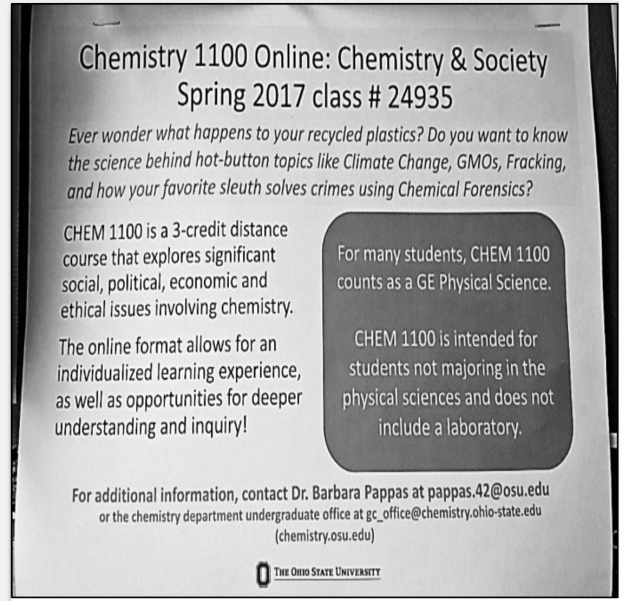
Visual Segmentation for Information Extraction from Heterogeneous Visually Rich documents

Ritesh Sarkhel*, Arnab Nandi
Department of Computer Science and Engineering

Outline

- Visually Rich Documents : Lots of useful information
- Challenges and observations
- Our contributions
  - **VS2-Segment**
  - **VS2-Select**
- Experimental results
  - **Event information extraction**
  - **Outperforms text-based baseline by an average F1-score of 5.07%**

Visually Rich Document



- Posters
- Leaflets
- Banners etc.

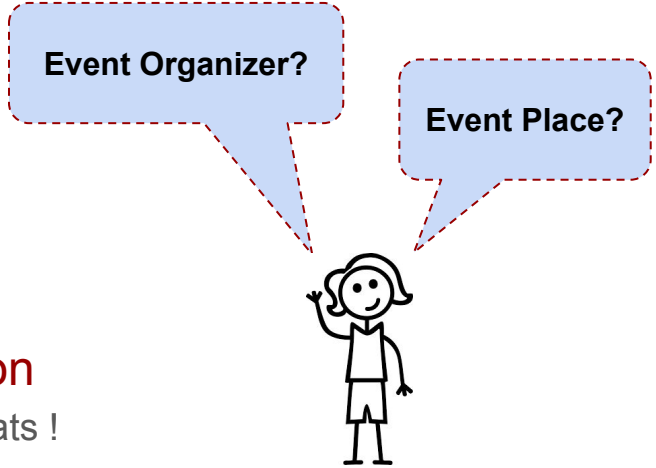
Properties of Visually Rich Documents

- Visual cues to highlight distinct semantic entities
  - Color
  - Font size
  - Text orientation and positioning
  - Negative-space
  - Symmetry etc.
- May be sparsely worded

Image source: Getty Images

Motivation

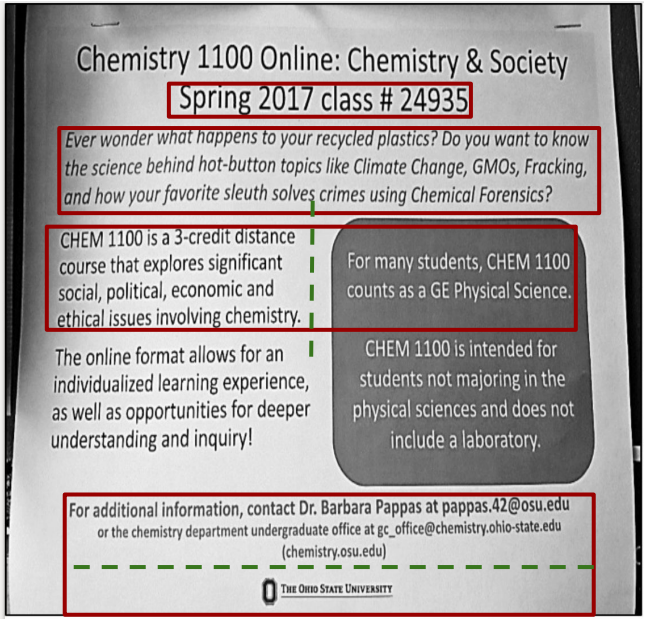
- Rich source of ad-hoc content
  - Contains a lot of useful information
  - Not easily available in an indexed database
- Motivating example:
  - Event information extraction
- Objective: Automated Information Extraction
  - **Recall**: Robust towards different document formats !



Outline

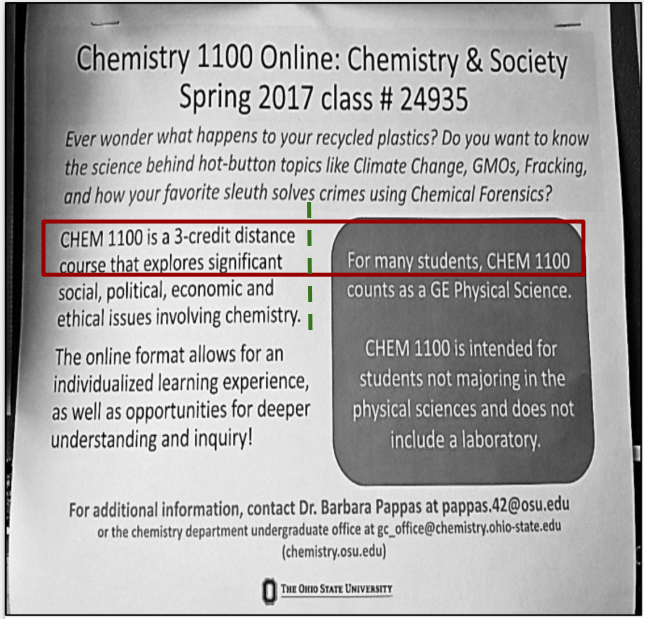
- Visually Rich Documents : Lots of useful information
- Challenges and observations
- Our contributions
  - VS2-Segment
  - VS2-Select
- Experimental results
  - Event information extraction
  - Outperforms text-based baseline by an average F1-score of 5.07%

Possible Solution 1: Text-based solutions?



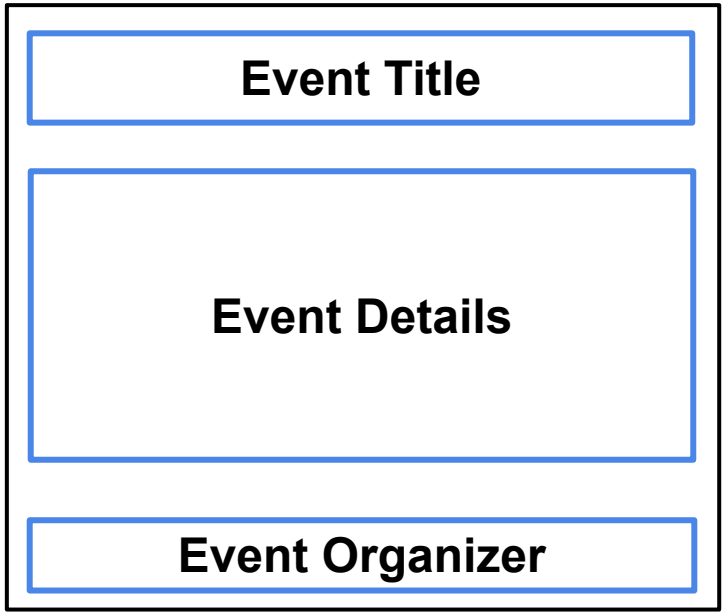
- Identifying the semantic entities
- Determining context boundaries
  - Semantic role played by visual features not considered by text-based solutions
- Transcription errors

Possible Solution 1: Text-based solutions?



Chemistry 1100 Online: Chemistry & Society ‘
1 Spring 2017 class # 24935
i 'er wonder what happens to your recycled plastics? Do you want to know i e science behind hot-button topics like Climate Change, GMOs, Fracking, and how your favorite sleuth solves crimes using Chemical Forensics?
CHEM 1100 is a 3-credit distance course that explores significant For many students, CHEM 1100 , social, political, economic and counts as a GE Physical Science. ethical issues involving chemistry. i r The online format allows for an CHEM 1100 is intended for ' individualized learning experience, Students not majoring in the 1 ' as well as opportunities for deeper physical sciences and does not understanding and ianifV! include a laboratory.
‘ 1
‘ For additional information, contact Dr. Barbara Pappas at pappas.42@osu.edu or the chemistry department undergraduate office at gc_offite@chemistryrohiastateedu (chemistryosuedu)
THE OHIO STATE UNIVERSITY

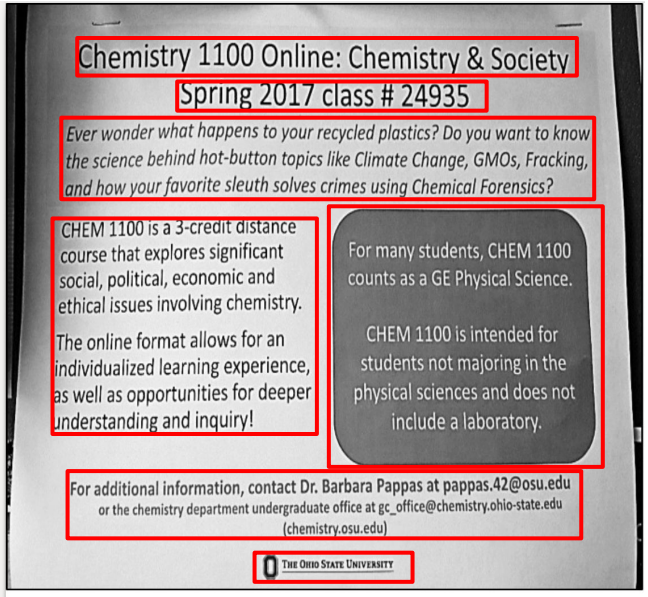
Possible Solution 2: Visual rule-based solutions?



- Template-based masks as rules
  - Create and store masks to identify where different named entities appear
  - e.g. ReportMiner

- Cost of scalability
  - Hard to maintain exact rules for all possible layouts
  - Expensive to deploy and maintain

Observations

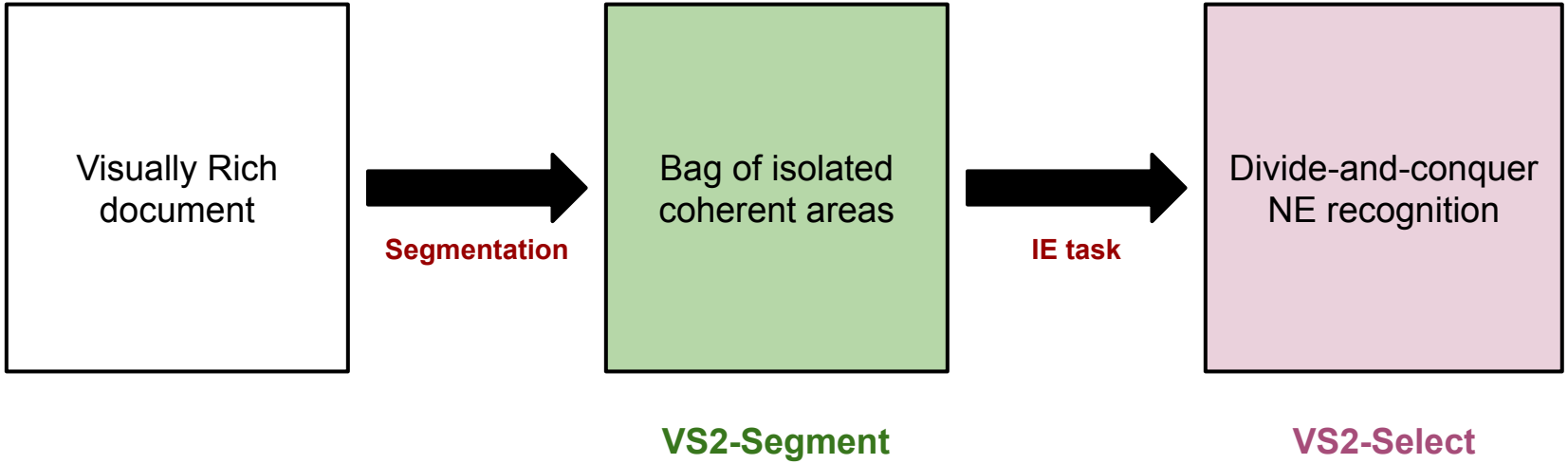


- Visually rich document = Bag of isolated coherent visual elements
  - Visual heterogeneity is significantly reduced in each of these coherent visual areas
- A few of these visual areas act as 'interest-points'
  - Areas with high semantic significance

Our hypotheses

1. Decompose a document into a bag of visually isolated coherent areas
   - Incorporate both visual and semantic features
2. Divide-and-conquer by text-based IE methods within each segmented area

Proposed IE workflow for Visually Rich documents



Outline

- Visually Rich Documents : Lots of useful information
- Challenges and observations
- Our contributions
  - **VS2-Segment**
  - VS2-Select
- Experimental results
  - Event information extraction
  - Outperforms text-based baseline by an average F1-score of 5.07%

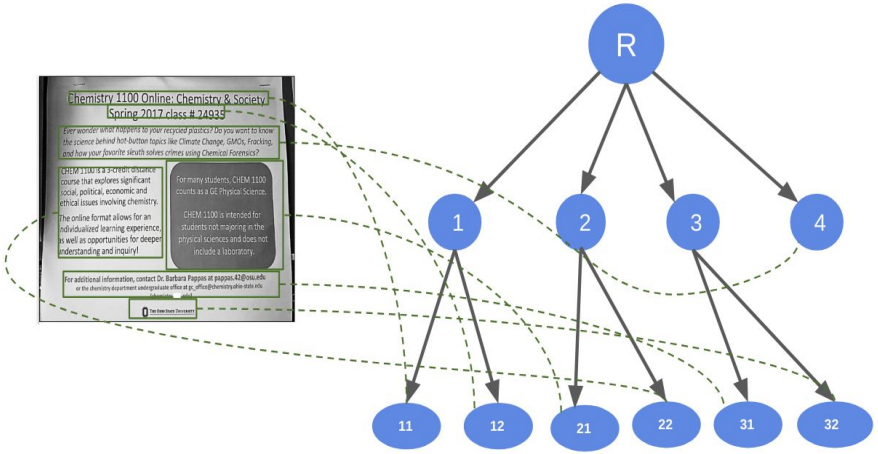
VS2-Segment: Visual Segmentation by layout analysis

Input: Visually Rich document (D)

Output: Bag of isolated coherent visual areas

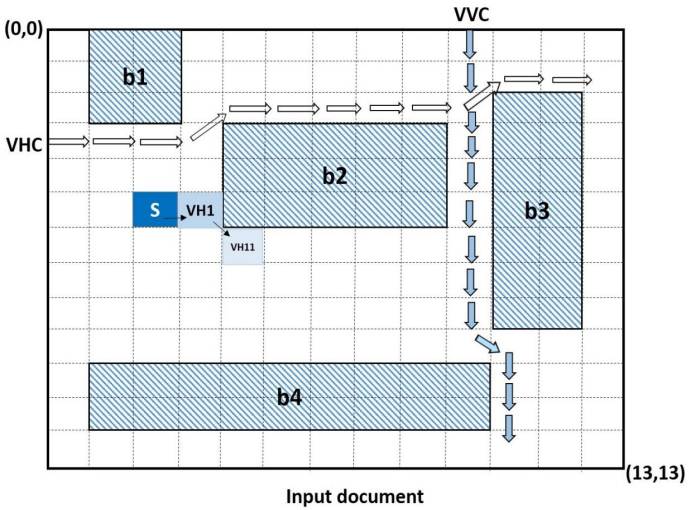
Constraint: Robust towards document formats

VS2-Segment: Segmentation by layout analysis



- A hierarchical layout-tree of D
  - Node → A visual area in D
  - Parent → Contains the visual area denoted by child
- Visual features: Explicit and Implicit
- Semantic features: Cosine similarity
- Return leaf nodes when terminates

VS2-Segment: Feature Descriptors

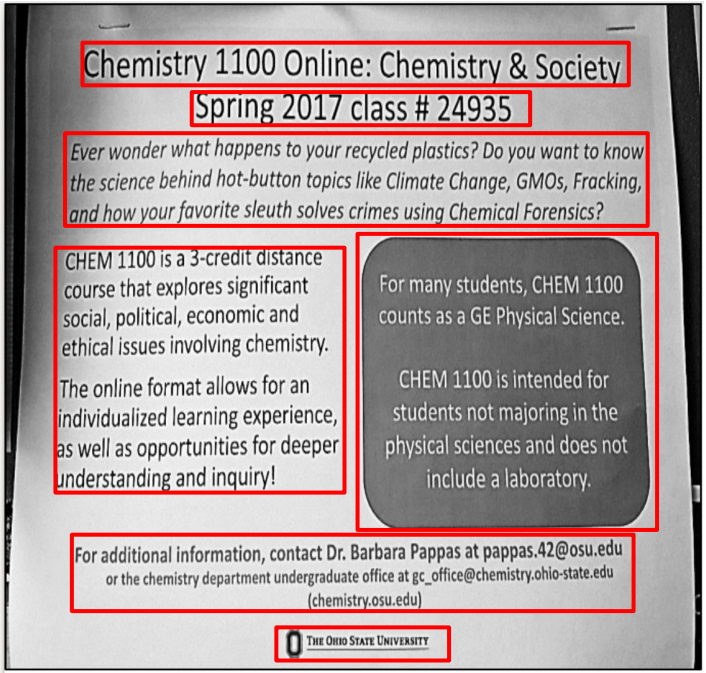


Index	Feature
1	Centroid coordinates
2	Height
3	LAB color
4	Angular distance from origin
5	Pairwise sum of angular distances

Explicit Visual Features

Implicit Visual Features

VS2-Segment: What did we learn so far ?



- A recursive algorithm to find semantically coherent areas
- Incorporates both visual and semantic features

Next up: IE by divide-and-conquer

Outline

- Visually Rich Documents : Lots of useful information
- Challenges and observations
- Our contributions
  - VS2-Segment
  - **VS2-Select**
- Experimental results
  - Event information extraction
  - Outperforms text-based baseline by an average F1-score of 5.07%

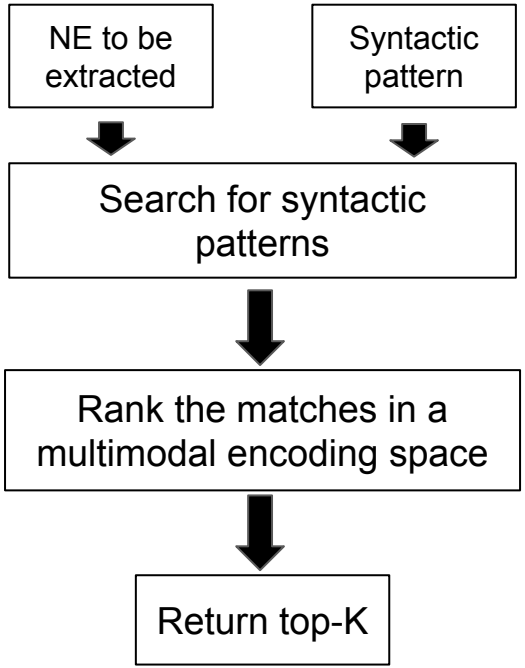
VS2-Select: Named entity recognition using distant supervision

Input: Bag of coherent visual areas **+** NE to be extracted

Output: Text extracted corresponding to the NE

Constraint: Robust towards ad-hoc content

VS2-Select: Named entity recognition using distant supervision



Divide

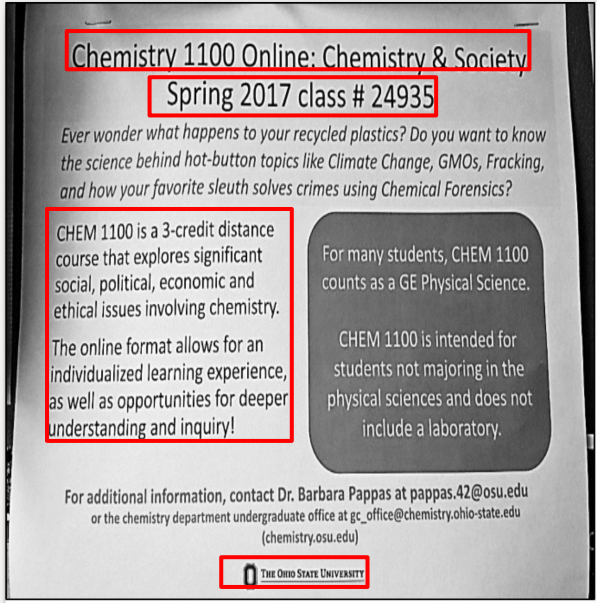
&

Conquer !

VS2-Select: Learning the syntactic patterns

- For each appearance **T** of an NE in a holdout corpus **H:**
  - Annotate and construct the dependency parse tree of **T**
  - Identify the patterns **P** represented by most frequent subtree(s)
  - **P** represents the NE
- Search for **P** within each segmented area

VS2-Select: Rank the matched patterns



Some Interest points of the poster

- Rank the matches based on relative semantic significance
  - Metric: Weighted L1 distance from the closest ‘interest point’ in the document
  - Return top-K
- ‘Interest Point’ → Higher semantic significance
  - Subset of visual areas obtained from segmentation
  - Pareto-optimal front in a multi-objective optimization problem
  - Each objective represents a design principle

VS2-Select: What did we learn so far ?

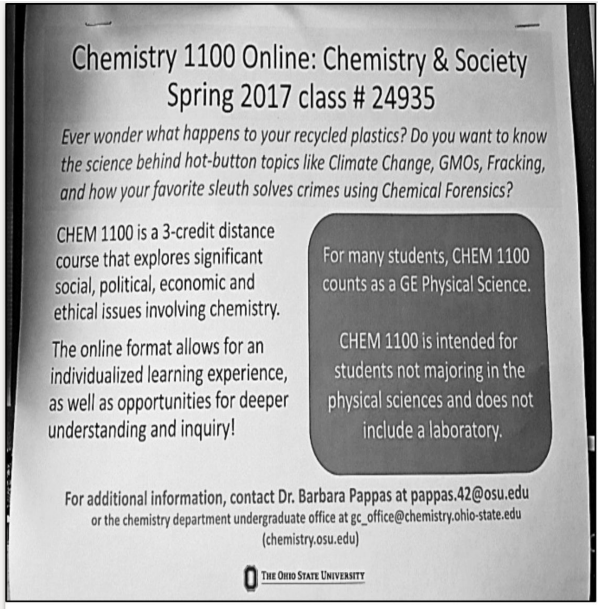
- Identifies NE's by leveraging the context boundaries determined by VS2-Segment
- Distantly supervised syntactic patterns represent each NE.

Outline

- Visually Rich Documents : Lots of useful information
- Challenges and observations
- Our contributions
  - VS2-Segment
  - VS2-Select
- **Experimental results**
  - **Event information extraction**
  - **Outperforms text-based baseline by an average F1-score of 5.07%**

Image source: Getty Images

Experiments

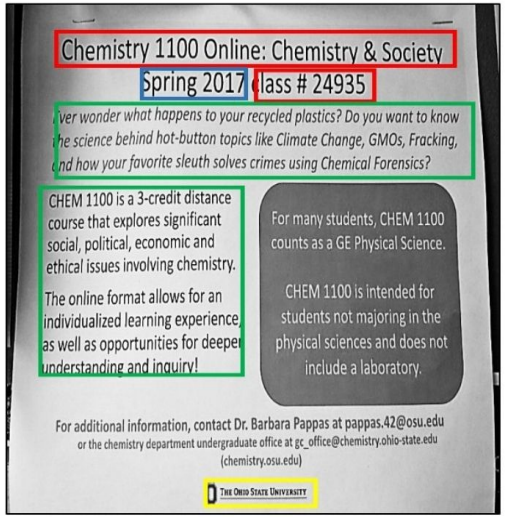


- 2190 event posters and flyers
  - Heterogeneous layouts
  - Digital images and PDF's
- 5 distinct named entities
  - Related to event information

Experiment: Event information extraction

Named entity	Definition
Event Title	Short description of the event
Event Place	Full address of the event
Event Time	Time of the event
Event Organizer	Host of the event
Event Details	Other details of the event

Evaluation metrics: Two-phase evaluation



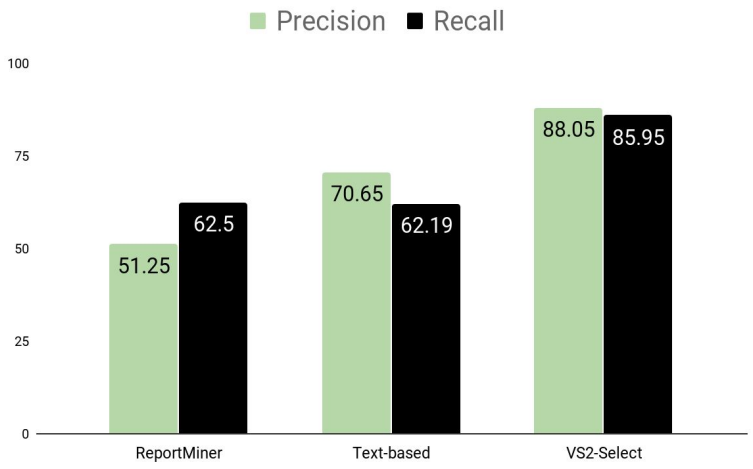
Color Legends

- Event title
- Event description
- Event organizer
- Event time
- Event place

- Text T identified as NE-type N is accurate:
  a. Its position is accurate **and** the named entity type N is accurate too
  b. Compared against human-annotated groundtruth

Quantitative evaluation: End-to-end

- Objective: End-to-end performance ?
- Accurate iff position coincides with ground-truth **and** inferred NE type is correct



Quantitative evaluation: End-to-end

- Takeaways:
  a. VS2 demonstrates robust performance for all NE types
  b. Both localization and classification capabilities are robust.
  c. Improves over ReportMiner[1], a visual template driven rule-based method
  d. Improvement over text-only baseline is significant

More experiments and analysis in paper!

Takeaways

- Prior segmentation helps IE from visually rich documents
  - Visual features need to be considered
- VS2: VS2-Segment and VS2-Select
  - Effectiveness of NE extraction explored
- Opens up a lot of exciting possibilities !
  - Explore the semantic relationship between the segmented visual areas
  - Utilize latent semantic information for more complex IE tasks -- QA, Summarization etc.

Thank you!

Image source: Getty Images

Questions?

Supplementary

Examples of learned lexical / syntactic patterns

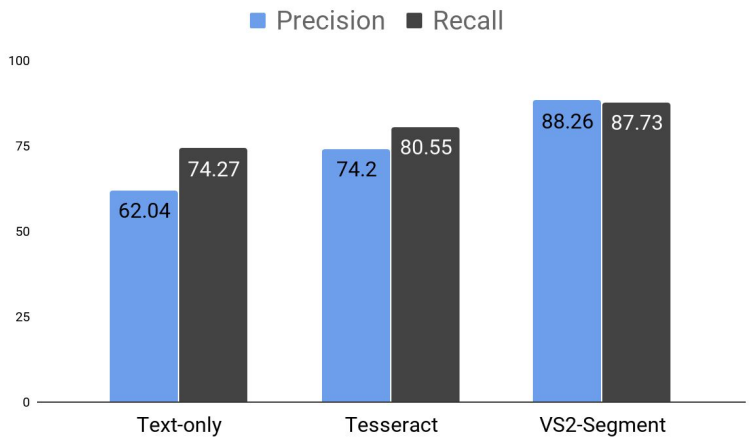
Named entity	Learned patterns
Event Title	{ *VP*, *NP* with numeric (*CD*) and textual (*JJ*) modifiers, *SVO* }
Event Place	*NP* with valid geocode tags
Event Time	*NP* with *TIMEX3* tags
Event Organizer	{ *Captain / Create / Reflexive-appearance* verb-senses, *NP* with *Person / Organization* named entities }

Learning the syntactic patterns

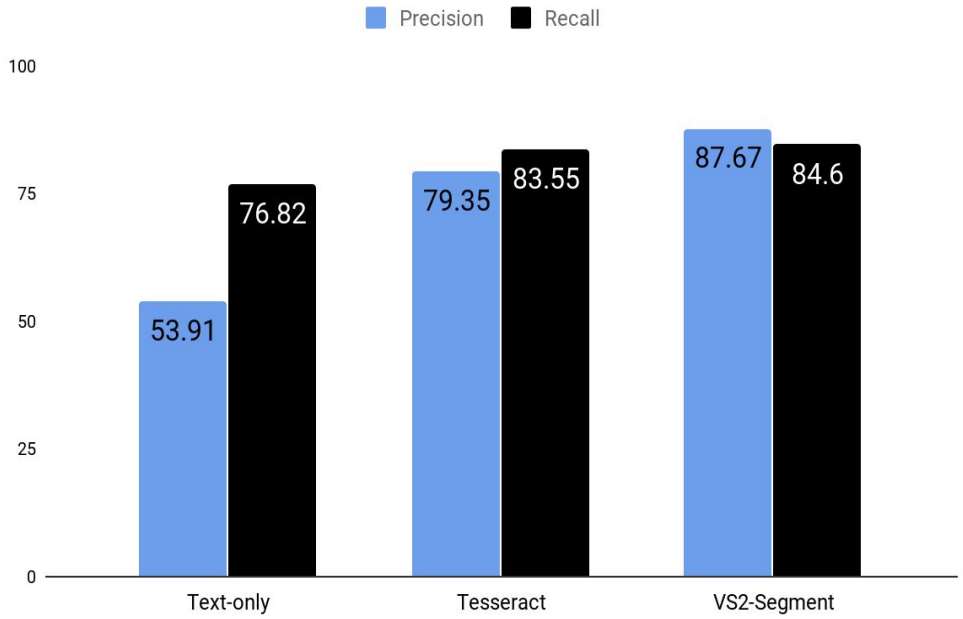
- Holdout corpus **H** created by scraping public-domain website
  - URL: allevents.in
  - Filters: Place and Date
- H contains 500 structured tuples
  - Tuple → {Named entity type, Corresponding text}
- Derive lexico-syntactic patterns for every named entity type in **H**

Quantitative evaluation: Localization

- Objective: Are the extracted named entities accurately localized ?
- Accurate if coincides with ground-truth

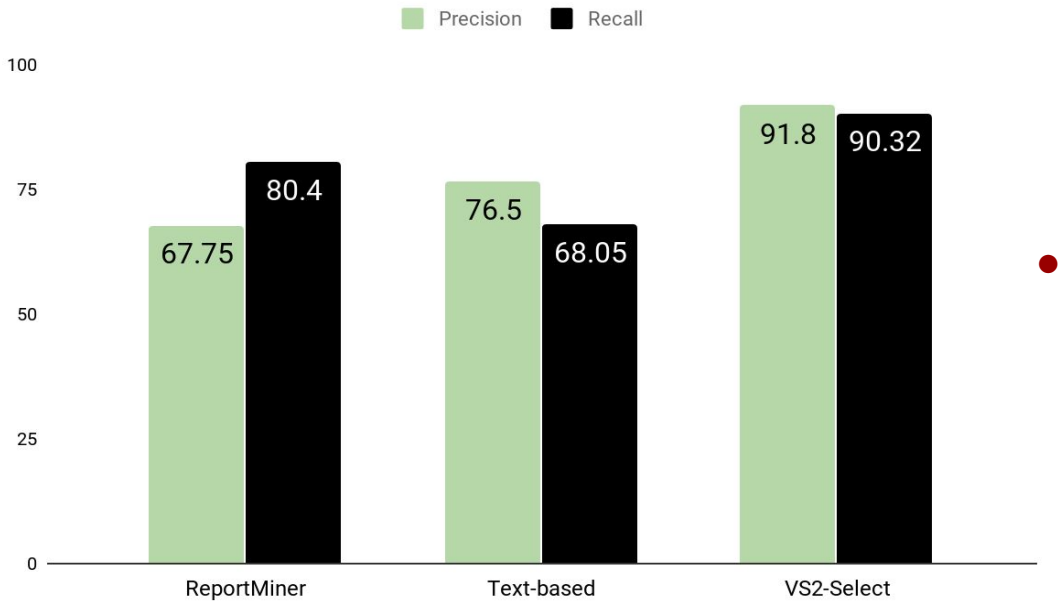


Property Information Extraction: Localization capability



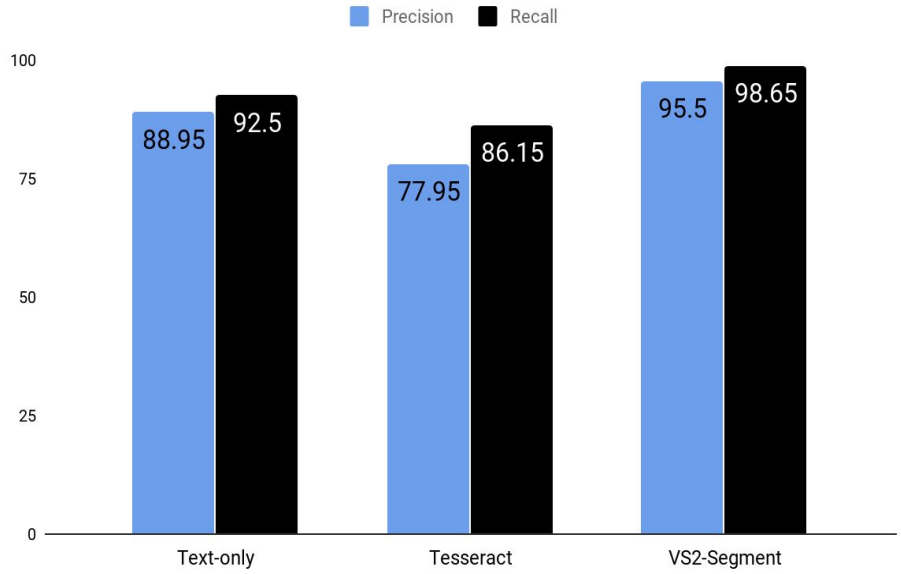
- Accurate iff IoU score against ground-truth ≥ 0.65

Property Information Extraction: Classification capability



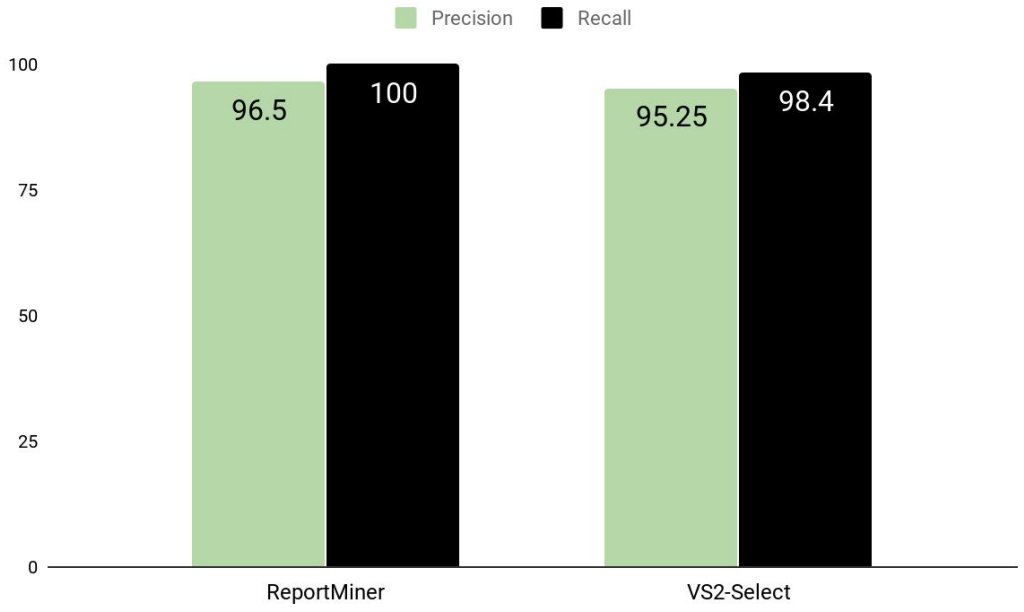
- Accurate iff IoU score ≥ 0.65 & inferred NE type is correct

Form field extraction: Localization capability



- Accurate iff IoU score against ground-truth ≥ 0.65

Form field extraction: Classification capability



- Accurate iff IoU score ≥ 0.65 & inferred NE type is correct

Ablation study

Index	VS2-Segment		VS2-Select	ΔF1 (%)		
	Visual feature	*Semantic feature based merging*	*Entity disambiguation*	*D1*	*D2*	*D3*
A1	✓	×	✓	0.80	2.55	3.37
A2	×	✓	✓	1.07	4.22	3.84
A3	✓	✓	×	0.95	6.78	7.05
A4	✓	✓	Text-only	0.42	4.55	3.96

Syntactic Patterns learned for the Real-Estate flyers dataset

Named entity type	Description	Syntactic patterns to search
Broker Name	Full name of the listing broker	A bigram/trigram of NE's with *Person / Organization* tags
Broker Phone	Contact number of the listing broker	A regular expression containing digits, characters and separators such as '-', '(', ')', and ' . '
Broker Email	Email address of the listing broker	An RFC-5322 compliant regular expression containing character and separators such as '@', and ' . '
Property Address	Full address information of the listing	Noun phrase with valid geocode tags
Property Size	Size-attributes summarizing the size of a listing (e.g. 4 beds, 2,465 acres)	(1) Noun phrase with numeric (*CD*) or textual modifiers (*JJ*) and (2) Noun POS tags with senses *measure / structure / estate* in the Hypernym Tree [42]
Property Description	Mentions of the property type (e.g. building,floor,land/lot) and other essential details (e.g. parking, grocery)	Noun phrases with numeric (*CD*) or textual modifiers (*JJ*)

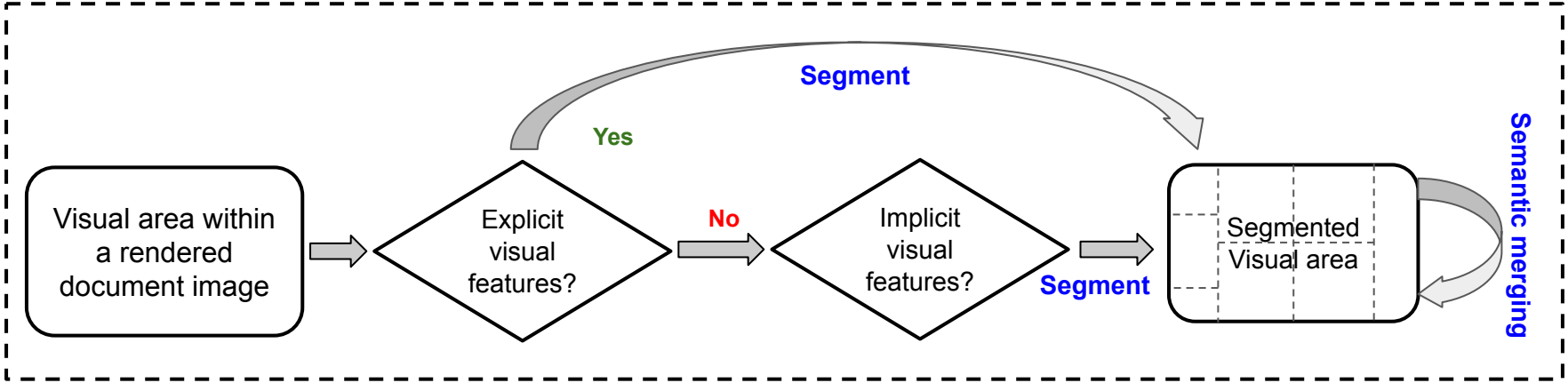
NE-specific breakdown for Event Information extraction

Index	Named Entity	Proposed method		
		Pr. (%)	*Rec.* (%)	$\Delta F1$(%)
N1	*Event Title*	84.88	81.09	8.98
N2	*Event Place*	76.68	86.37	3.76
N3	*Event Time*	94.67	84.70	0.49
N4	*Event Organizer*	72.56	74.41	10.50
N5	*Event Description*	76.59	86.00	1.60
	Overall	**81.08**	**82.51**	**5.07**

NE-specific breakdown for Property Information extraction

Index	Named Entity	Proposed method		
		Pr.(%)	*Rec.* (%)	$\Delta F1$ (%)
N1	*Broker Name*	94.72	90.85	10.18
N2	*Broker Phone*	96.15	82.25	1.63
N3	*Broker Email*	97.25	95.40	2.56
N4	*Property Address*	92.68	85.50	4.60
N5	*Property Size*	85.25	93.05	3.37
N6	*Property Desc.*	84.75	94.90	0.74
	Overall	**91.80**	**90.32**	**3.84**

VS2-Segment: Visual Segmentation by layout analysis



- Recursive operation to decompose a document into smaller visual areas