Joint Training for Open-domain Extraction

Presented by Aliaksandr Talaika

16.06.2011

Seminar “Probabilistic Models for Information Extraction”,

SS 11

Based on

1. *Rahul Gupta and Sunita Sarawagi*:
Joint Structured Models for Extraction from Overlapping Sources, CoRR'10

2. *Rahul Gupta and Sunita Sarawagi*:
Joint Training for Open-domain Extraction on the Web: Exploiting Overlap when Supervision is Limited, WSDM'11

Information on the web

Source 1

Saarland University (German Universität des Saarlandes) is a university located in Saarbrücken, the capital of the German state of Saarland, and Homburg. It was founded in 1948 in Homburg in co-operation with France.

Plain text
(unstructured)

Source 2

- Saarland University is located in Saarbrücken, It was founded in 1948 in Homburg.
- Stanford University was founded in 1891 in California.

HTML lists
(semi-structured)

Source 3

Name	Location	Year
Saarland University	Germany	1948
Stanford University	California	1891
MIT	Massachusetts	1861

Tables
(structured)

Regular structure

Efficient information extraction

Table augmentation problem

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

List source 1

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Saarbrücken, Germany

List source 2

CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

List source 3

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

extract

Augmented table

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**
Stanford University	**California**	**1891**
Princeton University	**Princeton, NJ**	**1746**
Florida State University	**Tallahassee**	**-**
…	**…**	**…**

Table augmentation (related example)

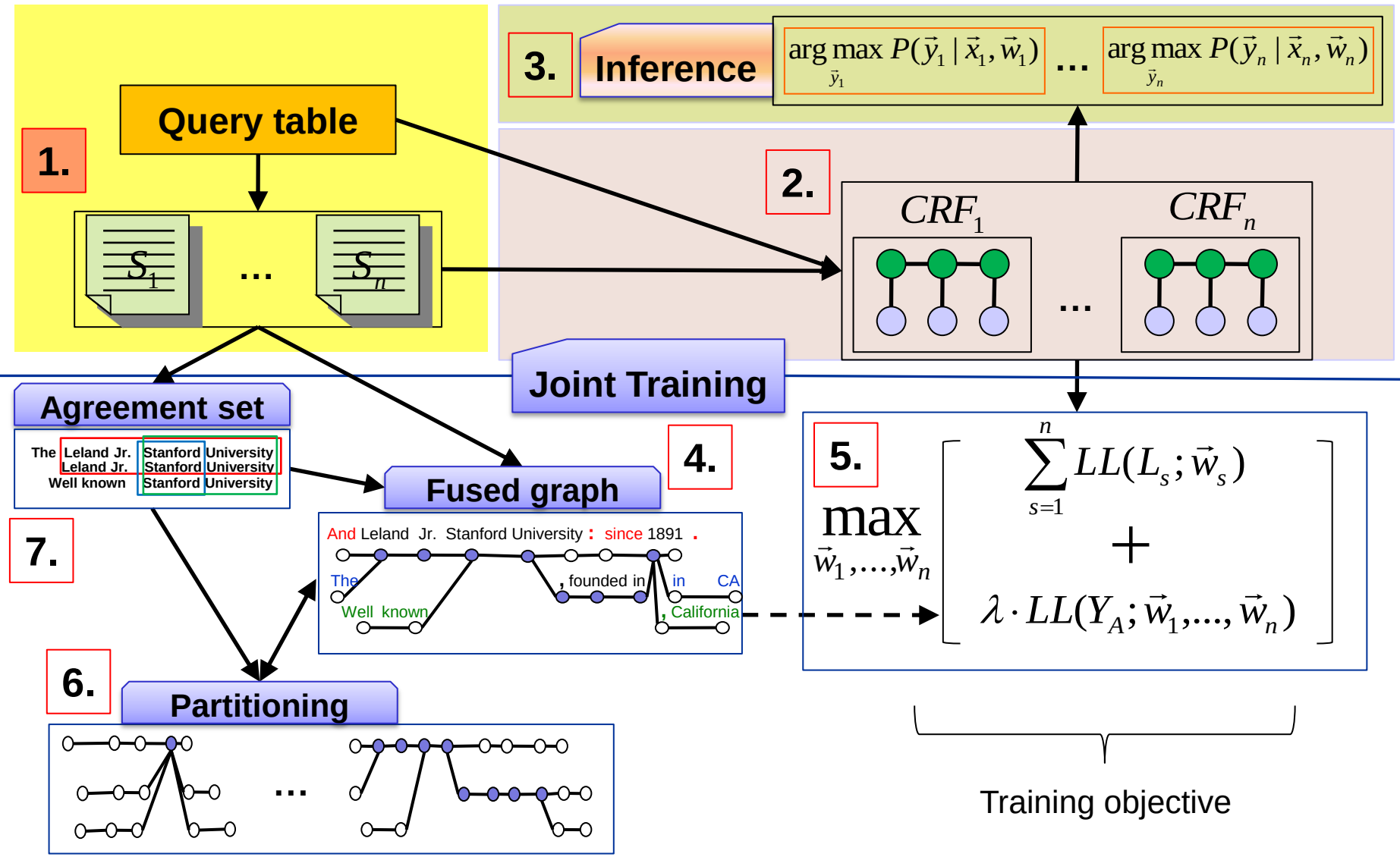
Google squared labs

universities | Square it

Item Name	Image	Location	Founded
Creighton University		Omaha, Nebraska	1878
University of Southern California		Los Angeles, CA	1880
Eastern Connecticut State University	EASTERN CONNECTICUT STATE UNIVERSITY	CT, Willimantic	1889
Stanford University		Stanford, California	1891
Seattle University		Seattle, WA	1891
California Institute of Technology		Pasadena, CA	1891

- Similar task
- Different approach

Architecture



Relevant lists

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

- Collect
- Index
- Offline

indexed repository of HTML lists (from web crawl)

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Saarbrücken, Germany

List source **2**

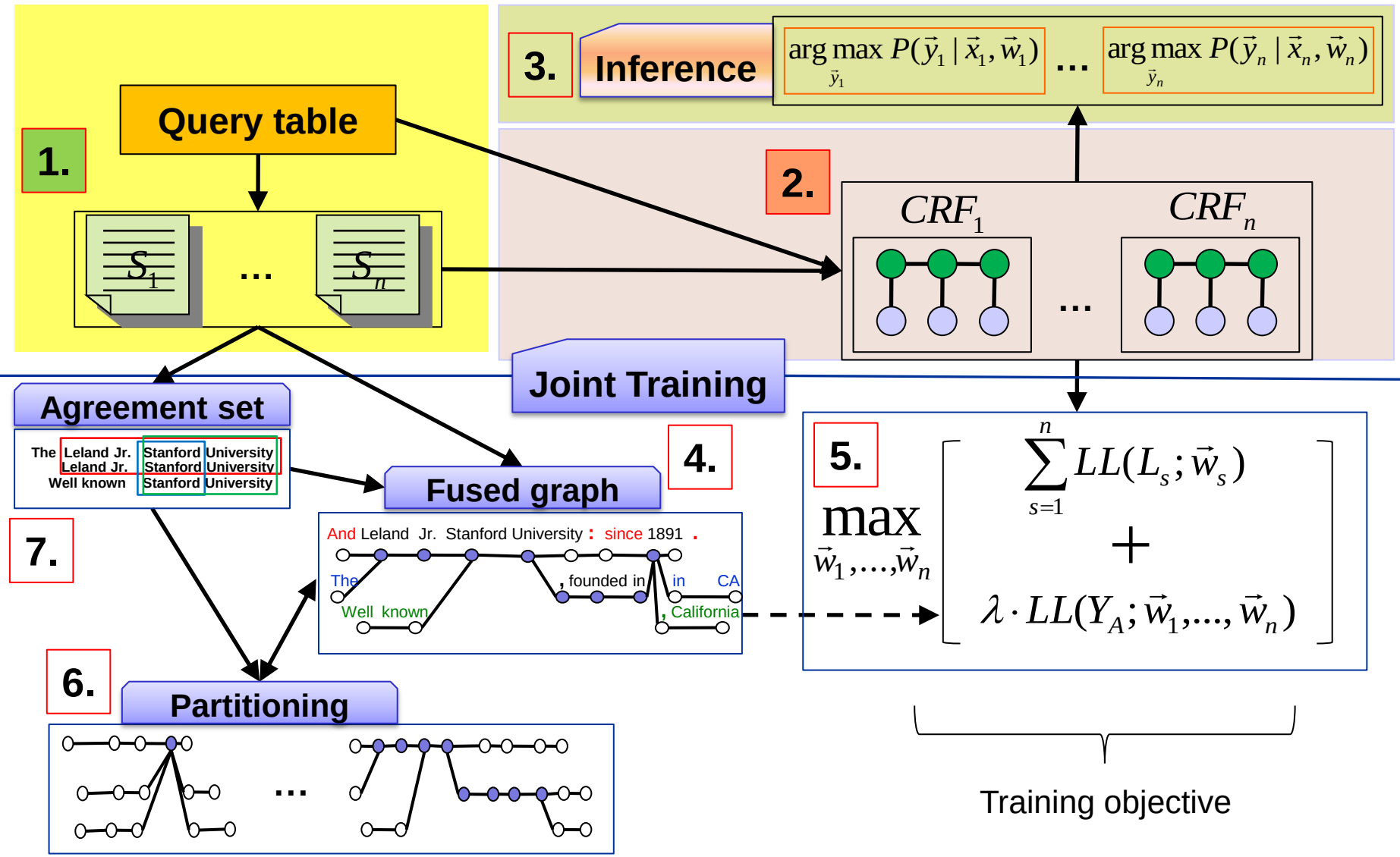
CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

…

List source **n**

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

Architecture



Convert lists to structured tables

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Saarbrücken, Germany

List source **2**

CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

…

List source **n**

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

Stanford University	**California**	**1891**
…	…	…

Florida State University	**Tallahassee**	**-**
…	…	…

Princeton University	**Princeton, NJ**	**1746**
…	…	…

Merge, de-duplicate, rank ⇨ one table

Constraints

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

1) *Few input rows*

Limited labeled set

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Saarbrücken, Germany

2) *Extraction over arbitrary domain (query about everything)*

No pretrained extractor

3) *Lists may not be machine generated*

no pattern-based extractor

…

List source **n**

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

?

Stanford University	**California**	**1891**
…	…	…

?

Florida State University	**Tallahassee**	**-**
…	…	…

?

Princeton University	**Princeton, NJ**	**1746**
…	…	…

Merge, de-duplicate, rank ⇨ one table

Convert lists to structured tables

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Saarbrücken, Germany

List source **2**

CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

…

List source **n**

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

Use!

CRF_1

- Different formats/styles
- Disjoint feature set

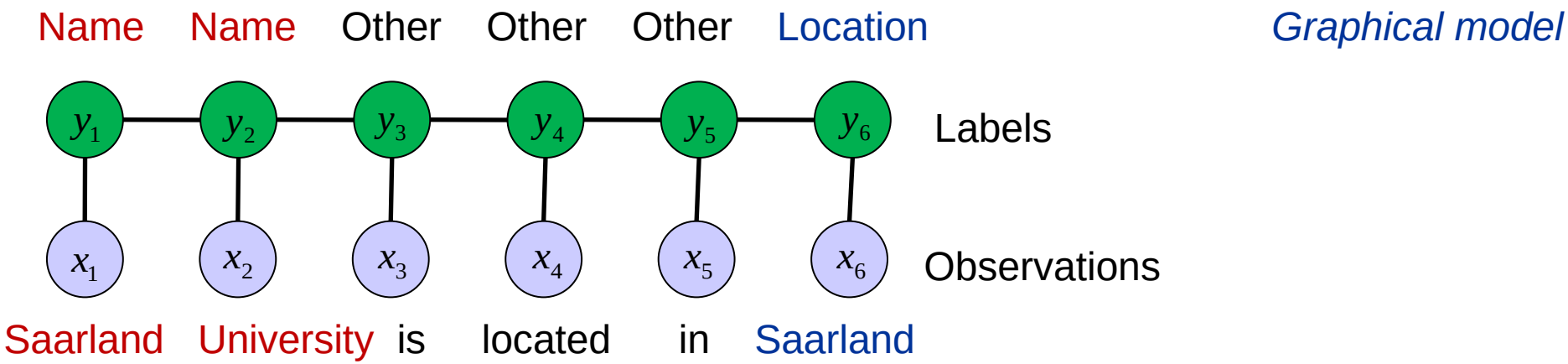
CRF_2 • • • CRF_n

Stanford University	**California**	**1891**
…	…	…

Florida State University	**Tallahassee**	**-**
…	…	…

Princeton University	**Princeton, NJ**	**1746**
…	…	…

Conditional random fields (reminder)



Model:

$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_x} \prod_i \exp\left(\sum_j \lambda_j f_j(y_i, y_{i-1}) + \sum_k \mu_k g_k(y_i, x_i) \right)$$

Weight

Transition feature function

Previous label is Name

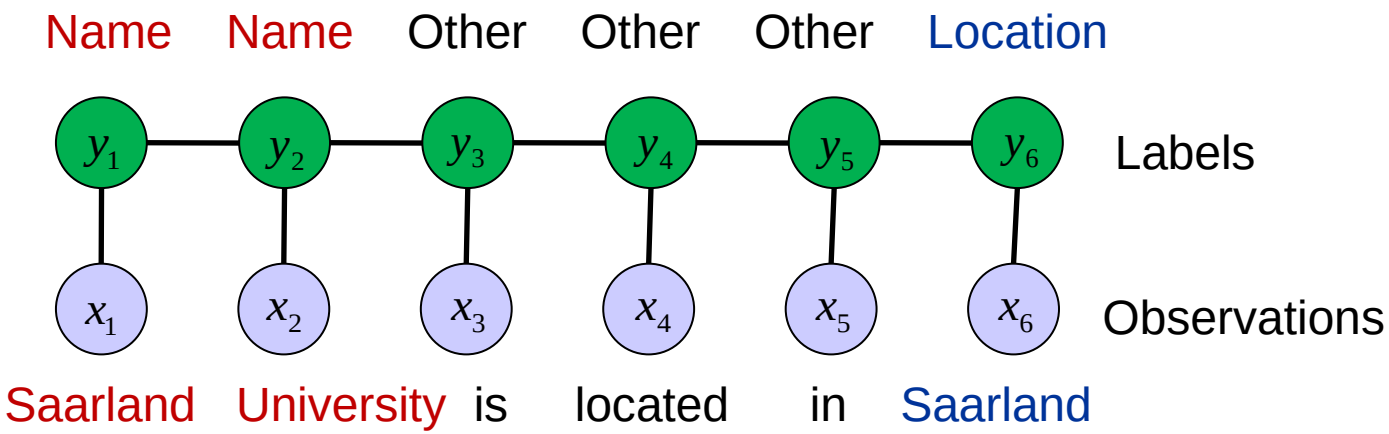
Normalization

Weight

State feature function

Word starts with upper case

Conditional random fields (CRF)



Model:

Weight vector

Feature function vector

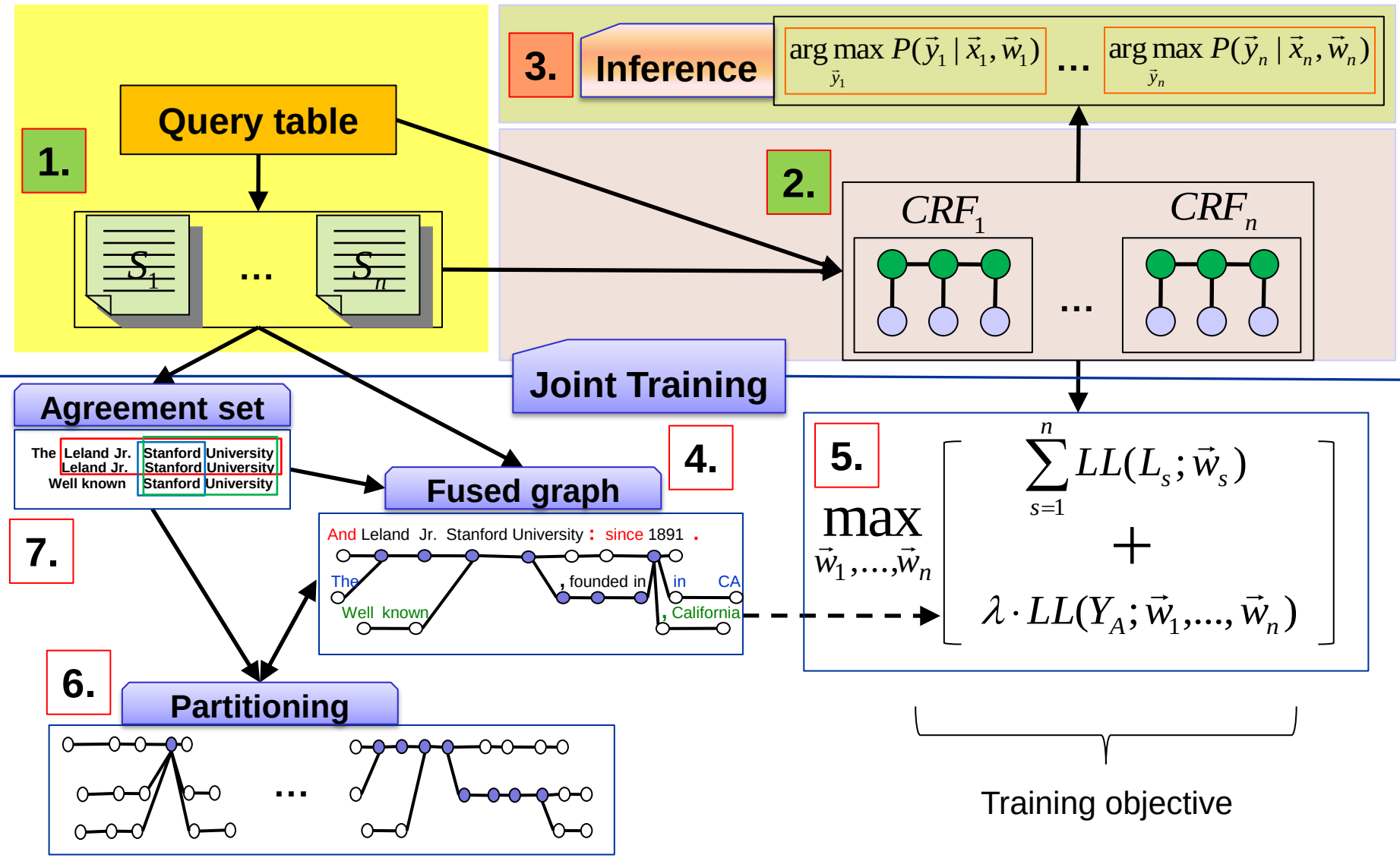
Word starts with upper case

Previous label is Name

$$P_s(\vec{y} \mid \vec{x}, \vec{w}_s) = \frac{1}{Z(\vec{x}, \vec{w}_s)} \exp(\vec{w}_s \cdot \vec{f}_s(\vec{x}, \vec{y}))$$

Normalization

Architecture



Inference

Maximization

$$\arg\max_{\vec{y}} P(\vec{y} \mid \vec{x}, \vec{w})$$

What is the most likely labeling y of x?

$O(|L|^n)$

Exponential in y

NP-hard

Marginalization

$$P_{marg}(\vec{y}_C) = \sum_{\vec{y}':\vec{y}'_C = \vec{y}_C} P(\vec{y}' \mid \vec{x}, \vec{w})$$

What is the marginal probability of a subset y_C?

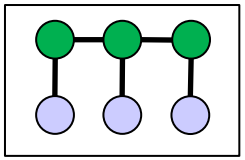
Viterbi ← **Forward-Backward** ← Special case ← **Belief Propagation**

$O(n \cdot |L|^2)$

exact

exact

$chain - CRF$

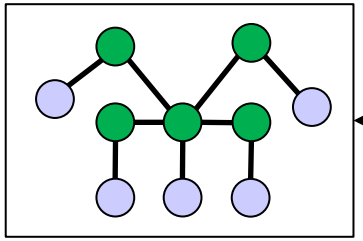


exact

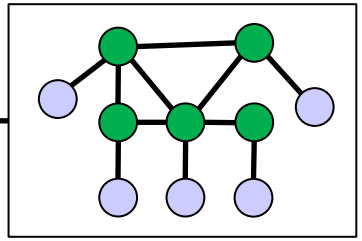
Approx.

$O(n \cdot |L|^T)$

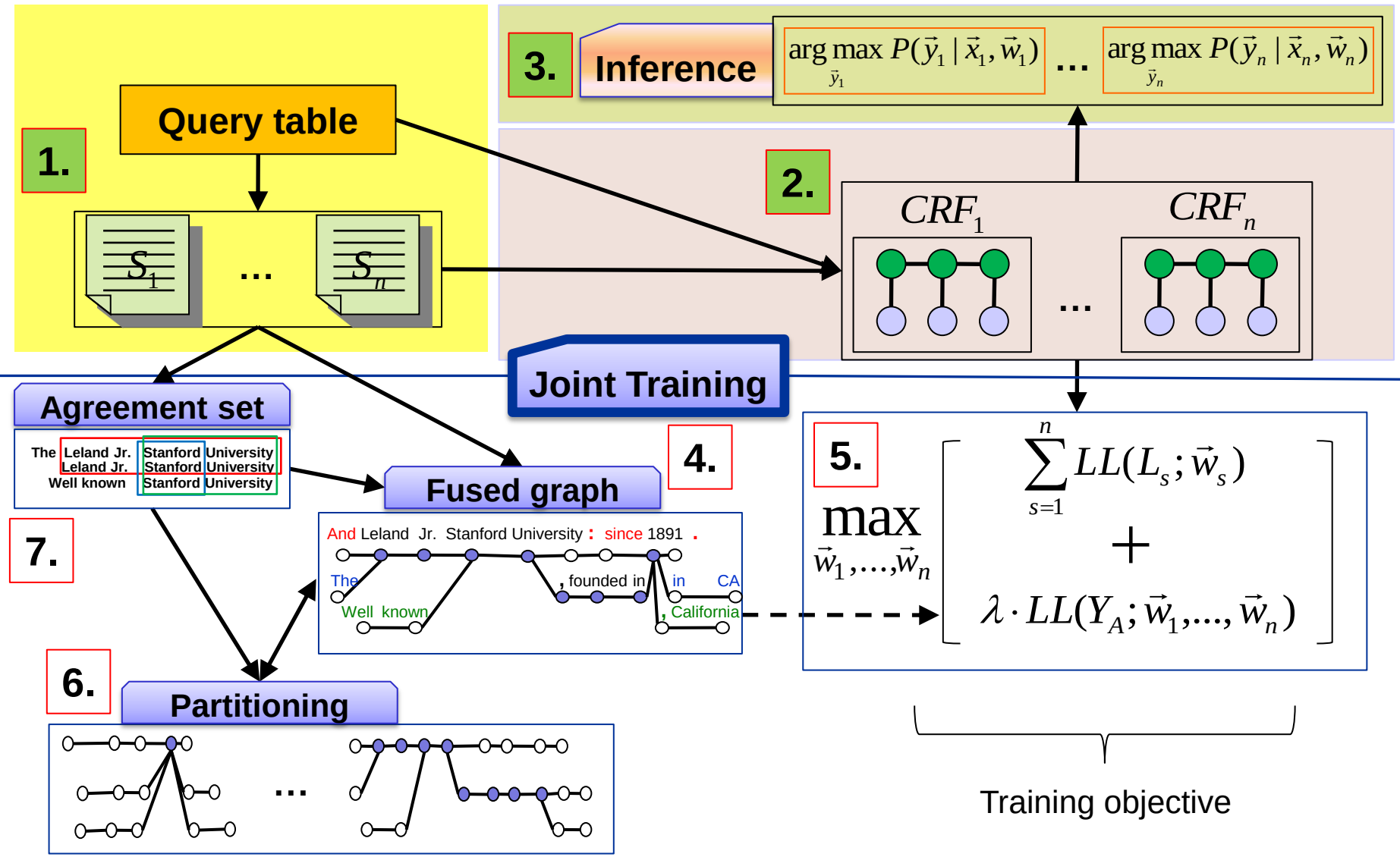
$tree - CRF$



$general - CRF$



Architecture



Why joint training?

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Germany

List source **2**

CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

…

List source **n**

▪Saarland University - 1948 - was established in Saarbrücken in cooperation with France.

▪Stanford University - 1891 - in California.

▪Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

➢Very few labeled data
➢Incomplete information

CRF_1

CRF_n

Stanford University	**California**	**1891**
…	…	…

Florida State University	**Tallahassee**	**-**
…	…	…

Princeton University	**Princeton, NJ**	**1746**
…		…

Overlap

Query table Q

Saarland University	**Saarbrücken**	**1948**
CALTECH	**Pasadena**	**1891**

List source **1**

1891
Stanford University was founded by Leland Stanford in California.
CALTECH is located in Pasadena, California.

1948
Saarland University is located in Germany

List source **2**

CALTECH, Pasadena, CA, in 1891
Stanford University, California, 1891.
University of Southern California, LA, California, 1980.
Florida State University, Tallahassee, 1851

…

List source **n**

- Saarland University - 1948 - was established in Saarbrücken in cooperation with France.
- Stanford University - 1891 - in California.
- Princeton University - 1746 - is a private institution that was founded in Princeton, NJ

Many *overlapping text segments*:

- *Any length*
- *Across any number of sources*

CRF_1

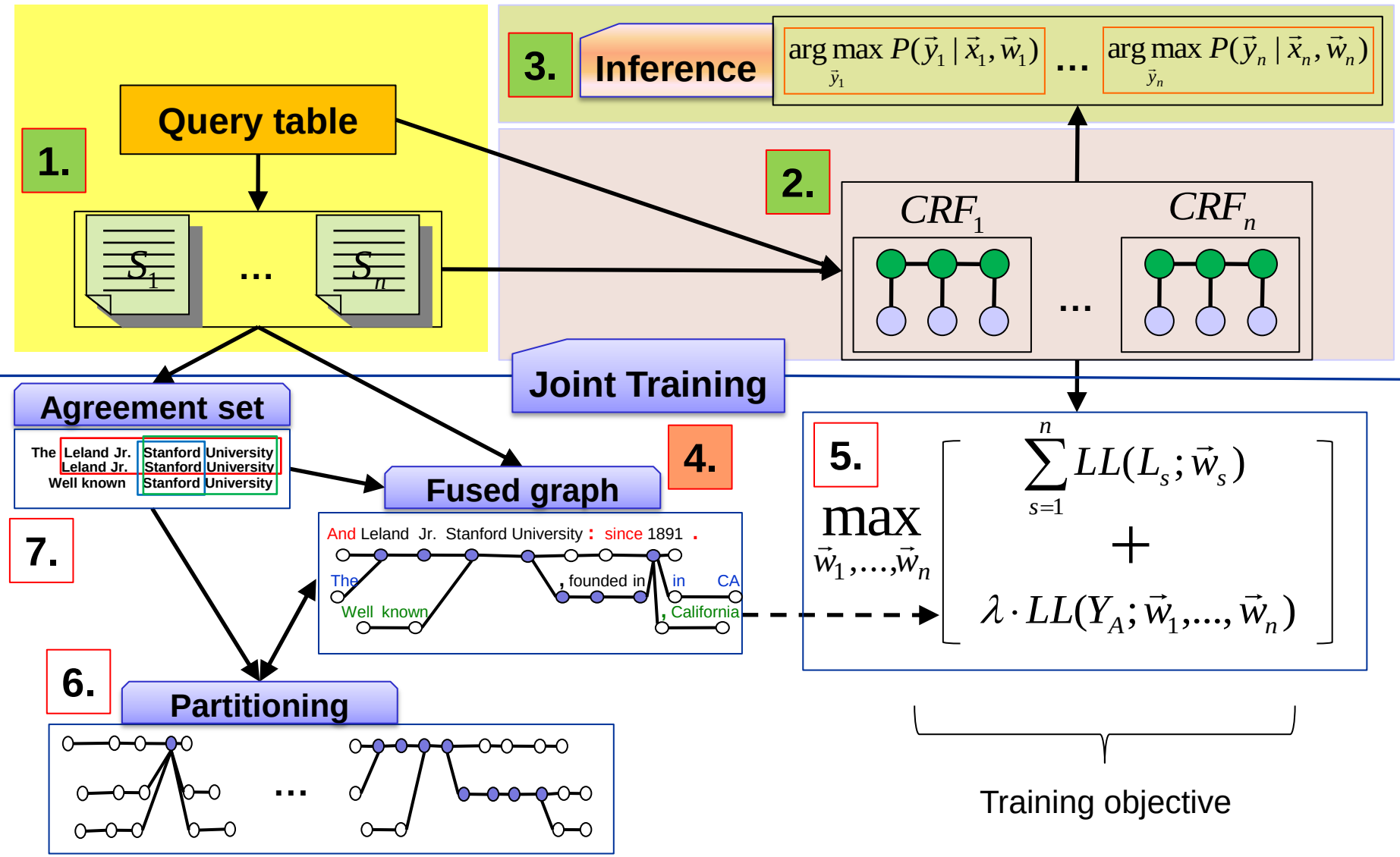
CRF_n

Stanford University	**California**	**1891**
…	…	…

Florida State University	**Tallahassee**	**-**
…	…	…

Princeton University	**Princeton, NJ**	**1746**
…	…	…

Architecture



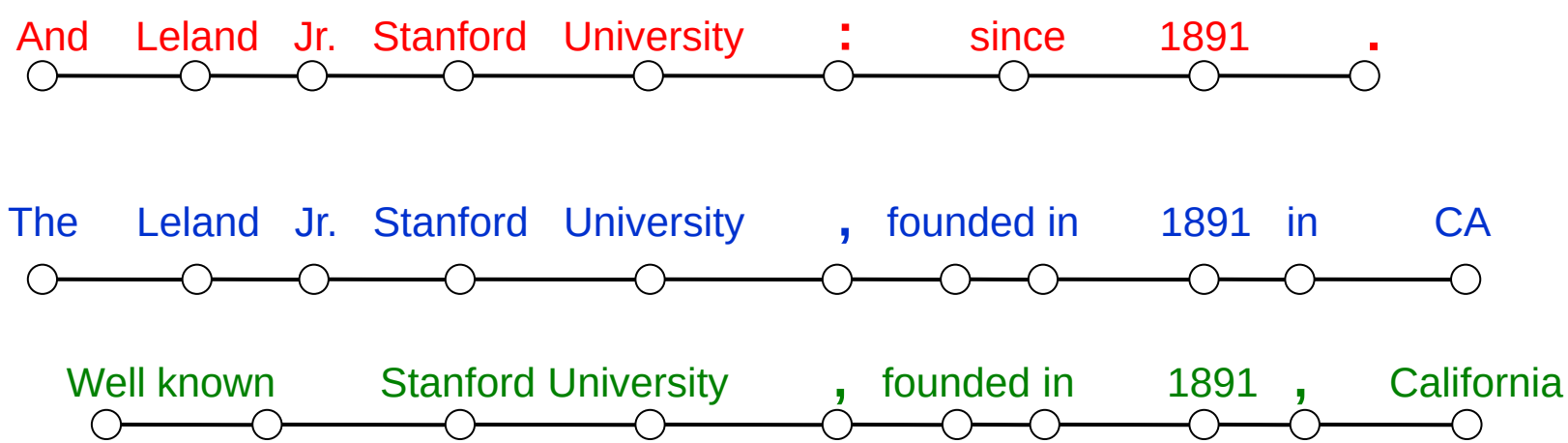
Agreement set

1) And Leland Jr. Stanford University : since 1891 .

2) The Leland Jr. Stanford University , founded in 1891 in CA

3) Well known Stanford University , founded in 1891 , California

A:= { “Leland Jr. Stanford University“. “Stanford University founded in 1891“, “Stanford University“, “1891“ }



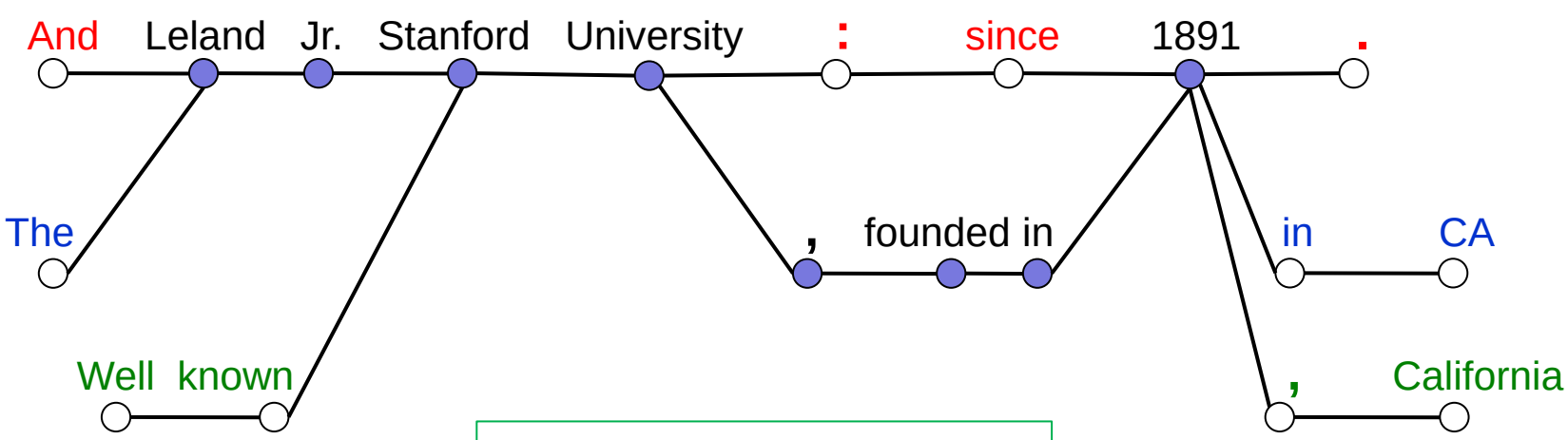
Fused graph

1) And Leland Jr. Stanford University : since 1891 .

2) The Leland Jr. Stanford University , founded in 1891 in CA

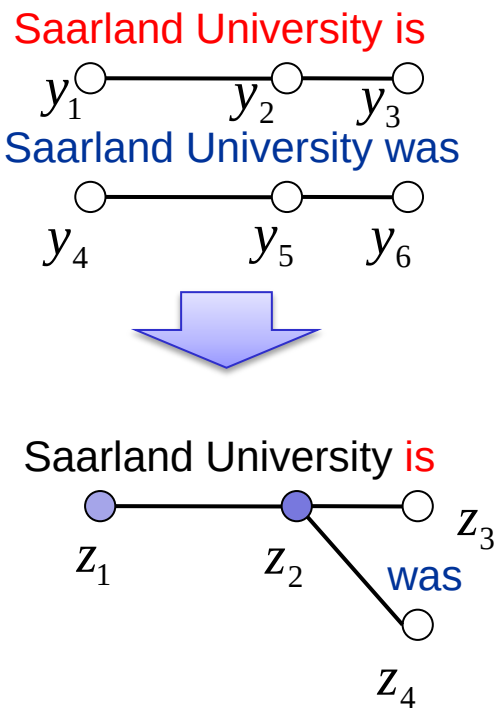
3) Well known Stanford University , founded in 1891 , California

A:= { “Leland Jr. Stanford University“. “Stanford University founded in 1891“, “Stanford University“, “1891“ }



Collapse on segments from A

Fused model



$$P_s(\vec{y}_s \mid \vec{x}_s, \vec{w}_s) = \frac{1}{Z(\vec{x}_s, \vec{w}_s)} \exp\left(\vec{w}_s \cdot \vec{f}_s(\vec{x}_s, \vec{y}_s)\right)$$

Sum up log-factors of same nodes

$$\theta(z_k) = \sum_{node(chain1)=node(chain2)} \vec{w}_s \cdot \vec{f}_s(\vec{x}_s, \vec{y}_s)$$

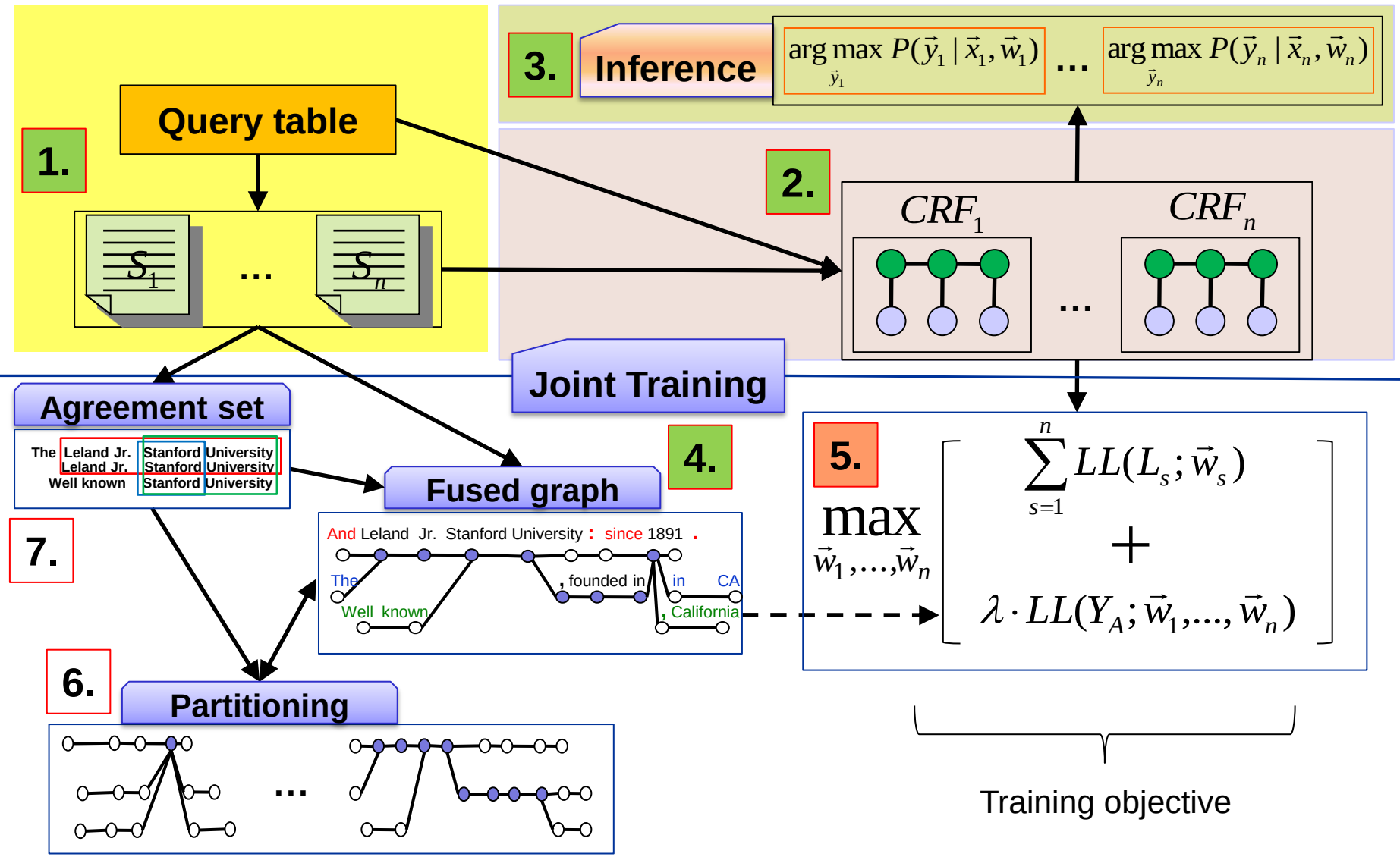
Sum up log-factors of same edges

$$\theta(z_k, z_l) = \sum_{edge(chain1)=edge(chain2)} \vec{w}_s \cdot \vec{f}_s(\vec{x}_s, \vec{y}_s)$$

Fused model:

$$P_A(\vec{z} \mid \theta) = \frac{1}{Z_{fused}} \exp\left(\sum_{fusednodek} \theta(z_k) + \sum_{fusededge(k,l)} \theta(z_k, z_l) \right)$$

Architecture



Joint training

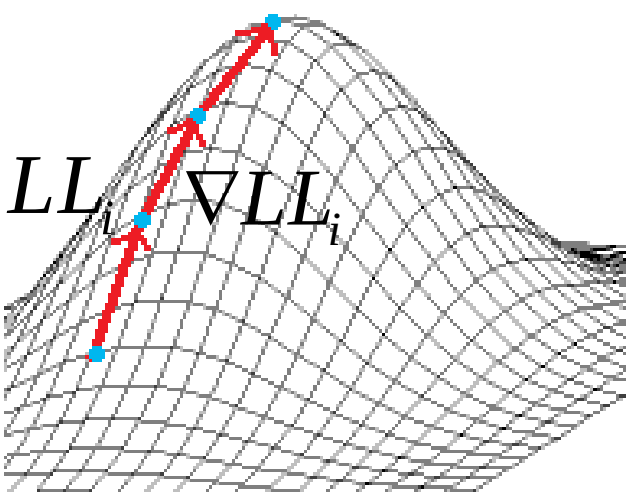
Training CRFs *(Gradient ascent):*

Balancing parameter

$$\max_{\vec{w}_1,\ldots,\vec{w}_n}\left[\sum_{s=1}^{n} LL(L_s;\vec{w}_s)+\lambda\cdot LL(Y_A;\vec{w}_1,\ldots,\vec{w}_n)\right]$$

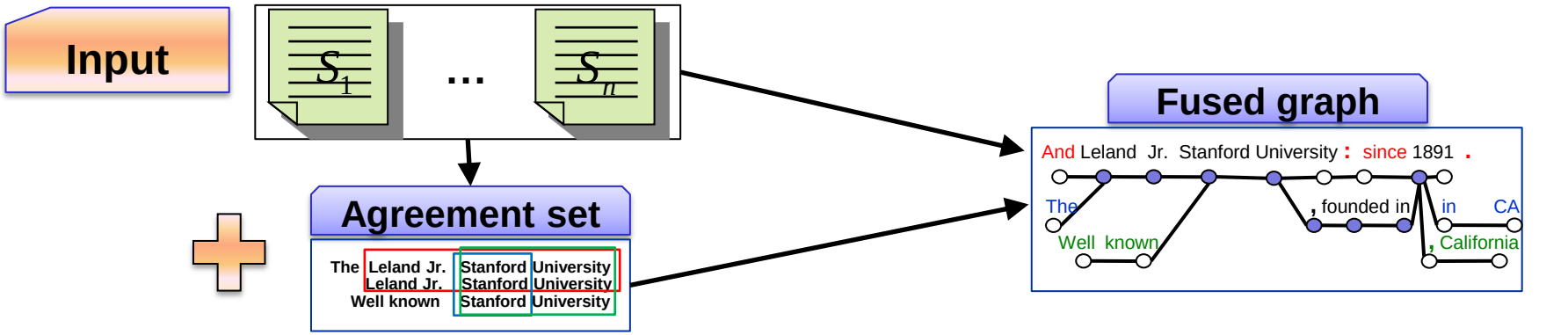
Base Log-Likelihood of labeled records

LogLikelihood of agreement



Goal: learn $\vec{w}_1,\ldots,\vec{w}_n$

Training algorithm (gradient ascent)



Initialize $\vec{w}_1, \ldots, \vec{w}_n$

Iterate:

Compute $LL(L_1; \vec{w}_1), \ldots, LL(L_n; \vec{w}_n)$

and $\nabla LL(L_1; \vec{w}_1), \ldots, \nabla LL(L_n; \vec{w}_n)$

Base CRF models

Compute $LL(Y_A; \vec{w}_1, \ldots, \vec{w}_n)$

and $\nabla LL(Y_A; \vec{w}_1, \ldots, \vec{w}_n)$

Agreement objective

Until convergence

Output $\vec{w}_1, \ldots, \vec{w}_n$

CRF base model training

$$P_s(\vec{y} \mid \vec{x}, \vec{w}_s) = \frac{1}{Z(\vec{x}, \vec{w}_s)} \exp(\vec{w}_s \cdot \vec{f}_s(\vec{x}, \vec{y}))$$

$$\log P_s(\vec{y} \mid \vec{x}, \vec{w}_s) = \vec{w}_s \cdot \vec{f}_s(\vec{x}, \vec{y}) - \log Z(\vec{x}, \vec{w}_s)$$

Feature function vector

Normalization

Weight vector

Training ***(with Gradient ascent):***

$$\max_{\vec{w}_s} LL(L; \vec{w}_s) = \sum_{(\vec{x},\vec{y}) \in L} \log P_s(\vec{y} \mid \vec{x}, \vec{w}_s) = \sum_{(\vec{x},\vec{y}) \in L} \vec{w}_s \cdot \vec{f}_s(\vec{x}, \vec{y}) - \sum_{(\vec{x},\vec{y}) \in L} \log Z(\vec{x}, \vec{w}_s)$$

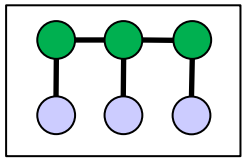
LogLikelihood

Supervised set

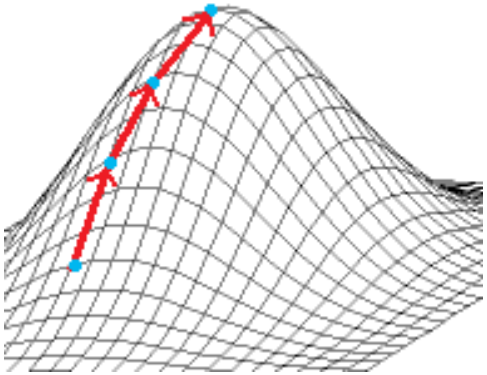
Compute with ***Forward-Backward***

(exact inference)

$chain - CRF$



(marginals of) $\nabla LL(L_i; w_i)$



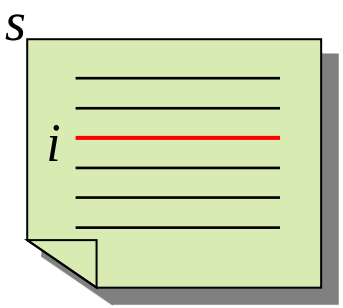
Likelihood of the agreement

Log-Likelihood of agreement

Joint propability, that all CRF agree on A

$$LL(Y_A; \vec{w}_1, \ldots, \vec{w}_n)$$

Probability that CRF for **s** agree on A



$$\log \sum_{\vec{y} \in Y_A} \prod_{(s,i)} P_s(\vec{y}_{(s,i)} \mid \vec{x}_{(s,i)}, \vec{w}_s)$$

$$Y_A := \{\vec{y} \mid \forall C \in A: \vec{y} \text{ labels } C \text{ consistent}\}$$

$$\log Z_{fused} - \sum_{(s,i)} \log Z(\vec{x}_{(s,i)}, \vec{w}_s)$$

Same as for base CRF (Forward-backward)

- *Sequences of labels are consistent in shared segments.*

Likelihood of the agreement

Training many CRFs ***(Gradient ascent):***

Log-Likelihood of agreement

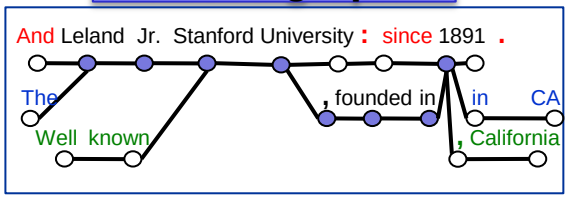
$$LL(Y_A; \vec{w}_1, \ldots, \vec{w}_n) = \log Z_{fused} - \sum_{(s,i)} \log Z(\vec{x}_{(s,i)}, \vec{w}_s)$$

Same as for base CRF (Forward-backward)

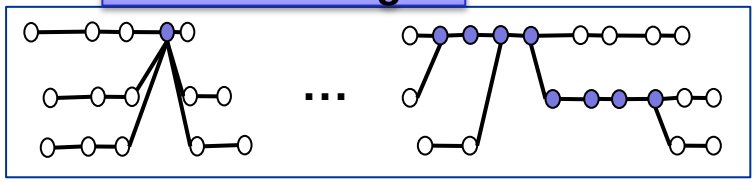
Belief Propagation

☹

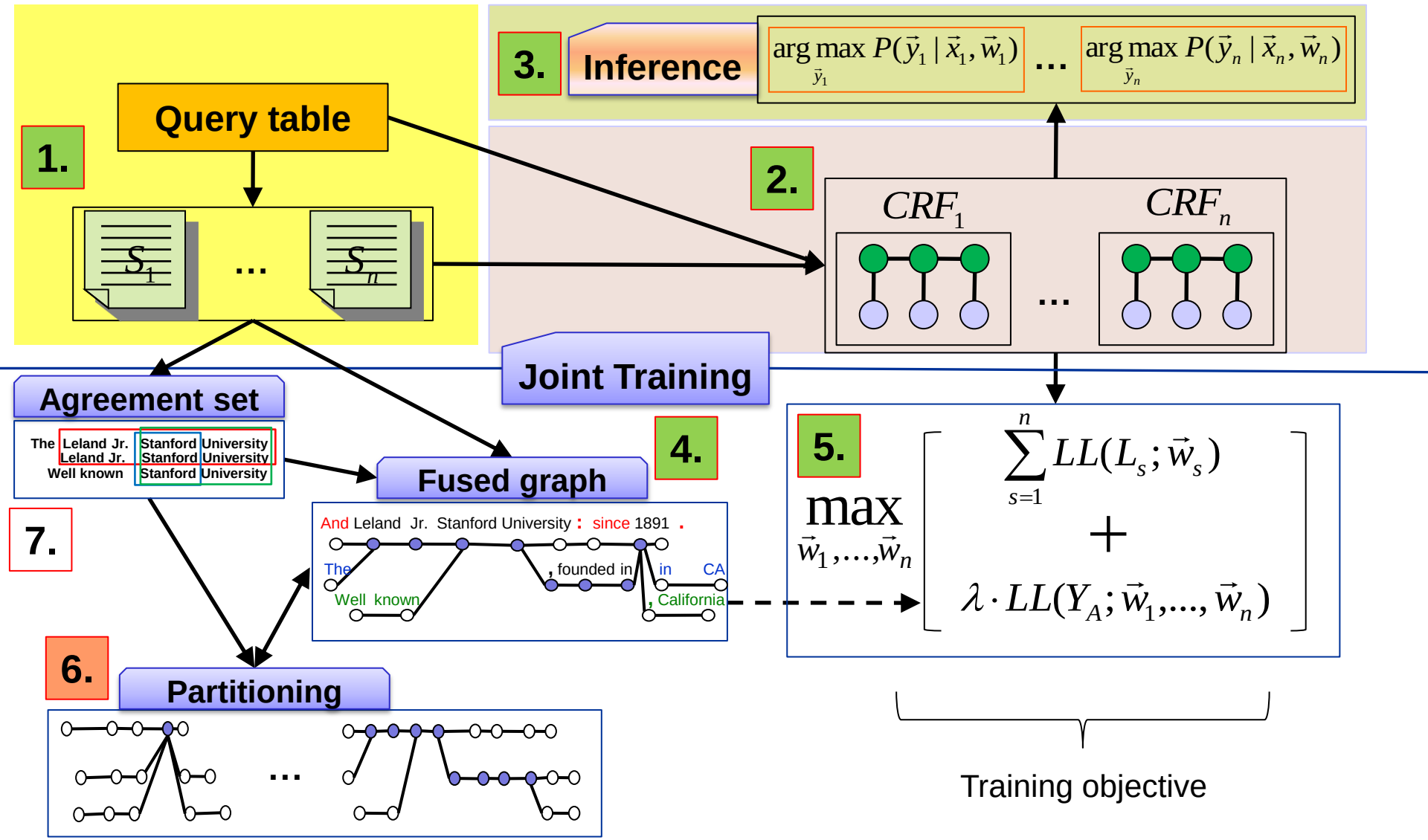
Fused graph



Partitioning



Architecture

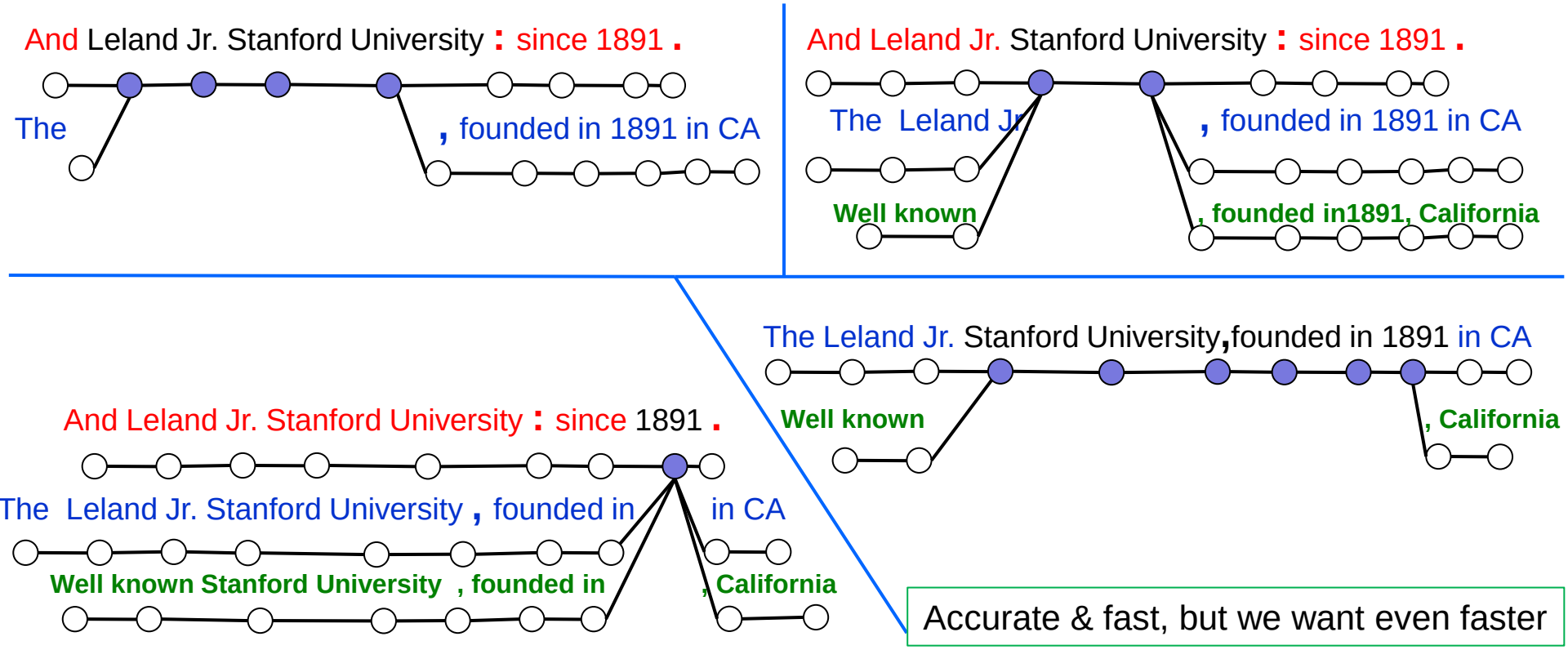


Per segment partitioning

Partition A into smaller sets $A_1, \ldots, A_n$

And Leland Jr.	Stanford University	: since	1891	.
The Leland Jr.	Stanford University	, founded in	1891	in CA
Well known	Stanford University	, founded in	1891	, California

A:= { “Leland Jr. Stanford University“. “Stanford University“, “Stanford University founded in 1891“, “1891“ }



Accurate & fast, but we want even faster

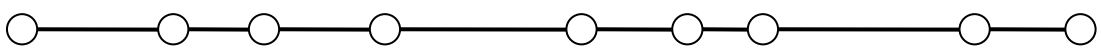
Tree-based partitioning

Partition A into smaller sets $A_1,\ldots,A_n$ and reduce #nodes.

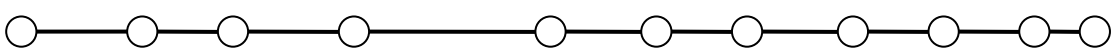
And Leland Jr. Stanford University : since 1891 .
The Leland Jr. Stanford University , founded in 1891 in CA
Well known Stanford University , founded in 1891 , California

A:= {“Stanford University“, “Stanford University founded in 1891“, “Leland Jr. Stanford University“. “1891“ }

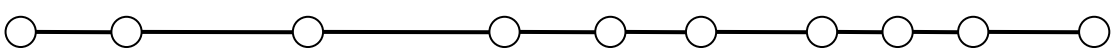
And Leland Jr. Stanford University : since 1891 .



The Leland Jr. Stanford University , founded in 1891 in CA



Well known Stanford University , founded in 1891 , California

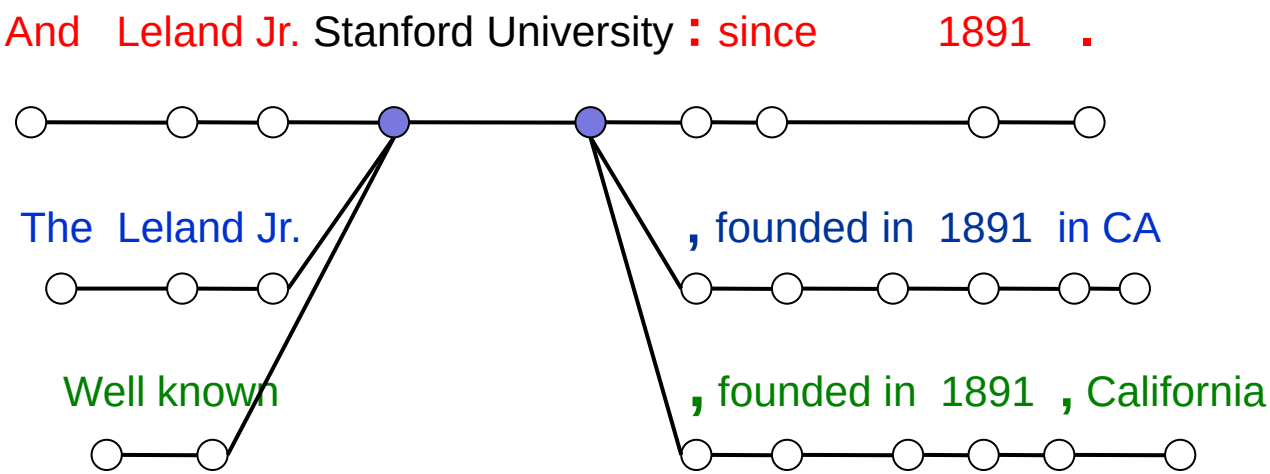


Tree-based partitioning

Partition A into smaller sets $A_1,\ldots,A_n$ and reduce #nodes.

And Leland Jr. Stanford University : since 1891 .
The Leland Jr. Stanford University , founded in 1891 in CA
Well known Stanford University , founded in 1891 , California

A:= {“Stanford University“, “Stanford University founded in 1891“, “Leland Jr. Stanford University“. “1891“ }



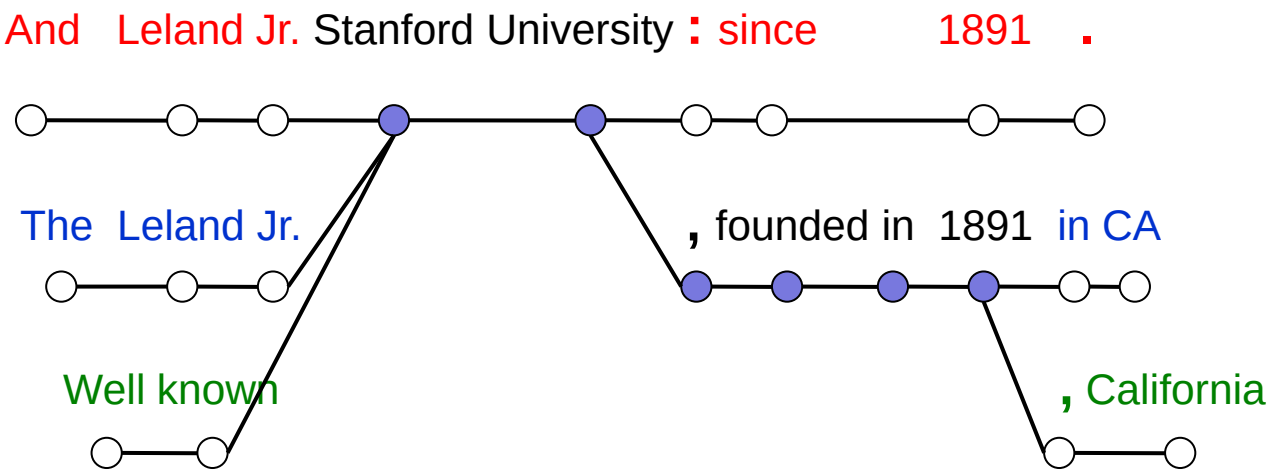
Still a tree? Yes!

Tree-based partitioning

Partition A into smaller sets $A_1,\ldots,A_n$ and reduce #nodes.

And Leland Jr.	Stanford University	: since	1891	.
The Leland Jr.	Stanford University	, founded in	1891	in CA
Well known	Stanford University	, founded in	1891	, California

A:= {“Stanford University“, “Stanford University founded in 1891“, “Leland Jr. Stanford University“. “1891“ }



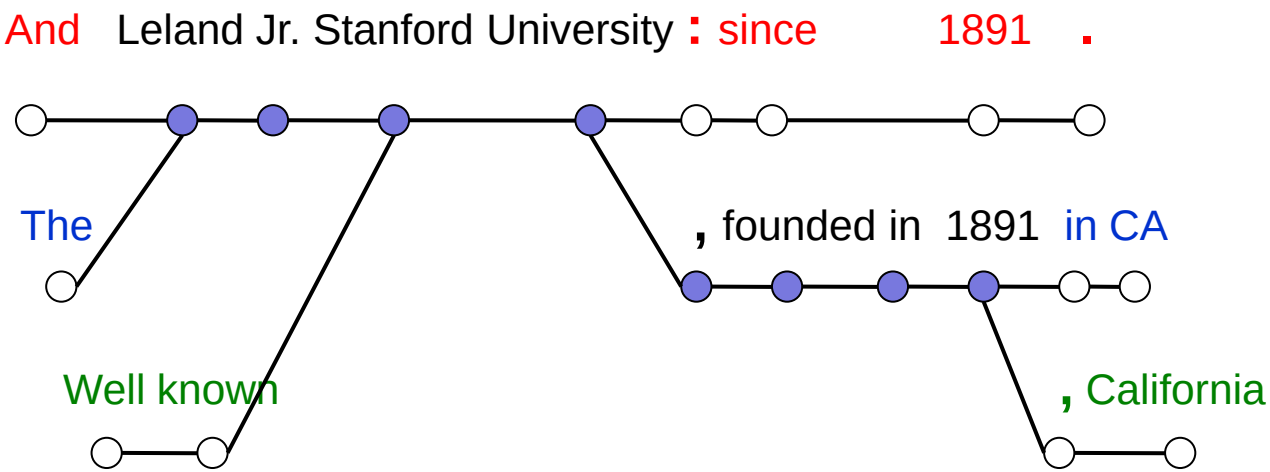
Still a tree? Yes!

Tree-based partitioning

Partition A into smaller sets $A_1,\ldots,A_n$ and reduce #nodes.

And Leland Jr. Stanford University : since 1891 .
The Leland Jr. Stanford University , founded in 1891 in CA
Well known Stanford University , founded in 1891 , California

A:= {“Stanford University“, “Stanford University founded in 1891“, “Leland Jr. Stanford University“. “1891“ }



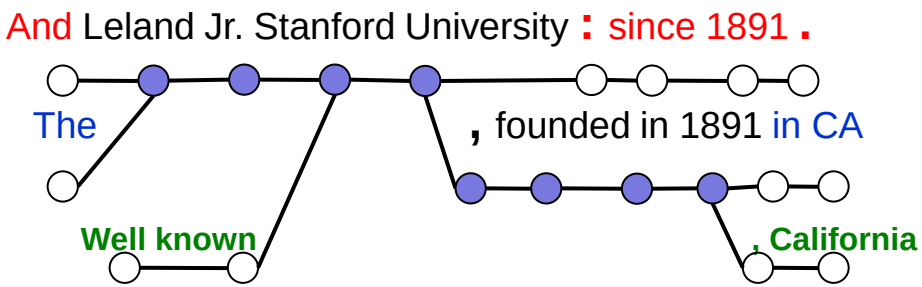
Still a tree? No!

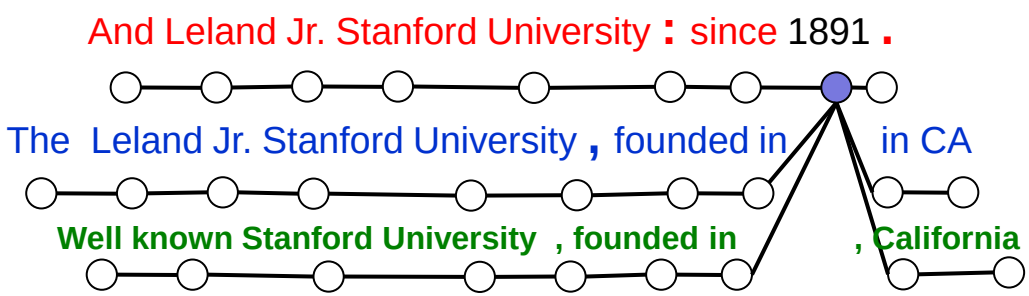
Tree-based partitioning

Partition A into smaller sets $A_1,\ldots,A_n$ and reduce #nodes.

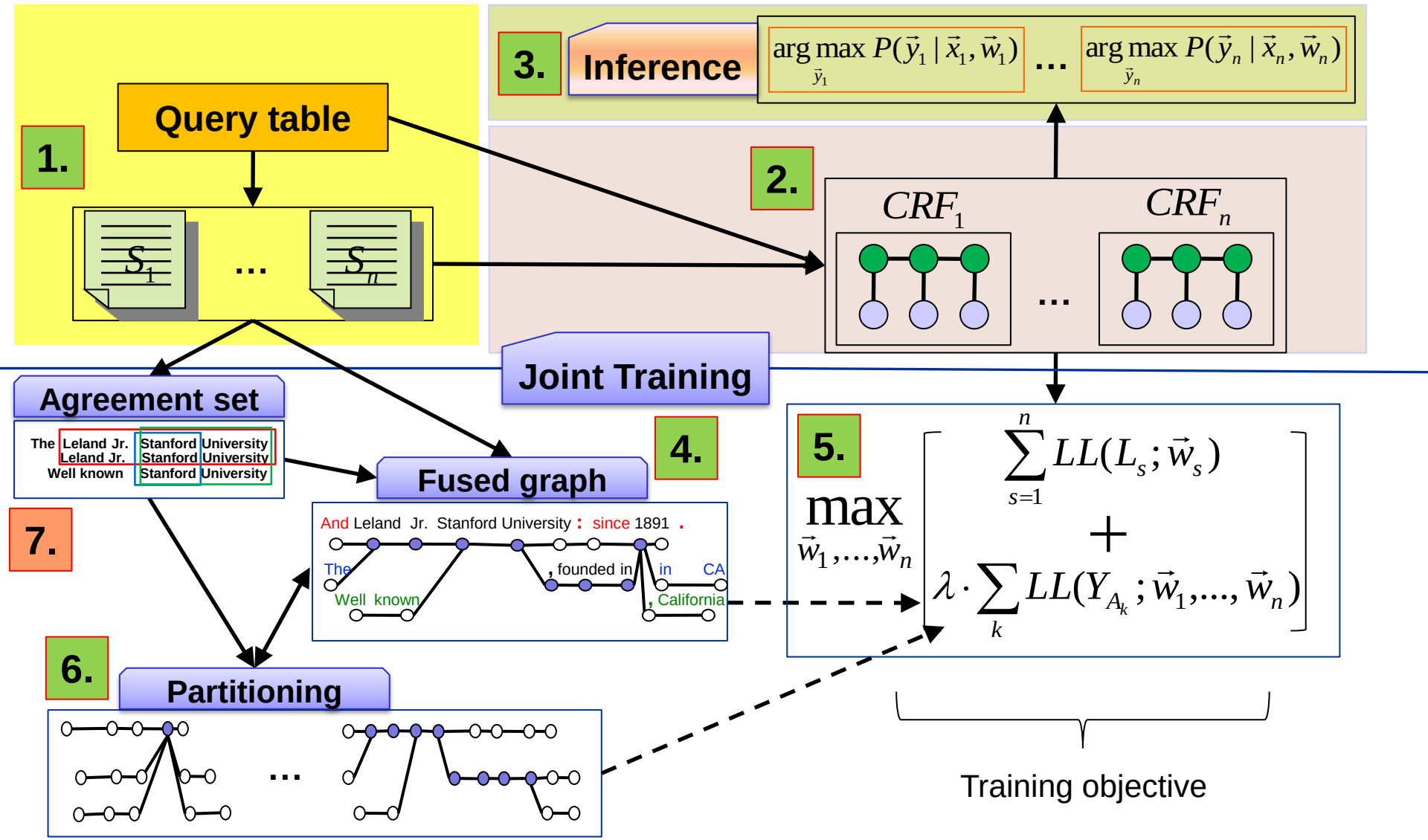
And Leland Jr.	Stanford University	: since	1891	.
The Leland Jr.	Stanford University	, founded in	1891	in CA
Well known	Stanford University	, founded in	1891	, California

A:= {“Stanford University“, “Stanford University founded in 1891“, “Leland Jr. Stanford University“. “1891“ }





Architecture



Noise & problems in agreement set

Saarland University is located in Germany.

Stanford University is founded in 1891 .

Mr. Stanford paid 1891 $ for tickets to New York.

Year | Price

A:= { “Stanford“. “1891“, “University“ }

University | Person | Unigram segment

- *Disambiguate!*
- *Take max segments*

	Item Name	Image	Orbital Period
	Jupiter		4331.572 days
	Pluto		90 613.305 days
	Saturn		29.46 yrs

Good agreement set (step 1)

Stanford, located in California
Saarland University, in Germany
Stanford, located in Palo Alto

Stanford - 1891 - California.
Saarland University - 1948 - Germany
CALTECH - 1891 - Pasadena

Stanford, Palo Alto (1891)
MIT, Cambridge (1861)

Jaccard / TF-IDF

Good agreement set (step 1)

Stanford, located in California
Saarland University, in Germany
Stanford, located in Palo Alto

Stanford - 1891 - California.
Saarland University - 1948 - Germany
CALTECH - 1891 - Pasadena

Stanford, Palo Alto (1891)
MIT, Cambridge (1861)

Cluster records within a source by similarity

Stanford, located in California
Stanford, located in Palo Alto

Saarland University, in Germany

Stanford - 1891 - California.

Saarland University - 1948 – Germany

CALTECH - 1891 - Pasadena

Stanford, Palo Alto (1891)

MIT, Cambridge (1861)

Jaccard / TF-IDF

Good agreement set (step 1)

Stanford, located in California
Stanford, located in Palo Alto

Saarland University, in Germany

Stanford - 1891 - California.

Saarland University - 1948 – Germany

CALTECH - 1891 - Pasadena

Stanford, Palo Alto (1891)

MIT, Cambridge (1861)

merge similar clusters together

Stanford, located in California
Stanford, located in Palo Alto
Stanford - 1891 – California.

Saarland University, in Germany
Saarland University - 1948 – Germany

CALTECH - 1891 – Pasadena

Jaccard / TF-IDF

Good agreement set (step 1)

Stanford, located in California
Stanford, located in Palo Alto
Stanford - 1891 – California.

Saarland University, in Germany
Saarland University - 1948 – Germany

CALTECH - 1891 – Pasadena

Stanford, Palo Alto (1891)

MIT, Cambridge (1861)

merge similar clusters together

Stanford, located in California
Stanford, located in Palo Alto
Stanford - 1891 - California.
Stanford, Palo Alto (1891)

Saarland University, in Germany
Saarland University - 1948 – Germany

CALTECH - 1891 – Pasadena

MIT, Cambridge (1861)

Jaccard / TF-IDF

Good agreement set (step 2)

Stanford, located in California
Stanford, located in Palo Alto
Stanford - 1891 - California.
Stanford, Palo Alto (1891)

Saarland University, in Germany
Saarland University - 1948 – Germany

CALTECH - 1891 – Pasadena

MIT, Cambridge (1861)

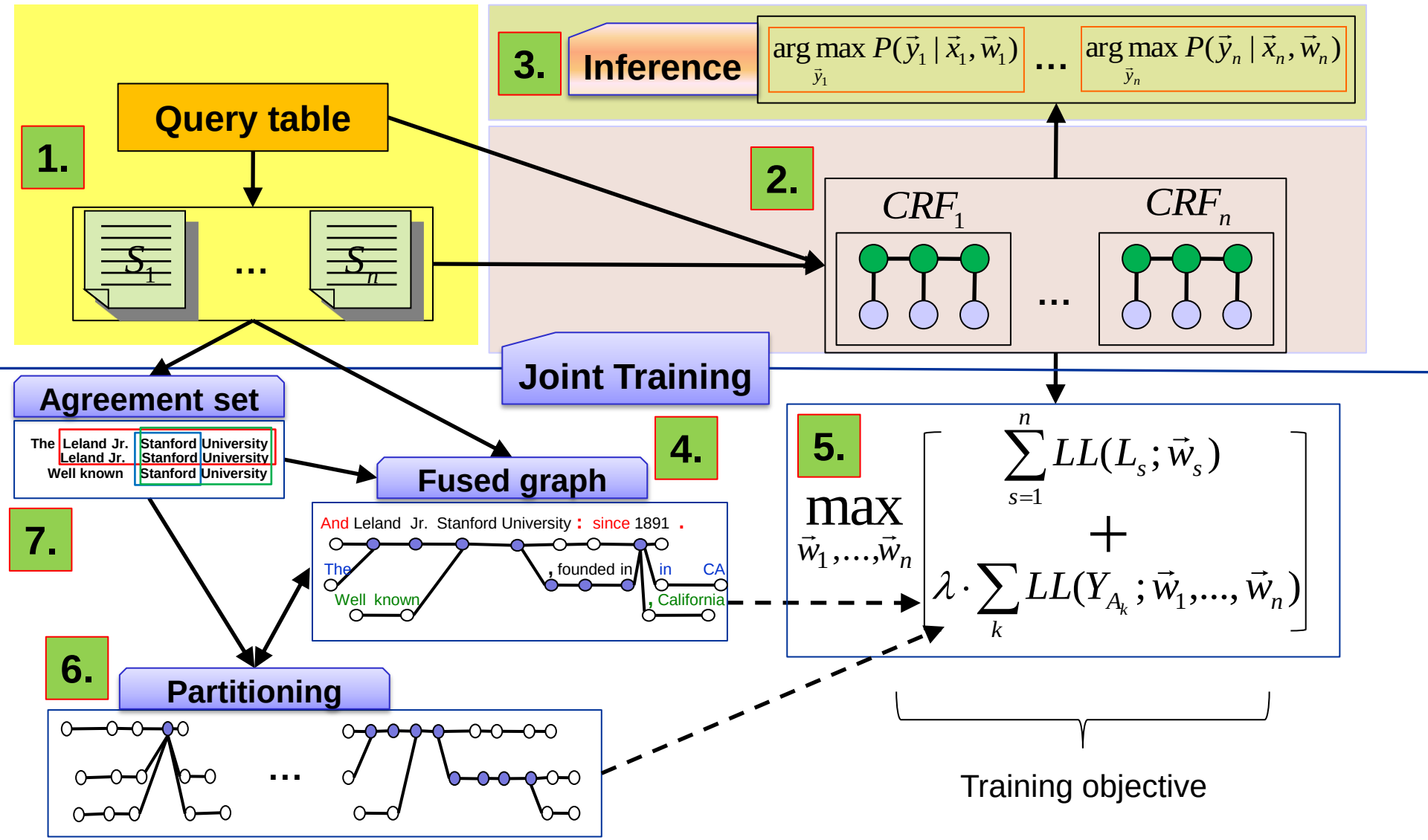
Longest common subsequence problem

- Few noise
- Maximally long segments

A:= { “Stanford located in“. “Palo Alto“, “Saarland University“ }

Jaccard / TF-IDF

Architecture



Other approaches

Collective inference (CI):

Train base CRFs.

Performed over all base CRFs <u>together</u> by running **loopy *max-product* belief propagation** over the **fused graph** constructed from the lists**.**

Label transfer (Staged):

Train CRF for most confident source first, then transfer labels to the next one.

Posterior Regularization (PR):

Different training objective. Trains the various models so as to minimize the distance between the posteriors over the labels of the unlabeled data.

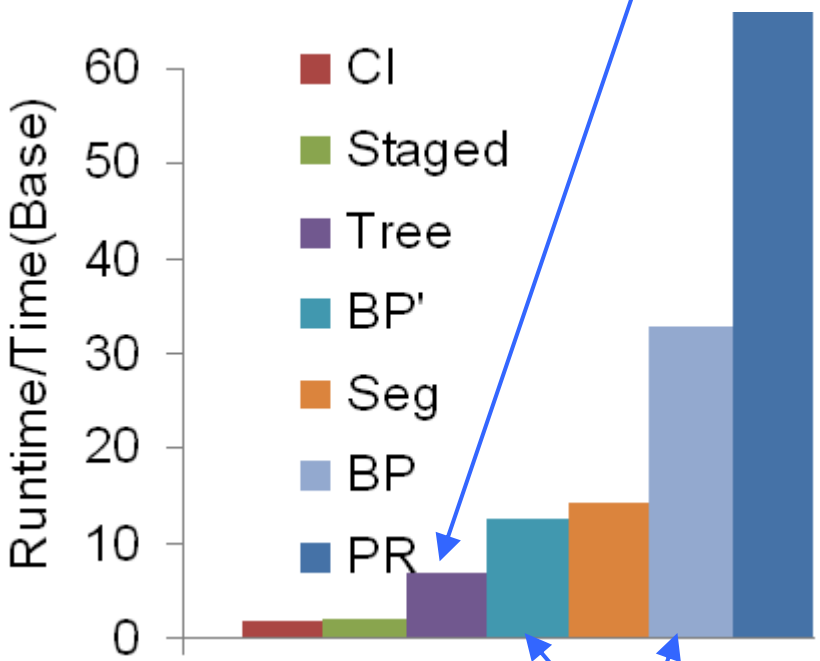
Experimental setting

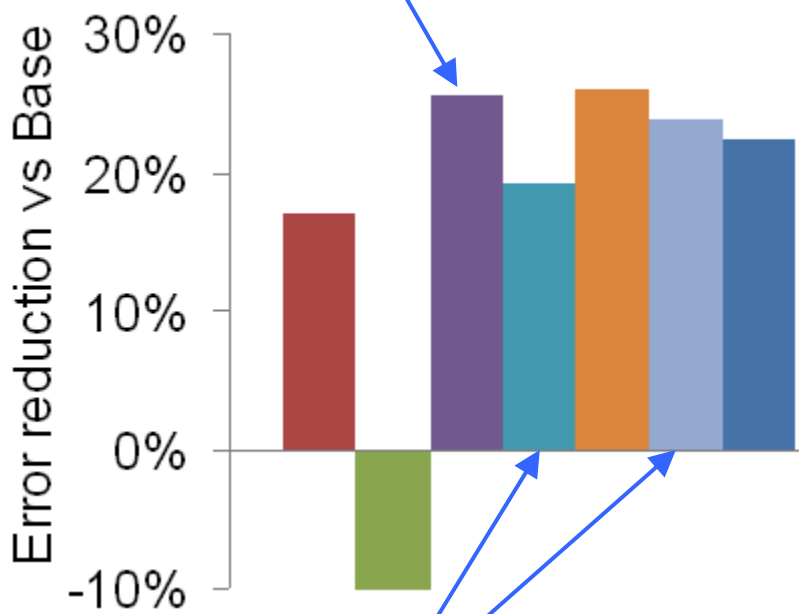
Extraction from 58 datasets (query+lists):

- *Query:* ~ 3 rows
- *Lists:* 2-20 corresponding lists
- *Different domains:* oil spills, Uni mottos, movies,…

Runtime/Accuracy of all methods

Tree-based partitioning has the best runtime/accuracy tradeoff





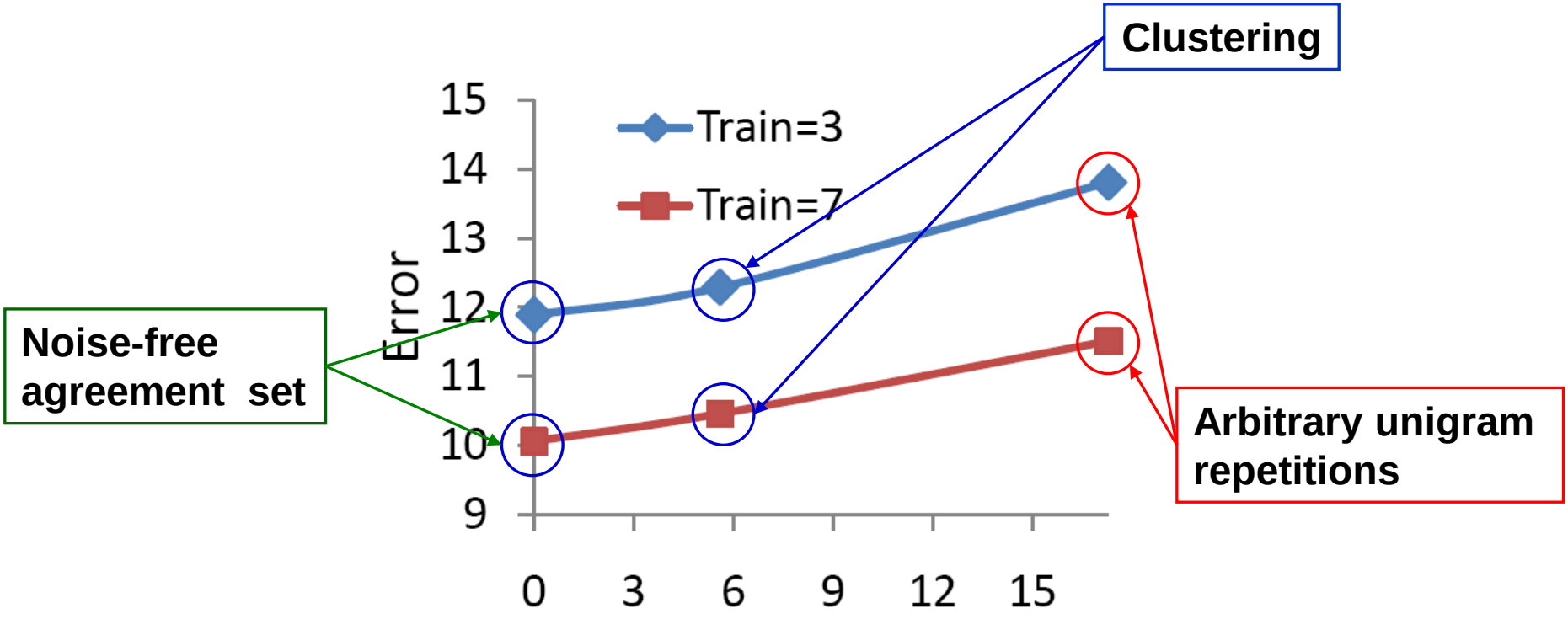
Belief Propagation (BP) quite slow,
Fast variant (BP') not as accurate

Relative error reduction

	50F	50M	40F	40M	30F	30M	20F	20M	10F	10M	All
Absolute F1 Error of Base											
Base	44.8	45.4	33.1	32.7	26.5	23.9	14.4	13.4	5.7	3.9	16.7
Percentage Error Reduction over Base											
CInfer	1.7	3.2	10.4	3.3	-2.9	16.4	31.3	28.2	10.1	13.1	17.0
Tree	6.0	2.3	11.2	9.5	4.4	28.0	38.0	40.6	43.4	13.8	25.5
Seg	6.6	0.6	14.3	9.8	4.5	31.5	38.8	42.7	36.2	9.3	26.8
BP	6.0	2.4	10.6	9.3	3.6	28.7	38.6	42.0	43.3	14.9	26.0
BP'	1.6	2.1	11.8	3.5	-3.1	18.6	34.3	35.0	13.2	-0.5	19.1
PR	2.3	7.9	4.7	10.3	4.1	28.7	30.5	33.3	30.2	9.3	22.4

Red: Increase in error
Green: Best method

Noise in agreement set



- Clustering: ~5% noise, small F1 rise
- Arbitrary unigrams: ~15% noise, significant F1 rise

Conclusion

Joint training:

- Use the *text overlap* to compensate for a *lack of supervision*
- Strategy to find *low-noise agreement set*
- Partitioning of *fused graph* for tractability.

Best accuracy/speed tradeoff with tree-based partitioning

Thank You!