Improving Information Extraction from Visually Rich Documents using Visual Span Representations

Ritesh Sarkhel, Arnab Nandi
Department of Computer Science and Engineering
The Ohio State University

TL;DR

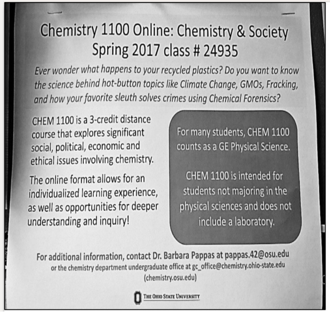
1. We formulate Information Extraction as a span classification problem
2. We show that learning a multimodal representation for a span of visual area in a document helps incorporate domain-specific knowledge in a IE pipeline
3. Our results show that this improves downstream performance on heterogeneous datasets
4. We present ML-based techniques on how to learn these representations with minimal human supervision

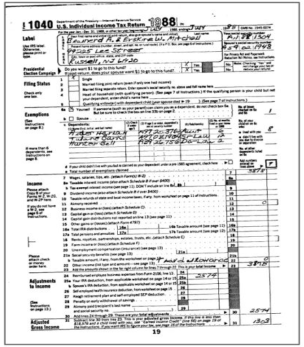
Overview

- ❖ Visually Rich Documents
- ❖ Problem definition
- ❖ Key challenges
- ❖ Overview of Artemis
- ❖ Experiments
- ❖ Takeaways

Visually Rich Documents

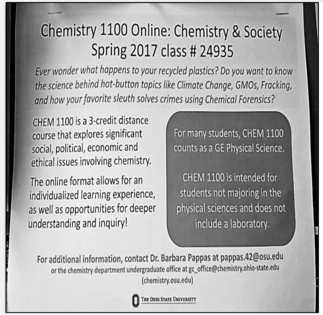
- ➢ Documents in which not only linguistic cues but <u>visual features also play a significant role in the semantics</u>
- ➢ **Visual features:** Font size, color distribution, whitespace balance, distance, orientation etc.

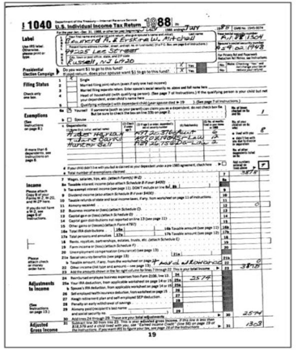




Visually Rich Documents

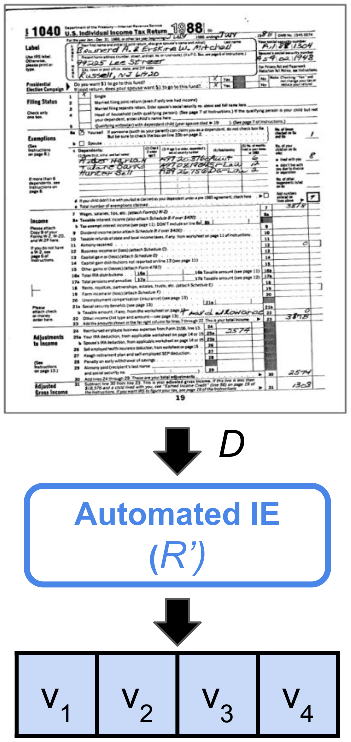
- ➢ We encounter visually rich documents everyday
    - ○ Posters, Banners, Forms, Magazine articles
    - ○ Can be sparsely or densely worded
- ➢ They are heterogeneous in nature
    - ○ Originate from various sources
    - ○ Diverse formats (e.g. PDF, HTML)
    - ○ Diverse layouts





Motivation

- ➢ Visually rich documents are rich source of ad-hoc information
- ➢ Extracting structured records makes it easy to search, index and query these documents using off-the-shelf analytical engines
- ➢ Reduces human effort, easier to gather insights



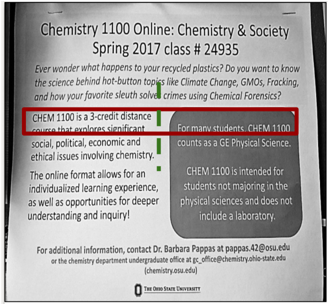
Problem Definition

Given a visually rich document *D* and a relational schema $R = \{a_1, a_2, \ldots a_n\}$, extract a structured record from *D* with schema *R*

- $a_1, a_2, \ldots a_n$ are various named entities we want to extract from *D*

Limitations of Existing Works

- Text-based extractors
  - Transcribe and apply off-the-shelf NLP solutions
  - Serialization error
  - Visual cues are not considered when identifying semantically distinct entities
- Rule-based extractors
  - Custom masks constructed for every entity to be extracted
  - Hard to maintain and update masks for all layouts
  - Expensive to deploy and maintain
- **A generalizable solution needs to be robust for diverse document types and reduce human-effort**


Chemistry 1100 Online: Chemistry & Society
Spring 2017 class # 24935
Ever wonder what happens to your recycled plastics? Do you want to know the 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 social, political, economic and ethical issues involving chemistry.
The online format allows for an individualized learning experience, as well as opportunities for deeper understanding and inquiry!
For many students, CHEM 1100 counts as a GE Physical Science.
CHEM 1100 is intended for students not majoring in the physical sciences and does not include a laboratory.
For additional information, contact Dr. Barbara Pappas at pappas.42@osu.edu or the chemistry department undergraduate office at gc_office@chemistry.ohio-state.edu (chemistry.osu.edu)

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.
‘ l
‘ 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

Formulation and Solution Overview

- Identifying named entities in a document is a span classification task
- The IE task boils down to a binary classification problem once we have identified the candidate visual spans
- We can leverage machine-learning algorithms to reduce human-effort at each step

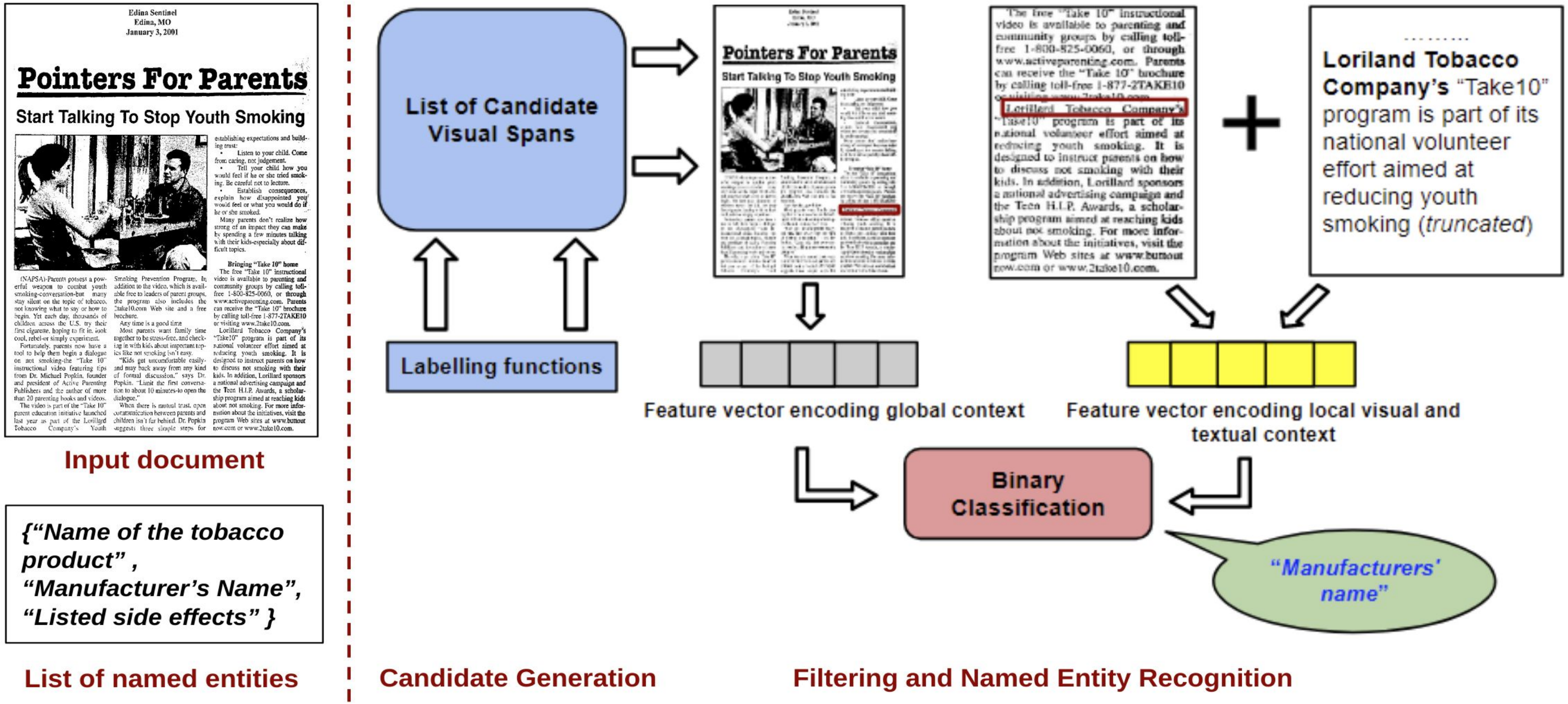
ARTEMIS: IE as Visual Span Classification

- Artemis extracts named entities a_i, $i = 1$ *to* n, as defined in a relational schema $R = \{a_1, a_2, \ldots a_n\}$ from a document D in two steps:
  a. Identify candidate visual spans in D using domain-specific knowledge in the form of weak supervision
  b. Identify the visual span containing a NE a_i, $i = 1$ *to* n using a supervised classifier

Input: Rendered image of a visually rich document (*D*), relational schema *R*

Output: A structured record with schema *R*

ARTEMIS: IE as Visual Span Classification



Step 1: Candidate Generation

- Artemis leverages domain-specific knowledge in the form of multimodal labelling functions to identify candidate visual spans
  - Labelling functions as weak supervision sources were first introduced by Ratner et al.[1]

```
def text_matcher(ne_lst,D,T){
    candidate_span_lst = []
    text = transcribe(D)
    for ne in ne_lst:
        if ne in text:
            span_coords = T.lookup(approx_match(text,ne))
            candidate_span_lst.append(span_coords)
    return candidate_span_lst
}
```

```
def position_matcher(ne_pos_lst,D,T){
    candidate_span_lst = []
    for ne_pos in pos_lst:
        text_line_coords = T.traverse(ne_pos)
        span_coords = pad(text_line_coords,50)
        candidate_span_lst.append(span_coords)
    return candidate_span_lst
}
```

[1] Ratner, Alexander J., Stephen H. Bach, Henry R. Ehrenberg, and Chris Ré. "Snorkel: Fast training set generation for information extraction." In *Proceedings of the 2017 ACM international conference on management of data*, pp. 1683-1686. 2017.

Step 2: Representation Learning and Classification

- After transcribing and preprocessing, we chunk each candidate visual span and classify them as an instance of the NE’s to be extracted

- **We represent each visual span using two fixed-length vectors**
  - Local context vector
    - Encodes invariant properties of a visual span from its local context
  - Global context vector
    - Encodes discriminative properties of the document

IE as Binary Classification

- The probability of a chunk within a candidate visual span containing a named entity depends on the output of an inference task
- We formulate it as a binary classification problem using local and global context vector for feautirization
- Repeated for every named entity in *R*
- More details in our paper

The Global Context Vector

- Some named entities are more likely to appear in certain document types
  - The named entity "SIDE_EFFECTS" is more likely to appear on a newspaper article on tobacco addiction than a real-estate flyer
  - We encode this corpus-level statistics i.e. the correlation between the named entities in *R* with discriminative properties of training documents using the global context vector
- We compute the global context vector of a visual span from the input document using a discriminative convolutional network [1]
  - Takes the rendered image of a document as input and outputs a softmax label
  - The global context vector is computed from the last fully-connected layer of the network

[1] Sarkhel, Ritesh, and Arnab Nandi. "Deterministic routing between layout abstractions for multi-scale classification of visually rich documents." In *28th International Joint Conference on Artificial Intelligence (IJCAI), 2019*. 2019.

The Local Context Vector

- **Local context:** A neighboring area enclosing the visual span that plays a significant role on its semantics
- To determine the local context boundary of a visual span, we segment the document into multiple isolated visual areas
- We develop an adversarial neural network that finds an optimal segmentation for each document using limited labeled examples
    - This reduces the effort required to carefully design handcrafted features for each unique layout

Identifying the Local Context Boundary

Edina Sentinel
Edina, MO
January 3, 2001

Pointers For Parents

Start Talking To Stop Youth Smoking

(NAPSA)-Parents possess a powerful weapon to combat youth smoking-conversation-but many stay silent on the topic of tobacco, not knowing what to say or how to begin. Yet each day, thousands of children across the U.S. try their first cigarette, hoping to fit in, look cool, rebel-or simply experiment.

Fortunately, parents now have a tool to help them begin a dialogue on not smoking-the "Take 10" instructional video featuring tips from Dr. Michael Popkin, founder and president of Active Parenting Publishers and the author of more than 20 parenting books and videos.

The video is part of the "Take 10" parent education initiative launched last year as part of the Lorillard Tobacco Company's Youth Smoking Prevention Program. In addition to the video, which is available free to leaders of parent groups, the program also includes the 2take10.com Web site and a free brochure.

Any time is a good time

Most parents want family time together to be stress-free, and checking in with kids about important topics like not smoking isn't easy.

"Kids get uncomfortable easily-and may back away from any kind of formal discussion," says Dr. Popkin. "Limit the first conversation to about 10 minutes-to open the dialogue."

When there is mutual trust, open communication between parents and children isn't far behind. Dr. Popkin suggests three simple steps for establishing expectations and building trust:

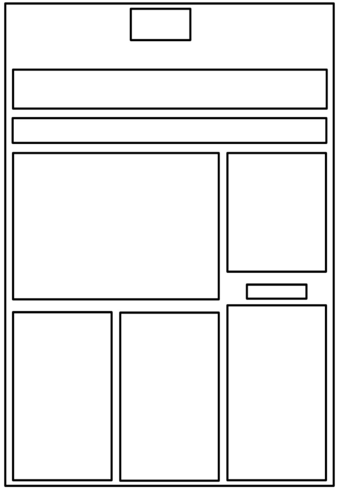
- Listen to your child. Come from caring, not judgement.
- Tell your child how you would feel if he or she tried smoking. Be careful not to lecture.
- Establish consequences, explain how disappointed you would feel or what you would do if he or she smoked.

Many parents don't realize how strong of an impact they can make by spending a few minutes talking with their kids-especially about difficult topics.

Bringing "Take 10" home

The free "Take 10" instructional video is available to parenting and community groups by calling toll-free 1-800-825-0060, or through www.activeparenting.com. Parents can receive the "Take 10" brochure by calling toll-free 1-877-2TAKE10 or visiting www.2take10.com.

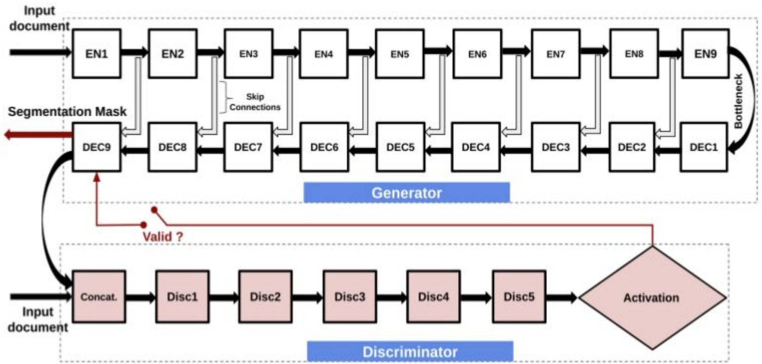
Lorillard Tobacco Company's "Take10" program is part of its national volunteer effort aimed at reducing youth smoking. It is designed to instruct parents on how to discuss not smoking with their kids. In addition, Lorillard sponsors a national advertising campaign and the Teen H.I.P. Awards, a scholarship program aimed at reaching kids about not smoking. For more information about the initiatives, visit the program Web sites at www.buttoutnow.com or www.2take10.com.



Adv. NN

Local context boundary

Identifying the Local Context Boundary



A snapshot of the adversarial neural network during inference

Block Type	Input	Operator	K	S
Input	$512^2 \times 3$+$512^2 \times 4$	*concatenation*	-	-
DISC1	$512^2 \times 7$	*discriminator-block*	2	2
DISC2	$256^2 \times 2$	*discriminator-block*	2	2
DISC3	$128^2 \times 4$	*discriminator-block*	2	2
DISC4	$64^2 \times 8$	*discriminator-block*	2	2
DISC5	$32^2 \times 16$	*discriminator-block*	2	2
Validity Matrix	$16^2 \times 1$	1×1 *convolution* + *sigmoid activation*	-	-

The Discriminator architecture

Operation Type	Symbol	Input	Operator	K	S
Encoder	EN1	$512^2 \times 3$	*encoder-block*	3	2
	EN2	$256^2 \times 2$	*encoder-block*	3	2
	EN3	$128^2 \times 4$	*encoder-block*	3	2
	EN4	$64^2 \times 8$	*encoder-block*	3	2
	EN5	$32^2 \times 16$	*encoder-block*	3	2
	EN6	$16^2 \times 16$	*encoder-block*	3	2
	EN7	$8^2 \times 16$	*encoder-block*	3	2
	EN8	$4^2 \times 16$	*encoder-block*	3	2
	EN9	$2^2 \times 16$	*encoder-block*	2	1
Decoder	DEC1	$1^2 \times 16$	*decoder-block*	2	1
	DEC2	$2^2 \times 32$	*decoder-block*	2	2
	DEC3	$4^2 \times 32$	*decoder-block*	2	2
	DEC4	$8^2 \times 32$	*decoder-block*	2	2
	DEC5	$16^2 \times 32$	*decoder-block*	2	2
	DEC6	$32^2 \times 32$	*decoder-block*	2	2
	DEC7	$64^2 \times 16$	*decoder-block*	2	2
	DEC8	$128^2 \times 8$	*decoder-block*	2	2
	DEC9	$256^2 \times 4$	*decoder-block*	2	2

The Generator architecture

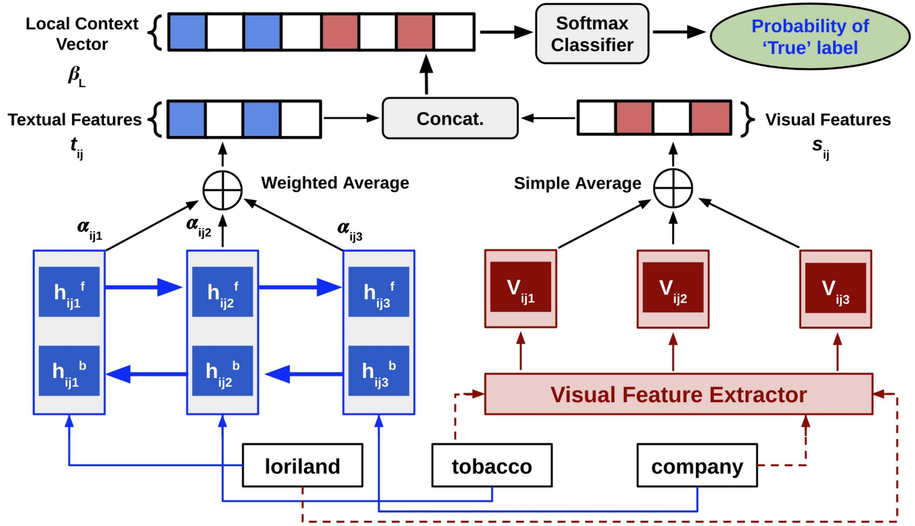
Block	Input	Operator	Output
encoder-block	$h \times w \times c$	conv2d	$h \times w \times c$
	$h \times w \times c$	Batch Normalization	$h \times w \times c$
	$h \times w \times c$	ReLU	$h \times w \times c^*$
decoder-block	$h \times w \times c$	conv2d-transpose	$h \times w \times c$
	$h \times w \times c$	Batch Normalization	$h \times w \times c$
	$h \times w \times c$	Dropout + Skip concat.	$h \times w \times 2c$
	$h \times w \times 2c$	ReLU	$h \times w \times c^*$
discriminator-block	$h \times w \times c$	conv2d	$h \times w \times c$
	$h \times w \times c$	ReLU	$h \times w \times c^*$

Computing the Local Context Vector

Edina Sentinel
Edina, MO
January 3, 2001

Pointers For Parents

Start Talking To Stop Youth Smoking



We obtain the local context vector of a visual span using a multimodal bi-directional LSTM network with attention

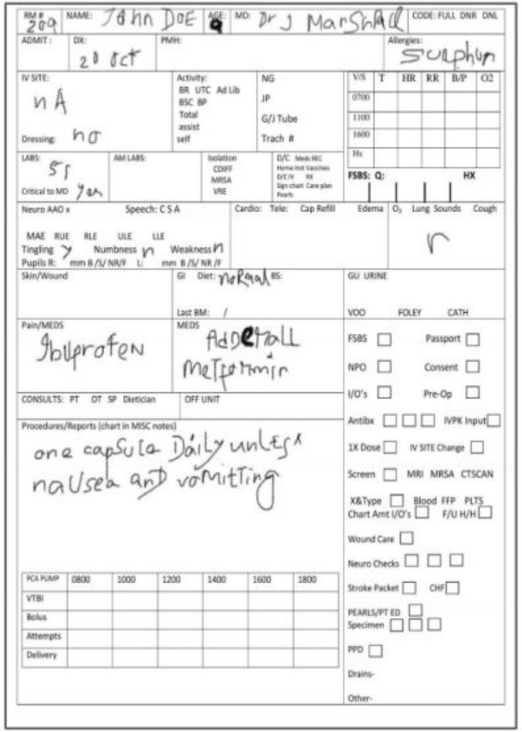
Experiments

- We evaluate Artemis on four visually rich datasets
  - NIST special dataset
    - 5595 scanned documents representing 20 different forms from the IRS-1040 package
  - Tobacco Litigation dataset
    - 1553 scanned front pages of biomedical journals from the National Library of Medicine
  - MARG dataset
    - 3482 documents from publicly available litigation records against US tobacco companies in ‘98
  - BRAINS dataset
    - Approx. 1M documents mimicking a record-keeping tool used by registered nurses for keeping up-to-date information about patients under emergency care

Experiments

- ➢ BRAINS dataset
  - ○ IE task targeting named entities related to patient identifiers used by RN's

Index	Named entity	Description
1	*"Patient's name"*	Name of the patient under medical care
2	*"Age"*	Patient's age when admitted
3	*"Gender"*	Patient's gender
4	*"Code"*	Resuscitation status of the patient
5	*"Admit date"*	The day when the patient was first admitted to the ER
6	*"Room number"*	Room number where the patient is now in the hospital
7	*"Diagnosis"*	Latest diagnosis of the patient made by the medical doctor responsible for the patient
8	*"Medical history"*	Past medical records of the patient
9	*"Dietary restrictions"*	Known food allergies
10	*"Consulting physician"*	Name of the medical doctor responsible for the patient



Experiments

- ➢ All of our datasets are heterogeneous (various sources of origin, layout, and format)
- ➢ IE tasks defined on them are also distinct
    - Complete list of named entities for all four datasets available here

Metrics

- We consider an output by our method accurate iff:
    - Its position in the document overlaps with the groundtruth with an IOU score >= 0.65
    - The NE type assigned to it by our classifier is the same as its groundtruth label
- We report both Accuracy@1 and F1-score for our method on all four datasets

Result Highlights

Dataset	Text-only (A1) Accuracy (%)	Text-only (A1) F1 (%)	ReportMiner (A2) Accuracy (%)	ReportMiner (A2) F1 (%)	Graph-based (A3) Accuracy (%)	Graph-based (A3) F1 (%)	Weak Supervision (A4) Accuracy (%)	Weak Supervision (A4) F1 (%)	Artemis Accuracy (%)	Artemis F1 (%)
NIST Dataset	89.75	86.33	**97.50**	**93.25**	95.50	91.86	95.50	92.0	95.55	92.60
MARG Dataset	69.45	67.50	67.70	62.25	72.0	70.95	71.25	69.07	**74.33**	**72.50**
Tobacco Litigation Dataset	51.20	49.65	59.70	55.25	65.25	62.90	63.50	61.35	**68.50**	**67.25**
Brains Dataset	68.50	64.33	62.07	56.50	74.25	70.96	74.50	69.42	**78.40**	**74.35**

More experiments and analysis in paper!

Result Highlights

- We compare our method against a number of baselines
  - Text-based (transcription + bi-LSTM)
  - State-of-the-art weakly supervised baseline
  - Graph-based method
  - A commercially available tool
- We perform better or comparably against all baselines
- Improvement of up to 17 F1 points against a text-based baseline
- Consistent performance on diverse datasets for separate IE tasks

Conclusion and Takeaways

- ➢ We described Artemis -- a visually-aware IE method for heterogeneous, visually rich documents
- ➢ It formulates an IE task as a visual span classification problem
- ➢ It represents each visual span in a multimodal embedding space
- ➢ Experiments on four heterogeneous datasets of visually rich documents for separate IE tasks show that our method is robust and generalizable