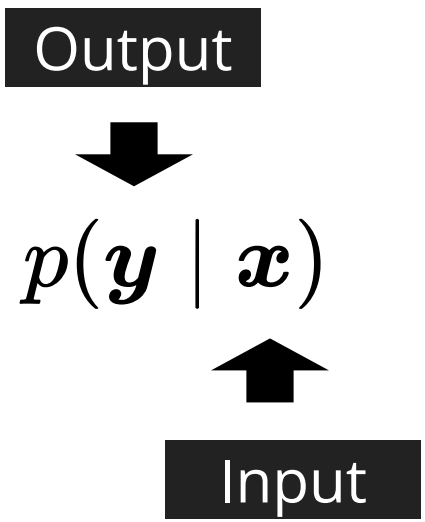
$$p(y_i \mid z_i) \propto \exp(\boldsymbol{f}(z_i, y_i) \cdot \boldsymbol{\theta})$$

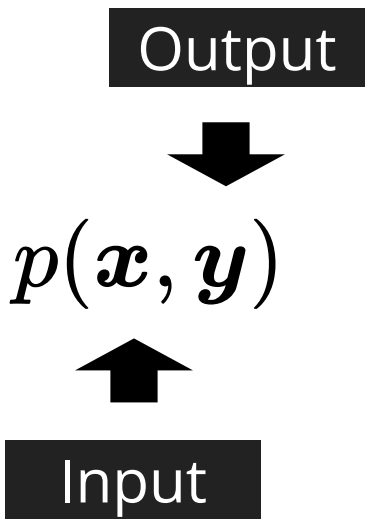
feature vector

Why can we parameterize the probability this way?

Two Types of Models

From now, let us use $\boldsymbol{x}$ and $\boldsymbol{y}$ for input and output respectively.





Discriminative

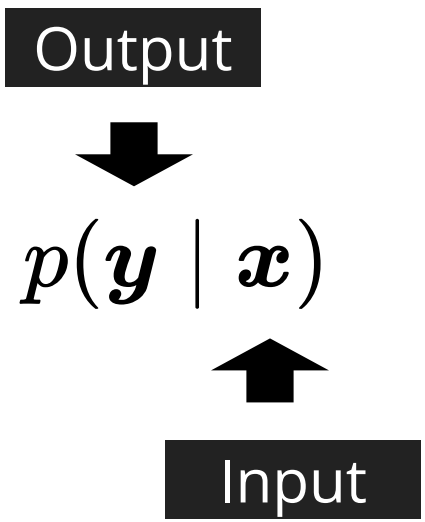
(no need to explicitly model input)

Generative

(need to explicitly model input)

Discriminative Model

From now, let us use $\boldsymbol{x}$ and $\boldsymbol{y}$ for input and output respectively.



Discriminative
(no need to explicitly model input)

Generative
(need to explicitly model input)

Maximum Entropy

A Maximum Entropy Approach to Natural Language Processing

Adam L. Berger†
Columbia University

Stephen A. Della Pietra‡
Renaissance Technologies

Vincent J. Della Pietra‡
Renaissance Technologies

The concept of maximum entropy can be traced back along multiple threads to Biblic
recently, however, have computers become powerful enough to permit the widesc
of this concept to real world problems in statistical estimation and pattern reco
paper, we describe a method for statistical modeling based on maximum entropy.
a maximum-likelihood approach for automatically constructing maximum entropy models and
describe how to implement this approach efficiently, using as examples several problems in natural
language processing.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra.
A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics 22, 1 (March 1996), 39-71.

Entropy

Degree of Uncertainty/Surprise

Entropy of a probability distribution $p(v)$
is defined as follows:

$$\mathcal{H}(p(v)) = -\sum_v p(v) \log p(v)$$

It measures the degree of uncertainty associated with the model.

Entropy
Degree of Uncertainty/Surprise

Entropy of a probability distribution $p(v)$
is defined as follows:

$$\mathcal{H}(p(v)) = -\sum_v p(v) \log p(v)$$

It measures the degree of uncertainty associated with the model.

When the number of discrete outputs is fixed, the uniform distribution gives the highest entropy!

Maximum Entropy

The entropy of the conditional model of our interest is defined as:

$$\mathcal{H}(p(\boldsymbol{y}|\boldsymbol{x})) = -\sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}) \log p(\boldsymbol{y}|\boldsymbol{x})$$

Our goal is to find the distribution that maximizes the entropy:

$$p^*(\boldsymbol{y}|\boldsymbol{x}) = \max_{p(\boldsymbol{y}|\boldsymbol{x})} \mathcal{H}(p(\boldsymbol{y}|\boldsymbol{x}))$$

What are the constraints for this optimization problem?

Maximum Entropy

$$p^*(\boldsymbol{y}|\boldsymbol{x}) = \max_{p(\boldsymbol{y}|\boldsymbol{x})} \mathcal{H}(p(\boldsymbol{y}|\boldsymbol{x}))$$

$$\max_{p(\boldsymbol{y}|\boldsymbol{x})} \mathcal{H}(p(\boldsymbol{y}|\boldsymbol{x}))$$

subject to:

1. It has to be a valid probability distribution:

$$\sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}) = 1$$

2. While we prefer a flat/uniform distribution, we shall respect the empirical features that we observed:

$$\mathbf{E}_{p(\boldsymbol{y}|\boldsymbol{x})}[f_k(\boldsymbol{x}, \boldsymbol{y})] = \mathbf{E}_{\hat{p}(\boldsymbol{y}|\boldsymbol{x})}[f_k(\boldsymbol{x}, \boldsymbol{y})]$$

The distribution we are looking for

The empirical distribution

Maximum Entropy

$$
\begin{aligned}
&\Lambda(p, \lambda) \\
&= -\mathcal{H}(p(\boldsymbol{y}|\boldsymbol{x})) \\
&- \sum_{k=1}^{K} \lambda_k \left(\mathbf{E}_{p(\boldsymbol{y}|\boldsymbol{x})[f_k(\boldsymbol{x},\boldsymbol{y})]} - \mathbf{E}_{\hat{p}(\boldsymbol{y}|\boldsymbol{x})[f_k(\boldsymbol{x},\boldsymbol{y})]} \right) \\
&- \lambda_0 \left(\sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}) - 1 \right)
\end{aligned}
$$

$$\text{where } \lambda_k \text{ and } \lambda_0 \in (-\infty, +\infty)$$

Take the partial derivative and set it to zero:

$$\frac{\partial \Lambda(p,\lambda)}{\partial p(\boldsymbol{y}|\boldsymbol{x})} = \log p(\boldsymbol{y}|\boldsymbol{x}) + 1 - \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y}) - \lambda_0 = 0$$

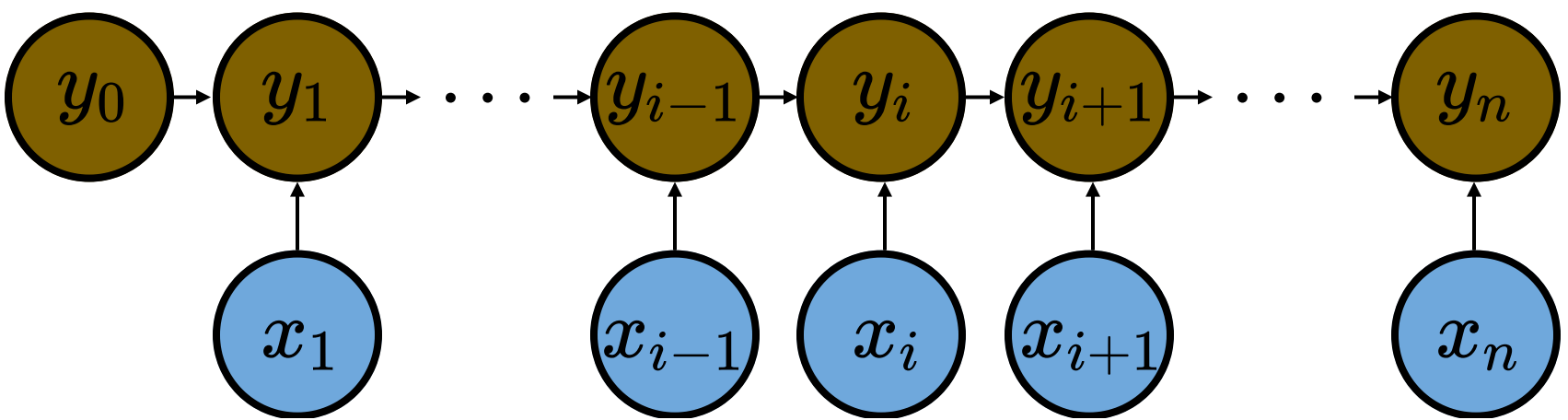
This gives us:

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left(\sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y}) \right)$$

Maximum Entropy Markov Model (McCallum et al., 2000)

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i \mid x_i, y_{i-1})$$

$$p(y_i \mid x_i, y_{i-1}) \propto \exp\left(\sum_{k=1}^{K} \lambda_k f_k(x_i, y_{i-1}, y_i)\right)$$



Part-of-Speech Tagging

Summary

Model	Pros	Cons
Transformation-based Model (Brill's Tagger)	Intuitive, interpretable, fast at testing time	Non probabilistic, training can be slow, accuracy not high enough
Hidden Markov Model (HMM)	Probabilistic, efficient training	Unable to exploit linguistic features
Maximum Entropy Markov Model (MEMM)	Probabilistic, able to exploit features, higher accuracy than HMM	Training is slower than HMM

Noun-Phrase Chunking

Question

Is it possible to perform chunking with tagging?

Chunking as Tagging

Essentially, we are assigning a tag to each individual word in the sentence

Can we just use the below tags for this task? Why?

Part of NP	Part of NP	Not Part of NP	Part of NP	Part of NP
NP			**NP**	
Fruit	*flies*	*like*	*a*	*banana*

Chunking

Tagging Schemes

BIO Tagging Scheme

B: Beginning of NP

I: Inner part of NP

O: Outside of NP

BILOU Tagging Scheme

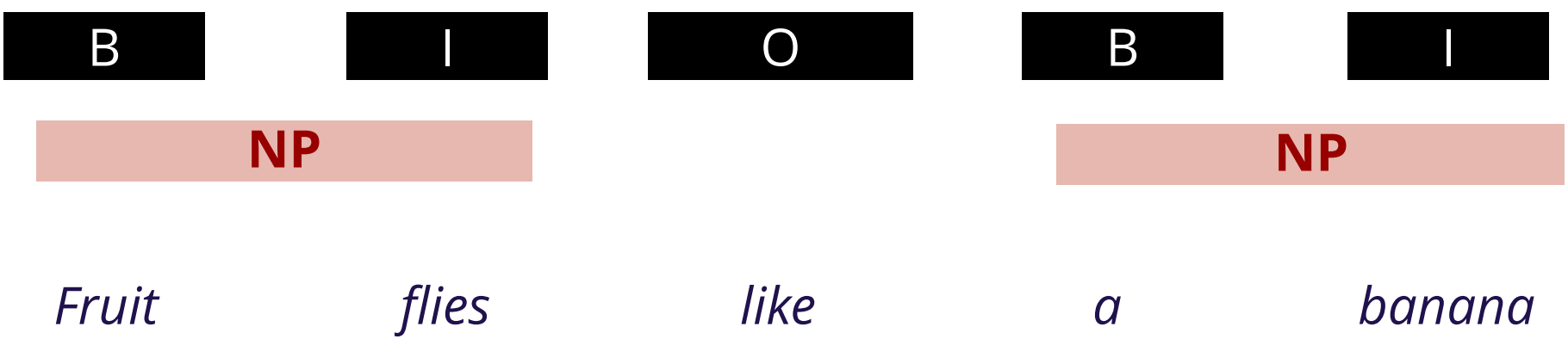
B: Beginning of NP

I: Inner part of NP

L: Last word of NP

O: Outside of NP

U: Unit word NP



Chunking

Tagging Schemes

BIO Tagging Scheme

B: Beginning of NP

I: Inner part of NP

O: Outside of NP

BILOU Tagging Scheme

B: Beginning of NP

I: Inner part of NP

L: Last word of NP

O: Outside of NP

U: Unit word NP

Chunking

Chunking as Shallow Parsing

Captures shallow, non-hierarchical syntactic information.

Identifies elementary constituent parts of a sentence.

As a very basic constituency parser with a flat representation

Named Entity Recognition

One of the most fundamental tasks within NLP, a sub-task within information extraction, that seeks to recognize and classify named entity mentions in text into pre-defined semantic categories (e.g., location, organization, geo-political entity, person, facility ...)

ORG

FAC

Singapore University of Technology and Design is close to the Changi Airport.

Question

How to perform named entity recognition with tagging?

Named Entity Recognition

Tagging Scheme

B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	L-ORG	O	O	O	O	B-FAC	L-FAC

ORG

FAC

Singapore University of Technology and Design is close to the Changi Airport.

Word Segmentation

近百名后港居民到白沙公园净滩

（早报讯）后港居民趁星期天早晨展开净滩活动，为绿化环境尽一份力。

后港基层组织年度净滩活动，号召近百名居民参加。他们上午8时30分齐聚白沙公园，每人一手拿钳子，一手拿垃圾袋，在沙滩上捡垃圾。

近 百名 后港 居民 到 白沙公园 净滩

（ 早报 讯 ） 后港 居民 趁 星期天 早晨 展开 净滩 活动 ， 为 绿化 环境 尽 一份 力 。

后港 基层 组织 年度 净滩 活动 ， 号召 近 百名 居民 参加 。 他们 上午 8时30分 齐聚 白沙公园 ， 每人 一手 拿 钳子 ， 一手 拿 垃圾袋 ， 在 沙滩 上 捡 垃圾 。

Word Segmentation

行动 (act/action)

党 (party)

行动党 (PAP)

首页 › 即时报道 › 新加坡即时

李总理：来届大选候选人背景多元

发布 / 2019年6月7日 11:43 PM
更新 / 2019年6月8日 5:26 PM
文 / 杨浚鑫

（早报讯）执政的人民行动党正在为来届大选备战，并已物色许多候选人，身为行动党秘书长的李显龙总理也亲自面试了好些人。

李总理今晚以主宾身份出席第九届“通商中国奖”颁奖晚宴，并在约50分钟的对话会中，谈及我国领导接棒过程和来届大选。他透露，新一批候选人来自不同种族、年龄和背景，其中一些人也曾走过不同的人生道路。

“他们也许在学校表现不佳，到了非政府组织发展事业，并对此深具热情。他们也可能在海外待过一段日子，更了解世界，回国后也更珍惜新加坡。”

Word Segmentation

中一
(secondary one)

一些
(some)

人
(people)

其中
(out of)

一些人
(some people)

首页 › 即时报道 › 新加坡即时

李总理：来届大选候选人背景多元

发布 / 2019年6月7日 11:43 PM
更新 / 2019年6月8日 5:26 PM
文 / 杨浚鑫

（早报讯）执政的人民行动党正在为来届大选备战，并已物色许多候选人，身为行动党秘书长的李显龙总理也亲自面试了好些人。

李总理今晚以主宾身份出席第九届“通商中国奖”颁奖晚宴，并在约50分钟的对话谈及我国领导接棒过程和来届大选。他透露，新一批候选人来自不同种族、年龄，其中一些人也曾走过不同的人生道路。

“他们也许在学校表现不佳，到了非政府组织发展事业，并对此深具热情。他们也可能在海外待过一段日子，更了解世界，回国后也更珍惜新加坡。”

Word Segmentation

非 (not/African)

政府 (government)

组织 (organization/tissue)

非政府组织 (NGO)

首页 › 即时报道 › 新加坡即时

李总理：来届大选候选人背景多元

发布 / 2019年6月7日 11:43 PM
更新 / 2019年6月8日 5:26 PM
文 / 杨浚鑫

（早报讯）执政的人民行动党正在为来届大选备战，并已物色许多候选人，身为行动党秘书长的李显龙总理也亲自面试了好些人。

李总理今晚以主宾身份出席第九届“通商中国奖”颁奖晚宴，并在约50分钟的对话[illegible]中，谈及我国领导接棒过程和来届大选。他透露，新一批候选人来自不同种族、年[illegible]背景，其中一些人也曾走过不同的人生道路。

“他们也许在学校表现不佳，到了非政府组织发展事业，并对此深具热情。他们也可能在海外待过一段日子，更了解世界，回国后也更珍惜新加坡。”

Word Segmentation

Tagging Scheme

Consider this Chinese phrase and consider the BILOU scheme:

他到了非政府组织发展事业

Assume we know the gold segmentation is as follows:

他 到了 非政府组织 发展 事业

How shall we tag each Chinese character?

他到了非政府组织发展事业

Word Segmentation

Tagging Scheme

Consider this Chinese phrase and consider the BILOU scheme:

他到了非政府组织发展事业

Assume we know the gold segmentation is as follows:

他 到了 非政府组织 发展 事业

U B L B I I I L B L B L

他到了非政府组织发展事业

Maximum Entropy Markov Model

One Theoretical Limitation

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i \mid x_i, y_{i-1})$$

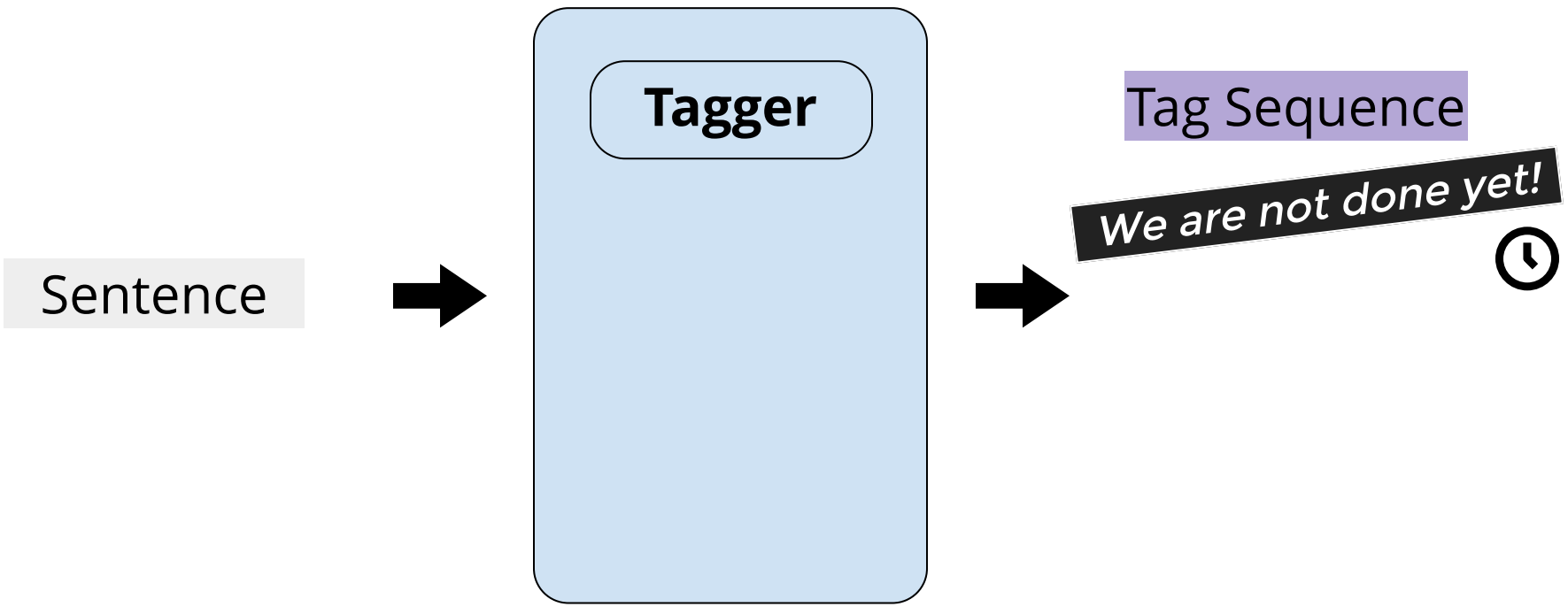
The model predicts one tag at a time (i.e., performs local normalization). Since we are modeling sequences, is it possible to predict a complete sequence at a time (i.e., perform global normalization)?

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n) \quad ?$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp(\mathbf{f}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{\theta})$$

Conditional Random Fields!

Structured Prediction



Structured Space $\mathcal{S}_i$ | Structured Prediction Model M_ℓ | Structured Space $\mathcal{S}_o$

$$\mathsf{M}_\ell : \mathcal{S}_i \rightarrow \mathcal{S}_o$$