Legal Claim Identification:
Information Extraction with Hierarchically Labeled Data

Mihai Surdeanu, Ramesh Nallapati and Christopher Manning

Presented by: Daniel Cer

...

31. On February 20, 2007, the USPTO duly and legally issued United States Pa\tent No. 7,179,046 B2 ("the '046 patent"), also entitled
"Fan array fan section in air-handling
8
systems." Huntair is the owner by assignment of all right, title and interest in and to the '046 patent. A copy of the '046 patent is attached
to the Complaint as Exhibit A.

FIRST COUNTERCLAIM INFRINGEMENT OF U.S. PATENT NO. 7,137,775 B2 32. Huntair repeats and realleges paragraphs 26-31 as though fully set forthn herein.

33. Upon information and belief, Plaintiff is and continues to be directly infringing, contributorily infringing, and/or inducing infringement of the '775 patent by, among other things, making, using,
offering to sell, selling and/or
importing, without authority or license from
Plaintiff, fan arrays in this district and elsewhere in the United States, which embody, incorporate, or otherwise practice one or more
claims of the '775 patent.

34. Upon information and belief, in its bid to obtain a contract to install an array of\
fans at facilities owned by Amcol in Chicago, Illi
nois, Plaintiff offered to utilize a fan system
that contains, embodies, and employs the invention described and claimed in the '775 patent.

35. Plaintiff's conduct constitutes infringement, as provided by 35 U.S.C. $ 271, of one or more claims of the '775 patent.

36. As a result of this infringement, Huntair has been damaged and deprived of the gains and profits to which it is entitled. Furthermore, Huntair will continue to be damaged unless this Court enjoins Plaintiff's infringing conduct.

...

Claim

...

31. On February 20, 2007, the USPTO duly and legally issued United States Pa\tent No. 7,179,046 B2 ("the '046 patent"), also entitled
"Fan array fan section in air-handling
8
systems." Huntair is the owner by assignment of all right, title and interest in and to the '046 patent. A copy of the '046 patent is attached
to the Complaint as Exhibit A.

FIRST COUNTERCLAIM INFRINGEMENT OF U.S. PATENT NO. 7,137,775 B2 32. Huntair repeats and realleges paragraphs 26-31 as though fully set forthn herein.

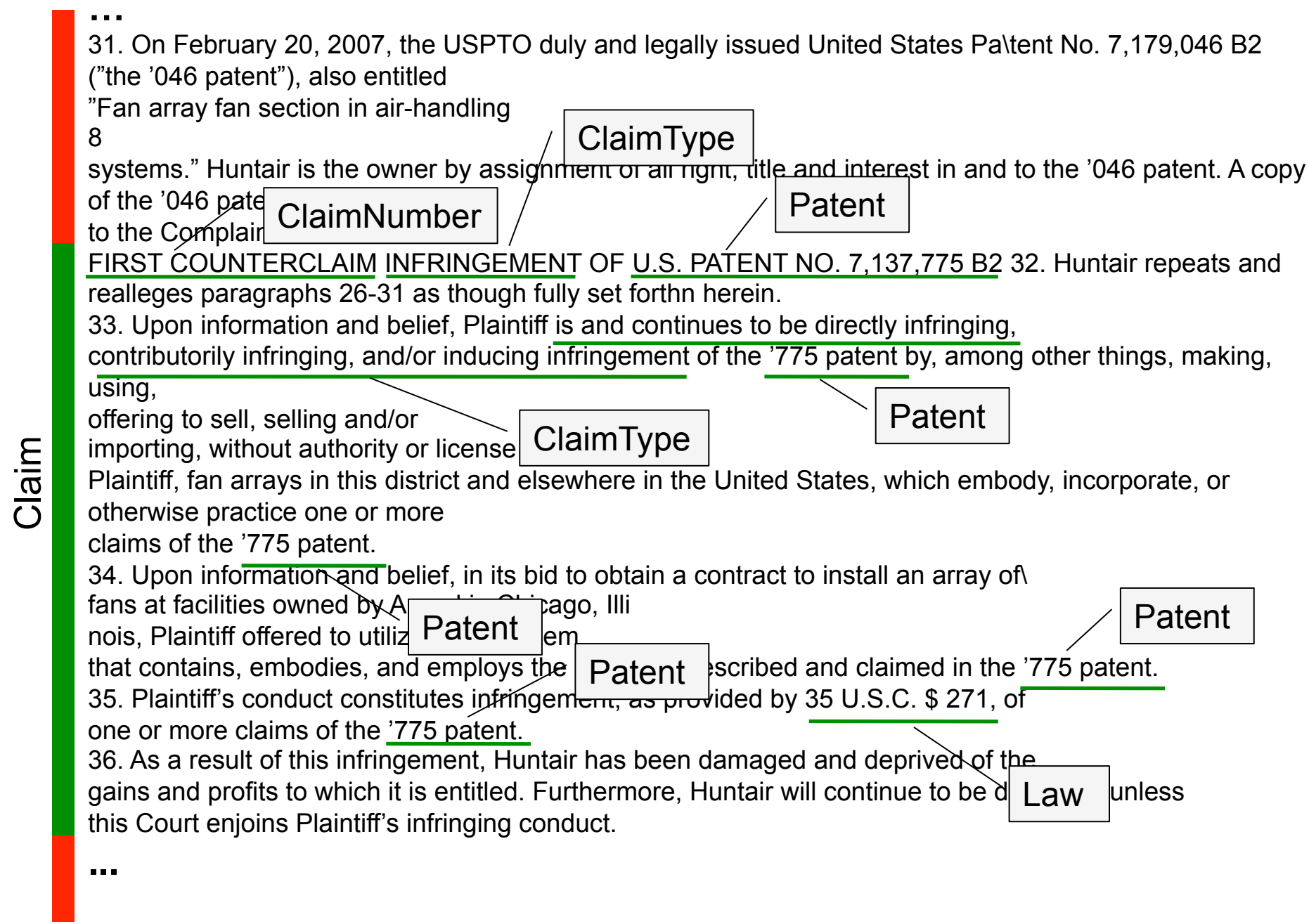
33. Upon information and belief, Plaintiff is and continues to be directly infringing, contributorily infringing, and/or inducing infringement of the '775 patent by, among other things, making, using,
offering to sell, selling and/or
importing, without authority or license from
Plaintiff, fan arrays in this district and elsewhere in the United States, which embody, incorporate, or otherwise practice one or more
claims of the '775 patent.

34. Upon information and belief, in its bid to obtain a contract to install an array of\
fans at facilities owned by Amcol in Chicago, Illi
nois, Plaintiff offered to utilize a fan system
that contains, embodies, and employs the invention described and claimed in the '775 patent.

35. Plaintiff's conduct constitutes infringement, as provided by 35 U.S.C. $ 271, of
one or more claims of the '775 patent.

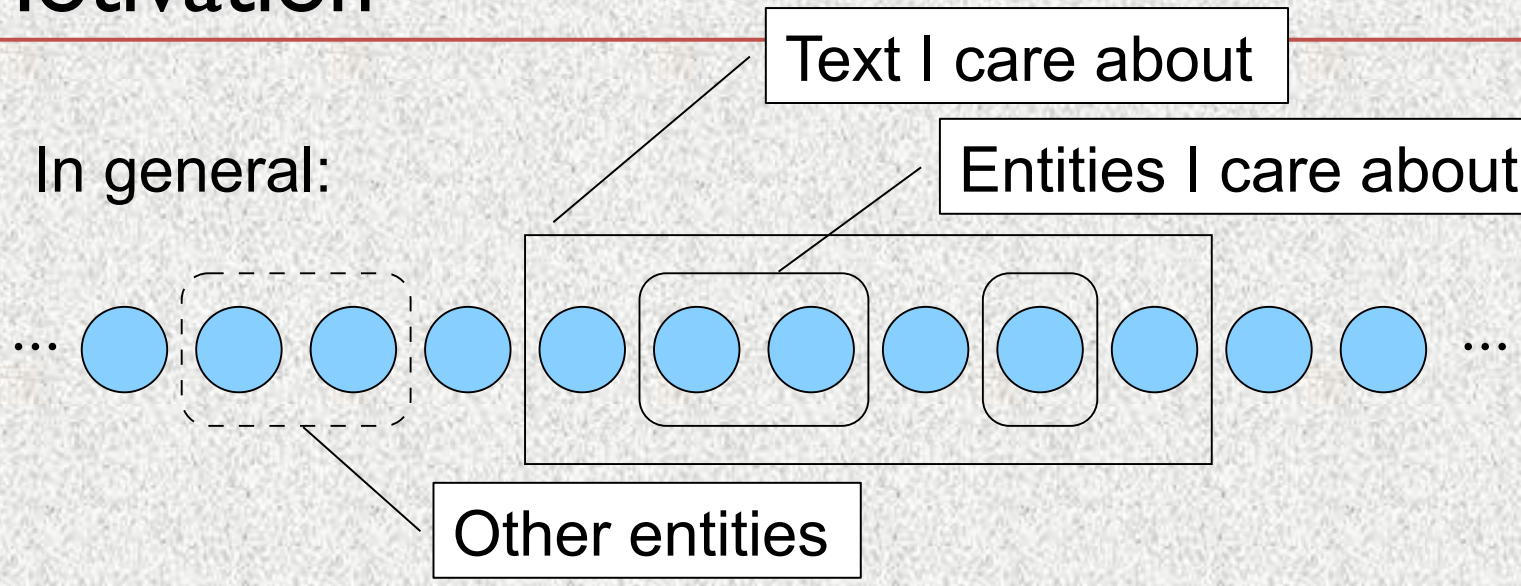
36. As a result of this infringement, Huntair has been damaged and deprived of the gains and profits to which it is entitled. Furthermore, Huntair will continue to be damaged unless this Court enjoins Plaintiff's infringing conduct.

...


...
31. On February 20, 2007, the USPTO duly and legally issued United States Pa\tent No. 7,179,046 B2 ("the '046 patent"), also entitled
"Fan array fan section in air-handling
8
systems." Huntair is the owner by assignment of all right, title and interest in and to the '046 patent. A copy of the '046 pate
to the Complair
FIRST COUNTERCLAIM INFRINGEMENT OF U.S. PATENT NO. 7,137,775 B2 32. Huntair repeats and realleges paragraphs 26-31 as though fully set forthn herein.
33. Upon information and belief, Plaintiff is and continues to be directly infringing, contributorily infringing, and/or inducing infringement of the '775 patent by, among other things, making, using,
offering to sell, selling and/or
importing, without authority or license
Plaintiff, fan arrays in this district and elsewhere in the United States, which embody, incorporate, or otherwise practice one or more
claims of the '775 patent.
34. Upon information and belief, in its bid to obtain a contract to install an array of\
fans at facilities owned by A cago, Illi
nois, Plaintiff offered to utiliz em
that contains, embodies, and employs the escribed and claimed in the '775 patent.
35. Plaintiff's conduct constitutes infringement, as provided by 35 U.S.C. $ 271, of one or more claims of the '775 patent.
36. As a result of this infringement, Huntair has been damaged and deprived of the gains and profits to which it is entitled. Furthermore, Huntair will continue to be d unless this Court enjoins Plaintiff's infringing conduct.
...
Claim
ClaimType
ClaimNumber
Patent
ClaimType
Patent
Patent
Patent
Patent
Law

Motivation

- In general:



- Generic IE problem, e.g.:
  - Someone interested only in information about the Olympic Games in documents about London
  - Blog readers interested only in passages about new gadgets
- Applications:
  - summarization, semi or structured search, visualization

Contributions

- Proposed a novel IE task motivated by a real-world application in the legal domain
- Experimented with multiple graphical models:
  - Conditional random fields
  - Semi-supervised
  - Hierarchical
  - Joint
- Evaluated on data from actual IP cases

Overview

Motivation

- Approach

Experiments

Conclusions

Problem Description (1/2)

- Recognize claim boundaries and relevant entities in case pleadings
- Text extracted from PDFs or OCRed
- Many data errors:
  - Incorrect pagination, missing or extraneous characters (“Pa\tent”), broken words (”Illi nois”), etc.

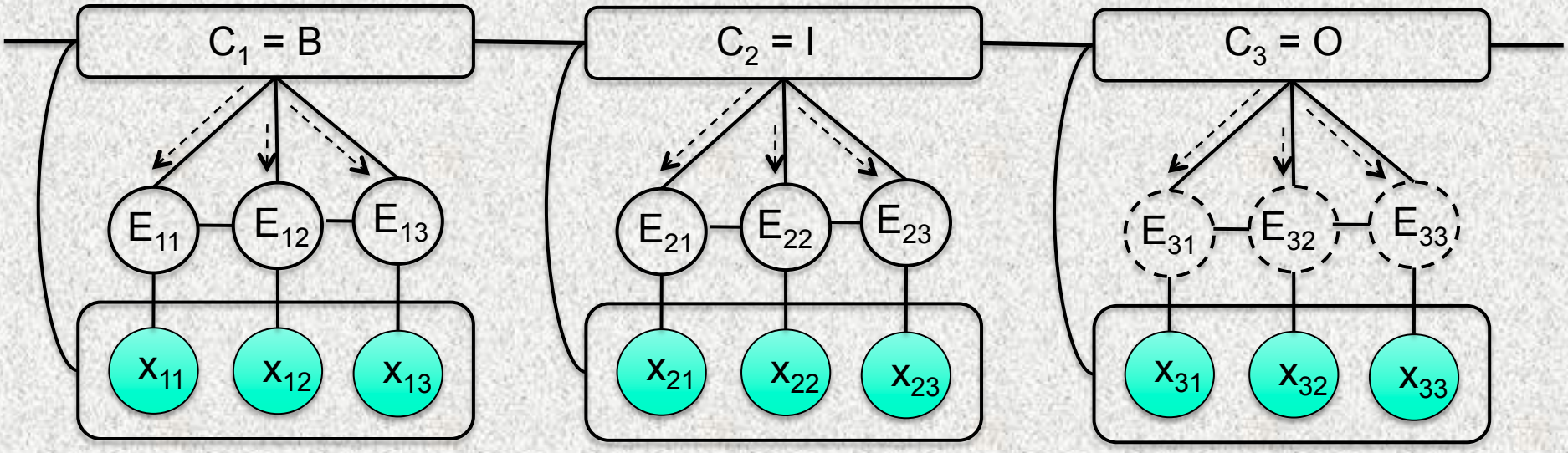
Problem Description (2/2)

- Two layers of annotations
- Top layer of annotation: *claim segments*
  - All text that is vital to understand the claim but no extraneous material
- Bottom layer of annotation: *claim entities*
  - **Patent**: “United States Patent No. 6,190,044”, “’044 patent”
  - **Law**: “35 U.S.C. $ 281, 283, 284, and 285” or “California 7 Business & Professions Code $ 17200, et seq.”
  - **ClaimNumber**: “First cause of action”, “Second claim for relief”
  - **ClaimType**: “INFRINGEMENT”, “is and continues to be directly infringing, contributorily infringing, and/or inducing infringement”

Models
Notations

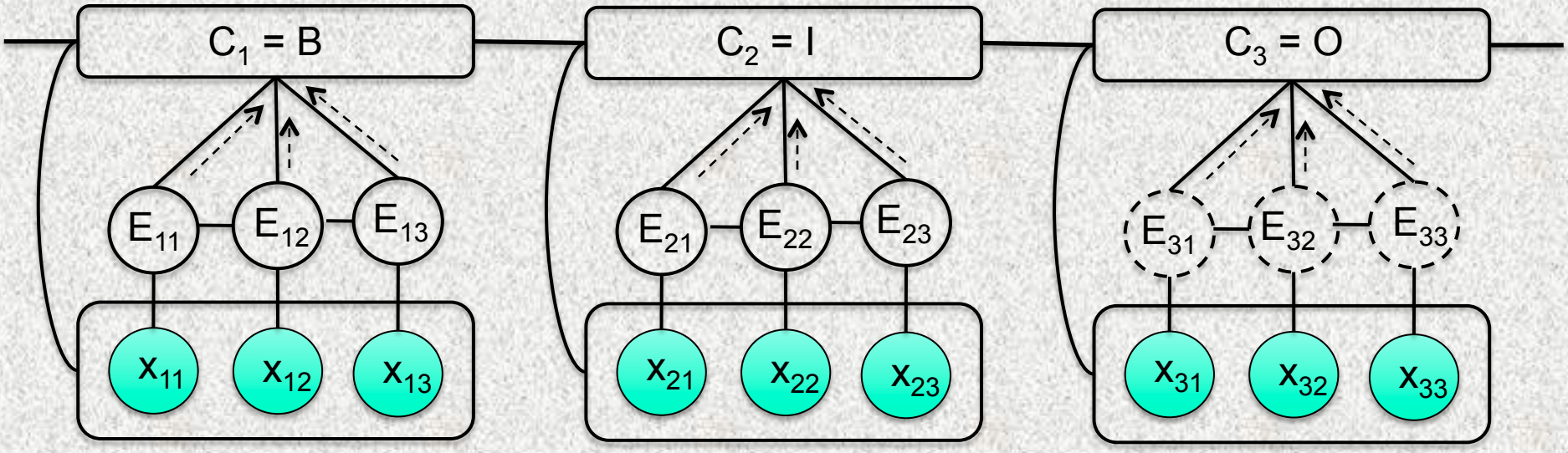
- Top layer annotates sentences; bottom layer annotates words
- We use the Begin (B) – Inside (I) – Outside (O) notation to mark relevant segments in both layers
- Top layer:
  - C_s – label of one sentence, in {B, I, O}
  - $\mathbf{C}_d$ – labels for an entire document
- Bottom layer:
  - E_i – label of word i, in {B, I, O} x {N, T, P, L}
  - $\mathbf{E}_s$ – labels for an entire sentence
- Surface text:
  - X_i – word at position i in a sentence
  - $\mathbf{X}_s$, $\mathbf{X}_d$ – labels for entire sentences or documents

Models
Top-Down CRF



Claim Model	Entity Model	Order of Inference
$P(\mathbf{C}_d \mid \mathbf{x}_d)$	$P(\mathbf{E}_s \mid \mathbf{x}_s, c_s)$	$\mathbf{c}^{(p)} \rightarrow \mathbf{e}^{(p)}$

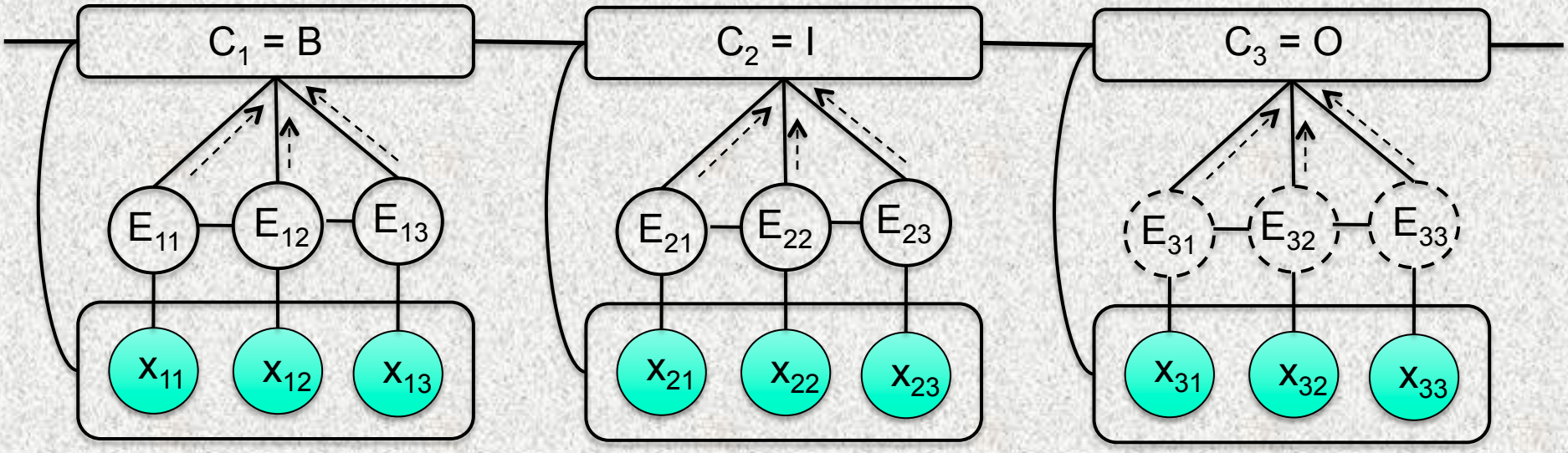
Models
Bottom-up CRF



Claim Model	Entity Model	Order of Inference
$P(\mathbf{C}_d \mid \mathbf{e}_d, \mathbf{x}_d)$	$P(\mathbf{E}_s \mid \mathbf{x}_s)$	$\mathbf{e}^{(p)} \rightarrow \mathbf{c}^{(p)} \rightarrow \mathbf{e}^{(\text{constraints})}$

Models

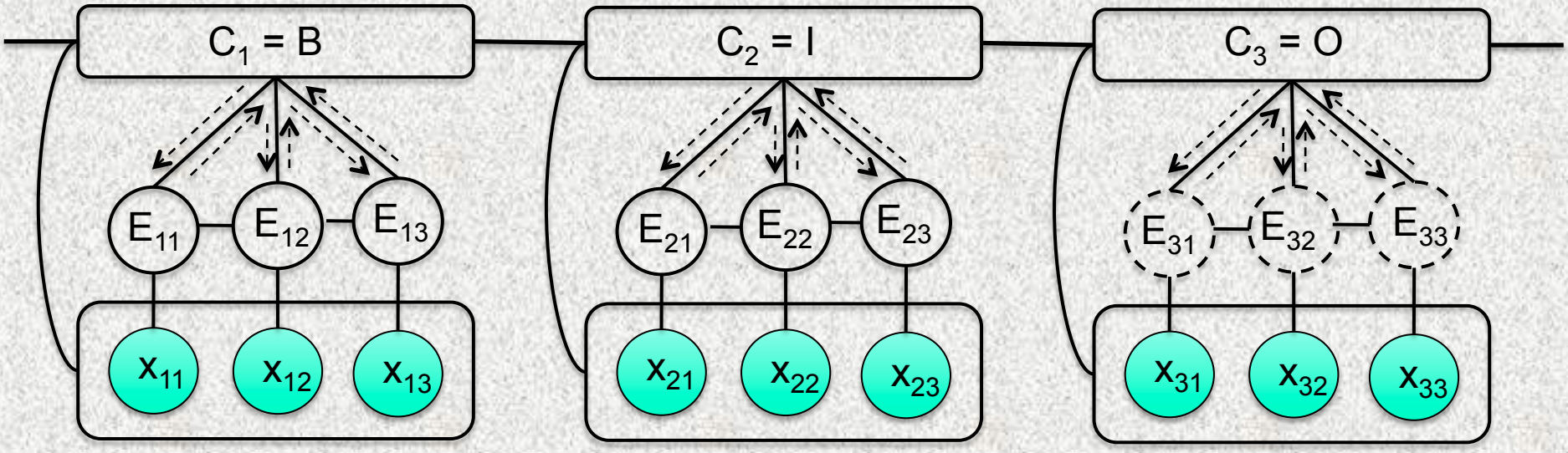
Semi-supervised Bottom-up CRF



Claim Model	Entity Model	Order of Inference
$P(\mathbf{C}_d \mid \mathbf{e}_d^{(semi)}, \mathbf{x}_d)$	$P(\mathbf{E}_s \mid \mathbf{x}_s)$	$\mathbf{e}^{(p)} \rightarrow \mathbf{c}^{(p)} \rightarrow \mathbf{e}^{(constraints)}$

Models

Semi-supervised Joint Hierarchical CRF



$P(\mathbf{C}_d, \mathbf{E}_d | \mathbf{x}_d)$

Claim Model	Entity Model	Order of Inference
$P(\mathbf{C}_d \mid \mathbf{e}_d^{(semi)}, \mathbf{x}_d)$	$P(\mathbf{E}_s \mid \mathbf{x}_s)$, $P(\mathbf{E}_s^{(semi)} \mid \mathbf{x}_s, \mathbf{c}_s)$	$\mathbf{e}^{(p)} \rightarrow \mathbf{c}^{(p)} \rightarrow \mathbf{e}^{(constraints)}$

Overview

Motivation

Approach

Experiments

Conclusions

Data and Evaluation Metrics

- Corpus:
  - 90 pleading documents from 49 IP cases
  - 70% training, 30% testing

Documents	Sentences	Words	Claims	Claim Numbers	Claim Types	Patents	Laws
90	25,250	548,402	362	319	579	1292	433

- Evaluation metric:
  - P/R/F1 using a strict matching criterion (CoNLL 2003)

Features

- Entity CRF:
  - Features extracted from previous, current, and next word. From each token we extract:
    - Word, POS tag, word shape
  - Claim tag of the current sentence (for top-down and joint CRFs)
- Claim CRF:
  - Sentence words
  - Number of new-line characters preceding the sentence
  - Entity tags in the sentence (for bottom-up and joint CRFs)

Overall Results

	Claims			Entities		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Top-down	80.00	54.05	64.52	**86.42**	52.63	65.42
Bottom-up	60.65	50.81	55.29*	48.1	**60.47**	53.58*
Semi-supervised Bottom-up	**89.74**	**56.76**	**69.54***	85.34	56.65	**68.09***
Semi-supervised Joint Hierarchical	88.89	56.22	68.87*	86.16	55.69	67.65*

- '*' indicates statistically significant differences w.r.t. the top-down model
- Bold-faced numbers correspond to the best performing model

Ablation Experiments

- Claim CRF (Semi-supervised Bottom-up):

Features	Precision	Recall	F1
All	89.74	56.76	69.54
- Lexicalization	61.84	25.41	36.02
- Pagination	88.33	57.3	69.51
- Entities	80.00	54.05	64.52

Entity information is the second most important feature

Lexicalization is the most important feature

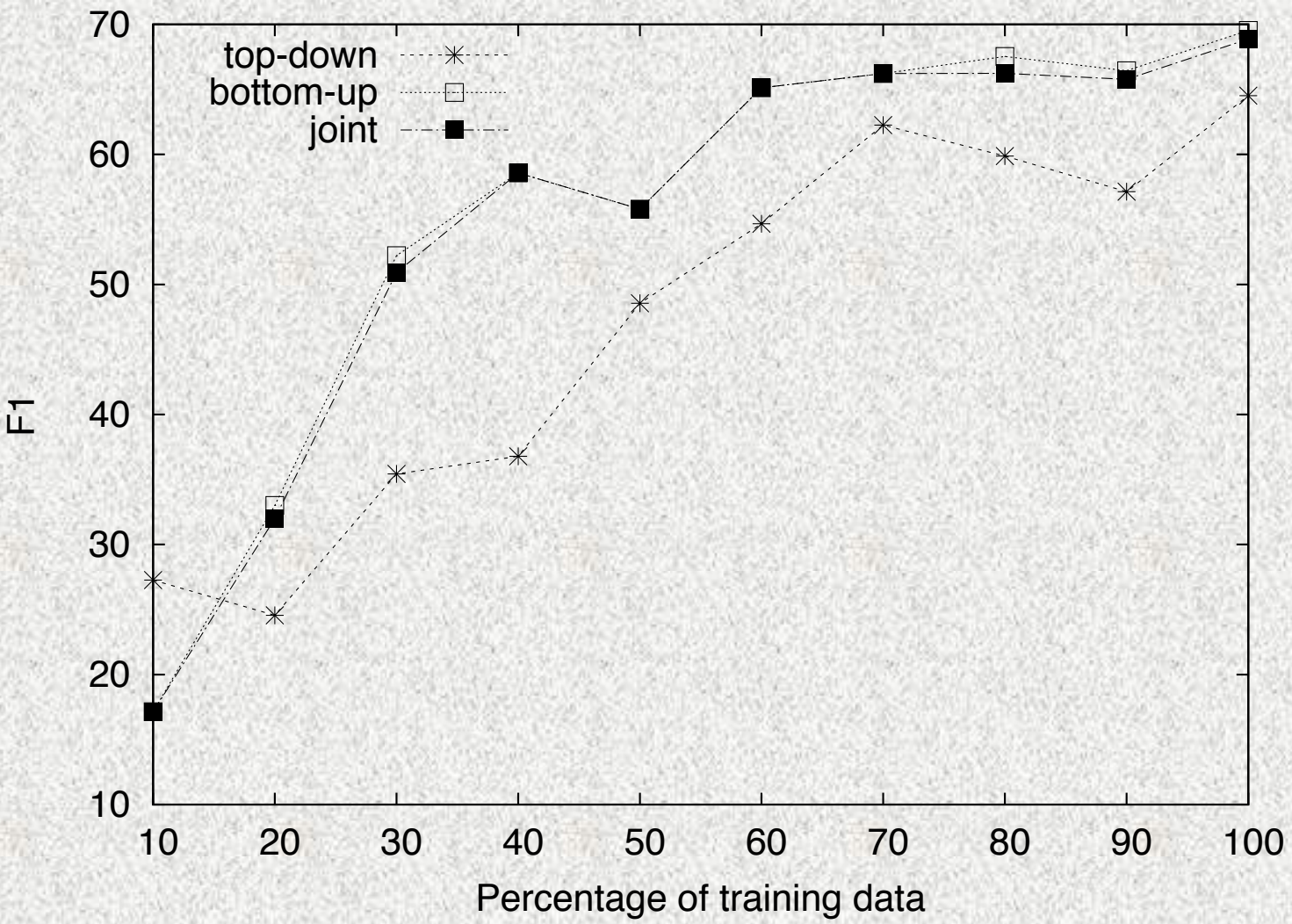
Ablation Experiments

- Entity CRF (Semi-supervised bottom-up):

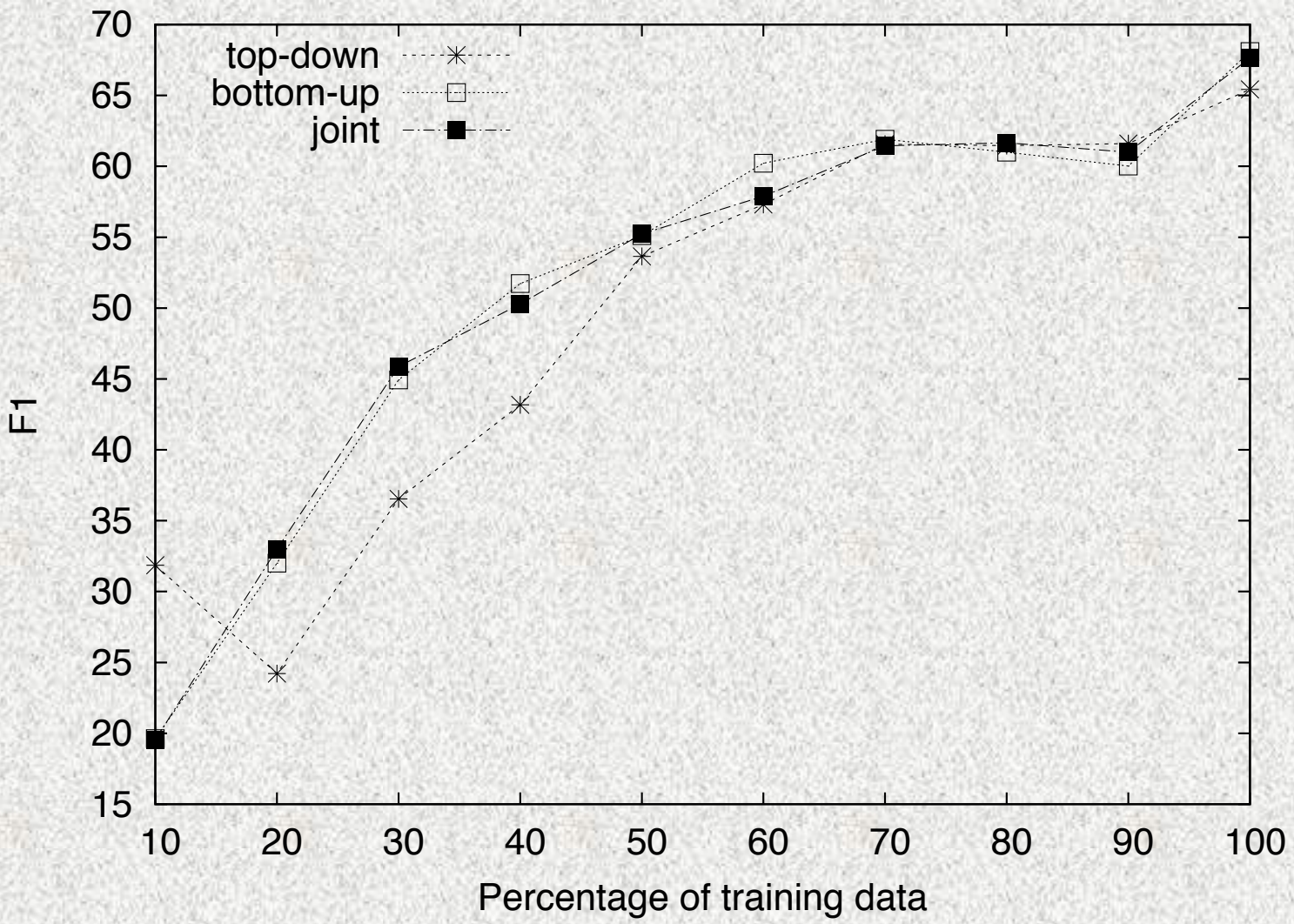
Features	Precision	Recall	F1
All	86.42	52.63	65.42
- Lexicalization	71.12	43.61	54.07
- POS tags	89.63	51.96	65.79
- Word shape	86.80	51.63	64.75
- Context	86.32	49.04	62.55

Lexicalization is the most important feature again!

Learning Curves
Claim CRF



Learning Curves
Entity CRF



Conclusions

- Introduced a novel hierarchical IE problem:
  - Only parts of documents are relevant
  - Linguistic annotations available only for those segments
- Investigated this problem on a novel IP Litigation domain
- Introduced two new approaches that outperform the traditional Top-down approach:
  - Semi-supervised Bottom-up
  - Semi-supervised Joint Hierarchical
- Showed that complex IE systems can be built training on hierarchical, partially labeled data
  - Reduces annotation work